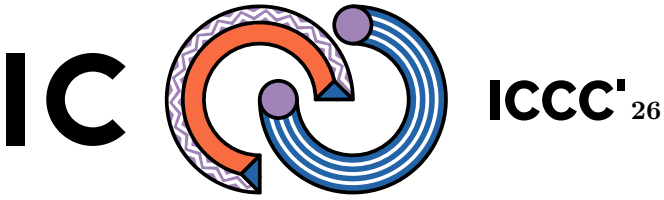
Proceedings of the 17th International Conference on Computational Creativity

June 29 – July 03, 2026 • Coimbra, Portugal

Editors: Hugo Gonçalo Oliveira • Anna Kantosalo • Hannu Toivonen • Penousal Machado • Ana Alves



Association for Computational Creativity
International Conference on Computational Creativity



Proceedings of the Seventeenth International Conference on Computational Creativity

ICCC'26
Coimbra, Portugal — June 29 - July 03

Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado and Ana Alves (Editors)

Published by the Association for Computational Creativity (ACC)

Association for Computational Creativity

University of Coimbra
Portugal

http://computationalcreativity.net/iccc26/

First published 2026

TITLE: Proceedings of the 17th International Conference on Computational Creativity

EDITORS: Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado and Ana Alves

ISBN: 978-989-54160-8-0
ISSN: 3051-6706

Published by the Association for Computational Creativity (ACC)

Copyright notice: all contents of these proceedings by ACC, the Association for Computational Creativity, are published under a Creative Commons Attribution (CC BY) license, which allows unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (https://creativecommons.org/licenses/by/4.0/).





Preface

Welcome to the 17th International Conference on Computational Creativity, ICCC'26, held from June 29 to July 3 in Coimbra, Portugal! The rich scientific program consisted of presentations of 27 peer-reviewed full papers and 38 peer-reviewed short papers, two invited keynotes, four workshops, and an Early Career Symposium.

The full paper track of ICCC'26 received 106 submissions. Each paper was assigned for review to three Program Committee members and one Senior Program Committee member, except for the 33 submissions that were desk-rejected by the Program Chairs. Based on the reviews and the recommendations of the Senior PC members, 28 full papers (26%) were initially accepted for publication and oral presentation in the conference but one accepted paper was retracted by its authors. Among the submissions that were peer-reviewed, the acceptance rate of full papers was 38%.

The short paper track of ICCC'26 received another 90 submissions, each of which was reviewed by three Program Committee members, including Senior members, except for 15 desk rejections. A total of 38 short papers (42%) were accepted for presentation as short talks or posters. Several of them were also presented as demos. Counting only the peer-reviewed submissions, the acceptance rate of short papers was 51%.

Overall, the diversity of submissions showcases a lively, growing community of Computational Creativity researchers. This year, the most popular topic categories were Human-Machine Co-Creativity, Applications of Computational Creativity, Computational Creativity Evaluation (full papers only), and Interdisciplinary Perspectives (short papers only). Domain-specific topic categories were less popular this year; the most frequent were Language and Visual Arts.

The diversity of the conference is also highlighted by the different paper types accepted to the conference. Among the accepted full papers we have eleven technical papers, six position papers, five system or resource description papers, four study papers, and two cultural application papers. The accepted short papers, in turn, are distributed into nine pilot studies, eight nuggets and gems, five bridges across fields, five debate sparks, four late-breaking results, four system demonstrations, two meta-perspectives on the field, and one field and event report.

Out of all of the excellent papers submitted to the conference, outstanding papers were recognised with best paper awards, namely:

- Best Paper Award winner: *Theory as Architecture: Why Aesthetic Frameworks Should Constitute, Not Merely Evaluate, AI Art Systems* by Atticus Sims.
- Best Student Paper Award winner: *What Makes Literary Quotes Stand Out? A Multidimensional Analysis of Literary Creativity* by Berke Yazan, Almila Akdag and Davide Dell'Anna.
- Best Short Paper Award winner: *Before the Prompt: Heteromorphic Imagination and the Cognitive Cost of Premature Crystallization in Human-AI Co-Creativity* by Emil Polyak and Rghad Balkhyoor.
- Best Poster Awards (two): *Back to Back with a Copy: A Computational Analysis of AI-Generated Visual Contemporary Art Pastiches* by Anca Dinu, Andreiana Mihail, Andra-Maria Florescu, Claudiu Creangă and Liviu P. Dinu; and *World-State Transformations to Foster Player Creativity in Neuro-symbolic Interactive Storytelling* by Santiago Góngora, Luis Chiruzzo, Gonzalo Mendez and Pablo Gervás.

Congratulations to all award winners!

In addition, the program featured two distinguished keynotes:

- Jennifer Jacobs, Associate Professor in Media Arts and Technology, University of California, Santa Barbara, gave a keynote titled *Materials and Machines: How the Physical World Can Inform Computational Tools for Creative Production.*
- Leonel Moura, a Portuguese artist who has pioneered the use of robotics and artificial intelligence in art, gave a keynote titled *Sapiens AI: The New Condition*, with the motto "Artificial Intelligence is not a tool. It is a partner."

The conference program also featured four workshops on varying topics and an Early Career Symposium. The workshops were: The Not-so-secret Workshop; Creative Practitioners Collaborating With AI; Computational Design and Computer-aided Creativity Workshop; and Workshop on Theoretical CS and Computational Creativity. The Early Career Symposium connected early career researchers with senior members of the community to support their growth through flash talks, mentoring discussions and a joint lunch.

As always, the scientific program of ICCC was the result of the efforts of numerous researchers. We would like to thank all the authors who submitted their work to ICCC'26, not just the 73 authors of the full papers and the 95 authors of the short papers accepted for the conference.

We would also like to thank the hard-working Program Committee for their peer reviews and discussions, and the Senior PC members for additionally coordinating the discussions and writing metareviews for full paper submissions. This year, the Program Committee consisted of 26 senior members and 61 regular members. PC members typically reviewed 3–4 full papers and 2–3 short papers. We extend our thanks to the additional reviewers who helped the committee in the review process.

ICCC'26 took place at the University of Coimbra, founded in 1290, and a UNESCO World Heritage site. We are happy and proud of the opportunity to organise the conference at one of the oldest universities in the world!

The rich social program of the conference included a welcome reception at the Botanical Garden of the University of Coimbra, a guided tour of the historical university area and the old town, a conference dinner at the charming Quinta do Encontro vineyard, as well as a guided tour to the Coimbra Biennial of Contemporary Art.

We would like to thank the local organizers at the University of Coimbra and all other members of the Organizing Committee for their tireless efforts to make ICCC'26 a successful and pleasurable event, in the spirit of the ICCC conference series since 2010.

Hugo Gonçalo Oliveira, Anna Kantosalo, and Hannu Toivonen — Program Committee Chairs
Penousal Machado — General Chair
Ana Alves — Proceedings Chair

Program Committee

Program Commitee Chairs

Hugo Gonçalo Oliveira — University of Coimbra
Anna Kantosalo — Haaga-Helia University of Applied Sciences
Hannu Toivonen — University of Helsinki

Senior PC Members

Alison Pease — University of Dundee
Ashok Goel — Georgia Institute of Technology
Brad Spendlove — Randolph College
Carlos León — Universidad Complutense de Madrid
Caterina Moruzzi — University of Edinburgh
Christian Guckelsberger — Aalto University
Dan Brown — University of Waterloo
Dan Ventura — Brigham Young University
Diarmuid O'Donoghue — Maynooth University
F. Amílcar Cardoso — University of Coimbra, CISUC, DEI and LASI
Geraint Wiggins — Vrije Universiteit Brussel / Queen Mary University of London
João Miguel Cunha — University of Coimbra
Kazjon Grace — The University of Sydney
Maria M. Hedblom — Jönköping University
Maria Teresa Llano — University of Sussex
Oliver Kutz — Free University of Bozen-Bolzano
Mary Lou Maher — University of Sydney
Maya Ackerman — Santa Clara University
Nada Lavrač — Jozef Stefan Institute
Oliver Bown — UNSW
Pablo Gervás — Universidad Complutense de Madrid
Pedro Martins — University of Coimbra
Penousal Machado — University of Coimbra, CISUC, DEI and LASI
Rafael Pérez Y Pérez — Universidad Autónoma Metropolitana at Cuajimalpa
Rob Saunders — Leiden University
Simon Colton — Queen Mary University of London
Tony Veale — University College Dublin

PC Members

Alena Denisova — University of York
Anna Jordanous — University of Kent
Antonios Liapis — University of Malta
Bob Sturm — KTH Stockholm
Colin G. Johnson — University of Nottingham
Enric Plaza — IIIA-CSIC
Fabrício Góes — University of Leicester
Filipe Calegario — Universidade Federal de Pernambuco
Filippo Carnovalini — Vrije Universiteit Brussel
Giovanni Pilato — ICAR-CNR
Ivan Guerrero — UNAM
Jason M. Reizner — IU International University of Applied Sciences
Jeba Rezwana — Towson University
Jer Hayes — Accenture
Jéssica Parente — University of Coimbra
Jon McCormack — Monash University
Juliana Shihadeh — Santa Clara University
Kıvanç Tatar — Chalmers University of Technology
Lakshmi Nair — Georgia Institute of Technology
Liane Gabora — University of British Columbia
Lucas N. Ferreira — Universidade Federal de Minas Gerais
Ludovica Schaerf — University of Zurich / Max Plank Society
Luís Espírito Santo — University of Coimbra / Vrije Universiteit Brussel
Lydia Chilton — Columbia University
Manex Agirrezabal — University of Copenhagen
Marcio Lima Inácio — University of Coimbra
Mark Sheppard — University of Kent
Mark Granroth-Wilding — AMD Silo AI
Matthew Guzdial — University of Alberta
Max Peeperkorn — University of Kent
Maximos Kaliakatsos-Papakostas — Aristotle University of Thessaloniki
Mei Si — Rensselaer Polytechnic Institute (RPI)
Mena Leemhuis — Free University of Bozen-Bolzano/Johannes Kepler University Linz
Solange Margarido — University of Coimbra
Michael Muller — Independent Researcher
Miguel Carvalhais — i2ADS / Universidade do Porto
Mihai Pomarlan — University of Bremen
Mohamad Elzohbi — University of Calgary

Nick Harley — Vrije Universiteit Brussel
Olga Vechtomova — University of Waterloo
Ori Amir — Fulbright University Vietnam
Paul Bodily — Idaho State University
Petra Jääskeläinen — KTH Royal Institute of Technology
Phil Lopes — Lusofona University
Piera Riccio — University of Amsterdam / ELLIS Alicante
Piotr Sawicki — University of Kent
Piotr Mirowski — DeepMind
Prashanth Thattai Ravikumar — Goldsmiths, University of London
Rafael Peñaloza — University of Milano-Bicocca
Richard Zhao — University of Calgary
Santiago Negrete — Universidad Autónoma Metropolitana
Sérgio Rebelo — University of Coimbra
Severi Uusitalo — Aalto University
Špela Vintar — University of Ljubljana / Jožef Stefan Institute
Simo Linkola — GIM Robotics Oy
Tapio Takala — Aalto University
Terence Broad — University of the Arts London
Thomas Winters — KU Leuven
Tiago Martins — University of Coimbra
Wendy Aguilar — UNAM
Xuechen Liu — Chalmers University of Technology
Ying-Ping Chen — National Yang Ming Chiao Tung University

Additional Reviewers

Anna Haverinen
Antti Salovaara
Ayon Mazumder
Giancarlo Colloca
Iikka Hauhio
Joana Rovira Martins
Konstantinos Soiledis
Konstantinos-Theodoros Tsamis
Lena Hegemann
Luca Resti
Nam Hee Kim
Niki Pennanen
Roberto Gallotta
Tero Jokela

Contents

Full Papers **1**

Measuring analogical creativity - An experiment
Stergos Afantenos, Leonardo Cortez Bernardes and Henri Prade . . 2

Generominator: Operationalizing the Computational Creativity Ideation Cards "Generonimos"
Gabriel Bacon, Rosalind Dixon, Max Kluiwstra-Descalle, Audrey Ostrom, Alexander Sherbrooke, Thomas Wessel, Jiangyue Wu and Markus Eger 13

Creative Transformation in Literary Texts: Modelling Change Across Representational Levels
Ioana-Roxana Boriceanu and Liviu P. Dinu 23

Who Can Scan a Verse? Evaluating Open and Closed Language Models for Portuguese Poetic Scansion
Filipe Calegario and André Valença 33

HumorSkills: Cognitive, Social and Creative Skills for AI Humor Generation
Lydia Chilton and Sean Kim 44

Experiments in Text-2-Tool-Use for Relatable Computational Creativity
Simon Colton 55

Reading Between the Lines: Can LLMs Discover the Question Behind the Text?
Claudiu Creanga and Liviu P. Dinu 65

Assistance, Not Usurpation: LLMs as Catalysts for Human Originality
Roberto Gallotta, Antonios Liapis and Georgios N. Yannakakis . . . 74

Artists' Strategies to Circumvent Representational Biases and Censorship in Text-to-Image Generation
Mijntje de Groote and Piera Riccio 85

Digital Weight: Towards an Ecological Evaluation Layer for Generative AI in Computational Creativity
Jekaterina Kaizere 96

Evolving to the Aesthetics of a Vision-Language Model
Stephen James Krol and Jon McCormack 104

Seeing Differently: Modeling Interpretive Perspectives in Computational Creativity using a Four-World Framework
Prerna Luthra 114

God, Servant, Master or Friend? How AI metaphors shape the way Creative AI is built and used
Alison Pease, Maya Ackerman and Dave Murray-Rust 124

World-Schema Alignment: Conditioning Data as a Compositional Strategy in Generative Audio
Frederick Rodrigues 135
Lessons Learned from Organizing a Literary Competition of Short Stories Co-created by Humans and AI Tools
Rudolf Rosa, Michaela Liegertová and Miloslav Linc 144
Conch(ious): A Tangible Generative AI Interface for Embodied Speculative Futures
Yaakulya Sabbani, Amina Magomedova, Ahmad Dahlan Hafizh and Nimrah Syed 159
Theory as Architecture: Why Aesthetic Frameworks Should Constitute, Not Merely Evaluate, AI Art Systems
Atticus Sims 169
What makes LLM-generated Answers Valuable for Research Problem Reframing?
Abhinav Sood, Stephen Wan, Cecile Paris and Kazjon Grace 179
Remember Me: Enhancing Long-term Character Persistence in LLM-Generated Stories using Knowledge Graphs
Alvin Tan, Jay Campanell, Richa Misra, Jessica Fu, Rida Faraz, Joel Walsh and Naina Panjwani 189
JEST: A Benchmark for Rating the Funniness of Short Texts
Joe Toplyn and Ori Amir 199
Generative Rules as Design Material: Constitutive Agency in Co-Creative Evolutionary Knit Design
Severi Uusitalo and Anna-Mari Leppisaari 209
IVIE: A Neuro-symbolic Approach to Incremental and Validated Generation of Interactive Fiction Worlds
Micaela Vaucher, Santiago Silveira, Santiago Góngora and Luis Chiruzzo 219
Snap, Crackle and ... Slop? Imbuing Generative AI with the Essence of Computational Creativity
Tony Veale 229
Activation Steering for Creative Estrangement: Generating Evocative Narrative Through Controlled Disruption of Language Models
Olga Vechtomova, Greg French, Xavier Schneider, Olivier Poulin and Jeff Bos 239
A Process-centric Analysis of Modern Generative AI As Computational Creativity
Dan Ventura 253
Who's driving? Exposing human self-deception in creative human-AI collaboration
Špela Vintar, Mojca Brglez, Damjan Popič and Jan Jona Javoršek 264

What Makes Literary Quotes Stand Out? A Multidimensional Analysis of Literary Creativity
Berke Yazan, Almila Akdag and Davide Dell'Anna 274

Short Papers **285**

Coherence-Oriented Dream Scene Visualisation
Azra Acil and Simon Colton 286

How are scientific concepts birthed? Typing rules of concept formation in theoretical physics reasoning
Omar Aguilar and Anthony Aguirre 294

From 31 Functions to 14 Categories: Grouping Propp's Narrative Structure for Co-Creative Writing Tools
Aman Alapati, Alaa Mohasseb, Ella Haig and Argenis Ramirez Gomez 299

Recipes for Creativity: Iterative Generation and Evaluation in Large Language Models
Rens Anderson, Tessa Verhoef and Amirhossein Zohrehvand 306

Creating and Evaluating Challenge in Generative Music Education Systems
Filippo Carnovalini and Geraint Wiggins 313

KYSS: A Framework for Evaluating Self-Knowledge Support in Human-AI Co-Creation
Giancarlo Colloca and Jeba Rezwana 318

Twenty-Five Years of The Painting Fool Project
Simon Colton and Blanca Perez Ferrer 324

Making Sense of "Cozy Chaos": Humans and LLMs Differences in Linguistic Creativity
Anca Dinu, Andra-Maria Florescu and Alina Resceanu 331

Back to Back with a Copy: A Computational Analysis of AI-Generated Visual Contemporary Art Pastiches
Anca Dinu, Andreiana Mihail, Andra-Maria Florescu, Claudiu Creangă and Liviu P. Dinu 340

Spot the Pipeline: Evaluating Perceptual Convergence Among Text-to-Image Tools
Martin Disley 349

In-Person First, Not In-Person Only: Reflections on Remote Participation through the case of ICCC'25 Watch Party
Luís Espírito-Santo 354

The Flow of LLM-Generated Poetry
Patrícia Ferreira, Catarina Silva, Ana Alves and Hugo Gonçalo Oliveira 361

International Conference on Computational Creativity: What If the Emperor's New Clothes Are A Straitjacket?
Pablo Gervás, Carlos León and Dan Ventura 370
World-State Transformations for Neuro-symbolic Interactive Storytelling
Santiago Góngora, Luis Chiruzzo, Gonzalo Méndez and Pablo Gervás 374
Cross-Modal Associations in Creativity: Evaluating Word-Image Pairing in Humans and Large Language Models
Eva Illion, Linxuan Wang and Liane Gabora 383
Gamification of Control in Co-Creativity through Design Cards
Alayt Issak, Jeba Rezwana and Casper Harteveld 389
Importing TRIZ into Narratology for LLM-Assisted Story
Milka Kaplan, Armin Shmilovici and Mark Last 398
The Affordance is the Message: Creative Media as Complex Systems
Ane Kristine Espeseth and Elias Najarro 407
Folded Bodies: Topological Control of an Untrained Spiking Instrument
Lucas Kuzma 412
Anchor-Conditioned Compositional Control for Landscape Image Generation
Gadha Lekshmi P, Govind Arun, Rohith Syam and Ahmed Elgammal 418
Transformation Is All You Need
Sean Luke and Matthew Cronin 423
Does AI Herald the End of Creative Coding?
Jon McCormack and Stephen James Krol 428
Co-Creative Vulnerability: Risks of Engagement Optimization in Creative AI
James Morgan and Margareta Ackerman 432
A Training-Free Novelty Detection Framework for Large-Scale Multi-Agent Social Creativity Simulations
Luuk Motz, Tessa Verhoef and Rob Saunders 437
Expressing our Authentic Voices within Co-Creative AI Partnerships: Insights from Augmentative and Alternative Communication
Alison Pease and Tom Griffiths 446
Before the Prompt: Heteromorphic Imagination and the Cognitive Cost of Premature Crystallization in Human-AI Co-Creativity
Emil Polyak and Rghad Balkhyoor 451
Harnessing Quantitative Creativity to Explore the Uncanny Valley of Synthetic Faces
Aditi Ramaswamy and Hana Chockler 456
I Was Scrolling and Then I Saw a Pregnant Strawberry
Piera Riccio 461

IRIDyOM: Exposing Musical Expectation as an Interactive Creative Space
Lluc Bono Rosselló, Julien Baudru and Hugues Bersini 466
An Initial Design for Generative Game Mechanics via LLM Function Emulation
Brad Spendlove and David Kline 472
ANTLER 1: A Multimodal Co-Creative Tutor for Stylistic Exploration in Tambour Embroidery
Anne Toomey, Zowie Broach, Mojtaba Kargar, Samira Valorian, Henrietta Dent, Aemika Zou and Chaonan Lin 478
Conceptualising Creative Autonomy in People, AI, and Their Interaction: A Scoping Review Protocol
Maija Välimäki and Christian Guckelsberger 483
Sketch-It: A Consent-First Creativity Support Tool for the Generation of Inspirational Sketches
Claudia Pàmies Vaqué, Garrit Schaap and Maria M. Hedblom . . . 490
Side-Splitter: A Visual Laboratory for Jokes with Verbal Humour and Physical Slapstick
Tony Veale 498
The Parallax View: An objective treatment of humorous subjectivity in joke generation
Tony Veale 503
Is It Happening? Lyotard's Sublime and the Indeterminate in AI-Generated Text
Olga Vechtomova 508
Introducing the Concept Collider: A White-Box Computational Framework for Targeted Creative Emergence
Sebastian Wahl 513
How Context Injection Shapes Creativity in LLM-Generated Alternative Uses
Theo Wong, Fabricio Goes, Marco Volpe and Paulo Mann 520

Full Papers

Measuring analogical creativity - An experiment

Stergos Afantenos, Leonardo Cortez Bernardes, Henri Prade
Institut de Recherche en Informatique de Toulouse (IRIT), Univ. Toulouse, Toulouse INP, CNRS
stergos.afantenos@irit.fr, bernardesleonardo0@gmail.com, henri.prade@irit.fr

Abstract

Analogical reasoning is said to be creative. After discussing views of computational creativity and its evaluation, we consider a particular form of analogical inference based on analogical proportions, which are quaternary relations comparing the differences inside two pairs of items. This inference mechanism is capable of generating a new item from three items described on the same set of features. This mechanism is used, in an experiment, to generate new prompts that are addressed to an LLM-based chatbot to generate constrained text. The prompts specify the emotions at work in a short text describing a mundane situation. Hellinger distance and Kullback-Leibler divergence are used for provisional assessments of creativity at the prompt level or the text level. The relative merits of these two measures of creativity assessment are discussed. The paper proposes an original creative mechanism coupling analogical inference and LLM-based text generation.

Introduction

Creativity refers to the ability to develop novel, and preferably valuable, ideas or productions. This clearly applies to a wide range of works. Human creativity ranges from arts, literature and humour to problem solving which itself covers a variety of situations encountered in mathematics and sciences as well as in problems of everyday life. Creativity as a major asset of the human mind, has been a challenging issue in artificial intelligence research for a long time (Boden 1994; McCormack and d'Inverno 2012; Colton and Wiggins 2012), with illustrations in particular in music (Wiggins and Forth 2018), in the visual arts (Cook and Colton 2011), or in literature (Van de Cruys 2020).

Drawing analogies has a reputation for being creative, especially in problem solving. This has been advocated by many authors, in cognitive psychology, e.g., (Dreistadt 1969; Halpern 1986), in artificial intelligence (Veloso 1993), or in mathematics (Polya 1954). Creativity seems all the more evident and striking as the analogy brings together two distant situations, or two quite different frames of reference.

Analogical reasoning can take various forms according to different modeling frameworks.[1] In structure mapping theory and related approaches (Gentner 1983; Falkenhainer, Forbus, and Gentner 1989; Hummel and Holyoak 1997), analogical reasoning is considered to be the ability to draw parallels between two universes (or, more simply, two situations) by matching the relationships between elements of the first universe (and their properties) to those applicable to elements of the second universe, and to draw potential conclusions in the second universe based on the similarity of relationships in the two universes (the second universe is supposed here to be less known than the first). This kind of analogical reasoning can be encoded in a classical logical framework (requiring second-order anti-unification) (Gust, Kühnberger, and Schmid 2006) where one can, for example, capture Rutherford's famous creative analogy, which compares the solar system (playing the role of the first universe) to Rutherford's atomic model (the second universe) (Schmidt et al. 2014).

Case-based reasoning (Kolodner 1993; Aamodt and Plaza 1994; Fuchs et al. 2020) is often considered as a simple form of analogical reasoning which uses a repertory of cases, made of pairs (problem, solution), and proposes a solution for a target problem by adapting the solution of a similar problem found in the repertory of cases. However, the creativity of case-based reasoning remains relatively modest, to the extent that the problems being compared are similar.[2] However, the creativity of case-based reasoning remains rather modest inasmuch the problems brought together are similar. A parallel is thus established between the pair (problem, solution) used and the target problem with its proposed solution. One can say that the proposed solution is to the target problem as the solution of the similar case is to the problem of this case.

The above statement that links four items is an analogical proportion and is of the form "a is to b as c is to d". In fact when one parallels a pair (a, b) in a first universe with the corresponding pair (c, d) in the second universe, one obtains

[1] Apart from the approaches reviewed in this introduction, we can also mention the iconic approach of analogical reasoning (Kunda, McGreggor, and Goel 2013) which has been applied to the solving Raven's Progressive Matrix tests, a problem which can also be dealt with an analogical proportion-based approach (Afantenos et al. 2024; Correa Beltran, Prade, and Richard 2016)

[2] Several works rely on case-based reasoning that have improved upon ideas on creativity that appear in (Schank and Kass 1990); let us mention, for example, (Ramalho and Ganascia 1994), (de Mántaras 1998) in the field of music creation.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

such proportions, as pointed out in (Hesse 1959). For instance, "the electron is to the nucleus as the planet is to the sun" in Rutherford's analogy. In such an example, the analogical proportion involves two (quite different) universes; in case-based reasoning, one has another type of split between problems and solutions.

There is a third type of analogical reasoning that is directly based on analogical proportions, where knowing that "a is to b as c is to d", one infers d (or an unknown part of the description of d) from the triplet (a, b, c). It requires that the four items be of the same kind in the sense that they can be described on the same set of attributes, which is not necessarily the case in the two previous approaches. It is worth pointing out (Barbot, Miclet, and Prade 2019) that analogical proportions may either involve four items describable by the same set of attributes, as in "a puppy is to a dog what a kitten is to a cat", or associate two universes (that of a and c, and that of b and d) which refer to two distinct sets of attributes, as in "wine is to the French as beer is to the English". In this paper, we only focus on the first kind of analogical proportions.

Analogical proportions appear to be creative in the sense that a, b, c and d can be all different, and where d — which can be obtained from a, b and c — is described by a recombination of the features of a, b and c. However the question of evaluating such a creativity is raised. In this paper, we propose an experiment that illustrates this creative process in the area of text generation, and try to measure the novelty of what is produced.

This paper investigates creativity using analogical proportion-based inference. It is organized into three main parts. The first part reviews different aspects of creativity and ways to assess it. It also considers means of measuring creativity that have been proposed, such as Hellinger distance or Kullback-Leibler divergence, and discusses their relative merits. The second main part provides a background on analogical proportions and analogical inference, and points out their potential for creativity, which is theoretically evaluated. The third main part describes the setting of an experiment: the analogical proportion-based inference is used to generate combinations of features pertaining to a short text that reports a sequence of ordinary facts. These combinations of features are then used as prompts that are addressed to an LLM-based chatbot to generate constrained texts. Thus the creativity can be both measured at the prompt level and at the text level. Then the results of the experiment are reported, which is followed by a final discussion.

Creativity and its evaluation

In this section, we first review the work on computational creativity aimed at formalizing or assessing it. We then recall two statistical divergence measures proposed to assess novelty, a major aspect of creativity.

Review of works on computational creativity

Computational creativity refers to AI systems capable of imitating human creative processes. This brings us back to human creativity. It is why it may be useful to start with the definition of (human) creativity given in a highly-cited paper (Rhodes 1961): "The word creativity is a noun naming the phenomenon in which a person communicates a new concept (which is the product). Mental activity (or mental process) is implicit in the definition, and of course no one could conceive of a person living or operating in a vacuum, so the term *press* is implicit. The definition begs the questions as how new the concept must be and to whom it must be new." This first proposed view is known for introducing four "p": person, product, process and press, involved in any creative processes.

There has been an active research trend in computational creativity for more than thirty years. Broadly speaking, in addition to work on specific computer programs capable of producing new works in the fields of art, literature or mathematics, a series of works have attempted to define, categorize and formalize computer creativity in general, or to propose general methods for assessing creativity.

This started with the influential work of Boden (Boden 1994; 2003), who distinguished between three kinds of creativity: combinational, exploratory, and transformational. In combinational creativity, the result is obtained by a mere combination of standard ideas, while the two other forms of creativity take place in a conceptual space, the first one explores this space, while the second transforms it (e.g., by relaxing some of the constraints defining the conceptual space). Transformational creativity is regarded as the highest form of creativity which goes beyond mere novelty.

Bundy (1994) points out that according to Boden's view (Boden 2003), an automated theorem prover solving an open conjecture in mathematics would not be considered truly creative, the new result being obtained by an exploratory procedure. In contrast, when the open conjecture is solved by a human, it is likely that the human work will be considered creative. Bundy emphasized the need to take into account the possible great complexity of exploring the conceptual space. Let's add that in mathematics, apart from the difficulty of the result to be established, an innovative proof based on a new way of looking at the problem can be considered an example of transformational creativity.

Besides, Boden (2003) distinguishes between p-creativity ("p" for psychological) and h-creativity ("h" for historical) where the former is judged in relation to the agent's productions, while the result of the latter creativity is regarded as new by society.

Boden's inspiring work has led some researchers to attempt to make more formal and refine her definitions of a conceptual space or transformational creativity, see, e.g., (Ritchie 2006) for a discussion. One noticeable line of research is the development of a 'Formal Creativity Theory' by Wiggins (Wiggins 2006a; 2006b; 2019), where the conceptual space is understood as a set of possible artefacts associated with a set of rules that select those that are acceptable (just playing the role of a quality checking), and where another set of rules attribute a value to a created artefact (Wiggins 2006b).

Wiggins views transformational creativity as exploratory creativity at the meta-level, taking up an idea defended in (Buchanan 2001). Creativity is viewed as a generalized

search process i) without closed-world assumption, ii) allowing for the combination of existing ideas, iii) distinguishing between heuristics that evaluate the quality of solutions and those that are used for finding a solution efficiently, and iv) preferring the notion of comparison to that of value. This view of creativity as search has been further discussed and restated in greater details in (Ritchie 2012). Recently, the setting of Formal Creativity Theory has been used for arguing that "while novelty is neither necessary nor sufficient for transformational creativity in general, when using an inspiring set, rather than a sequence of experiences, an agent actually requires novelty for transformational creativity to occur" (Santo, Wiggins, and Cardoso 2024).

Another line of research inspired by Boden's work, is the approach, more pragmatic, proposed by Ritchie (2001; 2007), which proposes a series of criteria for attributing creativity to a computer program, based on three indicators. Namely, in addition to the two recognized indicators for judging creativity that are *novelty* (the produced item should be different of what is known) and *quality* (the produced item must have value in its genre), Ritchie adds *typicality* to assess creativity (the produced item should be representative of its class). Moreover, the creative program may use an *inspiring set* that it is designed to recreate differently (as far as possible). The high-valued output artefacts of a creative program, which are not already in the inspiring set, constitute the *creative set* (Colton, Pease, and Ritchie 2001).

Maher (2010) advocates that evaluating creativity is a combination of the evaluation of three essential criteria for creativity: novelty, value, and *unexpectedness* (or surprise), where "unexpectedness is measured by how far the artifact is from the expected next artifact" (taking into account the sequential nature of expectation).

Schmidhuber (2010) takes a (reinforcement) learning perspective for modeling creative processes, where "the creativity reward of the action selector is the current surprise or novelty measured by the improvements of the world model". The idea is to use reinforcement learning for "creating / discovering / learning novel patterns in the growing history of actions and sensory inputs, where the theory formally specifies what exactly is a pattern, what exactly is novel or surprising, and what it means to incrementally learn novel skills leading to more novel patterns".

This is not to be confused with the use of modern machine learning tools for computational creativity, where, for instance, surprising images can be generated by a text-to-image model from prompts. In such a case, the creativity part relies in the (usually human-originated) prompt that specifies a requirement for an unusual *combination* of elements which are not encountered together in the real world.

Lastly let us mention some general ideas commonly associated with creativity. Creativity is often the result of a conceptual blending (Turner and Fauconnier 2003), which corresponds to the association of two different frames of reference that can be joined, connected together. Analogical reasoning, especially when making a parallel between two distinct universes through a structure mapping is often considered as a form of conceptual blending.

Lateral thinking (de Bono 1971), in problem solving, amounts to move from a familiar model to a new, unexpected one. This is close to the idea of "thinking out of the box", which itself sounds as a transformational operation for enlarging, modifying the conceptual space.

Potential measures of novelty

The previous survey has shown that everybody agrees with the idea that novelty is a central criterion in the evaluation of creativity. The quality, the value of the items produced seem more easy to maintain through constraints by the computer program. In this subsection we review two measures that have been used in practice for assessing novelty: the Hellinger distance and the Kullback-Leibler divergence.

These measures compare two probability distributions. Applied to creativity, one may consider that the first distribution describes the population of the items in the inspiring set and the second one the items produced by the system. These measures evaluate to what extent are the produced items dissimilar to the existing examples we start with.

Hellinger distance This distance is used in the assessment of creativity, e.g., in (Fele et al. 2022). It was first introduced in 1909 by Hellinger, a German mathematician. Given two discrete probability distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$ over the same space $X = \{x_1, \ldots, x_n\}$ (where $p_i = P(\{x_i\}), \sum_{i=1,n} p_i = 1 = \sum_{i=1,n} q_i$), their Hellinger distance is defined as

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1,n} (\sqrt{p_i} - \sqrt{q_i})^2}$$

By an easy calculation, it follows

$$H^2(P,Q) = 1 - \sum_{i=1,n} \sqrt{p_i q_i}$$

As can be seen the Hellinger distance is related to the Euclidean distance. We have $0 \leq H(P,Q) \leq 1$. $H(P,Q) = 0$ if and only if $p_i = q_i$ for all i. $H(P,Q) = 1$ if and only if P and Q are mutually singular, i.e., $\forall i, p_i > 0 \Rightarrow q_i = 0$ and conversely $\forall i, q_i > 0 \Rightarrow p_i = 0$.

The quantity $\sum_{i=1,n} \sqrt{p_i q_i}$ is known as the Bhattacharyya coefficient, a well-known measure of similarity between two probability distributions P and Q. So $H(P,Q)$ is a measure of dissimilarity between these distributions.

Kullback-Leibler divergence This divergence (Kullback and Leibler 1951) is a measure of relative entropy that evaluates how much the probability distribution Q is different from the probability distribution P. It is defined by

$$KL(P||Q) = \sum_{i=1,n} p_i \ln \frac{p_i}{q_i} = \sum_{i=1,n} (p_i \ln p_i - p_i \ln q_i)$$

As a natural generalization of the Shannon entropy this measure has been characterized in different ways and applied in different fields. It is positive : $KL(P||Q) \geq 0$, the divergence is non-symmetric $KL(P||Q) \neq KL(Q||P)$, so it is not a distance. The classical use of this divergence is to judge to what extent Q is an acceptable approximation of P, in that perspective, the smaller the divergence, the better, the

more similar P and Q since $KL(Q||P) = 0$ if and only if $p_i = q_i$ for all i.

Besides, $KL(P||Q)$ is finite only if the support of P is contained within the support of Q ($\forall i, p_i > 0 \Rightarrow q_i > 0$; but if we have $\exists i, p_i > 0$ and $q_i = 0$, then $KL(P||Q)$ will be infinite. Thus, when $KL(P||Q) > 0$ (and not close to 0), it indicates that Q is dissimilar with respect to P.

The Kullback-Leibler divergence has been advocated as a measure of "Bayesian surprise" (Itti and Baldi 2009). It has been used to detect surprise in short narratives, showing that a change in the model's prediction corresponds with a significant change in the participants' rating (Chieppe,Sweetser, and Newman 2022). Moreover, the use of Kullback-Leibler divergence has enabled the retrieval of key authors and works that shape the broad outlines of literary change in a vast corpus of French novels from the XIXth and early XXth centuries (Barré and Poibeau 2023). The use of this divergence seems appropriate for assessing novelty.

Lastly, some authors have proposed to use algorithmic complexity as a basis for attributing creativity (McGregor 2007; Mondol and Brown 2021; Brown and Ventura 2022). Conditional Kolmogorov complexity is proposed for evaluating novelty. For value, the idea is that an artefact is of high quality if it represents the output of a significant amount of computational effort. Although theoretically appealing, the approach is not so easy to use in practice, even if approximate evaluations could be obtained in principle. We have therefore not retained this option for our experiment, in spite of the fact that analogical proportions can be also thought in terms of Kolmogorov complexity (Bayoudh, Prade, and Richard 2012; Prade and Richard 2018).

Analogical inference and creativity

We first recall what are analogical proportions and what is their Boolean modeling, before explaining how they can be used for inference and for creative purposes. We end this section with a theoretical study of their creative power.

Analogical proportions

Analogical proportions are statements linking four entities of the same kind. They are of the form "a is to b as c is to d". In principle, these entities can be anything: numbers, word, images, sentences, ... Analogical proportions appeared for the first time (at least in Occident) with Aristotle, who takes lessons from works by mathematicians of the time just before him on arithmetic proportions ($a - b = c - d$) and geometric proportions ($\frac{a}{b} = \frac{c}{d}$). These numerical proportions equalize differences and ratios respectively. Aristotle extended such a pattern to words (or concepts).

As can be seen, such proportions operate both a comparison inside pairs, and a comparison between pairs. This is precisely what state "a is to b as c is to d". Such a proportion is denoted by $a : b :: c : d$. As suggested by the parallel with numerical proportions, analogical proportions are supposed to obey three postulates:

- $a : b :: a : b$ (*reflexivity*)
- $a : b :: c : d \Rightarrow c : d :: a : b$ (*symmetry*)

a	b	c	d
0	0	0	0
1	1	1	1
0	0	1	1
1	1	0	0
0	1	0	1
1	0	1	0

Table 1: Truth table of $a : b :: c : d$

- $a : b :: c : d \Rightarrow a : c :: b : d$ (*central permutation*)

While the two first postulates are not surprizing, the stability under central permutation seems unexpected. However, if we consider the following example "puppies are to dogs as kittens are to cats", we should admit that one may also state that "puppies are to kittens as dogs are to cats".

The application of the symmetry and central permutation (in sequence), shows that an analogical proportion satisfies

- $a : a :: b : b$ (sameness) ;
- $a : b :: c : d \Rightarrow d : b :: c : a$ (external permutation);
- $a : b :: c : d \Rightarrow b : a :: d : c$ (internal reversal);
- $a : b :: c : d \Rightarrow d : c :: b : a$ (complete reversal).

One can easily check that the above properties make sense on the "cats / dogs" example.

The entities we consider are vectors of Boolean values. Thus, any entity is represented by a vector $\vec{a} = (a_1, ..., a_n)$ where a_i is the Boolean value of a feature / attribute i. The analogical proportions between vectors (of the same size) is defined componentwise, namely

$\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ if and only if $\forall i \in \{1, ..., n\}, a_i : b_i :: c_i : d_i$.

We first deal with one component and four Boolean variables representing the value of some feature of interest for four distinct items.

The reflexivity postulate $a : b :: a : b$ forces a Boolean analogical proportion to be true for all a and all b, which entails that the four valuations $(0, 0, 0, 0)$, $(0, 1, 0, 1)$, $(1, 0, 1, 0)$, and $(1, 1, 1, 1)$ should correspond to true analogical proportions. Due to the central permutation postulate, the minimal Boolean model that satisfies the three postulates is true for the 6 following patterns for the Boolean variables a, b, c, d (and false for the $2^4 - 6 = 10$ other possible patterns) given in Table 1 (Prade and Richard 2018).

The following quaternary logical formula is true only for the 6 above patterns of values (and thus false for the 10 others) (Miclet and Prade 2009):

$$a : b :: c : d = ((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((\neg a \wedge b) \equiv (\neg c \wedge d)) \quad (1)$$

It precisely expresses that "a differs from b as c differs from d, and b differs from a as d differs from c" (and "when a and b do not differ, c and d do not differ").

Let us take the cat / dog example, "a puppy is to a dog as a kitten is to a cat" for illustrating the analogical proportion between vectors that represent the four animals on a set of 6 attributes, here `mammal`, `bird`, `canid`, `feline`, `young`, `adult`, as in Table 2.

As can be seen, the attributes in the example are such that each of them, vertically, illustrates one of the 6 patterns

	$mam.$	$bird$	$canid$	$felid$	$young$	$adult$
puppy	1	0	1	0	1	0
dog	1	0	1	0	0	1
kitten	1	0	0	1	1	0
cat	1	0	0	1	0	1

Table 2: A Boolean validation of $puppy : dog :: kitten : cat$

in Table 1. It validates the fact the analogical proportion $puppy : dog :: kitten : cat$ holds.

Analogical inference

Given a triplet $a, b, c \in \{0, 1\}^3$, one may look for $x \in \{0, 1\}$ such that $a : b :: c : x$ holds true. From truth Table 1, it can be checked that if a solution x exists, it is unique, but there are two cases where the equation has no solution. Indeed, there is no $x \in \{0, 1\}$ such that $1 : 0 :: 0 : x$ and $0 : 1 :: 1 : x$ are true analogical proportions. It can be checked that the Boolean analogical equation $a : b :: c : x$ is solvable if and only if $(a \equiv b) \vee (a \equiv c)$ is true, where $\equiv$ is the equivalence connective. In that case, the *unique* solution is given by $x = a \equiv (b \equiv c)$ (remember $\equiv$ is associative); thus $x = b$ if $a = c$ and $x = c$ if $a = b$.

Analogical proportion-based inference, when applied to classification, relies on a simple principle: if four (Boolean) vectors $\vec{a}$, $\vec{b}$, $\vec{c}$ and $\vec{d}$ make a valid analogical proportion component-wise for each 4-tuple of values pertaining to the same attribute, then it is expected that their class labels also make a valid proportion[3] (e.g., (Bounhas, Prade, and Richard 2017)).

$$\frac{\vec{a} : \vec{b} :: \vec{c} : \vec{d}}{cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : cl(\vec{d})} \quad (2)$$

If the class labels for vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$ are known (i.e., they belong to the sample set), the classification of a new Boolean vector $\vec{d}$ is possible using (2) only if i) the equation $cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : x$ is solvable, and ii) the analogical proportion $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ holds true on all components. If these two conditions are met, $cl(\vec{d})$ is the unique solution of $cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : x$. In practice there are usually several triplets $(\vec{a}, \vec{b}, \vec{c})$ leading to several solutions for $cl(\vec{d})$, which may be different. Then a majority vote decides the class assigned to $\vec{d}$. Such a procedure yields good results on benchmarks (Bounhas, Prade, and Richard 2017).

In (Afantenos et al. 2024), a slightly different inference pattern has been proposed for creativity purposes. From three items $\vec{a}$, $\vec{b}$, $\vec{c}$, one can build another item $\vec{d}$, provided that for each attribute i the Boolean analogical equation $a_i : b_i :: c_i : x_i$ is solvable, which requires that

$$(\vec{a} \equiv \vec{b}) \vee (\vec{a} \equiv \vec{c})$$

holds true component-wise. This can be summarized by the following inference pattern

$$\frac{\vec{a}, \vec{b}, \vec{c}}{\vec{d} = (\vec{a} \equiv \vec{b} \equiv \vec{c})} \quad (3)$$

It can be shown that $\vec{d}$ is different from $\vec{a}, \vec{b}, \vec{c}$, provided that $\vec{a} \neq \vec{b}$ and $\vec{a} \neq \vec{c}$. The next section studies how much different $\vec{d}$ can be from the three other vectors.

A theoretical evaluation of analogical proportion-based creativity

An analogical proportion $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ establishes a parallel between the pairs $(\vec{a}, \vec{b})$ and $(\vec{c}, \vec{d})$. The 6 patterns of Table 1 making an analogical proportion true enables us to partition the n attributes on which the items are described ($\vec{a} = (a_1, ..., a_n)$, $\vec{b} = (b_1, ..., b_n)$, $\vec{c} = (c_1, ..., c_n)$, $\vec{d} = (d_1, ..., d_n)$), into 3 subsets:

- i) the subset of attributes i where $a_i = b_i = c_i = d_i = 1$ or $a_i = b_i = c_i = d_i = 0$, i.e., where the four items $\vec{a}, \vec{b}, \vec{c}, \vec{d}$ are equal;
- ii) the subset of attributes j where $a_j = b_j$ and $c_j = d_j$ but $a_j \neq c_j$, i.e., where the items are equal within the pairs $(\vec{a}, \vec{b})$ and $(\vec{c}, \vec{d})$, but not in the same way;
- iii) the subset of attributes k where $a_k = c_k$ and $b_k = d_k$, i.e., the subset of attribute(s) whose value(s) change(s), in the same direction, from $\vec{a}$ to $\vec{b}$ and from $\vec{c}$ to $\vec{d}$.

This is pictured in Table 3.

items	All equal	Equal by pairs	Same Differences
$\vec{a}$	1 0	1 0	1 0
$\vec{b}$	1 0	1 0	0 1
$\vec{c}$	1 0	0 1	1 0
$\vec{d}$	1 0	0 1	0 1

Table 3: The 3 types of patterns of an analogical proportion

Note that the central permutation of $\vec{b}$ and $\vec{c}$ exchanges the last two columns in Table 3. It allows to state the result:

$$\vec{a} : \vec{b} :: \vec{c} : \vec{d} \Leftrightarrow \begin{cases} Equ(\vec{a}, \vec{b}) = Equ(\vec{c}, \vec{d}) \\ Dif^{10}(\vec{a}, \vec{b}) = Dif^{10}(\vec{c}, \vec{d}) \\ Dif^{01}(\vec{a}, \vec{b}) = Dif^{01}(\vec{c}, \vec{d}) \end{cases} \Leftrightarrow \begin{cases} Equ(\vec{a}, \vec{c}) = Equ(\vec{b}, \vec{d}) \\ Dif^{10}(\vec{a}, \vec{c}) = Dif^{10}(\vec{b}, \vec{d}) \\ Dif^{01}(\vec{a}, \vec{c}) = Dif^{01}(\vec{b}, \vec{d}) \end{cases}$$

where, given two vectors $\vec{x} = (x_1, ..., x_n)$, $\vec{y} = (y_1, ..., y_n)$,

$Equ(\vec{x}, \vec{y}) = \{i \mid x_i = y_i\} = Equ^0(\vec{x}, \vec{y}) \cup Equ^1(\vec{x}, \vec{y})$,

$Equ^0(\vec{x}, \vec{b}) = \{i \mid x_i = y_i = 0\}$,

$Equ^1(\vec{x}, \vec{y}) = \{i \mid x_i = y_i = 1\}$,

$Dif^{10}(\vec{x}, \vec{y}) = \{i \mid x_i = 1, y_i = 0\}$,

$Dif^{01}(\vec{x}, \vec{y}) = \{i \mid x_i = 0, y_i = 1\}$.

The Hamming distance between $\vec{a}$ and $\vec{b}$, $Ham(\vec{a}, \vec{b})$, is equal to the cardinality of $Dif^{10}(\vec{a}, \vec{b}) \cup Dif^{01}(\vec{a}, \vec{b}) = Dif(\vec{a}, \vec{b})$. We have $Ham(\vec{a}, \vec{b}) = |Dif(\vec{a}, \vec{b})| \triangleq e$. Note that $Ham(\vec{a}, \vec{b}) = Ham(\vec{c}, \vec{d})$. Let $r = |\{i \text{ s.t. } a_i = b_i = c_i = d_i = 1 \text{ or } a_i = b_i = c_i = d_i = 0\}|$. Thus $Ham(\vec{c}, \vec{d}) = e$; $Ham(\vec{b}, \vec{d}) = n - e - r$; $Ham(\vec{a}, \vec{d}) = n - r$.

[3] Similarly to Boolean attribute values, the only valid analogical proportions between two class labels, possibly identical, s and t, are $s : t :: s : t$ and $s : s :: t : t$, while $s : t :: t : x$ has no solution (except for for $s = t = x$).

If $e = 0$, $\vec{a} = \vec{b}$ and $\vec{c} = \vec{d}$; if $r + e = n$, $\vec{a} = \vec{c}$, $\vec{b} = \vec{d}$. In both cases, the proportion $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ is trivial: we get $\vec{a} : \vec{a} :: \vec{b} : \vec{b}$ or $\vec{a} : \vec{b} :: \vec{a} : \vec{b}$. If $r = n$, the 4 items are equal.

Thus $\vec{d}$ is all the more distant from $\vec{a}$ and $\vec{c}$ as r is small and e is large. But if $r = 0$, $e = n$, we get $Ham(\vec{b}, \vec{d}) = 0$. *So we have to equally (or almost equally, if n is odd) balance e and $n - e$, with $r = 0$, if we want to maximize the distance of $\vec{d}$ to $\{\vec{a}, \vec{b}, \vec{c}\}$* equal to $\min(e, n - e - r, n - r)$.

If n is even this maximal distance is just $n/2$. So the vector $\vec{d}$, entirely obtained from $\vec{a}, \vec{b}, \vec{c}$ can be very different from these three items, thus quite novel with respect to them.

Thus the analogical inference (3) appears to be a controlled manner to explore a universe of potential possibilities from existing options. This is better than a random selection (as early proposed in (Kaufmann 1969)).

Framework of the experiment

In the following we report an experiment illustrating the effectiveness of the creative power of analogical proportion-based inference. This is a two-step process: first prompts are generated, and then texts are produced. The novelty is checked both at the prompt level and at the text level. We first present our source of inspiration for the experiment, before explaining how analogical proportion-based inference is used for creating prompts, in turn used for generating texts.

The source of inspiration The source of inspiration is a well-known book by Queneau, *The Exercises in Style* (Queneau 1947), where the author rewrites the same story 99 times, a short text where the narrator gets on a bus, witnesses a dispute between a man and another passenger, and then sees this person again two hours later getting advice on adding a button to his overcoat. Each time the story is written in a new style. This can be viewed as an example of (repeated) exploratory creativity in Boden's sense.

Queneau is also famous for another book (Queneau 1961) published later, that exemplifies combinational creativity. This book is based on ten sonnets (a poem of 14 verses) where each line is printed on a separate strip of thick paper. The ten sonnets have the same rhyme scheme and the same rhyme sounds. Thus any line from a sonnet can be combined with a line from the nine others, allowing for 10^{14} (= 100,000,000,000,000) different poems. This then led Queneau, together with a mathematician, Le Lionnais, to start the "Ouvroir de littérature potentielle" (Oulipo) gathering a group of people interested in experimental literature.

Creating prompts Queneau, in his book *The Exercises in Style* (Queneau 1947) varies the style each time he rewrites the story, making use of a wide variety of rhetorical devices. For our experiment, we transpose the idea, by choosing a limited number of features that can apply to a text. We use 10 Boolean features corresponding to the presence or absence of the following 10 emotions: joy, fear, anger, tranquility, nostalgia, shame, confusion, curiosity, love, and resignation.[4] This has the advantage of providing vectors that: i) can be processed by analogical inference; and ii) can be associated with prompts sent to an LLM-based chatbot able to generate a constrained text.

For building an inspiring set of 100 vectors (with 10 components) in our experiment, we start with an initial set of 11 vectors: these vectors are the vector $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and the 10 vectors with a single 1 (and 9 '0') corresponding to each feature. There are $2^{10} = 1024$ distinct Boolean vectors with 10 components. In order to obtain the $100 - 11 = 89$ missing vectors of the inspiring set, we proceeded as follows. To a set of vectors we can associate a probability distribution describing the probability of obtaining any combination of feature values using this set as a sample set. For completing the inspiring set we add vectors one by one to the current set. The principle is to use the probability distribution of the current set and to add a vector (not yet in the set) such that the Shannon entropy of the resulting set is minimized. This procedure stops when the current set includes 100 vectors. This is our inspiring set. Let us call it $\mathcal{I}$. Note that $\mathcal{I}$ is rather "homogeneous" in the sense that $\mathcal{I}$ gathers vectors with a small number of '1'.

From this set $\mathcal{I}$ we generate an output set $\mathcal{O}$ by means of analogical inference. In order to do so, we use triplets of vectors $(\vec{a}, \vec{b}, \vec{c})$ in $\mathcal{I}^3$ in order to apply the pattern of inference (3). We apply a "semi-guided" procedure by selecting the candidate pair $(\vec{a}, \vec{b})$ then used to form the triplets, as explained in the next paragraph. Then we generate all the possible outputs $\vec{d}$ that can be obtained from a selected pair and a compatible vector $\vec{c}$ (i.e., a vector $\vec{c}$ for which the proportion equation $\vec{a} : \vec{b} : \vec{c} : x$ has a solution), by applying the analogical inference (3). This forms $\mathcal{O}$.

The selection of pairs $(\vec{a}, \vec{b})$ makes sure that the number of '1"s in $\vec{b}$ increases with respect to $\vec{a}$: this means that $\forall i, a_i = 1 \Rightarrow b_i = 1$ and $\exists j, a_j = 0$ and $b_j = 1$. The idea is that if when going from $\vec{a}$ to $\vec{b}$ a property is gained, it will also be gained when going from $\vec{c}$ to $\vec{d}$ (since 01 can be only completed by 01 for making a valid proportion). Moreover the pattern 1010 corresponding for $\vec{b}$ and $\vec{d}$ to the repeated loss of a property, is then forbidden. Still it does not mean that $\vec{d}$ inherits all the properties of $\vec{b}$ since 11 can be completed by 11 but also by 00 for obtaining a valid analogical proportion pattern. This last point shows that the analogical mechanism is not a pure property enhancement procedure, but it also manages to build a renewed set of outputs from the inspiring set, even if we privilege particular pairs and we forbid the use of pattern 1010 in the inference. Altogether, we favor a collection of outputs made of new enriched vectors, thus distinct and independent from the initial collection.

Starting with a set $\mathcal{I}$ of size 100, the analogical inference yields a set $\mathcal{O}$ of size approximately 150. Each of these two

[4] This choice may look a bit arbitrary, even if it substantially overlaps the 16 basic emotions advocated in (Ekman 1999). Note that here it is not so much a matter of emotions experienced by people, but just a text that is expected to communicate some feelings to the reader. The number (10) was chosen as being large enough for having variety, but not too large for limiting computation cost.

sets can be associated with a probability distribution. We are then in position to compute the dissimilarity or the divergence between the distributions $p_{\mathcal{I}}$ and $p_{\mathcal{O}}$ associated with $\mathcal{I}$ and $\mathcal{O}$ respectively.

Exploiting prompts The vectors in the sets $\mathcal{I}$ and $\mathcal{O}$ can be used as a basis for building prompts to be processed by a Large Language Model (LLM) for generating a text associated with each vector. Due to the one-to-one correspondence between the elements of the vectors and the presence or absence of an emotion, prompt construction is carried out by pre-defining a text containing ten variables to be instantiated. These variables are replaced by the strings "`true`" or "`false`" based on the binary value of the corresponding element. An example of a prompt is presented on page 7. Submitting the prompt to a language model subsequently generates a text associated with each vector.

The texts produced can be lemmatized and n-grams can be extracted (for fixed values of n) in order to compute a probability distribution of the n-grams for the texts generated from $\mathcal{I}$, and another probability distribution of the n-grams for the texts generated from $\mathcal{O}$, in order to test at the text level if some novelty can be detected in the texts issued from $\mathcal{O}$ wrt those issued from $\mathcal{I}$.

For the text generation, we used LLaMA 3. LLaMA 3 is a latest-generation language model developed by Meta AI, designed to push back the boundaries of natural text generation thanks to a sophisticated architecture and a high number of parameters (8 billions). This model is part of a series of LLaMA (Large Language Model Meta AI) language models distinguished by their ability to understand and generate text in a coherent and contextually appropriate way.

As a warning, we want to emphasize that our aim was not to test the quality of the texts that LLaMA 3 can produce in any way. LLaMA 3 is just used here as a tool for producing short texts that are constrained to follow the story line of an unremarkable story—freely inspired from the story Queneau starts with, in *The Exercises in Style*—which includes or not a series of 10 emotions, in order to have two collections of texts associated with the collections of vectors in $\mathcal{I}$ and $\mathcal{O}$ respectively. The aim is to check if the potential novelty at the vector level is reflected or not at the text level. The story line is specified to LLaMA 3 through the following exchange:

user: I will give you a story and you'll have to tell me the main route of the story in a paragraph (don't jump lines keep it all together) okay?

assistant: Sounds like fun! I'm ready to hear your story. Please go ahead and share it, and I'll do my best to summarize the main route of the story in one paragraph. Go!

user: An individual boards a public transportation vehicle. His figure blends into the mass of anonymous passengers. His gaze wanders, observing the familiar faces distantly. Suddenly, a peculiarity catches his attention: a man adorned with an unusual headgear, a felt hat of austere gray, perched majestically atop his head. And there, an irregularity, a disturbance in the usual landscape: a fly, intrepid explorer, ambles nonchalantly on the nape of this impassive man. A shiver of curiosity runs through him. A question nags at his mind: should he intervene? A decision takes shape in the meanders of his thought. Does the man, this bearer of felt, deserve his attention? Is it a mere accident or a sign from destiny? An irrepressible impulse propels him to act. He rises, approaches with determination. A brief, incisive exchange, punctuated by abrupt gestures, suffices to draw the attention of the man, but in a markedly different register. A reprimand escapes his lips, cutting, imbued with thinly veiled arrogance. And the fly? It continues its journey, indifferent to this human theater, oblivious to its involuntary role in this fleeting scene of daily life

assistant: As an individual boards a public transportation vehicle, he becomes distracted by a peculiar sight: a man wearing an unusual gray felt hat and sporting an intrepid fly on his neck. The individual's curiosity is piqued, and he begins to wonder if the situation warrants intervention. As he rises from his seat and approaches the man, he delivers a reprimanding remark laced with arrogance, drawing the attention of the hat-wearing individual in a completely different register than anticipated. Meanwhile, the fly remains oblivious to its role in this brief and fleeting drama that plays out amidst the daily routines of the anonymous passengers.

Thus the prompt corresponding to the initial vector $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ is

user: Redo the main route you mentioned before transforming it in a way that dimension Joy is false, Fear is false, Anger is false, Tranquility is false, Nostalgia is false, Shame is false, Confusion is false, Curiosity is false, Love is false, Resignation is false. (don't write any of this words in the paragraph (Joy, Fear, Anger, Tranquility, Nostalgia, Shame, Confusion, Curiosity, Love, Resignation) and don't write (True or false) either).

Experimental results

First of all, the comparison of the collections $\mathcal{I}$ and $\mathcal{O}$ in the experiment yielded the following results in terms of Hellinger distance and of Kullback-Leibler divergence:[5]

$H(p_{\mathcal{I}}, p_{\mathcal{O}}) = 0.8625$; $KL(p_{\mathcal{I}}||p_{\mathcal{O}}) = 0.6778$

As can be seen, the novelty of the output collection $\mathcal{O}$ with respect to the inspiring set $\mathcal{I}$ is clearly assessed by the high value of Hellinger distance (whose maximum is 1). This is confirmed by the divergence evaluation. However note that here the vectors $\vec{d}$ in $\mathcal{O}$ do not necessarily maximize their difference with the vectors $\vec{a}, \vec{b}, \vec{c}$ in $\mathcal{I}$ from which they are obtained, in the sense of subsection .

The text produced by LLaMA 3 from the prompt associated with the initial vector $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ is

assistant: As the individual steps onto the public transportation vehicle, his gaze drifts across the familiar faces, but one peculiarity stands out: a man clad in an austere gray felt hat, with an intrepid explorer of a different kind - a fly - perched

[5]Code is available here: `https://github.com/Sasufox/AnalogyStories`

nonchalantly on the nape of his neck. The individual's thoughts begin to weave a tapestry of inquiry, and he rises from his seat with purpose, approaching the hat-wearer with determination. A brief, incisive exchange ensues, punctuated by swift gestures, as the two individuals intersect in a fleeting dance of daily life.

This text is quite short, and is devoid of any particular emotion. As an example here is the text produced when Confusion, Curiosity, and Resignation are required (the prompt corresponds to vector $(0, 0, 0, 0, 0, 0, 1, 1, 0, 1)$):

As he boards the public transportation vehicle, his gaze drifts across the sea of anonymous faces, his mind a maze of contradictions. A peculiar sight captures his attention: a man's austere gray hat, a fly's leisurely stroll on the nape of his neck. His thoughts twist and turn, unsure whether to intervene or simply observe. The air is thick with uncertainty as he approaches the man, his words laced with an unexpected tone that belies the calm exterior. Meanwhile, the fly continues its aimless flight, seemingly unaware of the complex emotions swirling around it, leaving only a lingering sense of disorientation in its wake.

As we can see the writing of the new text has been transformed, keeping the same story line. As expected, the collection of texts associated to $\mathcal{O}$ differs from the collection associated to $\mathcal{I}$. This can be checked using Hellinger's distance (which has the advantage to be valued on a bounded scale) between the probability distributions of n-grams in the two collections. The results obtained for $n = 2$, $n = 3$, $n = 4$ are:
$H(p_{\mathcal{I}}^{2-gram}, p_{\mathcal{O}}^{2-gram}) = 0.3947$: $H(p_{\mathcal{I}}^{3-gram}, p_{\mathcal{O}}^{3-gram}) = 0.6279$; $H(p_{\mathcal{I}}^{4-gram}, p_{\mathcal{O}}^{4-gram}) = 0.7690$.

This confirms the novelty of texts in $\mathcal{O}$ with respect to those in $\mathcal{I}$. However, it has been noted (Chen and Ding 2023) that advanced large language models obtain very good results on divergent semantic association tasks, suggesting a capacity for creativity. Since the production of texts by LLMs is based on an stochastic mechanism, one may measure the distance between stories generated from the same vector, in the following way. From a representative sample (of size 280) of the 1024 vectors, 100 texts are generated for each vector. For each of these collections, a text is taken as a reference and compared with the other texts. The Hellinger distance between the collection of reference texts and the collection of other texts can then be calculated. The results suggest that LLaMA generates a large number of n-grams that are not present simultaneously in two documents generated from the same vector. Thus the stochastic nature of LLM makes it difficult to determine exactly how much novelty is brought about by the use of a new prompt.

Final discussion

In the introduction we have distinguished three kinds of analogical reasoning: i) the one, sometimes termed 'conceptual blending' that brings together two situations belonging to different universes, as in the analogy between the atom and the solar system; ii) case-based reasoning that extrapolates to a new situation what was observed in a similar case; iii) the solving of analogical proportion-based equation that produces a fourth item from three others, as in analogy puzzles. The first kind corresponds to transformational analogy, the second kind can hardly be called creative since it's essentially a question of adapting something already known, the last kind is a form of exploratory analogy, where a universe (the one describable by means of the attributes used for representing the considered items) is explored.

Unlike a purely random path through the universe of Boolean features, this universe can be explored by means of the analogical proportion-based inference in a controlled way, for instance by favoring the accumulation of valuable properties. Generally speaking, the analogical proportions $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ make sure that the differences between $\vec{a}$ and $\vec{b}$ are preserved between $\vec{c}$ and $\vec{d}$. The experiment has shown that even if the inspiring set (where $\vec{a}, \vec{b}, \vec{c}$ are taken) is made of items that are relatively close of each other, the novelty of $\vec{d}$ can be established in terms of Hellinger distance. In case vectors would describe options at a meta level, the inference mechanism might be used for transformational creativity (viewed as exploratory creativity at the meta-level).

In our experiment, we have used Shannon entropy minimization in order to build an inspiring set that is not too disparate. That does not prevent the analogical inference mechanism to produce novel vectors clearly outside the inspiring set. It is clear that the choice of the inspiring set has an influence on the outputs that can be produced. The use of selection procedures on top of the inference mechanism in order to start with "inspiring" items is an important issue to further explore in future research. Other constraints could be considered for building the inspiring set.

The experiment has illustrated how the vectors generated can be the basis of prompts addressed to an LLM-based chatbot, leading to the production of "constrained" texts where the novelty inherited from the novelty of the vector generated is somewhat preserved. As already suggested in (Afantenos et al. 2024), analogical proportion-based inference can be used as a tool for proposing novel items. The combination of analogical reasoning and LLM-based capabilities to produce text constitutes a kind of creative device.

Queneau's *The Exercises in Style* provides a very good example of exploratory creativity, like his *A Hundred Thousand Billion Poems* (Queneau 1961) perfectly exemplifies the idea of combinational creativity, in the literary domain. However, the remarkable creativity of Queneau is at the conception level of these two books, i.e., in the device itself. Interestingly enough the h-creativity of Queneau for *The Exercises in Style* is perhaps debatable, since Desiderius Erasmus in his 1512 educational rhethoric textbook *De Copia* instructed on how to give variety to writing; thus in the 33rd chapter he gave 195 variations on the same (Latin) sentence whose meaning was "Your letter delighted me greatly." Queneau was also highly interested by one of the earliest known published volumes of computer poetry (Baudot 1964) where creativity was combinational and the combination of words, belonging to a restricted vocabulary, aleatory. This is shown by his commentaries gathered at the end of the book among many other reactions by writers and linguists. In particular, he wondered about the possibility for humans to discern

sentences produced by the machine from sentences written, e.g., Surrealists poets – a Turing-like test!

Conclusion

The paper has focused on the creative ability of analogical proportion-based inference. This paper has made several contributions: first, we have highlighted and illustrated the creative power of the analogical proportion-based inference. Second, the assessment of the novelty of what is produced has been confirmed in the experiment by the divergence measures applied to the generated prompts. Notably, Hellinger distance that takes its values in a bounded interval seems especially appropriate. Moreover, the experiment also showed the possibility of combining analogical inference and LLM-based chatbots, while preserving creative power.

Besides, several lines for further research would be worth investigating. First, the analogical mechanism used is only based on Boolean proportions. It might be of interest to use graded analogical proportions extended to numerical data (Dubois, Prade, and Richard 2016) or fuzzy logic (Colombo, D'Onofrio, and Portmann 2020), for, e.g., introducing a graded view of emotions in our experiment. In addition, in this paper we have focused on the assessment of the novelty of what is produced, but one may also checked the quality of the texts obtained by estimating if the texts effectively contain words in relation with the emotions required by the prompts, using an appropriate lexicon (Mohammad 2018). Finally, one may like to compare the results obtained using analogically generated prompts with a direct request of generating texts on the basis of triplets of examples. In the absence of direct experimentation, any speculation regarding the results that would be obtained remains hypothetical, especially since the triplet of examples used can significantly influence the outcome. However, it should be noted that our methodology constrains the large language model in leveraging all terms located within a specific region of the embedding space associated with an emotion, while increasing diversity, as evidenced by the increase in the Hellinger distance.

The experimental process described in this paper should be viewed as a tool to aid creation rather than as a creative machinery in itself. Indeed, just as Queneau's creativity in *Exercices de style* lies primarily in the idea of producing multiple versions of a mundane text—acting as a reference constraint while simultaneously employing different stylistic constraints—here, the user decides which parameters to vary in the texts to be generated. They then use analogical inference to explore the space of possibilities to determine the constraint specifying the parameters that will induce the prompt to be applied to the LLM.

Acknowledgments

This research was supported by the ANR project "Analogies: from theory to tools and applications" (AT2TA), ANR-22-CE23-0023.

References

Aamodt, A., and Plaza, E. 1994. Case-based reasoning; foundational issues, me- thodological variations, and system approaches. *AICom, 7, 39-59*.

Afantenos, S.; Prade, H.; Richard, G.; and Bernardes, L. 2024. Analogical proportions and creativity: A preliminary study. In Grace, K.; Llano, M. T.; Martins, P.; and Hedblom, M. M., eds., *Proc. 15th Int. Conf. on Computational Creativity (ICCC'24), Jönköping*, 289–298.

Barbot, N.; Miclet, L.; and Prade, H. 2019. Analogy between concepts. *Artif. Intell.* 275:487–539.

Barré, J., and Poibeau, T. 2023. Beyond canonicity. modeling literary change in french novelist production: Canon vs. archive. In Sela, A.; Jannidis, F.; and Romanowska, I., eds., *Proc. Computational Humanities Research Conf., Paris, Dec. 6-8*, CEUR Workshop Proc., 3558, 814-830.

Baudot, J. 1964. *La Machine à Ecrire mise en marche et programmée par Jean A. Baudot*.Editions du Jour, coll. les poètes du jour, Montréal. 'Le premier recueil de vers libres rédigés par un ordinateur électronique',96p.

Bayoudh, M.; Prade, H.; and Richard, G. 2012. Evaluation of analogical proportions through Kolmogorov complexity. *Knowl.Based Syst.* 29:20–30.

Boden, M. A., ed. 1994. *Dimensions of Creativity*. MIT Press.

Boden, M. A. 2003. *The Creative Mind - Myths and Mechanisms (2. ed.)*. Routledge. 1st ed., Basic Books, 1990.

Bounhas, M.; Prade, H.; and Richard, G. 2017. Analogy-based classifiers for nominal or numerical data. *Int. J. Approx. Reason.* 91:36–55.

Brown, D. G., and Ventura, D. 2022. Ethics, aesthetics and computational creativity. In Hedblom, M. M.; Kantosalo, A. A.; Confalonieri, R.; Kutz, O.; and Veale, T., eds., *Proc. 13th Int. Conf. on Computational Creativity (ICCC'22), Bozen-Bolzano, June 27 - July 1*, 150–158.

Buchanan, B. 2001.Creativity at the meta-level.*AI Magazine* 22:13–28.

Bundy, A. 1994. What is the difference between real creativity and mere novelty? *Behavioral and Brain Sciences* 17:533–534.

Chen, H., and Ding, N. 2023. Probing the creativity of large language models: Can models produce divergent semantic association? *CoRR* abs/2310.11158. Published in Findings of EMNLP 2023.

Chieppe, P.; Sweetser, P.; and Newman, E. 2022. Bayesian modelling of the well-made surprise In Hedblom, M. M.; Kantosalo, A. A.; Confalonieri, R.; Kutz, O.; and Veale, T., eds., *Proc. 13th Int. Conf. on Computational Creativity (ICCC'22), Bozen-Bolzano, Jun.27-Jul. 1*, 126–135.

Colombo, M.; D'Onofrio, S.; and Portmann, E. 2020. Integration of fuzzy logic in analogical reasoning: A prototype. In *Proc. 16th IEEE Int. Conf. Intel. Comp. Communic. and Proc. (ICCP'20), Cluj-Napoca, Sept. 3-5*, 5–11.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In Raedt, L. D.; Bessiere, C.;

Dubois, D.; Doherty, P.; Frasconi, P.; Heintz, F.; and Lucas, P. J. F., eds., *Proc. 20th Europ. Conf. on Artificial Intelligence (ECAI'12), Montpellier, Aug. 27-31*, volume 242 of *Frontiers in Artif. Intellig.and Appl.*, 21–26. IOS Press.

Colton, S.; Pease, A.; and Ritchie, G. D. 2001. The effect of input knowledge on creativity. In *Proc. Workshop on Creative Systems: Approaches to Creativity in AI and Cognitive Science, at 4th Int. Conf. on Case-Based Reasoning, Vancouver, July 31*.

Cook, M., and Colton, S. 2011. Automated collage generation - With more intent. In Ventura *et al.*, D., ed., *Proc. 2nd Int. Conf. on Computational Creativity, Mexico City, April 27-29*, 1–3.

Correa Beltran, W.; Prade, H.; and Richard, G. 2016. Constructive solving of raven's IQ tests with analogical proportions. *Int. J. Intell. Syst.* 31(11):1072–1103.

de Bono, E. 1971. *The Use of Lateral Thinking*.Avon Books.

de Mántaras, R. L. 1998. It don't mean A thing (if it ain't got that swing). In Prade, H., ed., *Proc. 13th Europ. Conf. on Artificial Intelligence, Brighton, Aug. 23-28*, 694–696. John Wiley and Sons.

Dreistadt, R. 1969. The use of analogies and incubation in obtaining insights in creative problem solving. *The J. of Psychology* 71:159–175.

Dubois, D.; Prade, H.; and Richard, G. 2016. Multiple-valued extensions of analogical proportions. *Fuzzy Sets Syst.* 292:193–202.

Ekman, P. 1999. Basic emotions. In Dalgleish, T., and Power, M. J., eds., *Handbook of cognition and emotion*. John Wiley Sons Ltd. 45–60.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. The structure-mapping engine: algorithm and examples. *Artif. Intell.* 41(1):1–63.

Fele, B.; Babic, J.; Pollak, S.; and Znidarsic, M. 2022. Evaluation of curriculum learning algorithms using computational creativity inspired metrics. In Hedblom, M. M.; Kantosalo, A. A.; Confalonieri, R.; Kutz, O.; and Veale, T., eds., *Proc. 13th Int. Conf. on Computational Creativity (ICCC'22), Bozen-Bolzano, June 27 - July 1*, 364–373.

Fuchs, B.; Lieber, J.; Miclet, L.; Mille, A.; Napoli, A.; Prade, H.; and Richard, G. 2020. Case-based reasoning, analogy, and interpolation. In *A Guided Tour of Artificial Intelligence Research: Vol. I: Knowledge Representation, Reasoning and Learning*. Springer. 307–339.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2):155–170.

Gust, H.; Kühnberger, K.; and Schmid, U. 2006. Metaphors and heuristic-driven theory projection (HDTP). *Theo. Comput. Sci.* 354(1):98–117.

Halpern, D. 1986.Analogies as a critical thinking skill. In Berger, D.; Pezdek, K.; and Banks, W., eds., *Applications of Cognitive Psychology: Problem Solving, Education, and Computing*.75-86, Routledge.

Hesse, M. 1959. On defining analogy. *Proceedings of the Aristotelian Society* 60:79–100.

Hummel, J. E., and Holyoak, K. J. 1997. Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review* 104(3):427–466.

Itti, L., and Baldi, P. 2009. Bayesian surprise attracts human attention. *Vision Research* 49:1295–1306.

Kaufmann, A. 1969. L'imagination artificielle - (heuristique automatique). *Revue française d'Automatique, d'Informatique et de Recherche Opérationnelle (R.A.I.R.O.) - Recherche opérationnelle* 3(3):5–24.

Kolodner, J. L. 1993. *Case-Based Reasoning*. Morgan Kaufmann.

Kullback, S., and Leibler, R. A. 1951. On information and sufficiency. *Annals of Math. Stat.* 22:79–86.

Kunda, M.; McGreggor, K.; and Goel, A. K. 2013. A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research* 22:47–66.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In Christensen, B. T.; Kristensen, T.; and Boztepe, S., eds., *Proc. 1st DESIRE Network Conf. on Creativity and Innovation in Design, Aarhus, Aug. 16-17*, 22–28. ACM.

McCormack, J., and d'Inverno, M., eds. 2012. *Computers and Creativity*. Springer.

McGregor, S. 2007. Algorithmic information theory and novelty generation. In *Proc. 4th Int. Joint Workshop on Computational Creativity, London, June 17-19*, 109–112.

Miclet, L., and Prade, H. 2009. Handling analogical proportions in classical logic and fuzzy logics settings. In *Proc. 10th Eur. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09),Verona*, 638–650. Springer, LNCS 5590.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Gurevych, I., and Miyao, Y., eds., *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*, 174–184.

Mondol, T., and Brown, D. G. 2021. Incorporating algorithmic information theory into fundamental concepts of computational creativity. In de Silva Garza, A. G.; Veale, T.; Aguilar, W.; and y Pérez, R. P., eds., *Proc. 12th Int. Conf. on Computational Creativity, (ICCC'21), Mexico City, (virtual), Sept. 14-18*, 173–181.

Polya, G. 1954. *Mathematics and Plausible reasoning. Volume 1: Induction and Analogy in Mathematics*. Princeton University Press.

Prade, H., and Richard, G. 2018. Analogical proportions: From equality to inequality. *Int. J. Approx. Reason.* 101:234–254.

Queneau, R. 1947. *Exercices de Style*. Gallimard. Exercises in Style, tr. B. Wright, Gaberbocchus Press, 1958.

Queneau, R. 1961. *Cent mille milliards de poèmes*. Gallimard. Afterword: F. LeLionnais.Trans.: S. Chapman, A Hundred Thousand Billion Poems.

Ramalho, G. L., and Ganascia, J. 1994. Simulating creativity in jazz performance. In Hayes-Roth, B., and Korf, R. E., eds., *Proc. 12th National Conf. on Artificial Intelligence (AAAI'94), Seattle, Jul. 31 - Aug. 4, Volume 1*, 108–113. AAAI Press / The MIT Press.

Rhodes, M. 1961An analysis of creativity*Phi Delta Kappan* 42:305–310.

Ritchie, G. 2001. Assessing creativity. In *Proc. of the AISB Symp. on Artificial Intelligence and Creativity in Arts and Science, York*, 3–11.

Ritchie, G. 2006. The transformational creativity hypothesis. *New Generation Computing* 24:241–266.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds Mach.* 17:67–99.

Ritchie, G. 2012. A closer look at creativity as search. In Maher, M. L.; Hammond, K. J.; Pease, A.; y Pérez, R. P.; Ventura, D.; and Wiggins, G. A., eds., *Proc. 3rd Int. Conf. on Computational Creativity (ICCC'12), Dublin, May 30 - June 1*, 41–48.

Santo, L. E.; Wiggins, G. A.; and Cardoso, A. 2024. Towards a formal creativity theory: Preliminary results in novelty and transformativeness. *CoRR* abs/2405.02148.

Schank, R., and Kass, A. 1990. Explanations, machine learning, and creativity. In Kodratoff, Y., and Michalski, R. S., eds., *Machine learning: an artificial intelligence approach, volume III*. Morgan Kaufmann Publishers, San Francisco. 31–59.

Schmidhuber, J. 2010. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Trans. on Autonomous Mental Development* 2(3):230–247.

Schmidt, M.; Krumnack, U.; Gust, H.; and Kühnberger, K. 2014. Heuristic-driven theory projection: An overview. In Prade, H., and Richard, G., eds., *Computational Approaches to Analogical Reasoning: Current Trends*, Studies in Comput. Intelligence, 548, 163-194. Springer.

Turner, M., and Fauconnier, G. 2003. *The Way We Think : Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Van de Cruys, T. 2020. Artificial intelligence and literature. In Marquis, P.; Papini, O.; and Prade, H., eds., *A Guided Tour of Artificial Intelligence Research: Volume III: Interfaces and Applications of Artificial Intelligence*. Springer. 487–501.

Veloso, M. M. 1993. Prodigy/analogy: Analogical reasoning in general problem solving. In Wess, S.; Althoff, K.; and Richter, M. M., eds., *Topics in Case-Based Reasoning, 1st Europ. Workshop (EWCBR'93), Kaiserslautern, Nov. 1-5, Selected Papers*, LNCS, 837, 33-52. Springer.

Wiggins, G. A., and Forth, J. 2018. Computational creativity and live algorithms. In McLean, A., and Dean, R. T., eds., *The Oxford Handbook of Algorithmic Music*. Oxford University Press.

Wiggins, G. A. 2006a. A preliminary framework for description, analysis and comparison of creative systems. *Knowl.based Syst.*19(7):449–458.

Wiggins, G. A. 2006b. Searching for computational creativity. *New Gener. Comput.* 24(3):209–222.

Wiggins, G. A. 2019. A framework for description, analysis and comparison of creative systems. In Veale, T., and Cardoso, F. A., eds., *Computational Creativity - The Philosophy and Engineering of Autonomously Creative Systems*. Springer. 21–47.

Generominator: Operationalizing the Computational Creativity Ideation Cards "Generominos"
Paper type: System or Resource description paper

Gabriel Bacon* and **Rosalind Dixon*** and **Max Kluiwstra-Descalle*** and **Audrey Ostrom*** and **Alexander Sherbrooke*** and **Thomas Wessel*** and **Jiangyue Wu*** and **Markus Eger**
{gbacon,romdixon,mkluiwst,aostrom,asherbro,twessel,jwu454,meger}@ucsc.edu
UC Santa Cruz
Department of Computational Media

Abstract

To facilitate the exploration of computational creativity in all its aspects, the concept of "Casual Creators" arose to describe software that allows the playful expression of creativity in some creative space. Separately, Compton et al. presented a set of (physical) ideation cards for computational creativity they called "Generominos". In this paper, we present a digital implementation of these ideation cards that allow users to compose a wide range of different algorithms in a visual, graph-based web interface to generate digital computational creativity artifacts. We discuss the process of giving formal execution semantics to the previously loosely-defined ideation cards, additions and limitations we had to impose on the system, and demonstrate the wide range of possibilities enabled by our system, which include generative processes for sound, images, text, 3D graphics, etc., as well as transformations between them. Additionally, we evaluate our system against a set of desirable design patterns from the literature.

Introduction

Computational Creativity can involve a wide range of different media (Graves et al. 2021), operations (Compton, Kybartas, and Mateas 2015; Eger 2022), output artifacts (Mazeika and Whitehead 2017), and means of expression (Tuffs 2017; Mathewson and Mirowski 2017). Exploring such a vast space may seem intimidating to more casual users, as even an overview of what *may* be done is challenging to come by. Various authors have thus explored ways of making computational creativity more approachable, with Kate Compton arguably being the most vocal proponent of systems called "Casual Creators" (Compton and Mateas 2015). These systems present an approachable way to generative methods, following the tradition of thousands of years of human play with assembling building blocks into larger structures, and have also found adaption in computational domains, such as with Tanagra (Smith, Whitehead, and Mateas 2010). One way of enabling such exploration is through ideation cards, like Compton et al.'s (2017) Generominos, which present computational concepts as a possibility space which users can explore through imagination, with the option to later implement a system ideated this way. While this allows a more free-form exploration, it also does not provide any direct feedback of what a system may actually produce as output. In this paper, we present an operationalization of aforementioned Generominos in the form of our system, aptly named the Generominator, which allows users to explore a wide range of computational creativity algorithms and get real time feedback in the process. Our core contribution is this system, consisting of formal semantics for roughly 100 different cards which can process sound, images, 3D meshes, curves, and other data types and combine them in a wide variety of ways, and a web-based editor which allows users to construct graphs that combine these operations in a guided way, and supports them in finding cards compatible with their current processing pipeline. We will discuss our implementation in detail, provide several sample output artifacts showcasing the richness of the possibility space enabled by our system, and evaluate it against criteria outlined by Kate Compton in her work on Casual Creators.

*These authors contributed equally.

Related Work

Our work exists in the space Kate Compton and Michael Mateas (2015) termed "Casual Creators", i.e. software that allows the playful exploration of computational creativity, as an expression of everyday creativity (Kaufman and Beghetto 2009), or, as Compton and Mateas so aptly put it, because *it is fun*. While there may often be an emphasis on (computational) creativity in the pursuit of some more "productive" goal, such as engineering (Masood et al. 2024) or industrial design (Heisserman 1994), creation for creation's sake has also been explored by prior work. Kreminski et al. (2020) present the Germinate system, based on the Gemini game generator by Summerville et al (2018). Germinate allows users to define parameters for a game using a web-based interface, which are then passed to Gemini to generate a playable game artifact. While our system targets a different type of target artifact, and has a very different underlying formalism, it follows a very similar design philosophy.

More broadly, our work exists within a long tradition of visual, graph-based programming tools that span many genres of systems (Morrison 2011). Sood et al. (2025), for example, presented a graph-based system that supports creative writing. These graph-based editors also commonly appear within procedural content generation and game devel-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

opment work, and can be seen in tools such as Blender[1], Unreal Engine[2] or Unity ShaderGraphs[3]. Our system takes inspiration from these designs, enabling users to add, remove, and connect various types of nodes in a graph to form a completed and executable program.

The core inspiration for our work, though, are the Generominos ideation cards by Compton et al. (2017). Each card represents an input, transformation, or output operation that can process data of different data types, including sound, images, text, 3D meshes, etc. Our system represents an operationalization of these cards, together with a web-based interface that supports casual exploration and reflection (Kreminski and Mateas 2021). Since the Generominos were designed to encourage flexible interpretation across many possible implementations, some operational details such as data representations, evaluation order, and cross-type conversions were intentionally underspecified in the original design. Our implementation of these cards gives one of many possible definitions as a way to produce immediately working prototypes. We would like to note that Compton maintains an existing Generominos editor[4] that supports browsing and connecting cards in a web-based user interface, similarly to how the physical cards are used, but lacks the execution semantics of our system. Additionally, other community versions of Generominos exist, such as a port to the game TableTop Simulator[5]. In the next section, we discuss the concrete operational choices we made, and how our interface encourages graph exploration for artifact generation.

Operationalizing Generominos

As presented by Compton et al., each of the Generominos ideation cards can have various inputs and outputs that can have a variety of different data types. The entire generative process is driven by cards consuming their inputs, transforming them in some way and producing the corresponding outputs. Cards may be sources, which feed data into the system, transformations, which apply some algorithm (or just a simple operation) to data, or outputs, which produce an artifact consumable by an audience. While the original paper briefly describes the data types, its goal was not any particular implementation, and thus it does not provide concrete detail of how each is represented. For our system, we therefore first had to provide a concrete definition for each data type, and how cards can interface to consume and produce this data. We also implemented an execution engine that can take a graph consisting of connected cards and evaluate them in a defined order. To facilitate the design of these graphs, we also developed a web-based user interface in which users can intuitively connect the various cards to explore the creative capabilities of the system. In this section, we will describe each of these parts of our system, starting with the data types.

[1]https://www.blender.org/

[2]https://www.unrealengine.com/

[3]https://unity.com/

[4]https://www.galaxykate.com/generominos/editor-dev/

[5]https://steamcommunity.com/sharedfiles/filedetails/?id=1396406737

Data Types

Generominos are best understood as defining a data flow graph, where data can have one of several data types. Our system supports:

- Atomic types: value (number), text (string), and state (boolean).
- Array types: vector (array of numbers), vectorfield (array of vectors), shape (array of curves).
- Curve: A parametric curve that maps a value between 0 and 1 to points along the curve. Curves may also have a color associated with them. We define an abstract Curve interface for this purpose, and our system currently contains Bezier Curves and Polygons as concrete curve implementations.
- Graph: Node positions represented by vectors with edges that are defined in an adjacency matrix.
- Color: Stored as red, green, blue, and alpha values.
- Image: ImageData object, which represents the image as a flat array of bytes with an associated width and height.
- Waveform: Audio sample data stored as a flat array of numbers with an associated sampling rate.
- Depthmap: An array of numbers, going from top left to bottom right, storing a depth value for each point.
- Geolocation: Longitude and latitude stored as two values.
- Trimesh: A representation of 3D geometry as an instance of the THREE.Mesh class from the threejs library[6].
- Generic: A wrapper that is used as a placeholder by cards that may operate on any type (e.g. conditionals).

We will now describe how cards can process this data.

Cards

To process data in our system, we define two interfaces: Card and Port. The Card interface outlines what a card itself may do. It includes a title, a list of input ports, a list of output ports, and an optional description, as well as 3 functions: `evaluate`, `init`, and `cleanup`. The title is the name of the card displayed in the UI, as well as the card's identifier for operations such as saving to JSON. The description, like the title, is printed on the card in the UI and provides the user with more information on what the card does and how to use it. The input and output ports define what data types the card operates on and make use of the Port interface.

The lifecycle of a card consists of three stages:

1. Before execution begins, the `init` method is called, which can prepare any data needed to run the card, connect to API services, download assets, etc.
2. `evaluate` is called by the execution engine to process data and defines the cards behavior. It both takes in and outputs a list of DataTypes, which need to match the input and output port list of the card, respectively.
3. `cleanup` is run when execution ends and should be used to dispose of any resources the card used.

[6]https://threejs.org/

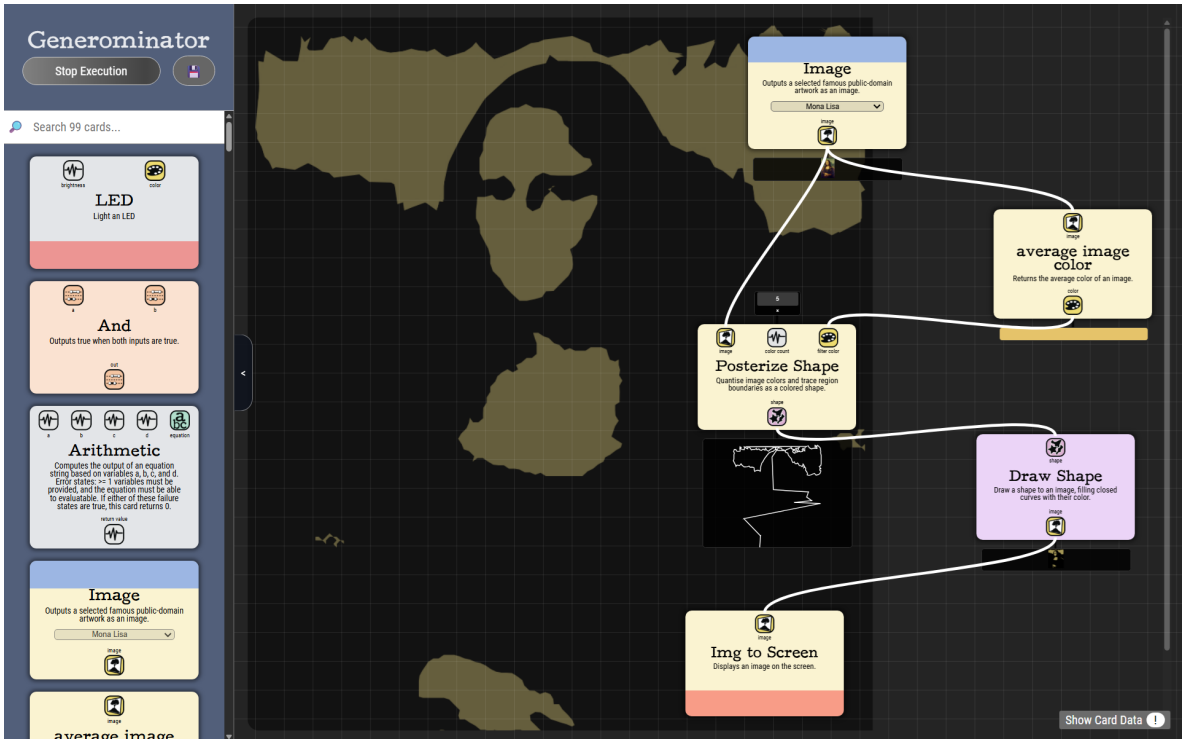


Figure 1: The user interface of Generominator during a system run. Using the card bank on the left, users can create systems by connecting cards together. Users can then run their system to see the outputs of each card, and the system itself.

Most simple cards will only implement `evaluate`, but more complex operations involving API calls or hardware access may require initialization or cleanup.

The Port interface outlines a single data port on a card. It includes four variables: a label, a flag indicating if the (input) port is optional, a data type and an optional default value. The label is defined per instance of each port and is printed above or below each data port in the UI. The optional flag marks whether or not the data port will require input for the execution graph to run. The data type defines which data can be passed through this data port, and the default value will be used in place of data from another card if no connections are made to the port during execution.

Using input- and output-ports as connecting points, we can define a graph consisting of multiple cards that connect some or all of the outputs of a card to another card, which in turn produces outputs that are passed to another card, and so on, until we reach a card that does not have any output ports. In our system, such cards will typically have side effects that display data on the screen, or play something on the speakers. Next, we will describe how our system takes a graph defined in such a way to actually produce an output.

Execution Model

At the heart of the Generominator is the execution engine, responsible for both managing the card life cycle and graph execution. When cards are dragged onto the editor canvas, an entry for them is created within the engine's internal Directed Acyclic Graph. As connections are made between cards in the interface, nodes in the graph are also linked. This gives an internal representation of where data should flow, and the order in which card execution must occur.

At the start of each engine run, the graph is first sorted topologically via Kahn's Algorithm (Kahn 1962). This results in a sorted list that guarantees any card dependent on another's output will be placed after it in the list. This ensures that execution flow is handled in the correct order and data can be propagated through the graph correctly. Our implementation supports running multiple disconnected graphs, but does not currently support cycles. If one is found, it may be dropped from the dependency tree and result in undefined behavior. To prevent this, the editor halts execution if a cycle is found.

Normal runs of the engine traverse the entire list in order, starting from the eldest parent nodes and passing their results through each child until every node in the list has been visited. In some cases, however, it may be desirable to execute only a subset of the full graph. This is most prominently on display with another engine feature: card events. Events enable cards to inform the engine that their data has changed and cause it to re-execute the graph in order to remove any stale data left in the result cache from the previous state. In other words: Using events, cards can cause the graph to be reevaluated when their data changes, which may include when certain external events have occurred that were queried over an API, the user performs an input action, or through

cards that provide a continuous stream of input, like a webcam, or time. When such cards are near the beginning of the graph, this is not much of an issue as all downstream cards will need to evaluate anyway to update their information, but can become a problem if they are near the end. With every run request, card data is needlessly re-evaluated.

To solve this, our engine supports starting execution from any arbitrary node with a feature that we call "subgraph execution." With this feature, when a card's data changes and the graph moves to re-evaluate it, any parent cards to that one are skipped and execution begins directly with the node that changed. Any data that the parent cards would have generated is instead pulled from the engine's internal cache of results from the previous execution. Executing from that node onwards completes like a standard run, and the resulting data is cached for later use.

Events and subgraph execution enable the creation of card types that would otherwise be impossible to implement, ranging from the potentiometer which updates its state with a GUI element, to the Bluesky card, which continually polls the API and outputs any strings that match its input. In the case that a card updates very frequently, however, such as the mouse coordinate card, the engine can still be overwhelmed with re-execution requests. Moving the mouse across the screen might push hundreds of requests into the queue which could take seconds to execute, leading to the user feeling as if the program is not responsive. Our solution to this problem is what we call "run merging": Each new execution request is compared with the one already in the run queue (if it exists). Whenever possible, new requests have their data merged into that request, overriding any data that might otherwise have been pulled from the cache. In the case that the same card is repeatedly updated, old states are overwritten by the new state which enables us to skip running steps that would otherwise appear to the user extremely briefly, and maintain realtime reactivity.

Finally, there are card interactions that might require complex routing of data and operations on that data. To address this, we have added the generic datatype and the ability for cards to trigger output on ports independently of each other. Generic ports have the ability to transform into any other type when connected to another card and can act independently or transform all other generic types on the card. Both of these features are employed by the branch card, where any data is taken in and is only pushed to the branch determined by a value port, enabling complex data flow compositions that can route data in a data-dependent way.

We will now describe how we allow users to manipulate the graphs that are run by our execution engine using our user interface.

User Interface

The Generominator is available as a web application accessible on both desktop and mobile devices[7], shown in figure 1 with our source code available on Github[8]. Users can drag cards from the card bank on the left onto the canvas on the right. On each card is a number of input and output ports that determine what data the card takes in or provides; by clicking and dragging from a port, users can create connections to other cards. Alternatively, users can set ports to accept constant inputs by interacting with the floating editors on the input ports.

When a user is done creating their system, they can click "Run Syste" to run their card graph through the execution engine. The outputs of each card appear under the output ports of each card, showing their system at every stage of the graph. The user is able to modify their graph and see their results update in real time. Additionally, some cards, such as the Button, Webcam, and Potentiometer, provide values that can change over time or through user inputs, allowing for dynamic systems to be created.

Unusual among graph-based programming systems, the Generominator allows users to save their systems as a PNG file, depicting a card, that can be shared online. Inspired by existing implementations, such as Spore's creature sharing system[9] or the PICO-8 game cartridge system[10], this file stores the system data within the pixels of the image using steganography. The image is then able to be uploaded to various social media platforms, such as Discord, Twitter and Tumblr[11]. Once shared, other users can save the image and load it into their editor to load the saved system.

Results

In this section, we present an evaluation of our system. A key goal during development was the creation of a system that is both flexible for developers and explorable for users. We will discuss how the very development process itself promotes this flexibility and extensibility, and then evaluate our system against design patterns set forth by Compton and Mateas (2015). Additionally, we will present a selection of actual graphs designed using our system, and the capabilities they demonstrate, while also discussing the limitations our system has.

Collaborative Coding and Extensibility

Generominator's strength as a tool lies in its modularity. Its appeal as creative support manifests in the possibility space each connection offers, and its reliability in that supportive pursuit is dependent on the fidelity of each card. This dynamic results in a massive capacity for meticulous, concentrated development. A natural extension of the tool is through the addition of cards. This is logically inherited from Kate Compton's original physical version. In modifying an analog set of cards, insertion is most straightforward. Without a technical framework, the path *between* cards is handled in the theater of the mind and modifying essential qualities of the card-base would necessitate reprinting

[7] Available online: https://generominator.dev

[8] https://github.com/Generominator/Generominator

[9] As described in Kate Compton's 2025 AIIDE keynote: https://www.youtube.com/watch?v=jlW8aaGcKzM

[10] https://www.lexaloffle.com/dl/docs/pico-8_manual.html

[11] Care must be taken, though, as some social media platforms, like Bluesky, may re-encode PNG files in a way that loses the embedded data

or posthumous editing. This leaves card-creation as the instinctive mode of alteration and is encouraged, even, by the original, where tools to create new cards were provided[12].

The Generominator technical framework endeavors to support card-creation with ease equal to its analog counterpart. For a collaborator in the card-creation process, only an understanding of input, output, and data type is required. Individual cards do not impact each other unless paired, and only function together if their input and output are compatible, significantly reducing the chance for dissonance between cards. Because incoming data is screened before being transformed, in a rare incompatibility scenario, problem data is surfaced to the user before contaminating outgoing data or its surrounding graph downstream.

This benefits the development process in a very practical sense. For a hopeful collaborator to contribute a card, they most likely do not need to alter major, load-bearing files like those in the card base or those associated with the execution engine. Instead, their card can be entered into the card folder, built in accordance with the card template, all without the risk of impacting any card but their own.

With our repository, we provide a detailed README file that contains a step-by-step guide for card contributions. In general, to add a new card, a developer only has to create a new subclass of the `Card` class, providing the title, as well as input and output ports, and implementing the `evaluate` function, and the new card will be found automatically by the system and be available for use in the system. In Generominator's development process, it was this exact disconnect between each card that allowed code contributions from a diverse set of developers (all eight authors of this paper contributed multiple cards to the system) to be swift, reliable, and safe. After the card base was established, each team member could develop their own selection of cards independently of each other. Without the risk of multiple parties significantly changing the same files, the likelihood of code conflicts was and remains low.

Additional opportunity for major changes lie in the addition of data types. This category of change was explored in this digital iteration of Generominator via the `generic` data type, which accepts all others, and is not present in Kate Compton's analog version of the tool, but future work can easily add new port types to process kinds of data we currently do not support.

Because of their detachment from each other, adding and altering data types bear the same benefits as adding and altering cards. Type additions do not require altering any other data type. This makes their contribution straightforward, while still being hugely valuable, compounding Generominator's dynamism and exponentially expanding each graph's possibility space, especially if conversion cards are provided to interface the existing cards with a newly added data type.

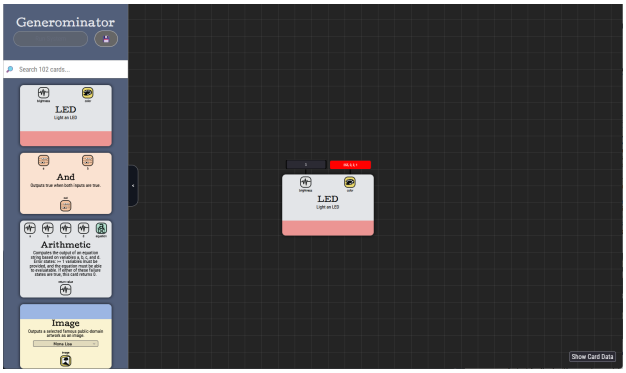
Generominator's inherent modularity, assumed from the card-based metaphor at its core, results in a code base that is easily improved, added to, and revised. This is exemplified by the authors' contribution patterns, which rarely overlapped, and by future contribution guidelines, which can be followed without risking the fidelity of the tool as a whole.

[12]This repository provides tools for users to create their own cards and suggests ways to get them printed: https://github.com/galaxykate/generominos

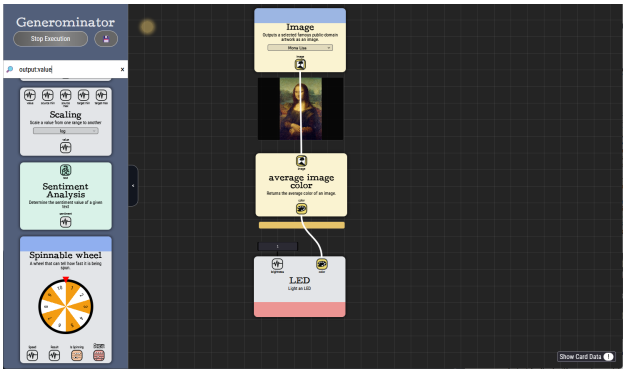
Design Patterns

Drawing upon the Casual Creators framework (Compton and Mateas 2015), our system is designed to support autotelic creativity and facilitate a state of flow. By integrating modularity into a card-based interface, we encapsulate complex functions within card units. This approach allows users to engage in reflection-in-action (Schön 2009) through direct interaction, where possible connections between the cards guide exploration of creative possibilities. By emulating physical-world object interactions, the system fosters a sense of Direct Engagement (Hutchins, Hollan, and Norman 1985). Instead of navigating through abstract menus, users experience a direct, unmediated connection to the generative components. This design challenges traditional command-based interaction, allowing users to "inhabit" a modular world where creative outcomes are the immediate result of physical-like manipulation. The Casual Creators framework specifically enumerates a list of desirable design patterns, of which our system directly employs several:

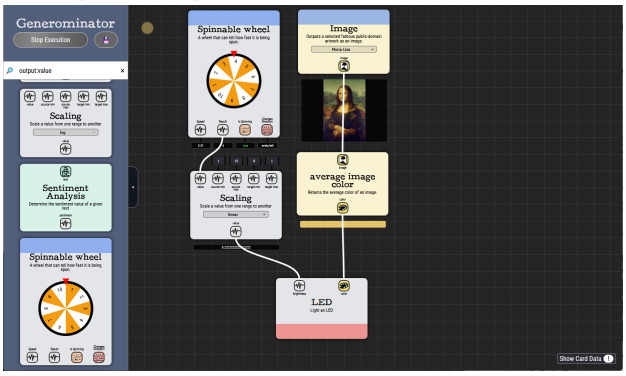
- **Instant feedback:** The core design philosophy centers on direct, card-based interaction. The Input/Output interfaces are engineered to provide "reactive electricity": instantaneous computational propagation that occurs the moment two cards are connected. Each card offers real-time feedback in the form of a direct preview of its outputs. On the other hand, by emulating the behavior of physical cards, the system significantly lowers the learning curve and effectively bridges the "Gulf of Evaluation" (Hutchins, Hollan, and Norman 1985), which is the cognitive gap between a user's intentions and the system's perceived state. In contrast, traditional creative suites like Photoshop or Maya bury their logic within deeply nested menus, forcing users into taxing mental simulation (Norman 2013). In the Generominator, the main interaction mode users are engaged with is adding cards to the graph by dragging new cards onto the canvas or establishing connections between cards.
- **Limiting actions to encourage exploration:** Instead of providing absolute freedom, the system deliberately restricts the user's action space. Each card is designed with a limited number of slots and specific data-type constraints, which a user can explore in a targeted way. Dragging a connection from an open port and releasing the mouse also shows a popup window listing only the cards that can connect to that port, and establishing such a connection immediately if one is selected, as shown in Figure 4. These constraints shift the user's cognitive focus from "what can I do" to "let me see which card fits here," effectively scaffolding the exploration process. By narrowing the possibility space, the system prevents the decision paralysis and anxiety often associated with excessive choice (Csikszentmihalyi 1990). Furthermore, as illustrated in Figure 3, the interface provides immediate negative feedback when an invalid connection is attempted.



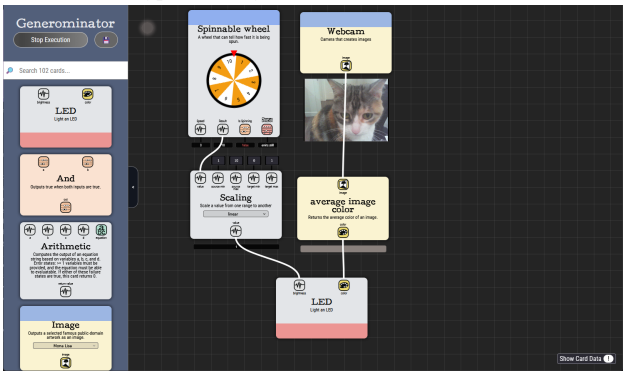
(a) The user can start their exploration by selecting any card, such as a simple LED output.



(b) The user can use the search function to browse cards that produce data of a type that they need.



(c) As the wheel is spun by the user, the halo effect (representing brightness) on the LED changes in real time.



(d) Cards can be replaced easily to explore other possible data sources or transformations.

Figure 2: A possible exploration sequence.

This mechanism ensures that the user remains within the "flow zone," where challenges are perfectly balanced with the system's structural guidance.

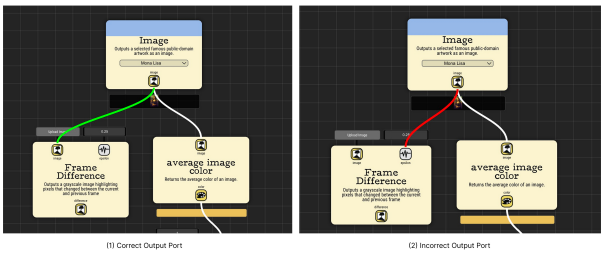


Figure 3: Attempting to connect invalid data types is disallowed and results in immediate feedback (red connection).

- **Mutant shopping:** Mutant Shopping serves as a mechanism to help users discover unexpected solutions within the vast possibility space. In our system, the card bank situated on the left side of the interface allows users to explore cards with diverse functionalities. Aforementioned suggestion box when the user drags from an unconnected port serves a similar purpose. Rather than drawing from a blank canvas, users "shop" from a series of generated functional cards with subtle variations. This shifts the creative act from active production to a process of browsing and recombination of existing processes. By leveraging the psychological pleasure and motivation inherent in non-directed discovery, this design alleviates the pressure to "produce" and encourages creators to find new opportunities in a state of freedom. This approach mirrors the collaborative discovery mechanism of Picbreeder (Secretan et al. 2011): users do not construct logic from scratch but instead "navigate" the possibility space, thereby significantly reducing cognitive load.

- **Saving and sharing:** As users explore our interface, they may design more and more complex graphs that incorporate a variety of different nodes in novel ways. We allow users to save and load graphs in a conventional JSON format, which can be edited in a text editor or from other scripts. However, we also support saving graphs to PNG files by using steganography, which embeds the graph data in the pixels of the PNG itself, while looking like an actual card with a label and (optional) description, as shown in figure 5. This allows graphs to be shared easily by posting the image on social media. To reimport the graph, the user just drags the image file onto the canvas,

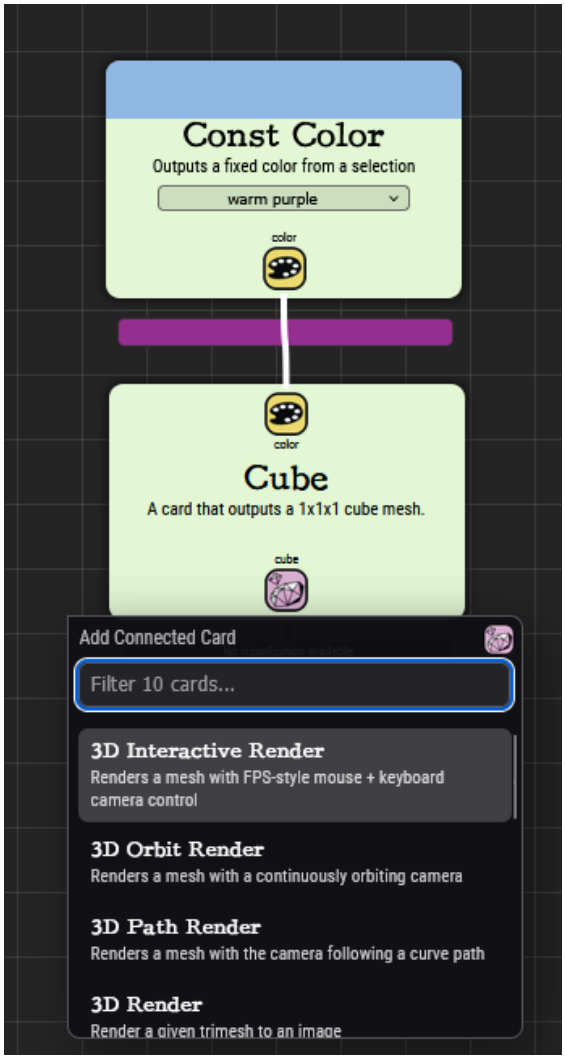


Figure 4: When the user drags a connection from an open port and releases it over the empty canvas, they are presented with the list of cards that are valid connection options. In this example, dragging from the unconnected mesh port will show the user all cards that can process mesh input. Filtering nodes on possible data types is a common feature in flow-based programming systems, and was directly inspired by how Compton's original editor filters nodes, as well as how Unity ShaderGraph presents the same filter operation.

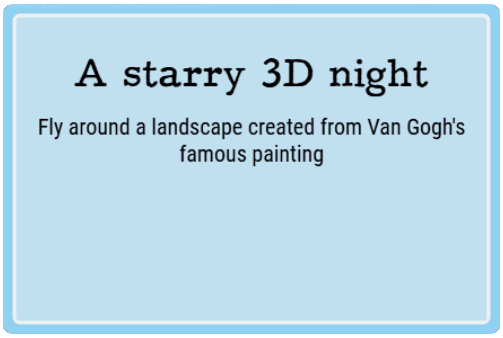


Figure 5: A graph exported to a PNG file looks like a card with a custom label, but the graph data remains embedded in the pixels of the image.

which will restore the embedded graph data.

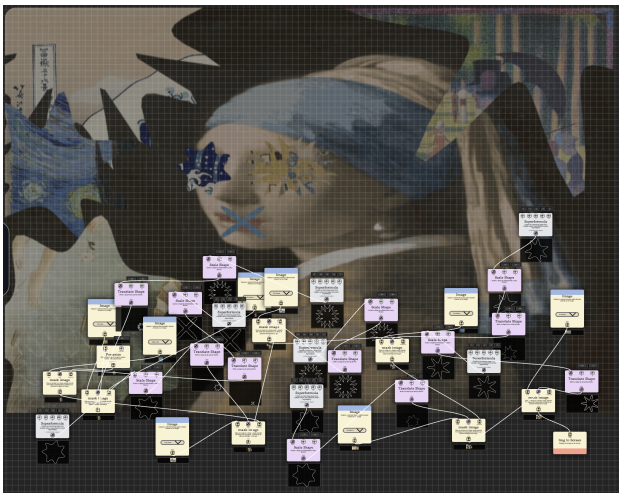
- **Modifying the meaningful:** While our system may not directly enforce this, the creation of complex graphs typically involves having some high level control parameters, which have ripple effects on other cards' behavior downstream. This way, the user can interact with the meaningful parameters of a generative system, and see their changes have an effect in real time.
- **Modding, hacking, teaching:** During the exploration process, user may discover that Generominator does not yet support every conceivable function. One core design goal was to make the system easily extensible, and advanced users can easily script their own card logic and share it with others. In fact, during the development process, we would often come up with new card ideas that expanded the creative space significantly but were trivial to implement, such as creating a duality between waveforms and vectors (handling sampling rate appropriately), which opens up the space of vector-processing cards for use with waveforms and vice versa. The card-based interaction tool is thus transformed from a "black box" into an extensible, dynamic ecosystem.

Examples

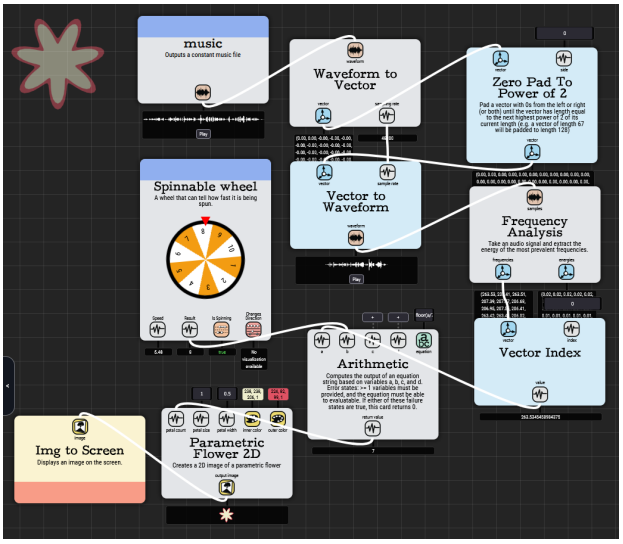
The raison d'être of our system is the design of graphs for creative expression. In this section, we want to showcase a variety of different graphs to demonstrate the range of artifacts that can be designed with our system. Figure 6 shows four example graphs and a preview of their output, where applicable. These graphs cover a wide range of data being processed, from images being masked, converted to heightmaps, or posterized, to audio-files driving parametric flowers. We also want to note that due to the arithmetic card and other capabilities of the system, it is possible to encode quite complex structures in our graphs, like the simple transformer model seen in figure 7 (with some modest custom additions to handle token-encoding).

Limitations

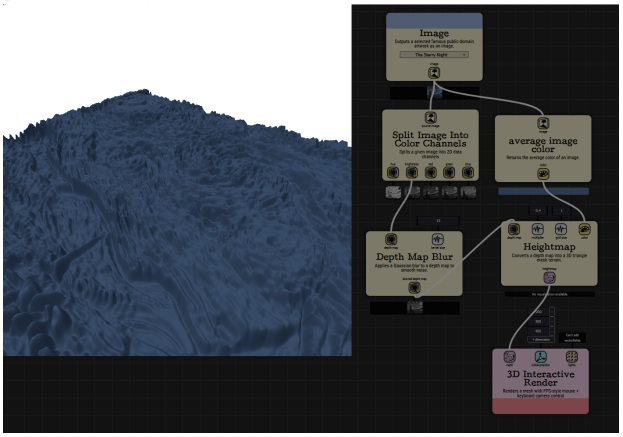
While our system enables a wide range of creative expression, several limitations exist within our work. As the most



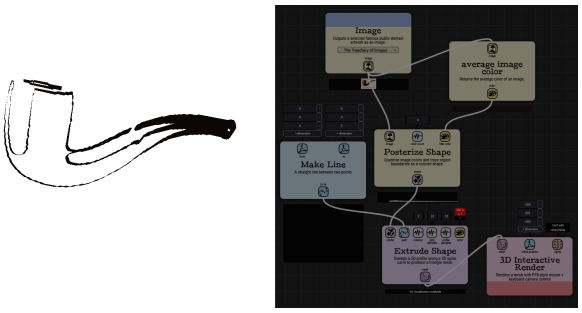
(a) Complex image masking and overlay operations can be composed arbitrarily.



(b) A parametric flower with a number of petals that depends on the predominant frequency in a given music file and the value the user spins on the wheel.



(c) Image color channels can be used to create heightmaps which can be rendered in 3D. In this case, the user can interact with the 3D scene using standard camera controls.



(d) Image posterization produces shapes which can be extruded to form 3D structures (and definitely not a pipe).

Figure 6: Four example graphs in the Generominator.

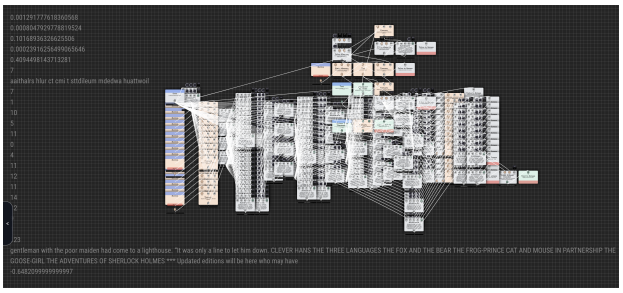


Figure 7: A transformer model implemented using the Generominator. Note that this example requires some custom nodes for token stream decoding.

notable deviation from the original Generominos, we currently do not support the *"voxel"* and *"particle"* data types from the original set. We purposefully omitted them due to their complex spatial requirements needed to represent these types of data visually, as well as the relative lack of cards that support them. Additionally, due to the nature of our system as a web-deployed application, there were several cards or modifiers with a physicality aspect described in the original work that we could not support. For example, Compton et al. (2017) describe modifiers that change which person performs a particular action, e.g. "the artist themself" or "a stranger", which does not have a meaningful interpretation in our execution environment. As another consequence of our web-based UI with the goal of being usable on a wide range of devices, we excluded cards that re-

quired specialized input or output devices (e.g. Kinect controllers). Overall, we implemented 64 of Compton et al.'s original cards, with the omissions falling into the aforementioned categories. However, we also added several cards that were necessary to construct meaningful graphs. The data types present in the Generominos are color-coded to indicate which types can be trivially converted between each other (Compton 2026). These trivial conversions themselves often need additional context, though, such as to define the sampling rate when converting an array of numbers to a waveform. To provide these transformations, and to make graph construction easier in general, we therefore also added several of our own cards to the system, bringing the total number of implemented cards up to over 100.

Additionally, due to resource constraints we were unable to conduct a formal user study to empirically evaluate artifacts. While earlier evaluations within an undergraduate student population exist for similar systems (Compton, Melcer, and Mateas 2017), we have not measured if a similar population would have differences in opinion regarding limitations in the previous study we have since resolved (e.g., missing construction cards, issues matching data types and conversion, etc.). Instead, we opted to evaluate our system against the Casual Creator design patterns, and present a wide range of potential output artifacts to showcase the flexibility of our system, and leave any claims regarding potential user interactions and perceptions of our system for future work.

Conclusion and Future Work

In this paper we have presented the Generominator, which is a web-based application that allows users to compose data flow graphs representing creative systems, based on the Generominos ideation cards by Compton et al. Each card in our system has several ports that may provide inputs and/or produce outputs, where a card can represent a data source like the keyboard, a random number, or predefined constants, a transformation, ranging from simple operations like extracting an element from a vector to more complex algorithms like a Fourier transform (Fourier 1822), sentiment analysis, or calculating Voronoi diagrams, or an output card that can show images, text, 3D meshes, etc. on the screen or play sound through the speakers. We provide a user interface that allows users to connect these cards in a guided fashion, and an execution engine to execute the graph and see the output on the screen in a live preview. Graphs can be saved, shared, and loaded as JSON files or even PNGs, allowing the easy exchange of graphs on social media.

While our implementation provides significant coverage of the originally presented Generominos, we also added several new cards that are necessary to actually be able to connect cards using different media by converting between them or extracting specific information. The way our system is designed makes it easily extensible, allowing rapid development of new cards to fill such gaps. Nevertheless, as users explore the system, the "wish list" for new cards keeps ever growing. While adding each card is straightforward for a developer, exploring how the creation of new cards could be made more approachable is an exciting avenue for future work. Additionally, our current approach is reliant on the UI for users to be able to preview their cards, but we would like to explore more direct ways for how users can share the artifacts they create as well.

Acknowledgements

The authors would like to extend special thanks to Kate Compton for creating the Generominos in the first place, and graciously providing additional context and feedback during the preparation of this paper for publication.

References

Compton, K., and Mateas, M. 2015. Casual creators. In *ICCC*, 228–235.

Compton, K.; Kybartas, B.; and Mateas, M. 2015. Tracery: an author-focused generative text tool. In *International Conference on Interactive Digital Storytelling*, 154–161. Springer.

Compton, K.; Melcer, E.; and Mateas, M. 2017. Generominos: Ideation cards for interactive generativity. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, 49–53.

Compton, K. 2026. Personal communication.

Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience*. New York: Harper & Row.

Eger, M. 2022. Instant architecture in minecraft using box-split grammars. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, 1–9.

Fourier, J. B. J. 1822. *Théorie analytique de la chaleur*. Firmin Didot.

Graves, J.; Royer, K.; Smith, G.; and Sullivan, A. 2021. Procedural patchwork: Community-focused generative design for quilting. In *Proceedings of the 13th Conference on Creativity and Cognition*, 1–3.

Heisserman, J. 1994. Generative geometric design. *IEEE Computer Graphics and Applications* 14(2):37–45.

Hutchins, E. L.; Hollan, J. D.; and Norman, D. A. 1985. Direct manipulation interfaces. *Human-Computer Interaction* 1(4):311–338.

Kahn, A. B. 1962. Topological sorting of large networks. *Commun. ACM* 5(11):558–562.

Kaufman, J. C., and Beghetto, R. A. 2009. Beyond big and little: The four c model of creativity. *Review of general psychology* 13(1):1–12.

Kreminski, M., and Mateas, M. 2021. Reflective creators. In *ICCC*, 309–318.

Kreminski, M.; Dickinson, M.; Osborn, J.; Summerville, A.; Mateas, M.; and Wardrip-Fruin, N. 2020. Germinate: A mixed-initiative casual creator for rhetorical games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, 102–108.

Masood, Z.; Usama, M.; Khan, S.; Kostas, K.; and Kaklis, P. D. 2024. Generative vs. non-generative models in engineering shape optimization. *Journal of Marine Science and Engineering* 12(4):566.

Mathewson, K., and Mirowski, P. 2017. Improvised theatre alongside artificial intelligences. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, 66–72.

Mazeika, J., and Whitehead, J. 2017. Solving for bespoke game assets: Applying style to 3d generative artifacts. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, 73–79.

Morrison, J. P. 2011. *Flow-based programming: a new approach to application development*. Unionville, Ont.: J. P. Morrison Enterprises, 2. ed edition.

Norman, D. 2013. *The Design of Everyday Things*. New York: Basic Books. Revised and expanded edition.

Schön, D. A. 2009. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.

Secretan, J.; Beato, N.; D'Ambrosio, D. B.; Rodriguez, A.; Campbell, A.; Folsom-Kovarik, J. T.; and Stanley, K. O. 2011. Picbreeder: A case study in collaborative evolutionary exploration of design space. 19(3):373–403.

Smith, G.; Whitehead, J.; and Mateas, M. 2010. Tanagra: A mixed-initiative level design tool. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, 209–216.

Sood, A.; Llano, M. T.; and McCormack, J. 2025. Do conversational interfaces limit creativity? exploring visual graph systems for creative writing. In *ICCC*.

Summerville, A.; Martens, C.; Samuel, B.; Osborn, J.; Wardrip-Fruin, N.; and Mateas, M. 2018. Gemini: Bidirectional generation and analysis of games via ASP. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, 123–129.

Tuffs, H. 2017. Poetry generation. In *Procedural Generation in Game Design*. AK Peters/CRC Press. 209–214.

Creative Transformation in Literary Texts: Modelling Change Across Representational Levels

Ioana-Roxana Boriceanu[1], Liviu P. Dinu[1,2]
[1]Faculty of Mathematics and Computer Science
[2]Human Language Technologies Research Center
University of Bucharest, Romania
ioana-roxana.boriceanu@s.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Creativity is often framed as the production of novelty, yet many cultural works emerge through transformation of earlier artifacts and not through isolated invention. Drawing on theories of imitation by Gabriel Tarde and James Mark Baldwin, this paper models creativity as selective transformation across multiple levels of textual representation. We introduce a multi-level framework that compares literary texts across lexical, semantic, conceptual, structural, and narrative dimensions using directional alignment and control calibrated similarity measures. Applying the model to historically documented literary relationships, we show that different pairs preserve source structure at different representational levels while diverging in others. These transformation profiles provide a quantitative method for characterizing how imitation persists and where creative divergence occurs within literary works.

Introduction

Creativity has long been a central concern in psychology, philosophy, and the arts, and has gained renewed urgency in the context of artificial intelligence, where questions of human distinctiveness and creative capacity are increasingly foregrounded (Boden 2016; Marcus and Davis 2019). A widely accepted formulation is the standard definition proposed in (Runco and Jaeger 2012), which holds that creativity requires both originality and effectiveness. Novelty alone is insufficient, as arbitrary variation may be original without being meaningful, while usefulness without originality reduces to routine competence.

Margaret Boden further clarified the relation between novelty and value in *The Creative Mind* (Boden 2004), distinguishing combinational, exploratory, and transformational creativity. In her account, creative production operates within structured spaces of possibility, and novelty emerges through transformations of those spaces instead of unconstrained deviation. Building on this perspective, computational creativity research has proposed formal frameworks for evaluating artificial artifacts (Maher 2010; Ritchie 2007; Franceschelli and Musolesi 2024). Many approaches measure novelty as distance from other artifacts within a class and value as performance relative to domain norms (Maher 2010). While these frameworks provide domain-independent evaluation criteria, they treat creativity as a property of an artifact relative to a distribution and not as a structured relation between specific works.

In many artistic domains, especially literature, new works do not simply occupy distant regions of a space of possibilities. They emerge through transformation of identifiable predecessors, retaining aspects of form, theme, or organization while reconfiguring others. Modelling creativity in such contexts therefore requires moving beyond class-level deviation toward an analysis of how representational structure is preserved and reshaped across works.

This relational view resonates with the social theory of Gabriel Tarde and the developmental psychology of James Mark Baldwin. In *The Laws of Imitation* (Tarde 1903), Tarde argued that social phenomena propagate through imitation, with variations introduced in transmission becoming the basis of innovation. Baldwin similarly described imitation as a recursive process through which repetition consolidates habits while enabling modification (Baldwin 1894). In both accounts, invention is not opposed to imitation but emerges from it. Novelty arises through selective transformation of inherited structure instead of originating from nothing.

Julia Kristeva, drawing on Bakhtinian dialogism, described every text as "a mosaic of quotations," framing intertextuality as the intersection of prior textual surfaces rather than the expression of a self-contained origin (Kristeva 1980). Dialogism here refers to the idea that every utterance is shaped by and answers to prior and anticipated utterances (Bakhtin 2010). Roland Barthes likewise emphasized the plurality of cultural codes that converge within a work, shifting attention from authorial origin to the network of textual traces that constitute it (Barthes 1977). Gérard Genette introduced analytic precision through his typology of transtextual relations, particularly his concept of hypertextuality, which foregrounds structured transformation between a later text and its precursor (Genette 1997). Across these accounts, literary production appears not as isolated invention but as patterned reconfiguration of prior forms.

We operationalise this insight through a layered framework for textual comparison. Instead of collapsing two texts into a single similarity score, we decompose comparison across five representational strata: lexical surface patterns, semantic similarity in embedding space, conceptual category distributions, structural organization, and narrative dynamics across the text. This design draws on

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

established traditions in stylistic analysis, stylometry, and influence detection (Burrows 2002; Reagan et al. 2016; Argamon et al. 2009; Burns et al. 2021; Xing 2025), which have long examined these features as distinct dimensions of textual relation. This framework integrates these dimensions into a unified relational model oriented away from authorship attribution and class-level similarity and toward modelling selective transformation between specific works.

We further hypothesise that creative transformation is characterised by non-trivial retention combined with substantive divergence. If retention is negligible, the relation between works is indistinguishable from random variation; if retention is near total, the result approaches reproduction. Creativity therefore occupies a peaked region of the retention spectrum, contingent upon the presence of meaningful divergence in one or more other representational channels. It arises not at the extremes of imitation or rupture, but where structured inheritance coexists with structured transformation.

This paper makes three contributions:

1. We introduce a multi-channel framework for modelling creative transformation between literary texts across lexical, semantic, conceptual, structural, and narrative levels.
2. We propose a directional alignment and comparison pipeline that measures retention and divergence between texts while accounting for structural variation in narrative progression.
3. We introduce a channelled aggregation mechanism that captures selective imitation by identifying the representational dimension in which structural inheritance is strongest while allowing other dimensions to diverge.

The framework is intended as a diagnostic model of literary transformation, not as a definitive classifier of literary influence. It combines directional alignment, multi-channel representation, control-calibrated comparison, and channelled aggregation to produce interpretable transformation profiles for historically motivated reference pairs.

Related Work

This work draws on two research areas: computational models of creativity, and computational approaches to stylometry and literary influence.

Computational Models of Creativity

Computational creativity research has undergone several conceptual shifts. Early work formalized creativity as structured search within constrained conceptual spaces, modelling combinational, exploratory, and transformational movement over rule-governed domains (Boden 2004; Wiggins 2006). Creativity was therefore treated primarily as a generative process operating within a space of possibilities. Later work shifted attention toward evaluation, proposing empirical criteria and standardized procedures for assessing novelty and value in creative artifacts (Ritchie 2007; Maher 2010; Jordanous 2012). Creativity increasingly became a property to be measured as well as a process to be described.

With the rise of large-scale machine learning, a distributional perspective emerged. Generative systems learn statistical regularities and produce outputs that deviate from learned patterns (Franceschelli and Musolesi 2022). Novelty has been framed as prediction violation (Grace and Maher 2019), as a balance between novelty and usefulness (Mukherjee and Chang 2023), or as a quantity difficult to stabilize across automatic metrics in large language models (Lu et al. 2025). In parallel, feature-based approaches operationalize creativity as deviation within a corpus, computing scalar scores from similarity networks, semantic distance, or temporal drift in feature space (Shrivastava et al. 2017; Orwig et al. 2025; Jimenez-Mavillard and Suarez 2022).

Stylometry, Authorship and Influence

Computational stylometry is grounded in the premise that authors exhibit statistically stable stylistic patterns. Early quantitative studies showed that high frequency function words can reliably distinguish authors (Mosteller and Wallace 1963). Later work expanded stylistic modelling to lexical, syntactic, and structural features such as n-grams, part-of-speech distributions, and vocabulary richness (Argamon et al. 2009). Distance-based measures, especially Burrows' Delta, further formalized stylistic similarity as deviation in frequency space (Burrows 2002).

With the adoption of machine learning, authorship attribution became increasingly framed as supervised classification. Stylometric features were combined with statistical and machine learning classifiers to scale across larger author sets and domains (Stamatatos 2009). Studies also examined the relative role of stylistic and topical signals (Sari, Stevenson, and Vlachos 2018), while later approaches combined stylometric features with classical models (Alshamasi and Menai 2022; Boriceanu and Băltoiu 2025) or used pretrained language models to detect stylometric differences (Barlas and Stamatatos 2020; Fabien et al. 2020).

Literary influence poses a more complex problem than authorship attribution. Instead of identifying the stylistic fingerprint of a single author, influence concerns relations between works produced by different authors and often separated by time, genre, or language. Early computational approaches focused on lexical overlap and text reuse, detecting quotation or phrase-level borrowing through n-gram similarity and related methods (Allen 2011; Forstall et al. 2015). Later work moved beyond surface overlap through semantic representations, embeddings, topic models, and latent semantic analysis, allowing texts to be compared through thematic or latent similarity even when wording differs (Burns et al. 2021; Manjavacasa, Karsdorp, and Kestemont 2020; Xing 2025; Blei, Ng, and Jordan 2003; Jockers and Mimno 2013). These approaches motivate the semantic channel of our framework, but our goal is not passage retrieval alone. We use semantic similarity as one component of a broader transformation profile.

Other work compares texts through stylistic, conceptual, and narrative structure. Function-word transition networks model style as relations between frequent grammatical markers (Segarra, Eisen, and Ribeiro 2015). WordNet supersenses and concreteness norms provide interpretable

Author	Work	Year	Literary Period	Tokens
Henry James	The Portrait of a Lady	1881	Late Victorian	233k
Henry James	Daisy Miller	1878	Late Victorian	22k
Edith Wharton	The House of Mirth	1905	American Realism	134k
Edith Wharton	The Age of Innocence	1920	Modernism / Realism	105k
George Eliot	Middlemarch	1871	Victorian Realism	324k
George Eliot	The Mill on the Floss	1860	Victorian Realism	214k
Jane Austen	Pride and Prejudice	1813	Regency	123k
Jane Austen	Emma	1815	Regency	165k
Thomas Hardy	Tess of the d'Urbervilles	1891	Late Victorian	155k
Oscar Wilde	The Picture of Dorian Gray	1890	Aestheticism	80k
Charles Dickens	David Copperfield	1850	Victorian	364k
Charles Dickens	A Tale of Two Cities	1859	Victorian	138k
Bram Stoker	Dracula	1897	Gothic / Late Victorian	163k
Gustave Flaubert	Madame Bovary (Eleanor Marx trans.)	1856	French Realism	118k
Gustave Flaubert	Salammbô (Eleanor Marx trans.)	1862	French Realism	107k
James Joyce	Dubliners	1914	Modernism	69k
James Joyce	Ulysses	1920	High Modernism	270k
Homer	The Odyssey (Alexander Pope trans.)	1725	Neoclassical Translation	115k
Homer	The Odyssey (Samuel Butler trans.)	1900	Victorian Translation	119k
Homer	The Iliad (Pope translation)	1715	Classical Epic	199k
Virginia Woolf	The Voyage Out	1915	Modernism	140k
Virginia Woolf	Mrs Dalloway	1925	High Modernism	65k
Mary Shelley	Frankenstein	1818	Romantic / Gothic	75k

Table 1: Corpus used in the experiments

Source Work	Target Work
The Portrait of a Lady	The House of Mirth
Pride and Prejudice	Middlemarch
Middlemarch	Tess of the d'Urbervilles
Odyssey (Pope)	Ulysses
Odyssey (Pope)	Mrs Dalloway
Odyssey (Butler)	Ulysses
Odyssey (Butler)	Mrs Dalloway
Odyssey (Pope)	Odyssey (Butler)
Madame Bovary	Dubliners
Madame Bovary	The Age of Innocence

Table 2: Reference text pairs based on historically documented literary relationships

conceptual categories (Ciaramita and Altun 2006; Brysbaert, Warriner, and Kuperman 2014), while digital humanities work has compared narrative dynamics through sentiment trajectories, pacing, and plot progression (Reagan et al. 2016). These strands motivate the structural, conceptual, and narrative channels of the model.

Finally, influence may involve compression, expansion, omission, or rearrangement, making direct passage-to-passage comparison insufficient. Dynamic time warping compares ordered sequences that differ in length or pacing while preserving trajectory structure (Berndt and Clifford 1994). Prior studies have usually examined lexical reuse, semantic similarity, stylistic structure, conceptual categories, or narrative trajectories separately. Our framework combines these signals in a single directional, control-calibrated comparison of source and target texts, using the resulting five-channel profile to characterize creative transformation as selective retention and divergence.

Dataset

The corpus consists of 23 literary works spanning the eighteenth to early twentieth centuries, including major texts from the English, American, and European literary canon. The dataset covers multiple literary periods, including Romanticism, Victorian realism, and high modernism, and contains approximately 3.5 million tokens in total. Table 1 lists the works included in the dataset together with their authors, publication dates, translators where applicable, literary periods, and approximate token counts after preprocessing.

The corpus was assembled to include both historically documented literary relationships and stylistically comparable but unrelated works. Instead of treating influence as a binary classification problem, the corpus serves as a comparative benchmark for analysing patterns of transformation between texts. Pairs of historically related works provide reference cases in which transformation is expected to occur, while unrelated comparisons provide a background distribution of stylistic similarity across the corpus.

To construct these reference comparisons, we define a set of text pairs based on well-established relationships discussed in literary scholarship. These include connections such as the influence of Henry James on Edith Wharton, Austen's influence on Eliot, Eliot's influence on Hardy, the influence of Flaubert on later realist and modernist writers, and the well-known engagement of Joyce and Woolf with Homeric epic. The dataset also includes comparisons between different translations of the *Odyssey*, allowing the framework to capture relationships that preserve narrative structure while substantially altering lexical expression. The selected reference pairs are shown in Table 2.

In addition to these reference comparisons, control pairs are constructed by pairing texts from the corpus that are not

historically linked through documented influence. These comparisons include both cross-tradition pairs and pairs drawn from similar literary periods. Together, these controls provide a background distribution of stylistic and structural similarity against which the transformation profiles of the reference pairs can be compared.

The texts were obtained from the Project Gutenberg digital library. Prior to analysis, non-content material was manually removed, including metadata such as source information, author descriptions, and editorial notes, retaining only the main body of the literary work. All texts are analysed in English. Non-English works are included through established English translations. The text was normalized through lowercasing, and tokenization was performed using a rule-based tokenizer.

For downstream comparison, texts were segmented into sequential units using one of two strategies: chapter boundaries or overlapping sliding windows. Overlapping sliding windows were constructed to reach a target size of approximately 2000 tokens, with successive windows advanced by a fixed step of 500 tokens. This segmentation allows structural patterns to be compared across works of different lengths while preserving local narrative organization.

Method: The Channelled Imitation Model

To operationalise the theoretical framework proposed by Gabriel Tarde and James Mark Baldwin, our pipeline models literary creativity as structured transformation of prior texts. In this view, new works emerge through selective preservation and modification of earlier textual structures. Instead of treating creativity as deviation from a statistical distribution of artifacts, the model measures how representational structures are retained or transformed between specific works.

The proposed framework therefore compares a source text and a derived text across multiple representational layers. Each channel captures a different level of textual organisation. Higher similarity within a channel indicates structural inheritance, while divergence reflects creative transformation. By examining patterns of retention and divergence across channels simultaneously, the model characterises how literary works inherit and reshape earlier forms.

Preprocessing and Directional Alignment

Before influence can be analysed, the texts must first be placed into a comparable structure. Because derived works frequently omit, reorder, or expand upon material from their sources, direct one-to-one alignment between passages is insufficient. A later work may compress an episode, elaborate on a minor event, or introduce entirely new sections while still maintaining a recognisable structural relationship to the original.

To account for these transformations, each text is segmented into sequential windows, as described in the previous section, and encoded using Sentence-BERT embeddings (specifically the `all-MiniLM-L6-v2` model) (Reimers and Gurevych 2019). Alignment between the two texts is then computed using Dynamic Time Warping (DTW) (Berndt and Clifford 1994), which allows flexible many-to-one or one-to-many correspondences between windows while preserving overall narrative order.

Channel	Representative features
Lexical	Token-frequency and content-word distributions, punctuation statistics, type–token ratio, inter-window lexical overlap
Structural	Function-word adjacency networks (density, degree distributions), part-of-speech transition patterns
Semantic	Sentence-BERT embeddings, LDA topic mixtures, LSA representations
Conceptual	Supersense categories, named-entity-type distributions, lexical concreteness
Narrative	Sentence-length pacing, sentiment trajectories, syntactic-complexity variation

Table 3: Feature layers used to capture each channel of textual transformation.

The alignment procedure is directional. Instead of treating the texts symmetrically, the algorithm anchors the derived text (B) to the source text (A), reflecting the theoretical assumption that influence flows from the earlier work toward the later one. To reduce matches between passages that occur at very different relative positions in the two texts, the DTW cost function incorporates a positional bias:

$$D(i,j) = d_{cosine}(A_i, B_j) + \lambda \left| \frac{i}{N} - \frac{j}{M} \right| \qquad (1)$$

where d_{cosine} is the cosine distance between embedding vectors for windows A_i and B_j, N and M denote the number of windows in the two texts, and λ controls the strength of the positional bias. The second term measures the difference between the normalized positions of the two windows within their respective texts. Alignments between windows that occur at very different relative narrative positions therefore incur a higher cost. This biases the alignment toward passages that occupy comparable positions in the two works while still allowing local expansions or contractions of narrative material. In all reported experiments we set $\lambda = 0.15$. Because both the cosine-distance and positional terms are bounded in $[0, 1]$, this caps the positional penalty at roughly 15% of the maximum semantic cost, ensuring it acts as a tie-breaker rather than overriding semantic correspondence. To keep alignments locally monotonic, the DTW search is restricted to a band proportional to the length of the longer sequence, and the traceback is biased toward diagonal steps so that each source window receives a corresponding match.

Multi-Level Representational Channels

Once the texts are aligned, their relationship is analysed across five representational channels corresponding to different levels of textual organisation. Table 3 presents the feature layers extracted for each channel.

The lexical channel captures surface-level stylistic similarity. Each window is represented as a normalized token

frequency distribution. Divergence between the source distribution (Q) and derived distribution (P) is measured using a dampened Kullback–Leibler divergence:

$$D_{KL}^{\text{damp}}(P \parallel Q) = \ln\left(1 + \max(0, S)\right) \tag{2}$$

$$S = \sum_{x} P(x) \ln \frac{P(x)}{Q(x)} \tag{3}$$

This formulation measures lexical novelty while preventing extreme divergence values when the derived text introduces entirely new vocabulary.

The structural channel models stylistic organisation independent of topical content. Function word adjacency networks (Segarra, Eisen, and Ribeiro 2015) are constructed from sequences of stop words within each window. Nodes correspond to function words, while edges represent transitions between successive function words in the text.

From these networks we compute structural metrics including network density and node degree distributions, capturing the global topology of function word usage. Structural similarity between aligned windows is evaluated using cosine similarity and Jensen–Shannon divergence over these distributions. Because function words are less topic dependent than content vocabulary, these measures provide a useful proxy for stylistic and structural inheritance.

The semantic channel captures thematic similarity through dense latent representations of meaning. Each window is encoded using contextual embeddings derived from Sentence-BERT, producing dense vector representations optimised for semantic similarity. Pairwise similarity between aligned windows is measured using cosine similarity between embedding vectors.

To complement embedding similarity, we also compute thematic similarity using topic modelling and latent semantic analysis (LSA). Topic distributions are obtained with Latent Dirichlet Allocation (LDA) using 50 topics, while LSA representations are derived from a truncated singular value decomposition of the term–document matrix with 100 dimensions. We use LDA because the corpus is relatively small and because topic mixtures provide an interpretable way to compare themes across aligned windows. The number of topics was fixed before evaluating the reference pairs, and the semantic channel combines LDA with other signals, so the results are not based on a single topic-model configuration. Together, these representations capture thematic similarity even when surface wording differs substantially.

The conceptual channel captures persistence of semantic categories that may remain stable even when lexical expression and sentence structure change. While the semantic channel models similarity in dense latent representations of meaning, the conceptual channel compares distributions of interpretable semantic categories across the text.

Three types of signals are used: lexical supersense categories (Ciaramita and Altun 2006), named entity type distributions, and lexical concreteness scores (Brysbaert, Warriner, and Kuperman 2014). These distributions provide a general representation of the conceptual landscape of a text, enabling the model to detect continuity of semantic domains even when vocabulary diverges substantially.

The narrative channel models large-scale narrative organisation by analysing how stylistic and emotional signals evolve across the text. This layer captures patterns that unfold across sequences of windows, including sentence-length based pacing, sentiment variation, and changes in structural complexity, instead of focusing only on individual sentences.

For each text, these signals are represented as time series defined over the ordered sequence of windows. For example, sentence-length statistics provide a proxy for narrative pacing, while sentiment scores capture shifts in emotional tone (Reagan et al. 2016). Similarity between the resulting narrative trajectories is then computed using correlation and cosine similarity between the corresponding time series. This representation allows the model to measure whether two texts exhibit similar narrative dynamics, such as comparable pacing patterns or sentiment arcs, even when the specific events or wording differ.

Discriminative Feature Weighting

Literary works produced within the same historical tradition often share stylistic conventions that do not reflect direct influence. To distinguish genuine textual inheritance from background similarity, the model applies a discriminative feature weighting procedure. For each feature f, a weight is computed as

$$w_f = \max\left(0.01, \frac{\sigma_{tradition}(f)}{\sigma_{global}(f) + \epsilon}\right) \tag{4}$$

where $\sigma_{tradition}$ measures feature variance among texts belonging to the same literary tradition, assigned by literary and national grouping such as English Victorian, French realist, or classical epic, and σ_{global} measures variance across the entire corpus. The constant ϵ is a small positive value added for numerical stability to prevent division by zero when global variance is extremely small. Features that exhibit higher variance within a tradition relative to their global variance receive higher weight.

These weights are used to compute standardized feature scores relative to tradition-matched control pairs. The resulting weighted score provides a calibrated estimate of transformation-based influence.

Channelled Aggregation of Influence

The per-channel scores z_c used in this section are computed from the discriminatively weighted features described above, standardized against tradition-matched control pairs. Because we hypothesise that creative transformation occurs selectively across representational levels, influence is not expected to appear uniformly across all channels. A derived work may preserve narrative structure while transforming lexical expression, or maintain conceptual themes while radically altering stylistic form.

Each representational channel therefore produces an influence signal measuring the degree to which that level of structure is retained between the aligned texts. Let z_c denote

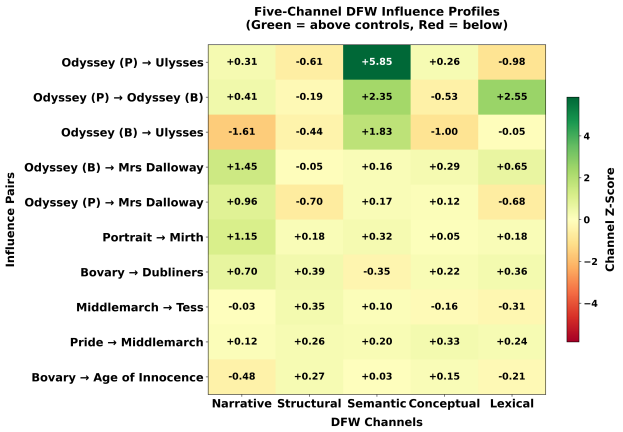


Figure 1: Five-Channel Transformation Profiles

the standardized similarity score for channel c. The resulting five-dimensional vector

$$\mathbf{z} = (z_{narr}, z_{struct}, z_{sem}, z_{concept}, z_{lex}) \tag{5}$$

represents the influence profile between the two works.

The model identifies the strongest retained structure as the primary locus of inheritance:

$$z_{peak} = \max_{c}(z_c) \tag{6}$$

Creative transformation often involves strong retention in one channel combined with divergence in others. To capture this pattern, we introduce a diagnostic scoring function that highlights channelled inheritance across representational levels:

$$I = \max(0, z_{peak})(1 + \alpha n_{neg}) \tag{7}$$

where n_{neg} is the number of channels whose standardized score falls below a divergence threshold of $z_c < -0.5$, corresponding to half a standard deviation below the control mean, and α is a scaling constant set to 0.20. The score is primarily determined by the strength of the dominant channel, with the second factor providing a bounded adjustment for the sharpness of the channelling pattern. Because at most four of the five channels can diverge while one is retained, the maximum bonus is $1+0.20\times4 = 1.80$, ensuring that the divergence term modulates the score without dominating it. This design reflects the theoretical expectation that creative transformation is characterised by a distinctive configuration in which strong inheritance in a specific representational dimension coexists with departure from the source in others, instead of by uniform similarity or uniform distance. The scoring function should therefore be interpreted as a diagnostic of selective transformation rather than as a measure of overall similarity between texts.

Experiments and Results

This section evaluates whether the framework produces distinctive transformation patterns for historically related texts compared to unrelated corpus pairs. Instead of treating influence as binary detection, we examine how retention and divergence are distributed across the five representational dimensions introduced above.

Group	n	Mean	Median	Std. Dev.
Reference pairs	10	1.94	1.25	2.39
Control pairs	73	0.29	0.19	0.30

Table 4: Aggregate channelled transformation scores for reference and control pairs

The model is applied to the 10 reference relationships in Table 2 and 73 unrelated control pairs constructed from source → target orderings. For instance, *Pride and Prejudice* → *Mrs Dalloway* is retained as a control because Austen predates Woolf, while the reverse direction is excluded as historically impossible. For each pair, the framework computes channel-specific similarity scores and the aggregate transformation score defined in Equation 7.

To assess robustness to segmentation, we run the full pipeline with both chapter-based segmentation and overlapping sliding windows. Reference-control separation is directionally consistent in both settings, but stronger with sliding windows ($r = 0.74$, $p < .001$) than with chapter segmentation ($r = 0.35$, $p = .07$). The remaining analysis therefore reports the sliding-window setting.

The evaluation asks whether historically related works produce stronger transformation signals than controls, and how retention differs across representational layers. Since the model allows selective inheritance, different pairs are not expected to preserve the same kinds of structure. We therefore consider both aggregate score magnitude and the distribution of channel-specific signals.

Results

Table 4 summarizes the transformation scores. Reference pairs obtain a higher mean than controls (1.94 versus 0.29) and show substantially greater variance. A Mann–Whitney U test confirms a significant separation ($U = 635$, $p < 0.001$, rank-biserial $r = 0.74$), indicating that reference pairs generally receive higher scores than controls. The larger variance among reference pairs suggests that influence does not manifest uniformly, but through different retained layers across different relationships.

Figure 1 presents the five-channel profiles for the reference pairs. Each row shows standardized channel scores relative to the control distribution, where positive values indicate retention and negative values indicate divergence. The profiles show that historically related pairs do not simply display uniformly higher similarity. Instead, they tend to activate one or two representational dimensions while diverging in others. This supports the central claim that creative transformation is best understood as selective inheritance across representational levels.

Comparison with Single-Channel Baselines

To assess whether channelled aggregation adds information beyond individual dimensions, we compare the full

Signal	r	AUC	p
Lexical overlap (text reuse) (Allen 2011; Forstall et al. 2015)	-0.04	0.48	.86
POS-trigram structure	0.09	0.55	.64
TF–IDF content similarity	0.10	0.55	.60
LDA topic mixture	0.30	0.65	.13
Sentence-BERT embedding (Burns et al. 2021; Xing 2025)	0.41	0.71	.04
Channelled model (ours)	$\mathbf{0.74}$	**0.87**	$< \mathbf{.001}$

Table 5: Reference versus control separation for single-channel baselines and the full channelled model under the sliding-window setting. The evaluation includes 10 reference pairs and 73 control pairs. r is the rank-biserial effect size, AUC is the probability that a reference pair receives a higher score than a control pair, and p is computed using a two-sided Mann–Whitney U test.

model with several single-channel baselines. These approximate common approaches to textual relation, including lexical overlap and content similarity used in text-reuse and intertextuality detection (Allen 2011; Forstall et al. 2015), as well as dense embedding similarity used in recent embedding-based intertextual analysis (Burns et al. 2021; Xing 2025). For each baseline, per-window scores are aggregated to the pair level and evaluated by their ability to separate reference pairs from controls.

Table 5 shows that no individual channel separates the groups as strongly as the full model. Lexical overlap, POS-trigram structure, and TF–IDF content similarity remain close to chance (AUC ≤ 0.55). LDA topic mixtures provide a stronger but non-significant signal, while Sentence-BERT embeddings are the strongest single-channel baseline ($r = 0.41$, AUC $= 0.71$). The channelled model achieves a larger separation ($r = 0.74$, AUC $= 0.87$), suggesting that the relevant signal lies in the configuration of retention and divergence across channels, not in high similarity along one dimension alone.

Reference Pair Case Studies

Several transformation patterns emerge among the reference pairs. The two English translations of the *Odyssey* show strong semantic and lexical retention, reflecting preservation of narrative content under different stylistic realizations. Pope's version renders the epic in heroic couplet verse, while Butler's presents it in unrhymed prose, introducing differences in rhythm, sentence segmentation, and pacing while leaving the underlying semantic structure largely intact.

The Homeric-modernist comparisons show a different configuration. In Joyce's *Ulysses*, which does not reproduce the *Odyssey* at the level of wording or surface style, the strongest signal appears in the semantic channel. This suggests mediated thematic continuity combined with substantial lexical and structural divergence. Since the aggregate score rewards strong retention in one channel together with divergence in others, this pair scores higher than the more uniformly similar translation pair.

The Woolf comparisons are more narrative-oriented. The Butler *Odyssey* $\rightarrow$ *Mrs Dalloway* pair produces a stronger signal than the corresponding Pope comparison, with a peak in the narrative channel. Because Butler's translation uses prose instead of heroic verse, its form is closer to the narrative conventions Woolf transforms. The stronger Butler signal may therefore reflect mediation through translation, corresponding to hypertextual transformation in Genette's terms.

Relationships within later prose traditions produce further configurations. The Henry James $\rightarrow$ Edith Wharton comparison shows elevated narrative and semantic signals, consistent with Wharton's engagement with Jamesian focalization and psychological interiority. The Eliot $\rightarrow$ Hardy comparison peaks in the structural channel, likely reflecting continuity in Victorian realist narration and syntactically complex prose without direct content borrowing.

The Flaubert $\rightarrow$ Joyce comparison is more heterogeneous, with narrative activation and moderate structural and conceptual signals. This fits Flaubert's role in the development of modern narrative technique, especially free indirect discourse and subjective consciousness, which Joyce extends within a modernist context. Other reference pairs, such as *Madame Bovary* $\rightarrow$ *The Age of Innocence*, show weaker but interpretable structural and conceptual continuity, suggesting that influence may also appear as distributed partial retention rather than a single dominant channel.

Reference Pairs and Controls

Table 6 presents representative profiles for reference and control comparisons. Reference pairs more often exhibit a pronounced dominant channel, while controls tend to show weaker or more diffuse similarity across channels.

This contrast is important because many controls are not maximally distant comparisons. They often come from similar periods, related prose traditions, or overlapping stylistic environments, and therefore share broad conventions of narration, syntax, and theme. Even against this stronger baseline, reference pairs more often display a clear dominant channel or coherent retained structure.

The full ranking across all 83 pairs shows the same tendency: the six highest transformation scores are all reference pairs, and seven of the ten reference pairs fall within the top 20. The three lower-ranking reference pairs have flatter profiles, suggesting that some realist influence may be distributed weakly across several channels. Overall, the comparison supports the view that creative transformation

Pair	Type	Score	Narr	Struct	Sem	Concept	Lex	Peak
Odyssey (Pope) → Ulysses	Influence	8.19	0.31	-0.61	5.85	0.26	-0.98	Semantic
Odyssey (Pope) → Odyssey (Butler)	Influence	3.06	0.41	-0.19	2.35	-0.53	2.55	Lexical
Odyssey (Butler) → Ulysses	Influence	2.56	-1.61	-0.44	1.83	-1.00	-0.05	Semantic
The Portrait of a Lady → The House of Mirth	Influence	1.15	1.15	0.18	0.32	0.05	0.18	Narrative
Middlemarch → Tess of the d'Urbervilles	Influence	0.35	-0.03	0.35	0.10	-0.16	-0.31	Structural
Pride and Prejudice → Middlemarch	Influence	0.33	0.12	0.26	0.20	0.33	0.24	Conceptual
Middlemarch → The Portrait of a Lady	Control	0.60	0.60	0.02	0.09	0.08	0.25	Narrative
Pride and Prejudice → The Voyage Out	Control	0.81	0.81	0.39	0.17	0.50	0.69	Narrative
Madame Bovary → Tess of the d'Urbervilles	Control	0.93	0.71	0.93	0.03	0.18	-0.14	Structural
Madame Bovary → The House of Mirth	Control	0.74	0.09	0.61	-0.10	-0.06	-0.53	Structural

Table 6: Representative transformation profiles for influence and control pairs under the chronologically filtered control setting

operates through selective retention and divergence, not uniform similarity.

Limitations

The present study is limited by the size and composition of its corpus, which focuses on a small set of canonical works from the eighteenth to early twentieth centuries. It should therefore be understood as an exploratory evaluation setting for developing the model, not as a representative sample of literary creativity in general. Other forms of influence, including diffuse genre conventions, multilingual transmission, unconscious stylistic inheritance, or mediation through broader cultural movements, may require larger datasets and additional representational channels.

The reference pairs are based on relationships discussed in literary scholarship rather than formally verifiable ground truth, and should be interpreted as historically plausible cases of transformation, not definitive instances of influence. Likewise, the model's representational channels are simplified proxies for complex literary phenomena: narrative structure, conceptual domains, and stylistic organization cannot be fully captured by the features used here. The resulting profiles should therefore be read as interpretable evidence of structural relation and as prompts for further literary analysis, not as replacements for close reading. Finally, because the framework assumes a plausible precursor, it is less applicable to works whose originality lies mainly in departing from a broader field; such cases may be better characterized by distributional novelty measures, making genealogical transformation and corpus-level novelty complementary perspectives.

Conclusion

This paper introduced a multi-channel framework for modelling creative transformation between literary texts. Instead of treating creativity as deviation within a corpus, the proposed approach analyses how representational structures are selectively preserved or transformed across lexical, semantic, conceptual, structural, and narrative levels.

Applying the framework to a corpus of canonical literary works shows that historically related texts tend to exhibit distinctive transformation profiles. Different relationships preserve different layers of structure, ranging from narrative organization in closely related prose traditions to semantic or conceptual continuity in more distant reinterpretations such as modernist engagements with classical epic. These findings support the view that literary creativity often operates through selective inheritance, with similarity concentrated in specific representational layers instead of appearing uniformly across texts.

More broadly, the results suggest that modelling influence as a multi-layered relation between specific works offers a complementary perspective to existing distribution-based approaches to computational creativity. By decomposing textual comparison across representational channels, the framework provides a way to analyse how creative works simultaneously inherit and transform earlier structures.

The framework may also be useful for evaluating creative text generation. For example, systems prompted to imitate an author or transform an existing narrative may reproduce surface style without preserving deeper semantic, conceptual, or narrative structure. A multi-channel transformation profile could help distinguish shallow stylistic imitation from more substantive transformation of source material.

Future work may extend this approach to larger corpora, additional literary traditions, multilingual settings, and other cultural domains where creativity emerges through structured transformation of prior artifacts.

Acknowledgements

The authors thank the Romanian writer Ștefan Agopian, whose reflections on literature and the creative process inspired the central idea of this work. They are also grateful to Professor Andra Băltoiu, whose suggestions on stylistic influence and creative transformation helped shape the initial direction of this work.

This work was supported by Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project SIROLA, number PN-IV-P1-PCE-20231701, within PNCDI IV.

References

Allen, G. 2011. *Intertextuality*. routledge.

Alshamasi, S., and Menai, M. 2022. Ensemble-based clustering for writing style change detection in multi-authored textual documents. In *CLEF (Working Notes)*, 2357–2374.

Argamon, S.; Koppel, M.; Pennebaker, J. W.; and Schler, J. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM* 52(2):119–123.

Bakhtin, M. M. 2010. *The Dialogic Imagination: Four Essays*, volume 1. University of Texas Press.

Baldwin, J. M. 1894. Imitation: A chapter in the natural history of consciousness. *Mind* 3(9):26–55.

Barlas, G., and Stamatatos, E. 2020. Cross-domain authorship attribution using pre-trained language models. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 255–266. Springer.

Barthes, R. 1977. *Image-music-text*, volume 6135. Macmillan.

Berndt, D. J., and Clifford, J. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, 359–370.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Boden, M. A. 2016. *AI: Its nature and future*. Oxford University Press.

Boriceanu, I., and Băltoiu, A. 2025. Style change detection using graph and structural-linguistic features for multi-author writing analysis. *Working Notes of CLEF*.

Brysbaert, M.; Warriner, A. B.; and Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.

Burns, P. J.; Brofos, J. A.; Li, K.; Chaudhuri, P.; and Dexter, J. P. 2021. Profiling of intertextuality in latin literature using word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4900–4907.

Burrows, J. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing* 17(3):267–287.

Ciaramita, M., and Altun, Y. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 594–602.

Fabien, M.; Villatoro-Tello, E.; Motlicek, P.; and Parida, S. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 127–137.

Forstall, C.; Coffee, N.; Buck, T.; Roache, K.; and Jacobson, S. 2015. Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching. *Digital Scholarship in the Humanities* 30(4):503–515.

Franceschelli, G., and Musolesi, M. 2022. Deepcreativity: measuring creativity with deep learning techniques. *Intelligenza Artificiale* 16(2):151–163.

Franceschelli, G., and Musolesi, M. 2024. Creativity and machine learning: A survey. *ACM Computing Surveys* 56(11):1–41.

Genette, G. 1997. *Palimpsests: Literature in the second degree*, volume 8. U of Nebraska Press.

Grace, K., and Maher, M. L. 2019. Expectation-based models of novelty for evaluating computational creativity. In *Computational creativity: The philosophy and engineering of autonomously creative systems*. Springer. 195–209.

Jimenez-Mavillard, A., and Suarez, J. L. 2022. A computational approach for creativity assessment of culinary products: The case of elbulli. *AI & SOCIETY* 37(1):331–353.

Jockers, M. L., and Mimno, D. 2013. Significant themes in 19th-century literature. *Poetics* 41(6):750–769.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Kristeva, J. 1980. Word, dialogue, and novel. desire in language: A semiotic approach to literature and art. *T. Gora et al*.

Lu, L.-C.; Liu, M.; Lu, P.-C.; Tian, Y.; Sun, S.-H.; and Peng, N. 2025. Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. *arXiv preprint arXiv:2508.05470*.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, 22–28.

Manjavacasa, E.; Karsdorp, F.; and Kestemont, M. 2020. A statistical foray into contextual aspects of intertextuality. *Proceedings http://ceur-ws. org ISSN* 1613:0073.

Marcus, G., and Davis, E. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books.

Mosteller, F., and Wallace, D. L. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association* 58(302):275–309.

Mukherjee, A., and Chang, H. 2023. The creative frontier of generative ai: Managing the novelty-usefulness tradeoff. *arXiv preprint arXiv:2306.03601*.

Orwig, W.; Luchini, S. A.; Beaty, R. E.; and Schacter, D. L. 2025. A 'sweet spot'for creative ideation: Non-linear associations between semantic distance and creativity. *The Journal of Creative Behavior* 59(3):e70041.

Reagan, A. J.; Mitchell, L.; Kiley, D.; Danforth, C. M.; and Dodds, P. S. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science* 5(1):31.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal* 24(1):92–96.

Sari, Y.; Stevenson, M.; and Vlachos, A. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th international conference on computational linguistics*, 343–353.

Segarra, S.; Eisen, M.; and Ribeiro, A. 2015. Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing* 63(20):5464–5478.

Shrivastava, D.; CG, S. A.; Laha, A.; and Sankaranarayanan, K. 2017. A machine learning approach for evaluating creative artifacts. *arXiv preprint arXiv:1707.05499*.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60(3):538–556.

Tarde, G. 1903. *The laws of imitation*. H. Holt.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-based systems* 19(7):449–458.

Xing, Y. 2025. Modelling intertextuality with n-gram embeddings. *arXiv preprint arXiv:2509.06637*.

Who Can Scan a Verse? Evaluating Open and Closed Language Models for Portuguese Poetic Scansion

Filipe Calegario, André Valença
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
{fcac, aaav}@cin.ufpe.br

Abstract

Automatic poetic scansion, segmenting verses into metrical syllables and identifying stress patterns, is a key component for meter-aware computational poetry systems, yet it remains particularly challenging for Portuguese due to complex phonological phenomena. Previous work fine-tuned GPT-3.5 on this task (88.6% syllable ratio) but was limited to a single proprietary model. This paper presents the first systematic evaluation of 24 language models for Portuguese scansion in terms of accuracy, latency, and cost, with an additional pilot study of 9 reasoning models. We benchmark proprietary, open-weight, and Portuguese-specialized models in zero-shot and few-shot settings, and compare them against fine-tuned GPT-4.1 variants. We also introduce a stricter exact-match metric and format-tolerant metrics that together reveal substantially more phonological competence than prior evaluation suggested. Our key findings are: (1) fine-tuning remains essential, with GPT-4.1 variants surpassing the GPT-3.5 baseline on the same test set, and diminishing returns above a minimum model capacity; (2) in a pilot study, all nine reasoning models achieve 0% exact match, failing through distinct mechanisms (output truncation, empty responses, incorrect notation); and (3) Portuguese-specialized models underperform general-purpose alternatives, a gap that reflects weaker instruction-following more than weaker phonological competence. We discuss implications for computational creativity and accessibility pathways for communities that cannot depend on centralized commercial infrastructure.

Introduction

Poetic scansion, the division of verse into metrical syllables and the identification of stress patterns, is a foundational task for computational poetry systems (Oliveira 2017). It underpins meter classification, rhyme detection, and poetry generation, yet remains particularly challenging for Portuguese due to the language's rich phonological phenomena: synalepha, elision, diphthong reduction, and other metaplasms that alter syllable boundaries between the written and spoken forms of verse (Bisol 2005; Ali 1999).

Recent work has shown that a fine-tuned GPT-3.5-turbo can reach 88.6% syllable-level accuracy on Portuguese scansion when evaluated against the Aoidos rule-based corpus (Valença and Calegario 2025). However, that result was limited to a single model family and did not explore the rapidly expanding landscape of open-weight, multilingual, and Portuguese-specialized alternatives.

This paper presents a systematic evaluation of 24 language models for Portuguese poetic scansion across three axes: *accuracy*, *latency*, and *cost*, complemented by a pilot study of 9 reasoning models. We benchmark proprietary, open-weight, and Portuguese-specialized models in zero-shot and few-shot settings, and compare them against fine-tuned GPT-4.1 variants. Beyond the original syllable-ratio metric, we introduce both a stricter exact-match measure (SYL_EM%) and format-tolerant evaluation measures, including syllable boundary F1 (BND_F1), stress accuracy (STR_ACC), and metaplasm detection rate (META%), enabling finer-grained comparison even when models produce outputs that differ in surface formatting.

This work is part of a broader effort to build reliable analysis components that can underpin meter-aware poetry generation: a system that can accurately scan a verse is a prerequisite for one that can compose metrically valid verses. This is especially relevant for traditions such as Northeastern Brazilian popular poetry, where strict metric and rhyme conventions govern improvised oral performance.

Our contributions are threefold. (1) We present the first comparative benchmark of language models for Portuguese scansion (24 models across accuracy, latency, and cost, plus a pilot of 9 reasoning models), and introduce a stricter exact-match metric (SYL_EM%) and format-tolerant metrics (BND_F1, STR_ACC, META%) that reveal substantially more phonological competence than the syllable-ratio metric used in prior work. This reusable evaluation framework establishes that fine-tuning remains essential (the best fine-tuned model reaches 93.9% syllable ratio versus 74.1% for the best zero-shot model), with GPT-4.1 variants confirmed as a genuine improvement over the re-evaluated GPT-3.5 baseline. (2) We provide preliminary empirical evidence that extended reasoning can be counterproductive for format-sensitive linguistic tasks: in a pilot study (N=10), all 9 reasoning models reach 0% exact match, and we characterize three distinct failure mechanisms (output truncation, empty or leaked responses, and non-standard notation) that inform whether and how reasoning models should be de-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

ployed in computational poetry pipelines. (3) We map the cost-accuracy trade-off across all model categories and find that Portuguese-specialized models underperform general-purpose alternatives, surfacing accessibility and linguistic-sovereignty concerns for communities that cannot depend on centralized commercial infrastructure.

Related Work

This section reviews prior work on automatic poetic scansion across languages, the use of large language models for poetry tasks, and the landscape of language models for Portuguese. Together, these strands reveal a significant gap: while computational scansion tools and LLM-based poetry evaluations have advanced considerably for several languages, Portuguese remains notably underserved.

Computational Scansion

Rule-based approaches have a long history in computational scansion, though unevenly distributed across languages. For English, early work dates back to Logan (1988) and Hayward (1996), who applied statistical and connectionist models to metrical analysis. Later systems such as Scandroid (Hartman 2005) and ZeuScansion (Agirrezabal et al. 2016), which combines finite-state technology with phonological dictionaries, achieved 86.78% per-syllable accuracy. Greene et al. (2010) used weighted finite-state transducers for rhythmic analysis with applications to poetry generation and translation.

For Spanish, Gervás (2000) proposed the first rule-based scansion system. Navarro-Colorado (Navarro-Colorado 2017) developed a fixed-metre scansion system and created the ADSO corpus of annotated Golden Age sonnets (Navarro-Colorado, Ribes Lafoz, and Sánchez 2016) with 96% inter-annotator agreement, enabling systematic benchmarking. De la Rosa et al. (2020) developed Rantanplan, the current state-of-the-art, achieving 96.23% accuracy on hendecasyllabic verse. Marco Remón and Gonzalo (2020) later demonstrated that scansion without syllabification can outperform traditional pipelines while running over 20 times faster.

For Portuguese, the landscape is notably sparse. Araújo and Mamede (2002) presented an early poem classifier in 2002, but the only comprehensive scansion system is Aoidos (Mittmann 2016; Mittmann and Maia 2017), which applies phonological rules to segment verses into poetic syllables and identify stress patterns. Validated on over 100,000 verses spanning three centuries, Aoidos provides the rule-based scansions that serve as ground truth. Building on this resource, our own prior work (Valença and Calegario 2025) organized these scansions into a structured dataset with stress annotations and fine-tuned GPT-3.5-turbo on it, achieving 88.6% syllable ratio and establishing the baseline and evaluation framework that the present work extends.

This asymmetry is striking: Spanish has at least three dedicated scansion systems, English has several, and even Basque (Agirrezabal et al. 2012) and German (Bobenhausen 2011) have specialized tools. The most comprehensive cross-lingual scansion studies (Agirrezabal's multilingual thesis (2017) and De la Rosa et al.'s transformer experiments across Spanish, English, and German (2023)) did not include Portuguese, despite it being spoken by over 250 million people.

LLMs for Linguistic and Poetic Tasks

Large language models have demonstrated remarkable few-shot capabilities across linguistic tasks (Brown and others 2020), and chain-of-thought prompting has been shown to improve performance on reasoning-intensive problems (Wei et al. 2022). However, phonologically demanding tasks like scansion require precise character-level manipulation that challenges autoregressive generation, especially for languages underrepresented in training corpora. Belouadi and Eger (2023) showed that token-free character-level models can outperform token-based LLMs on poetry tasks, precisely because token boundaries do not align with syllable or phoneme boundaries. Similar evidence comes from Chinese classical poetry, where, for the related task of format-controlled generation, token-free architectures substantially outperform token-based LLMs such as GPT-4 (Yu et al. 2024), underscoring tokenization as a cross-lingual bottleneck for strictly metered tasks.

Recent evaluation work has begun to assess LLM poetic capabilities systematically. Walsh et al. (2024) evaluated LLM recognition of over 20 English poetic forms. Koziev (2025) introduced a scansion tool and evaluation dataset for Russian poetry. Chatzikyriakidis and Natsina (2026) found that pure LLM generation for Greek poetry fails catastrophically (under 4% valid poems), but hybrid systems combining LLMs with phonological algorithms restore performance to 73.1%. This "Reasoning Gap" suggests that LLMs require specialized support for phonological tasks in non-English languages, a finding directly relevant to Portuguese scansion.

Poetry, Computational Creativity, and Accessibility

Computational poetry generation has explored meter-aware systems (Colton, Goodwin, and Veale 2012; Lau et al. 2018; Oliveira 2012; Ormazabal et al. 2022), and instruction-tuned models have been applied to collaborative poetry writing (Chakrabarty, Padmakumar, and He 2022). Gonçalo Oliveira (2024), in a recent survey of computational creativity systems for Portuguese, highlighted both the progress in platforms like PoeTryMe and the persistent scarcity of tools for this language. However, accurate scansion remains a prerequisite for meter-constrained generation, automatic verse evaluation, and human-AI co-creation of formal poetry. Without reliable automated scansion, such systems must rely on hand-crafted phonological rules or human evaluation, limiting their scalability and applicability to under-resourced languages like Portuguese.

The accessibility of these tools also matters. While the Sabiá series (Pires et al. 2023; Almeida et al. 2024; Abonizio et al. 2025) demonstrated cost-effective fine-tuning for Portuguese, these models remain behind commercial APIs. Open alternatives such as Tucano (2025) (up to 2.4B parameters, Apache 2.0) and TeenyTinyLlama (Corrêa et al. 2024) (up to 460M parameters, developed for under

$500) show that accessible Portuguese language models are emerging. These models were not included in our benchmark because they are base (non-instruction-tuned) models too small to follow zero-shot or few-shot prompts; however, they represent promising candidates for future fine-tuning experiments. The present study provides empirical evidence on how instruction-following models compare on a concrete, phonologically demanding task, informing decisions about building creativity support tools that do not depend on centralized commercial infrastructure.

Methodology

This section defines the scansion task, describes the dataset and models under evaluation, and introduces the metrics used to assess performance.

Task Definition

Given a Portuguese verse as input, the scansion task requires producing a string that segments the verse into *poetic* syllables separated by "/", marks stressed syllables with "⋆", and indicates the last metrically relevant stressed syllable with "#". For example:

Input	*Teus formosos quinze anos*
Output	`Teus / for- / ⋆mo- / sos / quin- / ⋆ze a-# / nos`

Poetic syllables differ from grammatical ones due to phonological phenomena called *metaplasms*. The verse above has 8 grammatical syllables (*Teus / for / mo / sos / quin / ze / a / nos*), but only 7 poetic syllables (6 metric + 1 post-tonic). The reduction occurs because the final vowel of *quinze* and the initial vowel of *anos* merge through *synalepha*, producing the fused syllable "ze a−". Stress markers identify the tonic syllables: "⋆mo−" (from *for**mo**sos*, position 3) and "⋆ze a−" (carrying the stress of ***a**nos*, position 6). The "#" on position 6 signals the metric boundary: Portuguese verse counts syllables only up to the last stressed one, making this a hexasyllable (6-syllable verse).

The notation thus captures three distinct linguistic phenomena: syllable segmentation (boundaries), stress assignment (tonic identification), and metaplasm detection (cross-word syllable fusion). Metaplasms are the most challenging aspect, as they require phonological knowledge to determine when adjacent vowels across word boundaries merge into a single poetic syllable (Bisol 2005; Ali 1999).

This task formulation targets stress-and-syllable-based prosody, characteristic of Romance and Germanic verse traditions, and is not intended to generalize to all poetic systems. Classical Arabic metrics (*ʿarūḍ*), for instance, are built on moraic units defined by vowel quantity rather than on lexical stress (Frolov 2000). Our notation, metrics, and conclusions are therefore language-specific by design.

Dataset

Our dataset derives from Aoidos (Mittmann 2016), a rule-based phonological system for Portuguese scansion available as an online platform.

In our prior work (Valença and Calegario 2025), we constructed the dataset by taking approximately 13,500 verses from the Aoidos test corpus (spanning three centuries of Brazilian Portuguese poetry, publicly available in the Aoidos repository), submitting them in batches to the Aoidos online platform, and parsing the resulting HTML output with a custom script to extract syllable segmentations, metrical classifications, and rhythmic patterns.

Stress markers (⋆ and #) were then added programmatically: for each verse, Aoidos reports the rhythmic pattern as the positions of the metrically stressed syllables, and a script maps these positions onto the segmented syllable sequence, prepending ⋆ to every stressed syllable and appending # to the last one to mark the metric boundary. The annotation script is provided in the supplementary material. Verses containing diérese were filtered out to ensure notational consistency, and the results were deduplicated, yielding 14,080 unique verses with ground-truth scansions.

Following the same split as our prior work, the dataset is divided 75/25 into 10,560 training and 3,520 test verses. Fine-tuned models are trained on 7,040 examples drawn from the training split (the same subset used in our prior work). From the 3,520 test verses, we sample 500 (random seed 42) as a canonical test set, balancing evaluation cost with statistical power.

All models are evaluated on the same 500 verses to ensure comparability. To prevent data leakage, we verified that the 500 test verses have zero overlap with the fine-tuning training and validation sets.

Models

Model selection was guided by API availability and practical accessibility: all models were accessed through cloud APIs (OpenAI, Anthropic, Google, DeepSeek, Maritaca AI, and OpenRouter for open-weight models), and no models were run locally. This reflects a deliberate design choice to evaluate the landscape as it is available to most researchers and practitioners; local deployment of open-weight models remains an avenue for future work. We evaluate 24 models in the main benchmark, organized into four categories.

Fine-tuned (3 models): GPT-4.1, GPT-4.1-mini, and GPT-4.1-nano, fine-tuned on the 7,040 training examples described above, using the same system prompt and output format as our prior work (Valença and Calegario 2025).

Proprietary (9 models): Claude Opus 4.6, Sonnet 4.6, and Haiku 4.5 (Anthropic 2026); GPT-4.1, GPT-4.1-mini, and GPT-4.1-nano (OpenAI 2023); DeepSeek V3 (DeepSeek-AI 2024); Gemini 2.5 Flash and Flash Lite.

Open-weight (7 models): LLaMA 4 Maverick and Scout (Meta AI 2025); Gemma 3 12B and 4B (Google DeepMind 2024); Ministral 14B, 8B, and 3B (Mistral AI 2025).

Portuguese-specialized (5 models): Sabiá 4, 3.1, and 3; Sabiazinho 4 and 3 (Pires et al. 2023; Abonizio et al. 2025).

Additionally, 9 reasoning/thinking models were evaluated on a 10-example pilot (see *Thinking Models* in Results): DeepSeek-R1, Gemini 2.5 Pro, GPT-5 Nano, Grok-4 Fast, Kimi-K2.5, Qwen3.5-27B, Qwen3-32B, Qwen3-14B, and Qwen3-8B. GPT-5 full and GPT-5-mini were tested in preliminary experiments but excluded due to prohibitive latency

(∼98s/verse) and cost ($0.042/verse for GPT-5 full); GPT-5 Nano was retained as a representative of the family at acceptable cost.

Model Settings and Prompting All evaluations use temperature 0.1, following our prior work (Valença and Calegario 2025). A near-zero temperature favors deterministic outputs while avoiding the degenerate responses that temperature 0 elicits from some providers; we quantify the resulting run-to-run consistency in the appendix.

The maximum output length is set to 512 tokens for standard models. For thinking/reasoning models, the token budget was increased to 8,192 completion tokens to accommodate the internal chain-of-thought (thinking mode), and the temperature parameter was omitted as these APIs do not support it. Note that, in some cases, extended thinking is a configurable mode rather than an inherent model property: Claude Opus 4.6, for instance, supports extended thinking but was evaluated with thinking disabled and is therefore classified as a standard proprietary model in this study.

The zero-shot system prompt was carefully revised to disambiguate formatting conventions (e.g., explicit rules for spacing around stress markers and hyphens), as preliminary experiments revealed that the original prompt's ambiguity unfairly penalized zero-shot models on exact match. The prompt specifies the output format but deliberately omits phonological rules, so that the evaluation measures the model's linguistic competence rather than its ability to follow detailed instructions.

Few-shot prompts include 3 worked examples that demonstrate the input/output format across diverse metaplasm types. Fine-tuned models use the original system prompt from our prior work. Throughout the paper we abbreviate these three settings as **ZS** (zero-shot), **FS** (few-shot), and **FT** (fine-tuned). To conserve space, Table 1 reports for each model only its best-performing configuration: fine-tuned models are labeled FT, while every non-fine-tuned model reached its best result in the few-shot setting and is therefore labeled FS. The corresponding zero-shot scores can be recovered by subtracting the few-shot gains reported in Table 2(a), and complete per-mode results are provided in the supplementary material. Open-weight models are accessed via OpenRouter; all other models use their respective provider APIs. All experiments were conducted between December 2025 and February 2026. Evaluation scripts, prompts, raw model outputs, and the test set are available as supplementary material (`https://github.com/filipecalegario/portuguese-poetic-scansion-benchmark`).

Evaluation Metrics

We employ seven metrics spanning four dimensions:

Syllable accuracy: We report two complementary metrics. *SYL_R%* (syllable ratio), from prior work (Valença and Calegario 2025), compares syllables position-by-position up to the last stressed syllable and computes the average proportion of matching syllables per verse. *SYL_EM%* (exact match) is a stricter metric introduced in this work: a verse scores 1 only if every syllable, stress marker, and formatting detail matches the ground truth exactly. The difference between these metrics is informative: a model scoring 88% SYL_R% but 62% SYL_EM% produces mostly correct syllables in each verse but rarely achieves a perfect transcription. SYL_R% enables direct comparison with the baseline from prior work; SYL_EM% indicates the proportion of scansion that is production-ready without post-processing.

Boundary quality: *BND_F1%*, the F1 score computed by aligning predicted and expected syllable sequences via edit distance; true positives are syllables with matching normalized text. Unlike SYL_R%, which uses positional alignment (via `zip`), BND_F1 uses edit-distance alignment and thus handles insertions and deletions more gracefully, providing a fairer comparison between fine-tuned models (which learn the exact corpus format) and zero-shot models (which may produce different syllable counts).

Stress quality: *STR_ACC%*, the percentage of correctly aligned syllables where the stress marker matches. When prediction and reference differ in length, syllables are aligned by the same edit-distance procedure used for BND_F1, and stress is compared only on pairs whose normalized text matches, so that stress accuracy is not confounded by segmentation errors. *STR_PAT%*, a binary stress pattern match up to the last stressed syllable.

Metaplasm detection: *META%*. We operationalize a metaplasm as a poetic syllable that spans two or more words through cross-word vowel fusion, which in our notation surfaces as an internal whitespace inside a syllable token (e.g., `que en` for synalepha, or `se há` for elision). For a verse, let M be the set of such fused syllables in the ground truth. A ground-truth metaplasm counts as recovered if some predicted syllable has the same normalized text (lowercased, with stress markers, trailing hyphens, and punctuation removed), regardless of its position in the line. META% is the number of recovered metaplasms divided by $\sum |M|$, aggregated over all verses that contain at least one metaplasm (verses with none are excluded). This operationalization locates cross-word fusion, the most linguistically challenging aspect of Portuguese scansion, but it does not classify the underlying phonological process (synalepha, elision, crasis) and requires the fused syllable to match the reference text exactly.

Efficiency: *Latency* (wall-clock seconds per API call, averaged over all 500 verses) and *Cost* (USD per verse). Cost is estimated as $(T_{\text{in}} \times p_{\text{in}} + T_{\text{out}} \times p_{\text{out}})/N$, where T_{in} and T_{out} are the total input and output tokens reported by the API, p_{in} and p_{out} are the provider's published per-token prices, and $N = 500$ is the number of verses. Input tokens include the system prompt, any few-shot examples, and the verse itself; output tokens cover the model's scansion response.

Results

Table 1 presents the main results across all model categories, showing the best configuration (zero-shot or few-shot) for each model. Table 2 details few-shot gains and thinking model performance. Figure 1 visualizes the accuracy-cost trade-off, and Figure 2 compares the best score per category across key metrics.

Table 1: Main results across model categories. Best mode (zero-shot or few-shot) per model. SYL_R% is the syllable-level ratio from prior work; SYL_EM% is our stricter verse-level exact match. Best value per column in **bold**.

Model	Mode	SYL_R↑	SYL_EM↑	BND_F1↑	META↑	STR_ACC↑	Lat(s)↓	$/verse↓
Fine-Tuned								
GPT-4.1-mini-FT	FT	**93.9**	81.8	**98.0**	**96.6**	98.0	2.2	0.00027
GPT-4.1-FT	FT	93.3	**82.4**	97.9	96.2	**98.4**	3.9	0.00102
GPT-3.5-FT	FT	88.0	62.2	95.7	92.5	96.0	1.7	0.00085
GPT-4.1-nano-FT	FT	76.1	34.8	90.4	82.0	91.4	0.8	0.00007
Proprietary								
Claude Opus 4.6	FS	74.1	25.8	86.8	67.6	94.4	2.8	0.00312
Claude Sonnet 4.6	FS	63.9	17.6	83.0	62.8	90.7	2.1	0.00187
Gemini 2.5 Flash	FS	63.9	15.4	71.9	47.5	90.5	1.8	0.00007
GPT-4.1	FS	55.0	9.2	76.5	53.9	89.0	1.2	0.00102
Gemini 2.5 Flash Lite	FS	52.8	6.4	80.3	49.5	84.1	0.7	0.00005
Claude Haiku 4.5	FS	51.2	7.4	79.6	48.2	81.6	1.0	0.00062
DeepSeek V3	FS	46.5	7.6	76.7	26.5	86.4	2.4	0.00012
GPT-4.1-mini	FS	40.6	3.6	72.9	27.9	79.0	1.2	0.00021
GPT-4.1-nano	FS	32.6	0.0	62.6	23.6	71.2	1.0	0.00005
Open-Weight								
LLaMA 4 Maverick	FS	45.9	7.4	74.8	34.7	85.5	1.3	0.00008
Ministral 14B	FS	40.0	2.6	73.1	29.9	78.7	0.8	0.00008
LLaMA 4 Scout	FS	37.5	1.8	70.8	28.9	75.7	1.2	0.00004
Gemma 3 12B	FS	33.0	0.8	68.6	22.6	72.3	1.4	0.00002
Ministral 8B	FS	30.7	1.4	65.9	13.5	70.2	0.6	0.00006
Gemma 3 4B	FS	23.7	0.0	50.4	16.2	66.8	2.4	**0.00002**
Ministral 3B	FS	22.4	0.0	52.3	21.2	68.0	**0.6**	0.00004
Portuguese-Specialized								
Sabiá 4	FS	45.8	3.4	74.8	43.3	75.9	1.5	0.00045
Sabiá 3	FS	33.5	0.8	63.7	36.7	68.8	2.6	0.00038
Sabiá 3.1	FS	31.5	0.6	66.2	44.3	65.8	2.3	0.00038
Sabiazinho 4	FS	30.5	0.8	63.8	31.9	67.1	0.7	0.00009
Sabiazinho 3	FS	26.6	1.0	55.6	25.8	64.4	1.1	0.00008

Fine-Tuning Dominates

Fine-tuned models dominate all metrics, as shown in the top group of Table 1. GPT-4.1-mini-FT leads on SYL_R% (93.9%), closely followed by GPT-4.1-FT (93.3%), both with BND_F1 $>$ 97% and META% $>$ 96%. The near-equivalence of these two models suggests that the scansion task may be saturated with fine-tuning and does not benefit from the additional capacity of larger architectures. However, GPT-4.1-nano-FT collapses to 76.1% SYL_R% (34.8% SYL_EM%), indicating a minimum model capacity threshold for this task.

A key finding is that re-evaluating the GPT-3.5-FT model from our prior work (Valença and Calegario 2025) on the same 500-verse test set yields 88.0% SYL_R%, reproducing the original 88.6% baseline and confirming that GPT-4.1 variants represent improvement (+5.3 percentage points on SYL_R%), not a regression. The difference between SYL_R% and SYL_EM% is also informative: GPT-3.5-FT scores 88.0% on the former but only 62.2% on the latter, meaning it produces mostly correct syllables but achieves a perfect transcription less often than GPT-4.1-FT (82.4% SYL_EM%).

As visible in Figure 2, the gap between fine-tuned and zero-shot models is stark across all metrics, with fine-tuned models achieving approximately 98% BND_F1 versus 87% for the best proprietary model.

Zero-Shot and Few-Shot Performance

Among non-fine-tuned models, Claude Opus 4.6 achieves the highest SYL_R% (74.1% few-shot) and SYL_EM% (25.8%), followed by Claude Sonnet 4.6 (63.9% / 17.6%) and Gemini 2.5 Flash (63.9% / 15.4%). While the exact-match scores are low, the syllable-ratio and BND_F1 metrics reveal substantially more competence: Claude Opus 4.6 reaches 86.8% BND_F1, meaning it correctly segments the majority of syllable boundaries but fails on stress placement or formatting details.

Table 2(a) quantifies the few-shot effect. All five selected models improve with 3-shot examples, with gains ranging from +3.0 to +9.2 percentage points on SYL_EM% and +0.7 to +13.4 on BND_F1. The largest relative gains are observed for META% (+7.3 to +23.3 points), indicating that few-shot examples can be particularly helpful for models to learn metaplasm conventions.

Figure 1 reveals the accuracy-cost trade-off landscape. The Pareto frontier traces from the cheapest models (Gemma 3 12B at $0.00002/verse with 68.6% BND_F1) through mid-range options (DeepSeek V3 at $0.0001 with

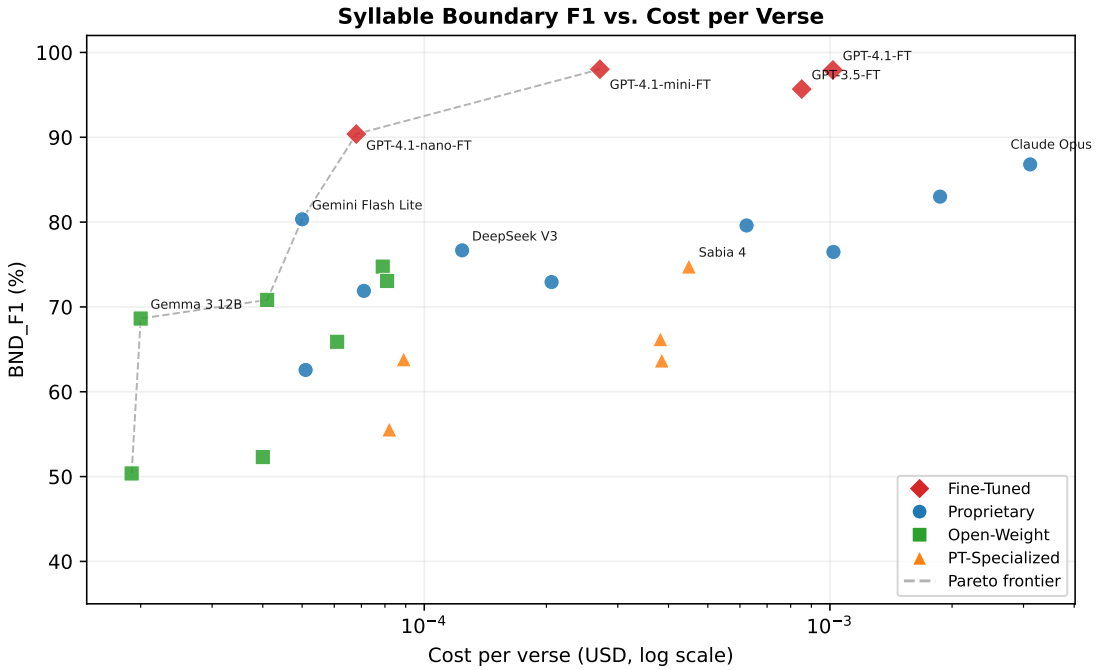


Figure 1: Syllable boundary F1 score vs. cost per verse (USD, log scale). Colors indicate model category: fine-tuned (red), proprietary (blue), open-weight (green), Portuguese-specialized (orange). The dashed line shows the Pareto frontier.

76.7%) to the premium tier (Claude Opus 4.6 at $0.003 with 86.8%). Fine-tuned models occupy the upper-right region, combining high accuracy with moderate cost.

Thinking Models: Preliminary Evidence of Failure

Table 2(b) reports results for 9 reasoning/thinking models evaluated on a 10-example pilot. The reduced sample size reflects the prohibitive cost of full evaluation: with latencies of 7–211s per verse (vs. <2s for standard models), a complete 500-verse evaluation across all 9 thinking models would require an estimated 40–200 hours of API time. Every model achieves 0% SYL_EM%, with costs up to $0.029/verse. Because exact match is unforgiving of partial or truncated outputs, we report it alongside two edit-distance-based measures, BND_F1 and normalized edit distance (NRM_ED, where 0 is identical), which together reveal that several models retain substantial syllabification content despite the 0% exact-match score.

Qualitative analysis of the outputs reveals that the 0% exact match stems from at least three distinct failure modes, not a single cause:

(1) Output truncation: Gemini 3.1 Pro allocates the vast majority of its token budget to internal reasoning (∼488 reasoning tokens) and produces only ∼20 completion tokens, cutting every scansion mid-output (e.g., `por / ⋆ vos- / sos / ⋆ fi- / lhos / cho- / ⋆ rar /`, missing the final syllables). The reasoning-to-completion ratio reaches 24.5×.

(2) Empty or leaked output: GPT-5 Nano produces empty responses for 80% of verses despite consuming ∼8,128 reasoning tokens per call (notably, with the default output limit of 512 tokens, 100% of responses were empty, as the reasoning phase alone exhausted the entire budget before any visible output could be generated). Similarly, Qwen 3.5 27B occasionally emits its internal chain-of-thought (e.g., "Thinking Process:") instead of the scansion, spending ∼14,323 reasoning tokens at 211s/verse without converging on an answer.

(3) Incorrect notation: Qwen3-32B and Qwen3-8B produce non-empty outputs but deviate from the required notation (e.g., placing stress markers after syllables instead of before: `vós ⋆` instead of `⋆ vós`, or merging syllables: `⋆vos-sos` instead of `⋆ vo- / ssos`).

Among the remaining models (DeepSeek-R1, Grok-4 Fast, Kimi-K2.5), the outputs follow the correct format, and the edit-distance metrics confirm that the failure is one of completion and formatting rather than phonology: DeepSeek-R1 reaches the highest BND_F1 (77.5%) at a normalized edit distance of only 0.25 (roughly three-quarters of the characters match the reference), with Grok-4 Fast and Kimi-K2.5 at NRM_ED 0.31 and 0.42. Errors in syllable boundaries and stress placement, not an inability to segment, are what prevent an exact match.

While the small sample size (N=10) limits generalizability, the pattern suggests that the extended reasoning process can be counterproductive for scansion: models spend thousands of tokens deliberating about phonological rules but fail to produce correctly formatted output, either by exhausting their token budget on reasoning, by not converging on a response, or by inventing non-standard notation conventions.

Table 2: (a) Few-shot gain on selected models. (b) Reasoning/thinking models (all N=10; all score 0% SYL_EM). NRM_ED is the normalized character-level edit distance to the ground truth (0 = identical; lower is better). Best per column in **bold**.

(a) Few-Shot Gain

Model	ΔR↑	ΔEM↑	ΔF1↑	ΔMETA↑
DeepSeek V3	**+13.6**	+6.8	+11.5	**+23.3**
LLaMA 4 Mav.	+8.8	+5.8	+7.0	+17.8
GPT-4.1	+4.5	+7.2	+9.6	+17.2
Cl. Opus 4.6	+4.1	**+9.2**	+0.7	+3.7
Sabiá 4	+1.7	+3.0	**+13.4**	+7.3

ΔR = ΔSYL_R, ΔEM = ΔSYL_EM, ΔF1 = ΔBND_F1

(b) Reasoning / Thinking Models

Model	SYL_R	BND_F1	NRM_ED↓	Lat	$/v
Gemini 3.1 Pro	**63.0**	59.6	0.53	**6.8**	0.0007
DeepSeek-R1	41.4	**77.5**	**0.25**	89	0.0015
Qwen 3.5 27B	35.4	66.6	0.35	184	0.0292
Grok 4 Fast	18.7	73.4	0.31	8.6	0.0006
Kimi K2.5	12.2	61.5	0.42	96	0.0118
Qwen 3 32B	11.7	45.7	0.59	87	0.0007
Qwen 3 14B	10.3	39.2	0.64	25	**0.0003**
Qwen 3 8B	9.9	34.0	0.70	54	0.0010
GPT-5 Nano	6.5	13.1	0.89	49	0.0033

Several of these failures are configuration-dependent rather than fundamental: raising the completion-token budget or adjusting the prompt to suppress chain-of-thought leakage would likely recover many exact matches, since the underlying syllabification (as reflected by NRM_ED and BND_F1) is often largely correct. We therefore present this pilot as preliminary evidence motivating a fuller study, not as a definitive verdict on reasoning models.

Portuguese-Specialized Models Underperform

Portuguese-specialized models (Sabiá family) do not outperform general-purpose alternatives on this task. The best Portuguese model, Sabiá 4, reaches 45.8% SYL_R% and 74.8% BND_F1 in few-shot mode. Notably, while its exact-match-oriented SYL_R% trails general-purpose models, its format-tolerant BND_F1 is comparable to several of them (e.g., GPT-4.1 at 76.5% and DeepSeek V3 at 76.7%, and above Gemini 2.5 Flash at 71.9%), indicating that the gap is concentrated in formatting and notation adherence rather than in syllable segmentation. This pattern is visible in both Table 1 (PT group vs. Proprietary group) and Figure 2.

Stress Is Easier Than Segmentation

An informative pattern emerges from the STR_ACC metric in Table 1: even models with low SYL_EM% achieve high stress accuracy on correctly segmented syllables. For example, Claude Haiku (7.4% SYL_EM%) reaches 81.6% STR_ACC, and DeepSeek V3 (7.6% SYL_EM%) reaches 86.4%. This indicates that models have implicit knowledge of Portuguese stress patterns but struggle with the more challenging syllable segmentation and metaplasm detection.

Table 3: Scansion outputs for the verse *"Verdade, que entre vós se há confundido,"*. Differences from the ground truth are underlined.

Source	Scansion
Ground truth	Ver- / *da- / de, / *que en- / tre / *vós / se há / con- / fun- / *di-# / do,
GPT-4.1-FT	Ver- / *da- / de, / *que en- / tre / *vós / se há / con- / fun- / *di-# / do,
Cl. Opus 4.6 (FS)	Ver- / *da- / de, / que en- / tre / *vós / se / *há / con- / fun- / *di-# / do,
Gemma 3 12B	Ver- / *da- / de, / que / en- / *tre / vós / se / *há / con- / *fun-# / di- / do
Sabiá 4	*Ver- / da- / de, / que en- / *tre / vós / se há / *con- / fun- / *di-# / do,

Qualitative Error Example

Table 3 illustrates how different models handle the verse *"Verdade, que entre vós se há confundido,"*, which contains two metaplasms: "que en-" (synalepha) and "se há" (elision). GPT-4.1-FT reproduces the ground truth exactly. Claude Opus 4.6 correctly detects "que en-" but splits "se há" into two separate syllables, a boundary error that costs the exact match. Gemma 3 12B misses both metaplasms and places the metric boundary marker (#) on the wrong syllable. Sabiá 4 detects the "se há" metaplasm but misplaces stress markers throughout. These examples illustrate the typical failure gradient: formatting and stress errors (Claude Opus 4.6) are less severe than boundary errors (Gemma, Sabiá), which the BND_F1 metric captures more faithfully than exact match.

Discussion

The results raise several questions about what drives scansion performance, what the metrics reveal beyond surface accuracy, and what the practical implications are for building accessible creativity support tools.

The Fine-Tuning Gap

Our results confirm that fine-tuning remains indispensable for production-quality Portuguese scansion. The gap between fine-tuned models (BND_F1 $\approx$ 98%) and the best zero-shot model (BND_F1 = 86.8%) is substantial and consistent across metrics, as illustrated in Figure 2. This aligns with the general finding that domain-specific tasks requiring precise formatting and linguistic conventions benefit disproportionately from supervised adaptation (Hu et al. 2022).

However, the near-equivalence of GPT-4.1-FT and GPT-4.1-mini-FT (differing by only 0.6 percentage points on SYL_R%) suggests that the task's complexity does not warrant the largest model architectures, a practically important finding for deployment cost. At $0.0003/verse, GPT-4.1-mini-FT offers the best cost-accuracy trade-off among fine-tuned models.

What BND_F1 Reveals

The introduction of format-tolerant metrics like BND_F1 substantially changes the picture painted by exact match

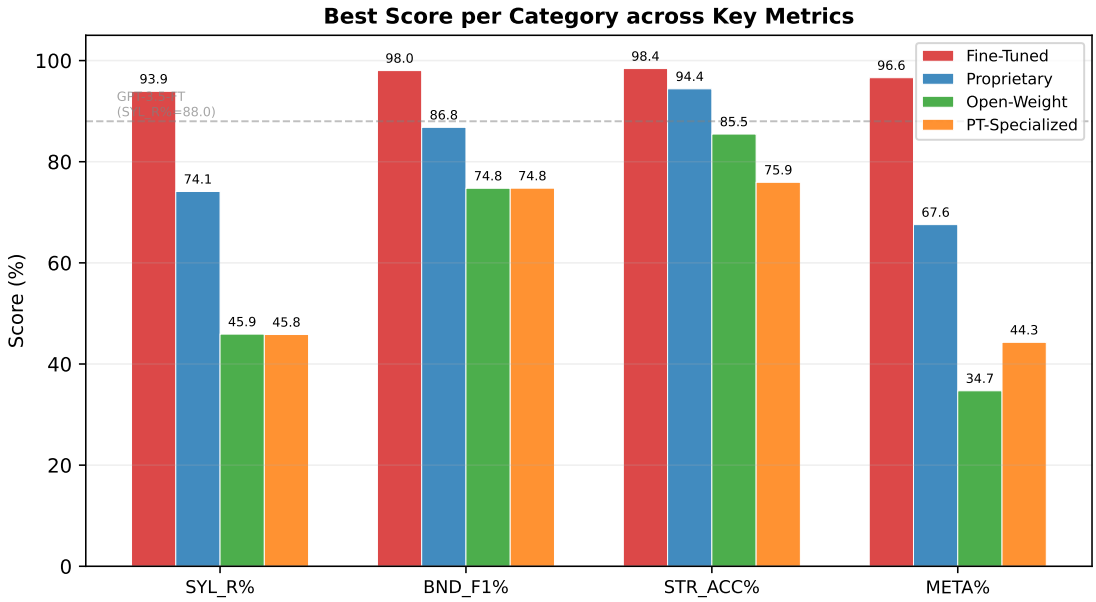


Figure 2: Best score per model category across key metrics. The dashed line marks the GPT-3.5 fine-tuned baseline (SYL_R% = 88.0%, measured on the same 500-verse test set).

alone. Under SYL_EM%, the best zero-shot model (Claude Opus 4.6 at 25.8%) appears to perform poorly. Under BND_F1, the same model scores 86.8%, revealing that it correctly identifies most syllable boundaries but fails on the precise formatting conventions (stress marker placement, hash notation) that the exact-match metric demands. This suggests that LLMs possess substantial implicit phonological knowledge of Portuguese but struggle to produce outputs that conform to a specific annotation standard without supervised examples.

This raises a natural question: does BND_F1 measure syllabification or merely formatting? By construction, it measures the former. Before alignment, both syllable sequences are lowercased and stripped of stress markers, hyphens, and punctuation, so two syllables match only when their underlying text is identical, irrespective of surface formatting. BND_F1 therefore credits correct *segmentation* of the Portuguese syllable string rather than adherence to the annotation convention, and a model gains nothing from English-style formatting habits, since the comparison is over Portuguese poetic syllables. The gap between SYL_EM% and BND_F1 isolates the formatting layer, while STR_ACC% and META% isolate stress assignment and metaplasm detection; read together, the metric suite separates genuine phonological competence from the ability to reproduce a specific annotation standard.

Implications for Computational Creativity

Our results have direct consequences for computational poetry systems. Meter-constrained generation (Lau et al. 2018; Ormazabal et al. 2022) requires a scansion module to verify whether candidate verses satisfy target meters (e.g., decasyllables). With the best fine-tuned model achieving 98% BND_F1, such a filter becomes practical: a generator can produce candidate verses and use the scansion model to accept or reject them based on metrical criteria. Even zero-shot models, with BND_F1 scores above 80%, could serve as approximate filters in interactive co-creation settings where a human poet refines the output.

This is particularly relevant to poetic traditions in which meter is not optional but constitutive. In Northeastern Brazilian popular poetry, *cantadores* (improvising poets) engage in *pelejas*, oral verse duels governed by strict metric and rhyme conventions. As a poet from this tradition once put it, there are three pillars a *cantador* must master: rhyme, meter, and subject matter (Caldas 2021). A computational system aspiring to participate in such traditions, whether as a creative partner, a training tool for aspiring poets, or an analytical aid for literary scholarship, must first demonstrate reliable metrical analysis. The scansion benchmark presented here thus serves as a foundation for our broader platform, Pajéu (Valença and Calegario 2025), aimed at meter-aware verse generation, literary games, and human-AI co-creation rooted in these living poetic traditions.

Accessibility and Linguistic Sovereignty

The cost-accuracy landscape in Figure 1 reveals a concerning pattern for accessibility. The most accurate models are either fine-tuned proprietary models (requiring API access to OpenAI) or expensive frontier models (Claude Opus 4.6 at $0.003/verse). Open-weight alternatives, while dramatically cheaper (Gemma 3 12B at $0.00002/verse), sacrifice significant accuracy.

More concerning is the linguistic sovereignty dimen-

sion, which is the non-dependence on foreign infrastructure. The highest-performing models are developed by organizations in the United States and China, whereas Portuguese-specialized models from Brazilian companies do not lead on this task. This gap, however, is less a deficit of Portuguese phonology than of instruction-following: the Sabiá-3 technical report itself notes that the model lags general-purpose alternatives, including cheaper ones, on instruction-following benchmarks such as IFeval (Abonizio et al. 2025), and our format-tolerant BND_F1 shows Sabiá's syllable segmentation to be comparable to general-purpose models. The underperformance is therefore concentrated in adhering to the scansion notation rather than in Portuguese phonology per se. The practical implication for sovereignty nonetheless stands: no accessible Portuguese-tuned model currently delivers production-quality scansion. A promising pathway is fine-tuning open-weight models on the Aoidos corpus, which could combine the accessibility of open models with the accuracy of supervised adaptation, though the infrastructure and expertise required for fine-tuning remain barriers for many poetry communities.

Limitations

Several limitations should be acknowledged. First, our ground truth is obtained from a rule-based system (Aoidos) rather than from human expert annotation. This means our evaluation measures agreement with Aoidos's phonological rules, not necessarily with all valid scansion interpretations (some verses admit multiple legitimate scansions depending on performance choices). Fine-tuned models may learn to replicate Aoidos's systematic biases. We adopt Aoidos as an *operational* ground truth, consistent with our prior work (Valença and Calegario 2025) and the original Aoidos validation (Mittmann 2016), while acknowledging that a human-validated gold set would strengthen future evaluations.

Second, thinking models were evaluated on only 10 examples due to prohibitive cost and latency; while the consistent 0% SYL_EM% across all nine models is an initial signal, a larger sample would strengthen this finding.

Third, we evaluated only one prompting strategy per mode; more elaborate prompt engineering (e.g., including explicit phonological rules) might improve zero-shot performance, particularly on formatting conventions.

Fourth, our cost estimates use published list pricing as of February 2026 and do not account for volume discounts or cached responses.

Fifth, training and test verses are drawn from the same corpus and thus share a metrical distribution dominated by decasyllabic meters (*heroico* and *sáfico*, together about 70% of verses) and heptasyllables (about 15%), spanning 20 classical metrical classes. Our results therefore measure in-distribution performance and do not test generalization to genuinely novel structures, such as haiku, *cordel* improvisation, or metrically free verse; evaluating fine-tuned models on such out-of-distribution forms is an important direction for future work.

Conclusion

We have presented the first comparative evaluation of language models for Portuguese poetic scansion, benchmarking 24 models across accuracy, latency, and cost, with an additional pilot of 9 reasoning models. Our key findings are:

1. **Fine-tuning delivers genuine improvement**: Re-evaluating the 2025 GPT-3.5-FT baseline on our canonical test set confirms 88.0% SYL_R%, while GPT-4.1-mini-FT achieves 93.9%, a +5.9pp improvement at a fraction of the cost of GPT-4.1-FT (93.3%). Performance saturates at the mini tier, while GPT-4.1-nano-FT (34.8% SYL_EM%) reveals a minimum capacity threshold.
2. **Format-tolerant metrics change the narrative**: BND_F1 scores reveal that zero-shot models possess substantial segmentation competence (up to 86.8%) masked by low SYL_EM% scores, suggesting that LLMs have implicit phonological knowledge that fine-tuning helps format correctly.
3. **Reasoning models struggle**: In a pilot study (N=10), all 9 thinking/reasoning models achieve 0% SYL_EM%, failing through three distinct mechanisms: output truncation (reasoning consumes the token budget), empty or leaked responses (models deliberate extensively without converging), and incorrect notation conventions. This suggests that extended deliberation is counterproductive for format-sensitive tasks like scansion.
4. **Portuguese specialization does not transfer to scansion**: general-purpose models outperform Sabiá-family models on this task, but the gap reflects weaker instruction-following (as reported for Sabiá-3 on IFeval (Abonizio et al. 2025)) more than weaker phonology, since their format-tolerant BND_F1 is comparable.
5. **The accessibility gap persists**: The most accurate scansion requires commercial fine-tuned models, while the cheapest alternatives sacrifice substantial accuracy. Fine-tuning open-weight models represents a promising but underexplored pathway.

Future work should explore fine-tuning open-weight models (e.g., LLaMA, Gemma) on the Aoidos corpus to evaluate whether the fine-tuning accuracy observed with proprietary models transfers to open architectures. A complementary direction is token-free, character-level modeling (Belouadi and Eger 2023; Yu et al. 2024): because the dominant errors in our results lie in syllable boundaries and metaplasms, exactly where subword tokens misalign with poetic syllables, such architectures could improve segmentation beyond what conventional fine-tuning achieves. With a reliable scansion module in place, the natural next step is meter-constrained verse generation, where a scansion model serves as a metrical filter or reward signal for a generator, enabling the production of formally valid poetry.

Beyond text generation, integrating low-latency speech synthesis with appropriate prosody would enable spoken-verse systems capable of participating in oral poetic traditions, such as the improvised verse duels of Northeastern Brazil, where meter is not a stylistic choice but a structural requirement.

Appendix: Run-to-Run Consistency

To assess how stable our single-run measurements are at temperature 0.1, we re-ran three representative models, one fine-tuned, one proprietary, and one open-weight, five times each on a fixed 30-verse subsample, and measured run-to-run agreement (Table 4). The fine-tuned model is nearly deterministic: 90% of verses yield byte-identical scansions across all five runs, syllable counts agree on 96.7% and stress patterns on 100%, with only 1.1 distinct outputs per verse on average. Determinism is not uniform across providers, however. Gemma 3 12B remains fairly stable (76.7% identical, 1.3 distinct outputs), whereas Gemini 2.5 Flash varies substantially (33.3% identical, 2.8 distinct outputs per verse) despite the low temperature. Two factors contribute: its few-shot outputs are frequently incomplete, with the truncation point shifting between runs, and the segmentation itself is unstable, since even on the common prefix preceding truncation only 56.7% of verses agree on syllable text. The headline metrics for the fine-tuned models, which are our primary results, are therefore highly reproducible, while single-run scores for high-variance proprietary models should be read as point estimates carrying run-to-run uncertainty.

Table 4: Run-to-run consistency at temperature 0.1: each model scanned the same 30 verses five times. *Exact* = identical output across all five runs; *Count* and *Stress* = identical syllable count and stress pattern across runs; *Distinct* = mean number of distinct outputs per verse (1 = fully deterministic).

Model	Exact%	Count%	Stress%	Distinct
GPT-4.1-mini-FT	90.0	96.7	100.0	1.1
Gemini 2.5 Flash (FS)	33.3	56.7	33.3	2.8
Gemma 3 12B (FS)	76.7	93.3	86.7	1.3

References

Abonizio, H.; Almeida, T. S.; Laitz, T.; Junior, R. M.; Bonás, G. K.; Nogueira, R.; and Pires, R. 2025. Sabiá-3 technical report. arXiv preprint arXiv:2410.12049.

Agirrezabal, M.; naki Alegria, I.; Arrieta, B.; and Hulden, M. 2012. Finite-state technology in a verse-making tool. In *Proceedings of the Finite State Methods and Natural Language Processing Conference*, 35–39.

Agirrezabal, M.; Astigarraga, A.; Arrieta, B.; and Hulden, M. 2016. ZeuScansion: A tool for scansion of English poetry. *Journal of Language Modelling* 4(1):3–28.

Agirrezabal, M. 2017. *Automatic Scansion of Poetry*. Ph.D. Dissertation, University of the Basque Country (UPV/EHU).

Ali, M. S. 1999. *Versificação Portuguesa*. São Paulo: EDUSP.

Almeida, T. S.; Abonizio, H.; Nogueira, R.; and Pires, R. 2024. Sabiá-2: A new generation of Portuguese large language models. arXiv preprint arXiv:2403.09887.

Anthropic. 2026. The Claude model card and system prompt. Accessed: February 2026. `https://docs.anthropic.com/en/docs/about-claude/models`.

Belouadi, J., and Eger, S. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7364–7381.

Bisol, L. 2005. *Introdução a estudos de fonologia do português brasileiro*. Porto Alegre: EDIPUCRS, 4 edition.

Bobenhausen, K. 2011. The Metricalizer: Automated metrical markup of German poetry. In Küper, C., ed., *Current Trends in Metrical Analysis*. Frankfurt am Main: Peter Lang. 119–131.

Brown, T., et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33:1877–1901.

Caldas, P. 2021. Ano passado eu morri, mas esse ano eu não morro: verso absurdo une paraibanos a novas gerações. `https://g1.globo.com/pb/paraiba/noticia/2021/01/31/ano-passado-eu-morri-mas-esse-ano-eu-nao-morro-verso-absurdo-une-paraibanos-a-novas-geracoes.ghtml`. Accessed: 2024-02-19.

Chakrabarty, T.; Padmakumar, V.; and He, H. 2022. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. arXiv preprint arXiv:2210.13669.

Chatzikyriakidis, S., and Natsina, A. 2026. LLMs got rhythm? Hybrid phonological filtering for Greek poetry rhyme detection and generation.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full-FACE poetry generation. In *Proceedings of the 3rd International Conference on Computational Creativity (ICCC)*, 95–102.

Corrêa, N. K.; Falk, S.; Fatimah, S.; Sen, A.; and Oliveira, N. D. 2024. TeenyTinyLlama: open-source tiny language models trained in Brazilian Portuguese. *Machine Learning with Applications* 16:100558.

Corrêa, N. K.; Sen, A.; Falk, S.; and Fatimah, S. 2025. Tucano: Advancing neural text generation for Portuguese. *Patterns* 6(11):101325.

de Araújo, P. A. M., and Mamede, N. J. 2002. Classificador de poemas. In *Conferência Científica e Tecnológica em Engenharia*.

de la Rosa, J.; Álvaro Pérez; Hernández, L.; Ros, S.; and González-Blanco, E. 2020. Rantanplan, fast and accurate syllabification and scansion of Spanish poetry. *Procesamiento del Lenguaje Natural* 65:83–90.

de la Rosa, J.; Álvaro Pérez; de Sisto, M.; Hernández Lorenzo, L.; Díaz, A.; Ros, S.; and González-Blanco, E. 2023. Transformers analyzing poetry: Multilingual metrical pattern prediction with transformer-based language models. *Neural Computing and Applications* 35:18171–18176.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Frolov, D. 2000. *Classical Arabic Verse: History and Theory of ʿArūḍ*, volume 21 of *Studies in Arabic Literature*. Leiden: Brill.

Gervás, P. 2000. A logic programming application for the analysis of Spanish verse. *Computational Linguistics and Intelligent Text Processing* 330–344.

Google DeepMind. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Greene, E.; Bodrumlu, T.; and Knight, K. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 524–533.

Hartman, C. O. 2005. The scandroid 1.1. Software for automated metrical scansion of English verse, `https://academic.hartman.digital.conncoll.edu/Programs.htm`.

Hayward, M. 1996. Analysis of a corpus of poetry by a connectionist model of poetic meter. *Poetics* 24(1):1–11.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Koziev, I. 2025. Automated evaluation of meter and rhyme in Russian generative and human-authored poetry. arXiv preprint arXiv:2502.20931.

Lau, J. H.; Cohn, T.; Baldwin, T.; Brooke, J.; and Hammond, A. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1948–1958.

Logan, H. M. 1988. Computer analysis of sound and meter in poetry. *College Literature* 19–24.

Marco Remón, G., and Gonzalo, J. 2020. Automatic scansion of Spanish poetry without syllabification. *arXiv preprint arXiv:2012.12799*.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

Mistral AI. 2025. Introducing Mistral 3. `https://mistral.ai/news/mistral-3`.

Mittmann, A., and Maia, S. R. 2017. Análise comparativa entre escansões manual e automática dos versos de Gregório de Matos. *Texto Digital* 13(1):157–176.

Mittmann, A. 2016. *Escansão automática de versos em português*. Ph.D. Dissertation, Universidade Federal de Santa Catarina, Centro Tecnológico.

Navarro-Colorado, B.; Ribes Lafoz, M.; and Sánchez, N. 2016. Metrical annotation of a large corpus of Spanish sonnets: Representation, scansion and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 4360–4364.

Navarro-Colorado, B. 2017. A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities* 33(1):112–127.

Oliveira, H. G. 2012. PoeTryMe: a versatile platform for poetry generation. In *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*.

Oliveira, H. G. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Natural Language Generation Conference*, 11–20.

Oliveira, H. G. 2024. Automatic generation of creative text in Portuguese: an overview. *Language Resources and Evaluation* 58(1):7–41.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ormazabal, A.; Artetxe, M.; Agirrezabal, M.; Soroa, A.; and Agirre, E. 2022. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation. arXiv preprint arXiv:2205.12206.

Pires, R.; Abonizio, H.; Almeida, T. S.; and Nogueira, R. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems: 12th Brazilian Conference, BRACIS 2023, Proceedings, Part III*, volume 14197 of *Lecture Notes in Computer Science*, 226–240. Springer.

Valença, A., and Calegario, F. 2025. Experimenting with large language models for poetic scansion in Portuguese: A case study on metric and rhythmic structuring. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.

Walsh, M.; Preus, A.; and Antoniak, M. 2024. Sonnet or not, bot? Poetry evaluation for large models and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 15568–15603.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

Yu, C.; Zang, L.; Wang, J.; Zhuang, C.; and Gu, J. 2024. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 315–325.

HumorSkills: Cognitive, Social and Creative Skills for AI Humor Generation

Sean Kim
Computer Science Department
Columbia University
New York, NY
ssk2245@columbia.edu

Lydia B. Chilton
Computer Science Department
Columbia University
New York, NY
chilton@cs.columbia.edu

Abstract

Humor is a social binding agent. It is an act of creativity that can provoke emotional reactions on a broad range of topics. Humor has long been an outstanding challenge in AI, probably because humans are complex, and humor requires our complex set of skills: cognitive reasoning, social understanding, a broad base of knowledge, creative thinking, and audience understanding. We explore whether giving AI such skills enables it to write humor. We target one audience: Gen Z humor fans. We ask people to rate image caption humor from three sources: 1) highly upvoted human captions, 2) basic LLMs, and 3) LLM captions with humor skills. We find that users like LLM captions with humor skills more than basic LLMs and almost on par with top-rated humor written by people. We discuss how giving AI human-like skills can help it generate communication that resonates with people.

Introduction

Producing humor is a difficult yet quintessential endeavor that underpins the everyday human experience. People use humor to connect with others, to impress others, to point out absurdities, to make light of a bad situation, and to recognize universal struggles - big and small. Despite massive amounts of training data and examples in large language models (LLMs), even state-of-the-art generative AI models are known for being disappointing in their humor abilities (Jentzsch and Kersting 2023). Humor is deeply complex as it depends on cognitive skills like observing details that others might overlook, creative skills such as diverging in multiple different ways (considering multiple viewpoints, multiple narratives, etc.) and social skills such as understanding your audience, understanding human relationship dynamics, and judging what is appropriate to make fun of.

We test whether AI-generated humor that uses social, cognitive and creative skills can come close to human performance. As a research context, we study the specific problem of generating humorous captions for images posted to Instagram (IG). This is a popular humor type within English-speaking Gen Z social media users that is widely made and widely appreciated by a well-defined audience. There is ground truth (in the form of upvotes) for what captions the audience finds funny. It is a fairly complex form of humor that uses both visual and language skills to generate, it applies to a wide variety of input images and requires implicit Gen Z cultural knowledge to generate.

Our method of generating humor with AI uses a divergent and convergent process, injecting it with three different human-like "skills" that current large language models (LLMs) may not emphasize. The input to the system is an image. The first phase of the LLM workflow is divergent thinking. The system starts with the **cognitive skill of observation**: an LLM examines the image and writes a detailed description of what it sees, giving the system a diverse set of potential joke targets. Next, the system generates multiple humorous angles from this description. Instead of having one pathway for creating humor, we use the **creative skill of divergent ideation**, generating jokes along two separate pathways that use different approaches. The first approach is simple: it creates humor strictly related to the image content. The second approach introduces the **social skill of narrative/conflict extrapolation**. Instead of relying solely on the information in the image to derive humor, it finds social conflicts that could be analogous to the image and uses those to drive the humor. Social conflicts could include arguments between a boss and their employees, romantic partners, political parties, parents/children, sibling rivalries, friend group drama, etc. These are all relatable social conflicts that are known to provide fodder for humor. Up to this point, the system will have generated 30 jokes. The workflow then enters a short convergent phase, where an LLM acting as a "Gen Z humor expert" rates the captions and selects the 5 best to return.

We discuss how a skill-based approach can be a useful technique for creative problem solving. For domains like humor where the goal is to create a rich, multidimensional artifact, it is likely that a multidimensional approach is necessary.

Background on Humor Theory

Humor is an intellectual challenge that has been studied by many great Western thinkers including Plato, Kant, and Freud (Morreall 2016). It is a multi-dimensional problem, but it clearly has both social and cognitive elements. Kant originated the incongruity theory of humor, which says that humor happens when we perceive a mismatch between what we expect and what actually occurs—a violation of our mental expectations that resolves in a surprising but sensible way. The act of perceiving an incongruity is a cognitive task

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

relating to logical reasoning. The benign violation theory of humor (McGraw and Warren 2010) builds on expectation violation by adding a social dimension. Typically the expectations that are violated are social norms—like somebody doing something crude, offensive, or inappropriate. However, not all violations are funny; we must have "psychological safety" from the violation expectation. A joke about someone else's country might be funny, but a joke about your country might be offensive. Thus, to be funny, the social norm being violated must be somewhat benign.

Theories of humor assert that humor has structure. At the most basic level, a joke has a setup and a punchline: the setup introduces an expectation, and the punchline delivers a violation. The semantic script theory of humor (Raskin 1985) adds more structure to this. It holds that a joke works by making two contradictory interpretations of the same text fit at once — the setup invites one reading, and the punchline reveals a second. Take Raskin's classic example: *" 'Is the doctor at home?' the patient whispered. 'No,' the doctor's young wife whispered back. 'Come right in.' "* At first, the reader assumes a sick-patient scene (script 1). After the punchline, the reader comes to a different interpretation of the setup: it's not a sick patient, it's someone having an affair (script 2). Both readings were available all along; the way the joke was told allowed the reader to detect the incongruity and discover the second meaning (Suls 1972).

In addition to structure, there are two levels of creativity needed for jokes: a *premise* and an *execution*. A premise is the core comedic idea—the angle, the unexpected connection, the thing the joke is "about." In script theory's terms, the premise is the collision of the two scripts; in incongruity terms, it is the violation itself. Execution is everything that makes that premise land: the phrasing, the brevity, the framing, the choice of words. A strong premise rendered in flat language falls flat, and flawless phrasing with no underlying premise is just noise. Both are needed. This division recurs across comedy traditions under different names. Improvisers distinguish the "game" of a scene—its core comedic premise—from the *heightening* that develops it (Besser, Roberts, and Walsh 2013). Joke writers separate a joke's underlying angle from the "maximizers" used to sharpen it (Toplyn 2014; Knospe 2014). Across these accounts, the premise supplies the core, but lands only through its execution. Writing a joke requires creative work at both levels.

Although there is no definitive formula for humor, many humorists have described a set of social principles for writing jokes. Key themes include:

Jokes should be relatable (Dean 2000; Holloway 2010; Vorhaus 1994; Kaplan 2013; Carter 2001), listeners have to relate and empathize with them in some way. Jokes typically engage human emotions - fear, hope, curiosity, cringing, and other heightened states that raise the stakes that get us to listen and relate to the material (Carter 2001). Often jokes are for an in-group (Hurley, Dennett, and Adams 2011) - using the shared knowledge and experiences of a group to create exclusive material which only that group relates to.

Jokes have details and observations. Observing the difficulties and absurdities of life is a good way to find relatable material (Kaplan 2013). But obvious absurdities tend not to be surprising, so jokes often come from observing details that others likely missed. (Carter 2001; Vorhaus 1994).

Jokes contain narratives that include a point of view. (Carter 2001) Like stories and narratives, jokes often use multiple points of view to see a situation in a new and surprising way. Understanding peoples' thoughts, actions, and behaviors helps fully "act out" (Carter 2001) the story. For example, *"Question: How do you get an Amish person to change a lightbulb? Answer: What's a lightbulb?"* The punchline takes the Amish person's point of view, which is unexpected. Sometimes the narrative can be based on a metaphor that helps the listener see something in a new way, like the many grad school memes comparing the negative "Reviewer #2" who rejected their paper to a movie villain (Weinberg 2017).

Jokes are creative and can be intentionally constructed Almost all humorists who write books about humor echo the design literature that divergent and convergent cognitive processes (Guilford 1950) are helpful processes in writing humor. Exploration is necessary to expand topics (Holloway 2010) (divergent thinking). Semantic Script theory can be used to structure jokes (Raskin and Hempelmann 2002). Many alternatives for punchlines should be explored to find the one that resonates the most (Dean 2000) (convergent thinking).

Overall, jokes require complex social understanding and logical construction. Methods of producing jokes are likely to need both skills intertwined.

Related Work

Computational Creativity and Humor

Computational humor is an outstanding challenge in AI. Even classifying whether or not something is funny is a computational challenge. This has been tried with some limited success by training deep learning classifiers on New Yorker Captions (Shahaf, Horvitz, and Mankoff 2015), and by "unfunning" jokes to create more training data (Horvitz et al. 2024). Wordplay is a common target for computational humor generation, and it relies on linguistic violations rather than social or cultural ones. Thus, many successful systems have produced computational word play (Binsted and Ritchie 1997; Kiddon and Brun 2011; Toplyn 2023c; He, Peng, and Liang 2019; Taylor and Mazlack 2004), but these techniques do tend to be limited to wordplay-based humor, and do not generalize more broadly.

LLMs have opened this as a possibility but humor continues to be a challenge for LLMs like ChatGPT (Jentzsch and Kersting 2023; Mirowski et al. 2024), even with fine tuning (Anonymous 2024). Recently, ChatGPT has been shown to be funnier than jokes written by Turkers (as rated by other Turkers), but not nearly as funny as professionally published humor (Gorenz and Schwarz 2024). Relatedly, in a study of live improvisational comedy (Winters and Van der Stockt 2025) found that chain-of-thought-prompted GPT-4 jokes were rated comparably to those of human comedians, with audiences only slightly favoring the human jokes. However,

the GPT-4 jokes were both selected and performed by the comedians, which undoubtedly added to their funniness.

A common technique for non-wordplay humor is finding and relating unexpected associations (Toplyn 2023a), using multiple humor strategies for associations (Toplyn 2023b). Adding multi-step reasoning can also be applied (Tikhonov and Shtykovskiy 2024), but is not sufficient. These attempts are largely in the right direction. Using associations, reasoning, and creative processes like divergent and convergent thinking make sense. But there are many more dimensions to humor. Considering more human-centric ideas like point of view, observation of details, social understanding, and understanding the audience is also essential for humor generation.

AI Generation Techniques

There are a number of existing techniques shown to improve LLMs' native abilities to generate text. LLMs' understanding of the goal can be improved by prompt engineering (Radford et al. 2019), fine-tuning (Radford et al. 2018), and few-shot prompting (Brown et al. 2020). LLMs can have more human-like production processes through prompt chaining (Wu, Terry, and Cai 2022) or by using a mixture of experts (Cai et al. 2024). LLMs' reasoning abilities can be increased by techniques like chain-of-thought prompting (Wei et al. 2024), thought experiments (Ma et al. 2023), and related approaches. LLMs can automatically close a feedback loop to improve outputs through reflection (Shinn et al. 2023) and self-evaluation (Brodsky 2024). Improving LLMs' theory of mind can also help with perspective taking (Strachan et al. 2024) and other social problems (Chen et al. 2025). When these techniques are used together as agents (Park et al. 2023), they can even combine these abilities to solve multi-faceted problems like software engineering (Yang et al. 2024) by taking different roles. All of these techniques are potentially relevant to get LLMs to use the right combination of social, logical, and creative skills necessary to make jokes.

System

HumorSkills is a system that takes an input image and outputs 5 image captions. The architecture has three key steps that mimic human skills needed for humor. *Visual Detail Extraction* is a step that describes the image in depth in order to make non-obvious observations about it. *Narrative and Conflict Extrapolation* is a step that finds narratives not in the image that could be related to it, to expand the topic of jokes to things that are not just in the image but also analogous to it. *Fine-tuning* the joke generator with examples of good Gen Z humor helps the jokes be more relatable to the target audience by using references, slang, topics, and insecurities that resonate with this group.

The system generates two types of captions: image-focused captions (which comment directly on the content in the image), and narrative-driven captions. Variety is important to humor. Humor relies on surprise, and jokes that are too similar start to become more predictable. Additionally, with an infinite set of input images with different subjects and situations, there are more strategies needed to find a humorous angle that fits the content.

AI Humor Generation Walkthrough

Figure 1 contains a visual diagram and example of intermediate outputs when generating captions for an image. We describe each phase and implementation in detail.

Most steps use GPT-4o for text and image processing such as describing the image, ideation, and narrative/conflict extrapolation. At the time of running the study, GPT-4o was generally the most capable and stable model. However, for one step we used a fine-tuned GPT-3.5. This is only for generating the captions in Gen Z language. The GPT-4o fine-tuning pipeline rejected our training data because it contained "sensitive content." The content is "edgy" in the way that humor often is, but passes Instagram's content policy. Our GPT-3.5 model was trained before such restrictions were put in place.

Visual Detail Extraction The first phase of the system's workflow involves the Visual Detail Extraction component, which utilizes GPT-4o's vision capabilities to analyze the input image. This system incorporates a prompt that asks for a detailed paragraph that explains the who/what/where of the image, distinguishing between identifying the subject of the image, the main action of the image if it exists, and the background elements of the image. This component is responsible for extracting key visual elements such as objects, human expressions, background settings, and any notable aspects that could serve as the foundation for humor.

For instance, in the demolition site example from the system diagram (figure 1), the system identifies a large industrial demolition excavator and a person with a hose spraying the demolition site.

Visual Humor Ideation On top of the visual detail extraction, the system ideates on possible humorous elements from the visual of the image. This incorporates an additional prompt using GPT-4o that intakes the image and asks it to identify and ideate on potential humorous visual elements in the image, whether they are directly humorous elements, such as funny facial expressions, or more analogous elements. For example, for the system diagram (figure 1) image, the system noted the visual contrast of the excavator and person, reminiscent of a David versus Goliath scenario, which provides a foundational metaphor for generating humorous captions.

Narrative and Conflict Extrapolation In this next step, the system generates a narrative and conflict framework by drawing upon common and relatable Gen Z experiences such as work, school, family, social interactions, relationships, and more. The system chains together the results of the previous steps, into a new prompt sent to GPT-4o. The prompt contains the visual details, the visual humor ideation, and a list of common Gen Z experiences, as well as the instruction to "generate narratives that reflect the essence of the image that is set within the framework of the Gen Z experience." This narrative generation adds depth to the humorous captions by applying relatable themes and conflicts



Figure 1: HumorSkills System Diagram. Given an image, the system first extracts visual details with a visual language model, then performs visual humor ideation to analyze the image and propose humorous angles. It then generates ten potential conflicts that could be used to extrapolate the image into a relatable experience. The system then generates humor with and without the narratives, for diversity. A separate instance of the LLM trained to rank Gen Z humor ranks all the captions and returns the top five.

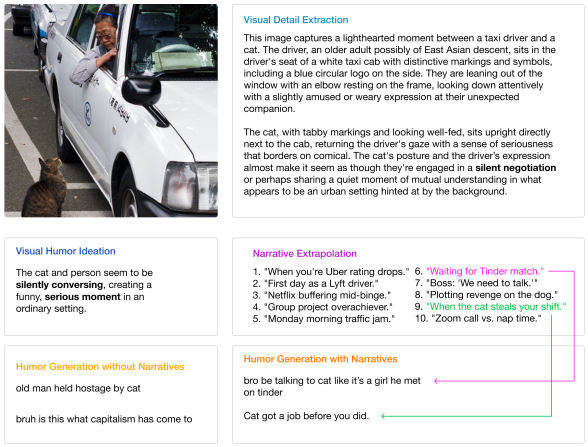


Figure 2: A diagram for how narrative and conflict extrapolation works

to the visual elements identified earlier. For example, in the system diagram 1, jokes like *"me mopping up my last relationship"* are clearly not about the image of a demolition site, it's an analogy to the metaphorical disaster left after a romantic relationship.

Figure 2 shows an example of how narrative/conflict extrapolation differs from more basic visual humor ideation. Both strategies start with a rich visual description of the image. Here, *"the image captures a lighthearted moment between a taxi driver and a cat...the cat's posture and the drivers expression make it seems as though they're engaged in a silent negotiation."*. Whereas the visual humor ideation without narratives lead to plausible jokes *"old man held hostage by cat"* they stay quite close to the image. In contrast, narrative extrapolation ideates multiple types of conflicts to generate jokes that are not about the image, but about social situations analogous to image. The two chosen by the LLM humor rater include: 1) a romantic conflict of waiting for a tinder match, *bro be talking to cat like it's a girl he met on tinder."* And 2) the workplace conflict of getting a shift stolen, *Cat got a job before you did."* Both successfully introduce a social conflict and relate it back to the image.

Humorous Caption Generation Following the narrative and conflict extrapolation, the system generates humorous captions using a fine-tuned GPT-3.5 model trained on Gen Z Instagram comments. It produces 15 image-focused captions (based on the visual description and humor ideation) and 15 narrative-driven captions (which additionally incorporate the narrative/conflict extrapolations), for a total of 30 candidates. The two pathways add variety: image-focused captions like *"bro out here getting paid $8 an hour to spray some water on some bricks"* stay close to the visual, while narrative-driven captions like *"The entitled bro you tried to make the group presentation with"* reach for analogous social situations.

Caption Ranking using Gen Z Agent The final component of the system architecture is the Caption Ranking and Filtering Agent, a GPT-4o-based agent fine-tuned to evaluate humor from a Gen Z perspective. This agent receives the list of 30 total captions from the narrative and visual humor-based caption generations and ranks the captions generated in the previous stage based on humor, relatability, and alignment with the image and narrative.

As illustrated in our system diagram, this agent ranks captions such as *"Me mopping up my last relationship"* and *"me pulling the emotional weight of the friend group"* based on their relevance to Gen Z humor. Captions that fail to meet the humor threshold are filtered out, such as *"Demolition worker really said 1v1 me bro,"* because although a phrase like "1v1 me bro" invokes Gen Z phrases, the content of the caption seems less relevant and relatable than a caption talking about school or relationships, ensuring that only the most effective and relatable captions are presented to the user.

Fine-tuning To make sure the LLM could generate content in the style of Gen Z, we fine-tuned a GPT-3.5 model on a dataset of 80 humorous Gen Z image captions. We first identified three popular Gen Z Instagram accounts that engaged in daily image-captioning contents (each account had over 400,000 followers). These IG accounts post images that are not inherently funny, then let the followers write potential captions for the image in the comment section. The most upvoted comments are typically the funniest captions for that image. We took the top 5 comments for 16 images (80 captions total) and used them to fine-tune GPT-3.5 to replicate their style.

Examples of the visual description of the images in addition to an explanation of potential humorous elements of the image were written in the fine-tuning prompts, then followed by the actual comment itself. This reflected the visual extraction and humor ideation being incorporated into the prompt of our current system.

Study 1: Humor Skills vs. GPT-4o vs. Top Humans

Methods

We ask people to rate humorous captions for images taken from 8 popular Instagram (IG) humor captioning accounts (separate from the training data accounts). For each image, users were shown 15 captions: 1) The top 5 most upvoted captions from IG 2) 5 captions from GPT-4o 3) 5 captions from HumorSkills

Data was collected from an online survey. Users would see an image and then rate 15 captions for it on a scale of 1 to 5, where 1 was "not funny", 3 was "somewhat funny", and 5 was "very funny". For each user, the images were presented in a random order, and the 15 captions for each image were also shown in a random order to avoid any possible ordering effects. Users were not told the condition of each caption.

The survey was distributed through email announcements to clubs and classes at a local university. Users were not told that AI was involved in the humor writing. The announcements advertised being for people who identify as "Gen Z" and who "like Instagram caption humor". These qualifications were added to ensure the raters were part of the target

Table 1: **HumorSkills vs. GPT-4o vs. Top Human Captions.** Comparison of user ratings for captions from HumorSkills, GPT-4o, and the top 5 Instagram comments, using a generalized linear mixed model (controlling for fixed effects of each caption). HumorSkills jokes were rated significantly funnier than GPT-4o jokes (.213 points higher on a scale of 1 to 5). HumorSkills jokes were as funny as the top Instagram posts (only 0.078 points less funny, which is not statistically significant).

Variable	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Intercept (HumorSkill rating)	2.273	0.040	56.601	0.000	2.194	2.351
GPT-4o rating	**-0.213****	0.040	-5.285	**0.000**	-0.291	-0.134
Top5 IG rating	**0.078**	0.040	1.934	**0.053**	-0.001	0.157
Group Var	0.323	0.025				

*Note: ** denotes significance at the $p < 0.001$ level.*

audience, and most university students fit this demographic. The survey took about 30 minutes and users were paid $10.

We test the following hypotheses:

H1. HumorSkills captions will receive higher humor ratings than GPT-4o captions.

H2. HumorSkills captions will receive humor ratings comparable to top-rated Instagram captions.

Results

In total, 32 people responded to the survey: 5 male, 9 female, and 18 declined to say their gender. Fourteen of 32 respondents were aged 18–25. One selected "25+", and 17 declined to say their age group.

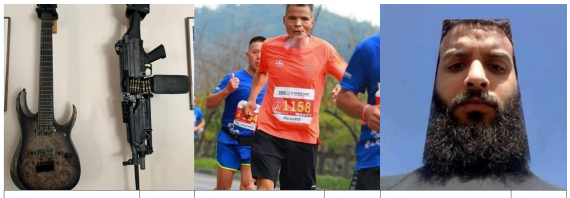
To test the relative humor score of the three conditions, we ran a Generalized Linear Mixed Model (GLMM) to estimate the funniness rating of each condition. The response types were ordinal numbers (ratings), the fixed effects were the caption type (IG, GPT-4o, HumorSkills), and the random effects were individual rater IDs and image IDs.

H1: HumorSkills vs. GPT-4o On the whole, captions from our system were rated as 2.27 (out of 5) for funniness. Captions written by GPT-4o were rated less funny (by 0.213 points), which is statistically significant at a $p < 0.001$ level. This indicates our system writes more humorous captions than a state-of-the-art VLM with prompt engineering. Results are shown in Table 1. Thus **H1 is supported. HumorSkills is rated funnier than GPT-4o.**

H2: HumorSkills vs. top IG captions On the whole, the top IG captions were rated only slightly better (0.078 points on a 5-point scale). But the difference was not statistically significant at the 5% level. However, it was very close ($p = 0.053$). This indicates that the system is competitive with the top IG captions. This effect could disappear with more data, but this dataset had over 6000 observations, and the 0.078 difference in average score is only 2% of the 1-5 scale. Thus **H2 is supported. Based on our sample, HumorSkills was rated as funny as the top 5 Instagram comments.** Or more precisely, HumorSkills was not statistically less funny than the top IG captions.

Qualitative Discussion

Whereas HumorSkills and Instagram comments tied for the highest rated captions, there are still qualitative differences between some of the jokes, and room for improvement. To analyze this, we identified the top-rated captions for each of the caption sources: Instagram, GPT-4o, and HumorSkills, and compared them to the highest-rated captions from the other two sources. We propose possible reasons for the top ratings, and ways to possibly improve the system. Figure 3 shows the images and the captions analyzed.



	Caption	Score	Caption	Score	Caption	Score
IG	American instruments	3.39	bro playing on max difficulty	3.14	Bro ready for his job interview at mojang	3.00
GPT-4o	Which one shreds harder? 🤔🤟	2.50	Running on fumes, literally	2.81	When you accidentally click "Crop" in real life	2.79
Humor Skills	me picking weapons in a video game	2.83	They really said "let's race to the hospital"	2.52	minecraft character looking ahh	3.62

Figure 3: Top rated image captions (marked in green) for Instagram, GPT-4o, and HumorSkills. (The top-scoring captions for the other sources for the same image are shown for comparison). From left to right, the figure includes three images: 1) The image with Instagram's top-rated caption (3.39/5). The image shows a guitar next to a machine gun. 2) The image with GPT-4o's top-rated caption (2.81/5). The image shows man running a race while smoking a cigarette and, 3) the image with HumorSkills' top-rated caption (3.62/5). The image shows a bearded man with his head cropped to a square.

Sharper social critique. The overall top-rated IG caption is "American instruments," for an image which juxtaposes an electric guitar and a machine gun. The joke critiques US gun culture. This caption is good because it's short but also a sharp social critique of a big problem. The HumorSkills caption for this image draws an analogy to video games, critiquing them for having silly weapon choices. It's relatable to Gen Z, but not nearly as sharp a critique as the top IG captions. Perhaps HumorSkills could be expanded to include more social critiques, from topics in the news. The GPT-4o caption is apt, but least funny: "Which one shreds harder?" It connects the two objects through unexpected word play

("shred"), but has no critique.

Learning the appropriate level of seriousness. The top-rated GPT-4o caption says *"running on fumes, literally"*, for an image with a runner smoking a cigarette. This is apt and connects the odd parts of the image (running and smoking), but it again uses wordplay. The IG caption was the highest rating for this image, *"bro playing on max difficulty."* It uses a video game metaphor to critique the runner smoking. The HumorSkills caption did least well for this image, it uses dark humor to critique the runner for smoking and point out the absurdity: *"They really said 'let's race to the hospital."'* Which is apt, but perhaps too serious and not benign enough for some readers. The system should do a better job of screening for jokes that are too "dark" or serious. The appropriate level of seriousness might have to be learned for each audience, or even individual audience member.

Mining more social references and in-group slang. The top-rated HumorSkills caption was the highest-rated caption in the entire study. It reads *"minecraft character looking ahh."* (The word "ahh" is Gen Z slang roughly meaning "ass".) Although this might not be broadly funny, it is funny to Gen Z because of this insider slang, because it's short, and because it is a critique of his looks related to Minecraft. The IG caption was rated less funny (*"bro ready for his job interview at mojang"*). It also relates the image to Minecraft (Mojang the developer of Minecraft). Gen Z grew up on Minecraft, and jokes about it resonate. The GPT-4o also makes fun of his hairstyle but relates it to cropping an image, rather than Minecraft, which is not as socially relevant. Overall, HumorSkills uses social references like Minecraft well, as well as in-group slang. Even more fine-tuning on Gen Z language could improve this ability. In the future, the fine-tuning doesn't necessarily need to come only from humor.

Study 2: HumorSkills performance on non-target images

The HumorSkills approach was designed and fine-tuned on a particular type of humor—Gen Z Instagram image captioning. It is possible that this approach is overfit to one type of image, especially since it uses fine-tuning to understand the types of images and captions this audience finds funny. Moreover, images in this dataset are not randomly selected—they are selected by people for being innately interesting or unusual. Other types of images that are not innately unusual could be harder to caption humorously.

To investigate how well the HumorSkills approach generalizes to other types of images, we compare how well HumorSkills and GPT-4o write captions for two types of images that are less likely to be inherently interesting.

1. **Camera roll images**. Camera roll images are the ordinary photos people take in their day-to-day lives. Typically they contain people, places, and things that are meaningful to the camera owner, but they are not generally interesting, noteworthy, meaningful, or funny.

2. **Museum Art**. Art is very different than photos. Although some of it depicts people doing things. Most of those things are historical, and not particularly relatable. Some paintings are abstract and do not concretely depict any discernible subject. Additionally, many of the art pieces are not paintings, but objects like vases or chairs, which could be challenging to make relatable and create a narrative for.

We randomly selected 30 images from the Flickr Image Dataset(Young et al. 2014) and 30 museum art images to test on our target audience. We take 15 images from the Museum of Modern Art (The MoMA) and 15 images of art from the Metropolitan Art Museum (The Met). For every image, we generate 5 HumorSkills captions and 5 GPT-4o captions. We create two separate surveys for rating the captions - one for art and one for camera roll images. As before, we randomize the order of the images and the captions to mitigate ordering effects. We recruited raters with the same process as the previous study.

We hypothesize that HumorSkills' captions will be rated more funny than GPT-4o captions for both new image sets. We did not compare to top-ranked human-written captions because these images were not taken from Instagram.

Results: Camera Roll Humor Ratings

As before, we ran a GLMM regression with image ID and user ID as random effects. There were a total of 4765 ratings. There were a total of 21 respondents: 10 female, 4 male, and 7 declined to say their gender. Twelve respondents were aged 18–25, 2 were aged "25+", and 7 declined to say their age group.

The results show that HumorSkills is 0.291 points higher than GPT-4o (2.19 vs. 1.9 out of 5). This difference is statistically significant with $p < 0.001$. Results are shown in Table 2.

Results: Museum Art Humor Ratings

There were a total of 4044 ratings. There were a total of 17 respondents, 7 female, 3 male, and 7 declined to say their gender. 8 respondents were aged 18-25, 2 were "25+", and 7 declined to say their age group.

The results show that HumorSkills is 0.18 points higher than GPT-4o (2.18 vs. 2.00 out of 5). This difference is statistically significant with $p < 0.001$. Results are shown in Table 3.

For both image sets, the image-level variance component was estimated at approximately zero, indicating that systematic differences between images explained negligible variance once caption-level fixed effects and rater random effects were accounted for; nearly all unexplained variance was attributable to individual rater differences.

This shows that HumorSkills can perform well on images that do not have obvious humorous qualities.

Qualitative Discussion

Benefits of Narrative/Conflict for Out-of-Domain Images
We expected the out-of-domain images to be harder to caption than the Instagram set, because Instagram humor accounts deliberately select for unusual or visually arresting subjects. Camera-roll photos, by contrast, are ordinary and offer no obvious joke target, and museum art is often abstract or historically distant from everyday life. With little inherent comedic material in the image itself, the system has to

Table 2: **Camera Roll GLMM Results.** Comparison of user ratings for captions generated by HumorSkills and GPT-4o on the Flickr image dataset of camera roll images. Using a generalized linear mixed model (controlling for fixed effects of each caption and random effect for individual differences in humor rating), HumorSkills captions were rated significantly funnier than GPT-4o captions (0.291 points higher on a scale of 1 to 5).

Variable	Coef.	Std.Err.	z	P> $\|z\|$	[0.025	0.975]
Intercept (GPT-4o rating)	1.901	0.054	35.357	0.000	1.796	2.006
HumorSkills rating	**0.291****	0.022	13.207	**0.000**	0.248	0.334
Group Var	0.000					
User_ID Var	1.260	0.116				

*Note: ** denotes significance at the $p < 0.001$ level.*

Table 3: **Museum Art GLMM Results.** Comparison of user ratings for captions generated by HumorSkills and GPT-4o on museum art images. Using a generalized linear mixed model (controlling for caption-level fixed effects and random effects for users), HumorSkills captions were rated significantly funnier than GPT-4o captions (0.175 points higher on a 1–5 scale).

Variable	Coef.	Std.Err.	z	P> $\|z\|$	[0.025	0.975]
Intercept (GPT-4o rating)	2.000	0.061	32.774	0.000	1.881	2.120
HumorSkill rating	**0.175****	0.023	7.677	**0.000**	0.130	0.219
Group Var	0.001					
User_ID Var	1.387	0.147				

*Note: ** denotes significance at the $p < 0.001$ level.*



	Caption	Score	Caption	Score	Caption	Score
Humor Skills	That's one way to deal with the **Space Force** budget cuts.	2.65	Medical staff having a discussion on whether it's **worth it to save the uninsured man**.	2.76	How your **charger** looks when you leave for 5 seconds.	2.76
GPT-4o	he really said "scope out the stars"	1.88	When the surgery team gets a crash course in the break room	2.53	When you attempt abstract art and accidentally create a web of confusion	1.64

Figure 4: Narrative Generation examples. Images across all 3 datasets with notable narrative generation. Instagram (left), Flickr (center), Museum Art (right)



	Caption	Score	Caption	Score	Caption	Score
Humor Skills	**Forehead so high**, it's getting its own zip code	2.39	dude looks like he's about to write a **manifesto**	2.87	bruh looks like they're waiting for a text back from the **shogun**	2.31
GPT-4o	That moment when you're caught off guard mid-snack, and now it's awkward	1.65	That "I came to the bar for solitude" vibe	2.13	That 'I'll start my homework in 5 minutes' energy.	2.15

Figure 5: Visual detail extraction examples. Images across all 3 datasets with notable visual extractions. Instagram (left), Flickr (center), Museum Art (right)

find a premise somewhere else. This is what the cognitive technique of narrative extrapolation does.

This shows up clearly in Figure 4. The first image, an unremarkable telescope, prompts a poorly-rated pun from GPT-4o; HumorSkills instead reaches for a topical political analogy: "That's one way to deal with the Space Force budget cuts." It use an external narrative with a rich/poor conflict to supply the humor the image itself doesn't. The second image, a break-room photo of doctors talking, again draws a pun from GPT-4o, while HumorSkills brings in a narrative of social conflict: *"discussing whether it's worth it to save the uninsured man."* The third, a fully abstract painting of intersecting colored lines, has nothing concrete to make fun of at all; HumorSkills reinterprets it through a relatable metaphor: *"how your charger looks when you leave for 5 seconds."* In each case, the joke comes from a narrative the image gestures toward but does not contain.

Benefits of Observation for Out-of-Domain Images A cognitive technique for finding jokes is observation: noticing details that others would miss. HumorSkills does this with an explicit visual detail extraction phase. GPT-4o is also a vision model and recognizes objects in images natively, but the explicit extraction step seems to make HumorSkills notice more, and notice differently.

Figure 5 shows examples from all three image datasets. In the first (from Instagram), the main subject has an unusually tall forehead; HumorSkills observes this and lands a well-rated joke about it, while GPT-4o reaches for a less interesting detail (eating snacks). In the second (from the Flickr camera-roll set), a man sits alone at a restaurant; HumorSkills picks up on his visual similarity to Hitler and

writes a sharp joke about composing a manifesto, while GPT-4o sees only that he "came for the solitude." In the third (from the Museum Art set), a Japanese woman in a kimono lies reading a book; HumorSkills registers the book, her forlorn expression, and the Japanese style, and writes *"bruh looks like they're waiting for a text back from the shogun."* GPT-4o's caption reads: *"That 'I'll start my homework in 5 minutes' energy"*. It is relatable but ignores the painting's visual cues entirely. In each case, the explicit observation step surfaces a detail the baseline either missed or underweighted, and that detail becomes the joke.

Discussion

Giving Human-like Skills to AI

This paper showed that for one form of humor — Gen Z style Instagram image captioning — HumorSkills was rated significantly funnier than GPT-4o's native humor. Additionally, HumorSkills was rated as funny as the top five most-upvoted human-written captions. This shows that while LLMs are not particularly funny out of the box, they can be made as funny as humans.

We attribute this to a set of explicitly added "skills," each addressing a different shortfall of native LLM humor generation. The cognitive skill of visual detail extraction found sharper, less obvious joke targets, closer to those of the human captions than GPT-4o's. The social skill of narrative and conflict extrapolation broadened the base of relatable targets, moving from the literal objects in the image to using them as metaphors for relatable situations such as relationship disasters, teamwork breakdowns, and the burden of student loans. The creative skill of divergent and convergent thinking allowed the model to expand the input into multiple joke premises and then select the best for the target audience, ranked by an LLM trained on Gen Z humor taste.

Each of these skills replicates a technique humorists have long described as useful: observing non-obvious details, finding relatable narratives and unexpected points of view, and treating joke-writing as a creative act of divergence and convergence. These are techniques current LLMs do not apply on their own. Asked to caption an image, a model like GPT-4o tends to name the most salient object and reach for wordplay; it does not, by default, scan for the non-obvious detail, search for an analogous social situation, or weigh what a specific audience will find funny. We show these skills can be added explicitly — and future foundation models may come to exhibit them inherently.

Humor as a Combination of Skills, Not a Single Trick

HumorSkills treats humor not as a single trick but as a combination of mechanisms working together. Recall from the background that humor operates on two levels: a premise — the core comedic idea — and an execution that makes the premise land. HumorSkills' skills span both. Visual detail extraction and narrative extrapolation expand the space of premises the system can draw on; narrative extrapolation also supplies material for heightening, such as exaggerated characters and emotionally charged situations. The fine-tuned Gen Z joke writer handles execution, rendering premises in the shorter, punchier language the audience rewards. No single one of these is the driver of a funny caption; the gains come from running them together to craft both premise and execution. This is why we are cautious about attributing the result to any single component: humor here looks less like a single capability than like several working together, mirroring the intertwining of logical construction and social understanding that humor theorists describe in human joke-writing.

Limitations and Future Work

This study targeted only one form of humor: Gen Z Instagram captions, which anchor the joke in an input image. Jokes can be anchored in many other ways: news headlines, mundane observations, or formats like stand-up that require generating both setup and punchline. Future work should explore what additional skills these forms demand.

The clearest near-term improvement is a stronger ranker. Many captions fail because they don't tie back to the image or because their logic never resolves into a coherent second script. A ranker that reasons explicitly about whether a joke's incongruity resolves could filter these out.

Most everyday humor is made within a friend group, leveraging insider knowledge — personality quirks, embarrassing moments, shared misery — to heighten relevance in benign ways. Future work could explore how to elicit such knowledge from a target group and personalize humor generation accordingly.

Our baseline used a simple GPT-4o prompt; more elaborate prompting might close some of the gap. Testing prompts with and without skill scaffolding would clarify how much of the gain comes from the skills themselves versus from prompting more carefully.

Finally, we did not measure joke originality. LLMs trained on internet-scale data risk producing jokes that are retellings of material in the training data (Jentzsch and Kersting 2023; Veale 2024). Because our system grounds each joke in the specific details of a given image and a social analogy chosen for that image, its outputs are more likely to be original to the input (Amir 2025) — though a rigorous measure remains future work.

Conclusion

Humor is a highly valued human ability. It has the power to teach, to entertain, and to connect people. Centuries of humor theory argue that humor is complex; it requires social, cognitive, and creative skills. In this paper, we study whether giving LLMs human-like skills for humor generation can improve their ability to write effective humor. Since humor is often about finding relatable connections between people, we focus on generating only one type of humor: Gen Z Instagram caption humor. We run two studies showing that LLMs with humor skills are rated as funnier than LLMs alone, and as funny as the top-rated human-written captions. This points to a future where LLMs can potentially write with similar social skills as people to produce emotional reactions and human bonding.

References

Amir, O. 2025. Are AI-generated jokes truly original? charting the "joke space". In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

Anonymous. 2024. FunLMs: Methods for fine-tuning LLMs to generate humor. In *Submitted to ACL Rolling Review - June 2024*. under review.

Besser, M.; Roberts, I.; and Walsh, M. 2013. *The Upright Citizens Brigade Comedy Improvisation Manual*. New York, NY: Comedy Council of Nicea.

Binsted, K., and Ritchie, G. 1997. Computational humor. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 97, 1089–1094.

Brodsky, S. 2024. Who watches the ai watchers? the challenge of self-evaluating ai. `https://www.ibm.com/think/news/ai-testing-advances`. Accessed: 2025-02-06.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2024. A survey on mixture of experts.

Carter, J. 2001. *The Comedy Bible: From Stand-up to Sitcom–The Comedy Writer's Ultimate "How To" Guide*. New York, New York, USA: Touchstone.

Chen, R.; Jiang, W.; Qin, C.; and Tan, C. 2025. Theory of mind in large language models: Assessment and enhancement.

Dean, G. 2000. *Step by Step to Stand-Up Comedy*. Portsmouth, New Hampshire: Heinemann Drama.

Gorenz, D., and Schwarz, N. 2024. How funny is chatgpt? a comparison of human- and a.i.-produced jokes. *PLOS ONE* 19(7):1–13.

Guilford, J. P. 1950. Creativity. *American Psychologist* 5(9):444–454.

He, H.; Peng, N.; and Liang, P. 2019. Pun generation with surprise. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1734–1744. Minneapolis, Minnesota: Association for Computational Linguistics.

Holloway, S. 2010. *The Serious Guide to Joke Writing: How To Say Something Funny About Anything*. Great Yarmouth, UK: Bookshaker.

Horvitz, Z.; Chen, J.; Aditya, R.; Srivastava, H.; West, R.; Yu, Z.; and McKeown, K. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models.

Hurley, M. M.; Dennett, D. C.; and Adams, Jr., R. B. 2011. *Inside Jokes: Using Humor to Reverse-Engineer the Mind*. Cambridge, MA, USA: The MIT Press.

Jentzsch, S., and Kersting, K. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (WASSA)*, 325–340. Toronto, Canada: Association for Computational Linguistics.

Kaplan, S. 2013. *The Hidden Tools of Comedy: The Serious Business of Being Funny*. Studio City, CA: Michael Wiese Productions.

Kiddon, C., and Brun, Y. 2011. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, 89–94. Stroudsburg, PA, USA: Association for Computational Linguistics.

Knospe, H. 2014. *Generating Humour: Analysing and Implementing a General Theory of Verbal Humour*. Doctoral dissertation, University of Leipzig.

Ma, X.; Mishra, S.; Beirami, A.; Beutel, A.; and Chen, J. 2023. Let's do a thought experiment: Using counterfactuals to improve moral reasoning.

McGraw, A. P., and Warren, C. 2010. Benign violations: Making immoral behavior funny. *Psychological Science* 21(8):1141–1149.

Mirowski, P. W.; Love, J.; Mathewson, K.; and Mohamed, S. 2024. A robot walks into a bar: Can language models serve as creativity support tools for comedy? an evaluation of llms' humour alignment with comedians. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 15. Rio de Janeiro, Brazil: ACM.

Morreall, J. 2016. Philosophy of humor. The Stanford Encyclopedia of Philosophy. `https://plato.stanford.edu/entries/humor/`.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Raskin, V., and Hempelmann, C. F. 2002. A computer model of joke generation based on script oppositions. In Behrendt, J., ed., *The April Fools' Day Workshop on Computational Humor*. Oldenburg, Germany: BIS-Verlag. 59–71.

Raskin, V. 1985. *Semantic Mechanisms of Humor*. Dordrecht, Netherlands: D. Reidel Publishing Company.

Shahaf, D.; Horvitz, E.; and Mankoff, R. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, KDD '15, 1065–1074. New York, NY, USA: ACM.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36. Publisher Copyright: © 2023 Neural information processing systems foundation. All rights reserved.; 37th Conference on Neural Information Processing Systems, NeurIPS 2023 ; Conference date: 10-12-2023 Through 16-12-2023.

Strachan, J.; Albergo, D.; Borghini, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* 8:1285–1295.

Suls, J. M. 1972. A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Goldstein, J. H., and McGhee, P. E., eds., *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*. New York: Academic Press. 81–100.

Taylor, J. M., and Mazlack, L. J. 2004. Computationally recognizing wordplay in jokes. In *In Proceedings of CogSci 2004*.

Tikhonov, A., and Shtykovskiy, P. 2024. Humor mechanics: Advancing humor generation with multistep reasoning.

Toplyn, J. 2014. *Comedy Writing for Late-night Tv: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-form Comedy*. Twenty Lane Media, LLC.

Toplyn, J. 2023a. Witscript 2: A system for generating improvised jokes without wordplay.

Toplyn, J. 2023b. Witscript 3: A hybrid ai system for improvising jokes in a conversation.

Toplyn, J. 2023c. Witscript: A system for generating improvised jokes in a conversation.

Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*, 236–244.

Vorhaus, J. 1994. *The Comic Toolbox: How to Be Funny Even If You're Not*. Los Angeles, CA: Silman James Press.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc.

Weinberg, J. 2017. How bad is reviewer 2, actually? data from a philosophy journal. Online; Daily Nous blog post. Accessed 20 May 2025.

Winters, T., and Van der Stockt, S. 2025. Evaluating humor generation in an improvisational comedy setting. *Computational Linguistics in the Netherlands Journal* 14:505–523.

Wu, T.; Terry, M.; and Cai, C. J. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery.

Yang, J.; Jimenez, C. E.; Wettig, A.; Lieret, K.; Yao, S.; Narasimhan, K.; and Press, O. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 50528–50652. Curran Associates, Inc.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, 67–78.

Experiments in Text-2-Tool-Use for Relatable Computational Creativity

Simon Colton

School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

s.colton@qmul.ac.uk

Abstract

We propose *relatable computational creativity* as an umbrella term for various theoretical and practical developments, such as explainable computational creativity, artificial authenticity, framing, embodiment, intrinsic motivation, artificial curiosity and creativity theatre. In overview, if a creative AI agent is relatable, then people can imagine themselves doing something similar to it for similar reasons. In this context, we explore the ability for AI systems to use the same tools that people use. Experimenting with a casual creator app for producing decorative art, we investigate how it can learn to employ the tools it makes available for users. Working with a vision-language model affords an interface where users can request that an image be altered in such a way as to achieve a particular look. The system responds with a series of tool actions shown offline to reliably alter images to achieve the desired look. We present experimental results showing that such tool usage can be learned, and describe the approach with respect to aspects of relatable computational creativity.

Introduction and Motivation

When a character in a novel has a similar backstory, attitudes, attributes or situation to the reader or someone they know, this can make the character more engaging and the story more resonant. Maslej, Oatley, and Mar (2017) put this succinctly:

> ... we know characters such as Sherlock Holmes, Anna Karenina, and Mrs. Dalloway in something like the way we know people in our daily interactions. We may think about these characters' intentions, respond emotionally to their actions, and invest ourselves in their outcomes, even though we know they are not real.

Such *relatable* characteristics can afford a bond between reader and character, so that when dramatic events occur, the reader cares about that character and is invested in the story.

When a visitor to an art gallery views a painting, in one sense, the experience is relatable: the artist has made marks with pigment on a canvas, as the visitor could. However, the visitor may marvel in elements of the painting or its production that seem beyond their own physical or mental abilities. As per Colton (2008), these could be many things, including *skill*, expressed through finesse or innovation in technique; *imagination* expressed through subject matter choices and perspective; or *appreciation* expressed through style, artistic treatment or political projections. So, just as relatable characters in novels can act as a platform for dramatic events to unfold, relatable characteristics in artists and processes can act as a platform for the projection of notions of creativity.

Generative AI techniques have recently become powerful enough to rapidly transition from university research laboratories and artist studios to industrial-scale worldwide adoption. Appropriately, these have focused on ease-of-use through prompting, quality of output and fidelity to specifications. However, end-to-end production via text-2-X techniques have, to some extent, excluded creative professionals. For instance, music lovers can now listen to audio files produced directly from natural-language prompts, with no need for composers or musicians. Likewise, visual art lovers can prompt generative AI systems to find imagery including content, styles and perspectives that satisfy their needs and tastes.

Left behind somewhat in the rush to industrialise end-to-end generative AI is the notion that creative AI systems can be entertaining, inspirational and educational to people interested in crafting artefacts rather than simply prompting for them (whether at an amateur or professional level). Within the context of casual creators (Compton and Mateas, 2015), we look here at ways in which a creative AI agent can be relatable while making artefacts, in order to educate and inspire amateur creators. In particular, we focus on user-centric generative systems which learn how to employ their own tools in order to better serve users in a relatable way.

In the next section, we expand on the notion of *relatable computational creativity*, by surveying research into implementation of human-like motivations, behaviours and evaluation techniques surrounding creative acts. We introduce a broad definition and suggest some refinements which could form a framework for implementing, describing and assessing future generative systems. Following this, we describe the context of casual creation and the *Art Done Quick* application used in our experiments. We then describe how we have used the MobileCLIP language-vision model (Vasu et al., 2024) in experiments to enable Art Done Quick to find sequences of tool usages to transform an image in response to a user's prompt. We present the results of the experiments which show that it is possible to discover tool usages which can be controlled via text. We then frame the approach in terms of aspects of relatable computational creativity, and end by describing some future directions for this work.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Relatable Computational Creativity

We start from the position that interacting with a generative AI system onto which we can genuinely project impressions of creativity could have similar benefits in terms of inspiration, education and entertainment that working with a creative person brings. If we draw from the literature on human creativity, Fink et al. (2012) showed that "cognitive stimulation via common or moderately creative ideas [from other people] was effective in improving [the] creativity" of participants asked to ideate alternative uses of everyday objects. Moreover, a study by Katz et al. (2022) showed "that a short narrative about the creativity of a person can produce implicit impressions of that person as creative," and Colton (2008) suggested that understanding the creative processes that went into producing an artefact can increase the value people attribute to it. More generally, Decety and Grèzes (2006) highlight the power of imagining the behaviour of others, arguing that "mental simulation may be a representational tool to understand the self and others".

When the creative counterpart in interactions switches from a person to a computational process, it can be difficult to understand the motivational, generative or evaluational processes the software employs. If we look at the two most successful generative AI techniques currently, we see that neither of them affords a particularly relatable process. Diffusion models produce artefacts by starting with a noisy image and a conditioning prompt and turning the noise into appropriate content over a series of iterations (Yang et al., 2024). In an equally unrelatable way, Transformer models take a partially formed artefact such as a text prompt and repeatedly calculate how to extend it by a single token (Lin et al., 2022).

In computational creativity research, there have been many projects leading to the implementation of human-like creative behaviours in AI systems. While these projects had diverse motivations, outcomes and applications, we propose to introduce the umbrella term *relatable computational creativity* to highlight that the behaviours implemented add to an overall picture of AI systems that people can interact naturally with, potentially as creative equals, and certainly as more than generative tools (Ventura, 2016). We prefer the term *relatable* to the term *human-like* because the latter might imply that simulating all elements of humanity is an aim, when subsets thereof could be beneficial. Just as relatable fictional characters in books are valuable to story telling even though they don't exist in the usual context of human life, relatable AI agents can be valuable to creative processes even though they don't exist in the usual context of human creative life. There are observer issues in computational creativity, and keeping a clear separation between human and artificial creativity may be beneficial (Colton et al., 2014).

In broad terms, a relatable computational creativity system employs some processes leading to the production of an artefact (or problem solution) which an average human observer could imagine him/herself undertaking personally. The processes could include any simulated motivations, generative behaviours and/or evaluations that the system undertakes, and we can imagine a system with more relatable processes being more relatable overall. With a partial survey of the literature on computational creativity, we can suggest the following five main areas where relatable processes were studied.

• **High Level Motivations:** Most people create artefacts for a reason, ranging from self expression to financial gain. This is such an important element of creative behaviour that a lack of intentionality in AI systems can make them seem unrelatable, firmly positioning them as just tools. The question of intent was raised as part of The Painting Fool project in (Krzeczkowska et al., 2010) and (Colton and Ventura, 2014). Guckelsberger, Salge, and Colton (2017) raised the question of why AI systems should be creative, providing empowerment maximisation as an intrinsic motivation for creative AI agents (Guckelsberger, 2020). A number of other relatable reasons for AI systems to be creative have been studied. These include exhibiting artificial curiosity and novelty-seeking, as explored in (Saunders, 2001). Saunders and Bown (2015) also introduced the notion of *computational social creativity*, studying creative AI agents that interact in a community, and Colton and Banar (2023) proposed adding to artistic cultures as a motivation for computational creativity.

• **Explainability:** Relating to someone or something involves understanding, empathizing and connecting, which is not always possible with generative AI systems. The field of explainable AI (Abusitta, Li, and Fung, 2024) acknowledges that some of the most powerful AI techniques, specifically those from deep learning research, have underlying processes that are difficult to understand. Llano et al. (2020) introduced the notion of *explainable computational creativity* to focus on creative AI systems, providing some design principles that can give computational creativity systems a voice for a deeper user interaction. This built on the study of how AI systems can add value to artefacts by *framing* their work in terms (amongst other things) of descriptions of the processes undertaken, internal evaluations and relation to the work of others (Cook et al., 2019). In generative deep learning, the process of chain of thought reasoning (Chen et al., 2025) has been introduced as a technique for interaction with LLMs, whereby a difficult problem (or generative task) is broken into stages, with separate prompts for each. One of the benefits this brings is more transparency and insight into the process.

• **Embodiment and Simulation:** There is perhaps no more relatable process than a demonstration where an AI agent performs actions in a way that is easily replicable by a person. When this is done robotically in physical reality, the effect is heightened, and there has been much research into *embodied computational creativity*. Moruzzi (2022) showed that "the physical dimension of artificial systems interacting with human artists contributes to the perception of the interplay between artificial and human agents as a creative collaboration." Moreover, Guckelsberger et al. (2021) provide a systematic review and prescriptive analysis of embodied computational creativity, suggesting directions for further study. There have been many successful robotic painters such as those described in Cohen (2017), Tresset and Fol Leymarie (2012) and Lindemeier (2018), which demonstrate how easy it is to relate to physical processes. In addition, other creative AI sys-

tems have simulated physical processes on-screen, e.g., The Painting Fool software simulates paints being applied to a canvas with an on-screen hand holding a paintbrush (Colton, 2012). Cook and Colton (2018) introduced the notion of *continuous computational creativity* in order to study how a creative AI system could have a presence in a community. Moreover, Colton et al. (2020b) investigated the notion of *creativity theatre* with a system demonstrating its abilities through human-like tool use and creative aims.

• **Personhood:** The process of 'othering' "consists of applying a principle that allows individuals to be classified into two hierarchical groups: them and us" (Kobayashi, 2020), often to exclude or ostracize them. It is natural to see AI systems as others in creative fields, as they are not people to start with, are perceived to threaten jobs, and are usually not relatable. There have been a number of studies in computational creativity research to explore ways in which communities might be encouraged to be more inclusive through advances in how generative AI systems operate and are deployed. One issue is that AI systems are not authentic voices, especially when the artefacts they produce reflect elements of humanity such as emotions or life experiences. Colton, Pease, and Saunders (2018) offer some solutions to this in terms of *artificial authenticity*. Ultimately, this line of thought led to the proposition of a framework called *the machine condition* in which an AI system could express elements of its existence through artistic output (Colton et al., 2020c). This was taken further in Pease, Colton, and Banar (2023) who proposed agency, self-expression and individuality as relatable traits for advancing an AI system towards gaining personhood in a creative domain. While agentic AI has a long history, it has recently been adopted in deep learning to describe autonomous systems that pursue complex goals with minimal human intervention (Acharya, Kuppan, and Divya, 2025). While no-one suggests such agents are given personhood, this trend highlights that autonomy – a quite relatable trait – is important in designing systems wrapped around LLMs.

• **Evaluation:** In creative communities, people often discuss the value of artefacts like paintings or songs, with the person who made them sometimes part of that discussion. Hence, if an AI system can appreciate the value of the artefacts it produces (and those of others), this may seem more relatable than if it cannot. Enabling generative AI systems to evaluate their own creations has been a mainstay technique in computational creativity research, which has been fed by studies into how people appreciate created artefacts. The idea that generated artefacts should evoke emotions of surprise through novelty has been extensively studied e.g., by Macedo and Cardoso (2017). This has been operationalized through techniques such as *novelty search* (Lehman and Stanley, 2008) and *surprise search* (Yannakakis and Liapis, 2016). Associated with the notion of surprise is that of unexpectedness, when surprising results come from a context where people have certain expectations. Grace and Maher (2014) generalise past the surprise evoked by artefacts, and develop a typology of expectations relevant to computational creativity evaluation. Returning to the question of autonomy, Colton, Charnley, and Pease (2011) investigated the idea that a computational creativity system can be evaluated in terms of how much responsibility has been handed over to it.

We argue that – while each project surveyed above had an individual goal, outcome and application – they contribute to an overall picture of effort in simulating relatable computational creativity, which can act as an umbrella term for this effort. Building on the projects of the authors cited above, we further argue that there are many benefits to relatable computational creativity, which depend on the mode in which creative AI systems interact with people. One mode of interaction might be that the AI system acts as a set of tools, but the system itself has the ability to control those tools. This is the context for the experiments described below, and enables the software to provide guidance to the user on tool use in a relatable way. If the software does something surprising, this could provide inspiration, with the relatable actions of the software providing expectations from which unexpectedness – in terms of artefacts and actions – can emerge.

Creative AI systems are generally multi-faceted, and it is unlikely that all the processes they employ would be relatable. This could be for a number of reasons, including: a process is naturally a black box, due to scale and/or complexity, e.g., forward inference in large neural models; or a process is purposefully kept hidden for some reason, perhaps to maintain mystique, or to simulate a subjective opinion or taste. Hence, our understanding of relatable computational creativity should admit the fact that there will be *knowable* and *unknowable* processes at work in creative AI systems, just as there are with creative people.

We also need to consider whether the AI agent is intended to be used in a realistic or more of a role-playing situation. For instance, with generative AI chat-bots, it is very easy for users to project personalities onto the systems, giving them a human-like role in doing so. In this situation, if the chat-bot is asked to provide explanations of its processes, it will likely provide a relatable scenario, even though it is not an accurate portrayal. LLMs are not reliable witnesses to their own creative processes, but we could say here that the system is *pseudo-relatable*, because users are at liberty to pretend it is a person, and take inspiration from its fictional portrayals of internal processes.

Another consideration for relatable computational creativity systems is *why* we relate to the human-like behaviours that they exhibit. One possible answer is because of demonstrations that they undertake, for instance with embodied AI systems creating artefacts in physically relatable ways or via on-screen creativity theatre demonstrations. Another possible answer is via explanations that the system gives, either as part of its normal processing or under scrutiny in post-hoc questioning. Naturally, a relatable system might provide a mixture of demonstrations and explanations. A final consideration for relatable computational creativity is in terms of where we seek relatable processes. As the general public become more aware of how generative AI systems are engineered, questions about their backstory may arise. We therefore need to consider how relatable both the generation-time processes and the development-time processes of an AI system are.

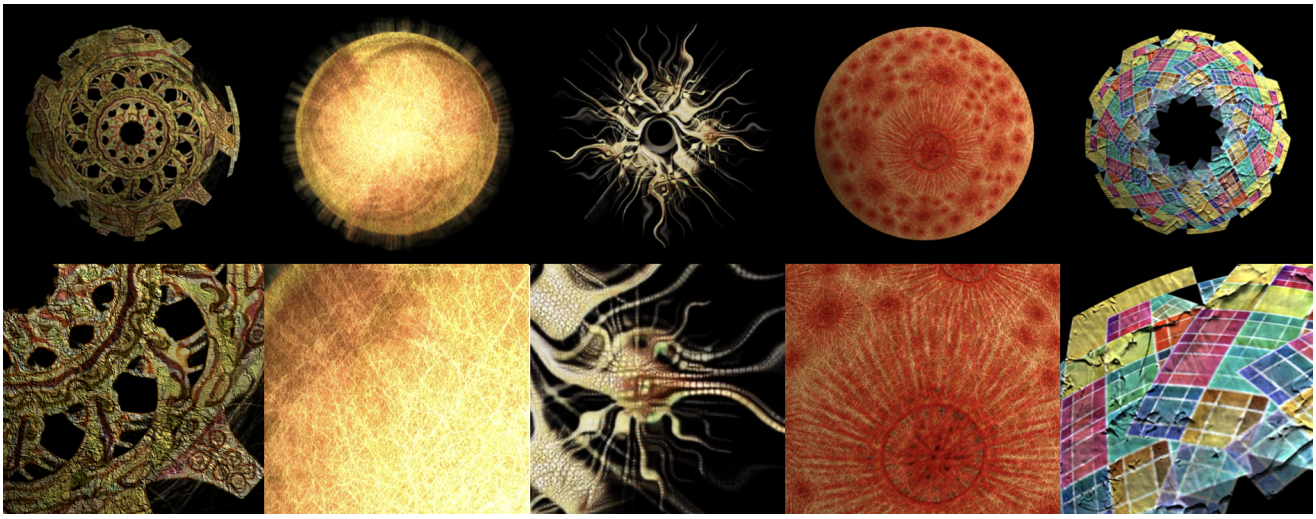


Figure 1: Example images produced with the editing interface of the Art Done Quick App, with details given below each.

Casual Creation and Art Done Quick

Compton and Mateas (2015) introduced the notion of a *casual creator* system, which they described as follows:

> A casual creator is an interactive system that encourages the fast, confident and pleasurable exploration of a possibility space, resulting in the creation or discovery of surprising new artifacts that bring feelings of pride, ownership, and creativity to the users that make them... The user of a casual creator is a *casual* user, and the system can expect no previous domain knowledge, no previous technical experience, or adherence to a long learning programme.

In the original paper and more extensively in Compton (2019), various design patterns, guidelines, typologies and worked examples are presented, which have formed essential reading for anyone implementing a casual creator application. Petrovskaya, Deterding, and Colton (2020) provided a further typology of commercial casual creator applications.

A casual creator app called Art Done Quick is used in the experiments described in the next section. This app enables users to quickly produce pieces of decorative visual art, as described in Colton et al. (2020a) and Colton (2020). Examples of the kinds of images users can produce are provided in figure 1. Users employ Art Done Quick in a two-stage process. Firstly, there is an evolutionary art interface, whereby users can double-tap and drag-and-drop to achieve mutation and crossover of underlying genomes which are rendered as images. In the first screenshot of figure 2, the user has double-tapped the central image to produce the eight variations that surround it. The images produced in this interface tend to be blocky patterns of shapes, which provide the raw material with which to produce more sophisticated images later.

Secondly, there is an editing interface with which users can alter a chosen image in various ways. There are ten editing screens in the app to provide comprehensive functionality in terms of fun things to do with images, as per the 'fun-first methodology' described in Colton et al. (2020a). The screens are as follows, with four of them portrayed in figure 2:

- **Art:** here the size, frequency, depth, grain and type of underlying shapes in the raw material image can be altered.
- **Colour:** here, tinting, greyscale, rainbow, spectrum and posterizing (reducing colours) image filters can be applied.
- **Liquify:** here, the user can add holes, glass beads, twirls, bulges/pinches and carnival mirror effects to images.
- **Montage:** here, the user can create layers which repeat and re-arrange an underlying image.
- **Draw:** here, the user can create a layer by drawing onto the image, using various simulated implements.
- **Effects:** here, the user can employ dozens of image effects, including various blurring/sharpening, dithering and neural style transfer (Jing et al., 2020) filters.
- **Texture:** here, the user can overlay one of 24 textures to add photo-realism to images.
- **Light:** here, the user can add spotlight, flood, glowing, shadow and sparkling lighting effects to images.
- **Stickers:** here, the user can personalise images by adding multiple stickers on top or below them.
- **Text:** here, the user can personalise images by adding text in a number of styles, fonts and sizes to the artworks.

Importantly for the experiments described below, as per the design pattern of 'modifying the meaningful' from Compton (2019), much of the editing functionality is limited to choosing which effect to add, and then tweaking two numerical parameters. As we see in figure 2, there is a top and bottom slider to apply these tweaks with, for each of the editing actions in the liquify, light, texture and effects screens.

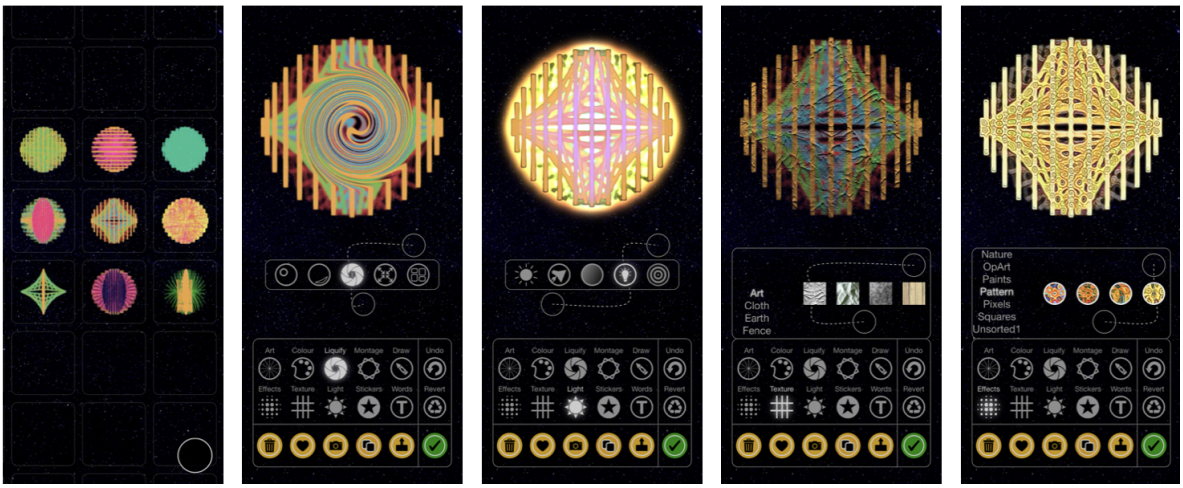


Figure 2: Screenshots of the evolutionary interface and the liquify, light, texture and effects editing screens in Art Done Quick.

Experiments in Discovering Tool Uses

Art Done Quick is an iOS/iPadOS application which has been used as a research platform rather than released commercially. Colton et al. (2020b) implemented an independent control on top of the app, which interacted with the evolutionary art interface and the editing screens to generate images which looked like certain real-world objects (according to the machine vision model ResNet (He et al., 2016)). While that project helped to develop notions of creativity theatre, it didn't bring any benefits to users other than entertainment in demonstration contexts.

For a more applicable study into relatable computational creativity here, we envisage a scenario where a user has an image that they like, but wants to improve it. The user knows to some extent what a better version of it would look like, according to their taste. However, as per the maxim of casual creation, they might not have enough experience of the app, and might not want to undertake trial and error tests to make the better image. Our experiments here support the implementation of an interface where the user chooses an image, then chooses a descriptive word from a given list, and indicates whether the altered image they require should look more or less like the chosen description. As examples, a user might say they want a 'more ghostly' or 'less spiky' version of their chosen image. In response to such a request, we envisage Art Done Quick demonstrating the use of a sequence of editing actions, producing an image which ideally still looks like the original, but reflects the description to an appropriately larger or lesser extent.

To achieve such a text-2-tool-use interface, we added a mobile version of the CLIP vision-language model (Radford et al., 2021) to Art Done Quick, so the software can assess images with respect to descriptive words. The MobileCLIP version we used is described in (Vasu et al., 2024), and consists of two neural models: the first encodes images into a latent space of 512-element vectors, and the second encodes text into the same latent space. Importantly, the loss function for training forced image/text pairs to be mapped to similar parts of the space if the text was an appropriate caption for the text. It also forced pairs to be further apart if the text did not reflect the image. Given a piece of text, T, and an image I, taking the cosine similarity of the CLIP encoding for T with the encoding for I gives a score for how well the text captions the image (or how well the image reflects the text).

With the experiments described here, we exhaustively try out a subset of the image editing actions Art Done Quick has, and determine which of them reliably increase/decrease significantly the cosine similarity of an image with each of a set of descriptive words. For instance, a particular image edit might, on average over a set of starting images, increase the cosine similarity with the word 'colorful', but might decrease the cosine similarity with 'crystalline'. Such information could be operationalised so that Art Done Quick can demonstrate using its tools to achieve a look that users want. We undertook the experiments in three stages: (i) some initial experiments to guide our usage of MobileCLIP (ii) a systematic evaluation of 612 single-image edits over 100 starting images, and (iii) constructing sequences of image edits by chaining together the most promising single-image edits.

Initial Guiding Experiments

Before fully committing to using MobileCLIP, we first tested whether it would be appropriate for our use case. Arias, Baldrich, and Vanrell (2024) highlight certain deficiencies in the ability of CLIP models to assign colours to images, with particular difficulties with achromatic colours such as grey, black and white. As colour is an important element of the images Art Done Quick produces, and a likely target for user requests, we decided to test MobileCLIP's abilities in labeling coloured images. We generated blank images in twelve standard colours including black, white and grey, and used MobileCLIP to encode them into its latent space. We then CLIP-encoded the words 'black', 'white', 'grey', 'blue' etc., and calculated the cosine similarity between every pair of word/colour-image. The results are shown in figure 3. We see that all but one colour matches its corresponding

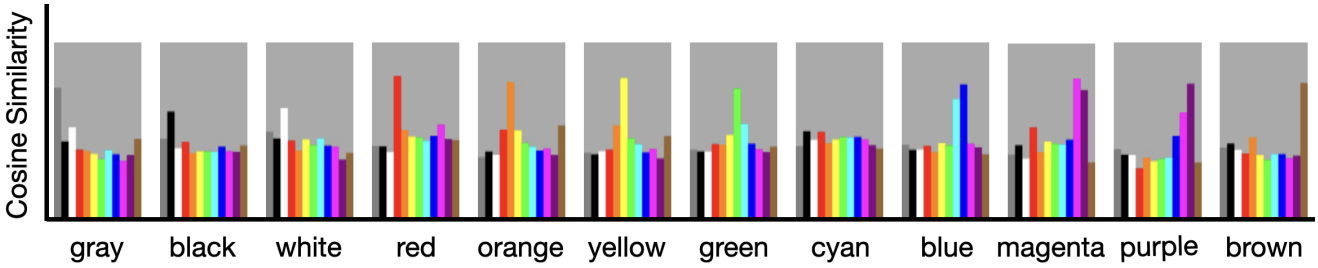


Figure 3: Cosine similarities of colour images (bars) with colour words (on the x-axis).

word more than any other. Importantly, achromatic colours do not cause a problem, with the appropriate colour word having higher cosine similarity to the grey, black and white colour images. However, the word 'cyan' has roughly the same cosine similarity with all the colour images. While these results indicate that there may be problems with lesser known colours like cyan, we felt that they weren't bad enough to necessitate using a different vision-language model.

Another question we addressed by using single-colour images was how to use CLIP encodings in cases where the user wants to alter an image to look *less* like a particular description. Given that CLIP can encode any text, not just single words, there are two alternatives for this: (a) Art Done Quick could suggest image edits which on average produces an image with smaller cosine similarity to a given description than the starting image, or (b) it could suggest edits which produce an image with a higher cosine similarity to a textual negation of the description, for instance the adjective 'icy' negated to 'not icy' or 'hardly icy at all'.

To determine which of these possibilities to use, we produced the same charts as in figure 3, but calculating cosine similarity to the colour words modified with 'not' in front. We found that, for instance, the phrase 'not black' had higher cosine similarity to the black image than any other images, and this was true for all the colours, and however we modified the descriptive word to negate it. From this we deduced that text modifiers don't move the encoded text very far from the unmodified text in the latent space. This meant that the only sensible way to find image edits which lower the reflection of a given descriptive word in an image was to find ones which lower the cosine similarity to the word.

Single-Action Experiments

Our aim here was to find single image edits among those Art Done Quick affords which can reliably move the CLIP encoding of an image towards or away from the encoding of a given descriptive word. In doing so, the edited image should look more/less like the description prescribes. To make the search more manageable, we restricted the image edits to just those editing functions in Art Done Quick in the Art, Colour, Liquify, Light, Texture and Effects screens, with the latter functions broken into Style Transfer and Other Effects. All of the edits in these screens have just two tweakable parameters, which we call the top and bottom parameters in line with the slider positions for them in the user interface. We made strategic choices in the top/bottom parameters to effect quite visible rather than subtle changes.

Action Type	Number of Actions	Top 20 Edits	
		Positive	Negative
Art	15 (2.5%)	87 (0.9%)	22 (0.2%)
Colour	57 (9.3%)	1059 (10.6%)	319 (3.2%)
Effects	153 (25.0%)	3589 (35.9%)	662 (6.6%)
Light	45 (7.4%)	417 (4.2%)	155 (1.6%)
Liquify	45 (7.4%)	3057 (30.6%)	107 (1.1%)
Style	96 (15.7%)	1048 (10.5%)	5967 (59.7%)
Texture	200 (32.7%)	743 (7.4%)	2768 (27.7%)

Table 1: Breakdown of single image edits overall and those used in the top twenty for positive and negative movements. Style is Style Transfer actions and Effects are Other Effects

In total, there were 612 single image edits – which we call *actions* – to search over, each represented as a tuple of type, identifier, top parameter and bottom parameter. A breakdown of these into the type of action is given in table 1. We applied each action to 100 different starting images, generated in the way Art Done Quick normally produces them randomly for the evolutionary art interface. This provided 61,200 pairs of (s)tarting and (e)dited images (s, e). For each pair, we generated the CLIP encoding of both s and e, then calculated the cosine similarity against 500 descriptive words. These descriptions were chosen to reflect: visual properties of abstract art images (e.g., dark, pointy, glistening); emotional projections (e.g., happy, melancholic, depressing); artistic terminology (e.g., layered, harmonious, gestural) and generally interpretable terms (e.g., haunted, technical, unruly).

For each descriptive word, d, we calculate the mean, $m(d)$ and standard deviation, $std(d)$, of the cosine similarities between its CLIP encoding and the encodings of the 61,200 edited images. Then, for each pair of description and action taking starting image, s, and producing edited image e, we calculate the *movement* w.r.t. d afforded by a as:

$$movement(a, d, s, e) = \frac{cos_sim(e) - cos_sim(s)}{std(d)}$$

Then, for a given description word, d, we calculated the following four measures for every action a, over all the starting, s, and edited, e, image pairs:

• $movement(a, d)$: the mean of $movement(a, d, s, e)$ over all pairs (s, e) that a produces.

• $fidelity(a)$: the mean cosine similarity between the CLIP encodings of images s and e, over all pairs (s, e) that a produces.

• $pConfidence(a, d)$: the proportion of pairs (s, e) that a produces where $movement(a, d, s, e) \geq 0$.

• $nConfidence(a, d)$: the proportion of pairs (s, e) that a produces, where $movement(a, d, s, e) < 0$.

As an overall score, $pScore(a, d)$, for an action, a, with respect to changing an image to look more like a descriptive word d, we calculated:

$$fidelity(a) + movement(a, d) + pConfidence(a, d)$$

Likewise, as an overall score, $nScore(a, d)$, for a with respect to changing an image to look less like d, we took:

$$fidelity(a) - movement(a, d) + nConfidence(a, d)$$

Note that by including the fidelity term, the scores penalise actions where edited images don't look enough like the starting images, as this is an unsatisfying outcome. Similarly, low probability in the action producing a desired movement is also penalised. Note, however, that the fidelity and confidence measures output positive integers less than 1, while the movements range from around -2 to +5, so this term is usually dominant in the overall score for actions which move images a lot. With the overall score for every action with respect to every descriptive word, we determined the best actions able to reliably move images in a positive or negative direction with respect to a user-chosen description word.

The spread of the top 20 actions across the action types is given in table 1. Comparing the percentages, we note that the actions involving changes of the raw art image shapes were rarely the best for positive or negative movements. We also note that style transfer actions are found extensively in the best 20 to move images away from descriptions, while the other effects such as dithering, blurring, etc., were disproportionately used to move images towards descriptions. In table 2, we recorded the number of adjectives (out of 500) that can be moved up to certain multiplicands of the standard deviation in both directions. It is promising to see that for 478 out of 500 description words, there is at least one action that can move an image (on average over 100 images) more than one standard deviation away from its starting point in a positive direction, with 468 words out of 500 having a similar action for the negative direction.

Action Sequence Experiments

The images in figure 1 were produced via Art Done Quick applying around five image edits to a starting art image in each case. In general, looking at the images after the single edits, we found them to be less sophisticated than those possible with a sequence of multiple edits. Hence, in a final round of experiments, we tested whether it is possible to discover longer action sequences which produce appropriate cosine similarity movements while retaining fidelity to the starting image.

Positive Movement (stds)	Num. Descs.	Negative Movement (stds)	Num. Descs.
$m > 4$	11	$m < -4$	0
$m > 3$	44	$m < -3$	1
$m > 2$	172	$m < -2$	100
$m > 1$	478	$m < -1$	468
$m > 0$	500	$m < 0$	500

Table 2: Number of descriptors for which there is at least one action achieving given movement distances.

For each of the 500 description words, we randomly took four actions from the top 20 (in terms of the scores described above) determined previously, for both positive and negative movements. Starting with a randomly generated raw art image, each action was applied in turn, which generated a sequence of five images. From these, the image with the highest movement measure in the given direction was chosen, but only if it had a fidelity with the starting image of 0.7 or above. This produced sequences of 1, 2, 3 or 4 actions, with the breakdown of how many for each given in table 3. Table 3 also shows that the fidelity of edited images reduces but the movement away from the starting image increases as sequence length grows, in both the positive and negative directions. Both of these is as expected, and indicates that we should be able to control the movement and fidelity of edited images while producing more sophisticated imagery through longer action sequences in Art Done Quick. Some illustrative sequences for positive and negative movements are given in figures 4 and 5 respectively.

As examples, the first sequence in figure 4 achieves a 'fiery' look through the application of two style transfer actions. This is also what the first sequence in figure 5 does to reduce the 'blue' in the starting image. To achieve a 'ghostly' look, the second sequence in figure 4 applies dithering, an ambient lighting change and a zoom blur. To reduce the 'detailed' nature of the starting image in the second action of figure 5, an Op-art (Houston, 2009) style transfer action is applied, followed by a texture overlay and finally a zoom blur again. The grey nature of the starting image in the third sequence of figure 5 is removed eventually with a style transfer action, achieving an altered image that is 2.5 standard deviations away from the description 'melancholy', while still having a fidelity with the starting image of 0.768.

Seq Len	Positive Direction Num	Move.	Fid.	Negative Direction Num	Move.	Fid.
1	92	1.14	0.86	95	-1.13	0.87
2	126	1.52	0.80	119	-1.45	0.82
3	145	1.77	0.76	135	-1.60	0.79
4	137	1.80	0.74	151	-1.57	0.78

Table 3: Sequence length, average movement and fidelity for action sequences in both positive and negative directions.

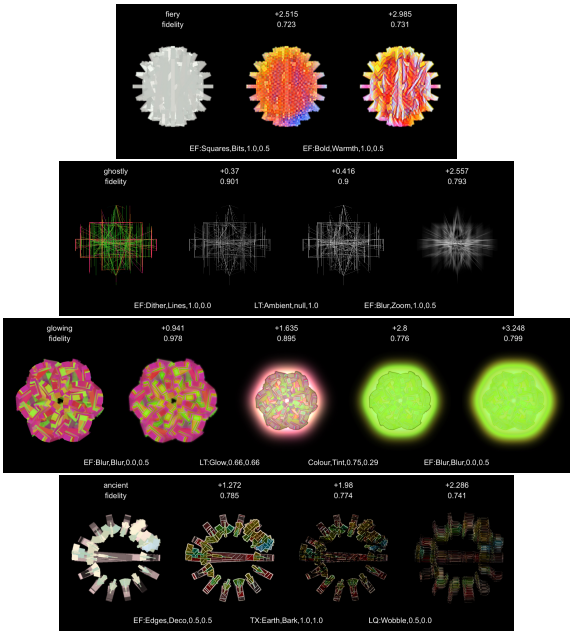


Figure 4: Action sequences which increase cosine similarity with the words ‘fiery’, ‘ghostly’, ‘glowing’ and ‘ancient’.

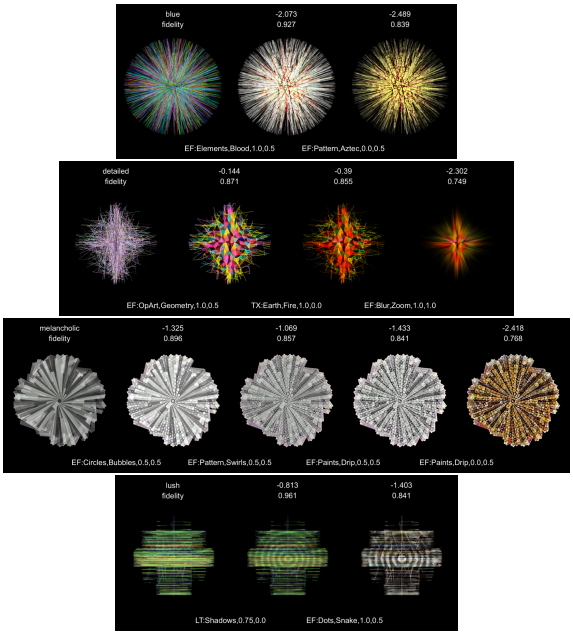


Figure 5: Action sequences which decrease cosine similarity with the words ‘blue’, ‘detailed’, ‘melancholic’ and ‘lush’.

Discussion

The top 20 image editing actions for positive and negative movements with respect to 500 descriptive words, and the action sequences resulting from the experiments reported here unlock new functionality in the Art Done Quick app. Given that the app now has a good understanding of how its own tools effect images, users can ask it for a sequence of tool uses which make changes to an image, so that the new version looks more (or less) like a given descriptive word. We believe this reflects aspects of relatable computational creativity as discussed above, because the demonstrations given are exactly how a person could do them. We hope this new mode of interaction will provide education to users of Art Done Quick, so they are exposed to editing possibilities, as well as surprising inspiration when the app chains together a sequence of actions that achieves unexpected results.

Looking in further detail, we would describe the computational creativity exhibited here as relatable rather than pseudo-relatable, as there is no reason to imagine the software is a person. Further, with respect to knowable and unknowable aspects of the creative acts, the sequence of actions is clearly knowable, and the motivation to change an image in accordance with a description with each edit action is likewise knowable. However, given that CLIP is used, the software is not able to explain why a particular image looks like a description or not, so this element is unknowable. Moreover, the backstory for how Art Done Quick gained its tool-use skills is quite relatable as an exhaustive search over single-edit actions followed by trial and error with action sequences. While users might not be able to imagine trying 61,200 image edits in order to master the app, they should understand the process, and perhaps relate to it.

It is beyond the scope of this paper to address the issue of so-called ‘AI slop’ (a derogatory term for AI generated artefacts that are deemed unworthy (Kommers et al., 2026)). However, it seems clear that one reason such artefacts are derided is the lack of effort (human or AI) that goes into their construction, as put succinctly by the phrase: “why bother reading something no one was bothered enough to write?” (x.com/sulamoon/status/1724888928866476417). AI systems like Art Done Quick which demonstrate tool use may encourage users to engage more with artworks they are making, rather than merely prompting for them, and we plan an investigation of relatable computational creativity tools in this context. Moreover, we hope that generalising text-2-tool-use from casual creators to creativity support tools, this mode for human-computer interaction may benefit professional creators whose creations are used (often without permission or copyright) to train the deep learning models which end-users employ to bypass their creative process.

Relatability is not the only consideration for human-computer interaction with creative AI software and casual creators in particular. Other factors for study include how much trust users have in the software, how easy/fun it is to use, its ethical deployment in various communities and how empowered people feel while using it (or watching it create autonomously). However, as argued above, the volume of literature on the simulation of human-like creative behaviours would indicate that relatability is of particular interest to computational creativity researchers. This new umbrella term colates and highlights decades of work in computational creativity research, and could act as a new axis for development and evaluation of generative AI systems.

Conclusions and Future Work

We hope to have made the case that the notion of relatable computational creativity suitably encapsulates a body of work covering aspects such as motivations, explainability, embodiment, personhood and evaluation. We have highlighted aspects of this term for describing relatable AI systems, including the interaction mode, opaqueness of processing, role-playing situations, explanations versus demonstrations and engineering backstory. We introduced new functionality into the Art Done Quick application, so that users can be given tool-use help and inspiration from the app in a relatable way.

While we have given the app 20 good single-action edits for each descriptive word and in both directions, we only provided a single sequence of actions for the same purpose. Hence, we plan to scale up the action sequences, so the software has an even greater understanding of its tools and higher potential for surprise. Having shown that Art Done Quick can find tool-uses to achieve requested image changes that CLIP deems appropriate, we can use this for enhanced functionality in the app, and we plan a prompting interface for this. A next step will then be to perform user testing to see if people are satisfied with the interaction style and the resulting images produced. A preliminary look at some of the action sequences generated in the above experiments indicates that people may agree with CLIP's assessment of movement for highly visual terms like 'bright' or 'crystalline' but disagree when the required description requires more interpretation, such as 'discordant' or 'dramatic'.

We hope to use the platform to raise and answer questions about relatable computational creativity in a practical way, for instance in this case, whether it is still possible to relate to the creative acts of a generative AI system if you disagree with its assessment of the outputs it produces. Addressing such questions will enable us to expand the general idea of relatability, with the ultimate aim to produce a suitable taxonomy and grounded study of how people relate to generative AI systems and any benefits this may have, and a typology of relatable AI systems in general.

To do this, we plan to undertake a more thorough survey of the computational creativity literature to identify concepts, commonalities, best practice and pitfalls in the building ad evaluation of relatable computational creativity systems. We will also draw upon other studies of performative relationships in human-computer interaction, such as HCI entanglement (Frauenberger, 2019). As the theory of relatable computational creativity advances, we plan to distill it into more sophisticated implementations of Art Done Quick. In particular, in addition to the software reacting to user's requests for image enhancements, we will employ ideas from *creativity theatre* Colton et al. (2020b) to enable the software to perform proactively with its tools, in ways which challenge assumptions of creativity in software, and – we hope – actively encourage, enthuse and inspire users to create.

Acknowledgements

We would like to thank the anonymous reviewers for their thoughtful and helpful comments and suggestions.

References

Abusitta, A.; Li, M. Q.; and Fung, B. C. 2024. Survey on explainable AI: Techniques, challenges and open issues. *Expert Systems with Applications* 255.

Acharya, D. B.; Kuppan, K.; and Divya, B. 2025. Agentic ai: Autonomous intelligence for complex goals – A comprehensive survey. *IEEE Access* 13.

Arias, G.; Baldrich, R.; and Vanrell, M. 2024. Color in visual-language models: CLIP deficiencies. In *Proceedings of the Color and Imaging Conference*.

Chen, X.; Zhao, A.; Heming, X.; Lu, X.; Wang, H.; Chen, Y.; Wang, J.; Li, W.; and Shen, X. 2025. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. arXiv preprint 2505.16782.

Cohen, P. 2017. Harold Cohen and AARON. *AI Magazine* 37:63–66.

Colton, S., and Banar, B. 2023. Automatically adding to artistic cultures. In *Proc. 12th Int. Conf. on Artificial Intelligence in Music, Sound, Art and Design (EvoMUSART)*.

Colton, S., and Ventura, D. 2014. You Can't Know my Mind: A festival of computational creativity. In *Proc. of the 5th International Conference on Computational Creativity*.

Colton, S.; Cook, M.; Hepworth, R.; and Pease, A. 2014. On acid drops and teardrops: Observer issues in computational creativity. In *Proc. AISB symposium on AI and Philosophy*.

Colton, S.; McCormack, J.; Berns, S.; Petrovskaya, E.; and Cook, M. 2020a. Adapting and enhancing evolutionary art for casual creation. In *Proc. 9th Int. Conf. on Computational Intelligence in Music, Sound, Art and Design*.

Colton, S.; McCormack, J.; Cook, M.; and Berns, S. 2020b. Creativity theatre for demonstrable computational creativity. In *Proc. 11th Int. Conf. on Computational Creativity*.

Colton, S.; Pease, A.; Guckelsberger, C.; Llano, M. T.; McCormack, J.; and Cook, M. 2020c. On the machine condition and its creative expression. In *Proceedings of the 11th International Conference on Computational Creativity*.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proc. 2nd Int. Conf. on Computational Creativity*.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. In *Proc. 9th International Conference on Computational Creativity*.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems*.

Colton, S. 2012. The Painting Fool: Stories from building an automated painter. In McCormack, J., and d'Inverno, M., eds., *Computers and creativity*. Springer. 3–38.

Colton, S. 2020. Casual creation, curation, captioning, clustering and crossover in the Art Done Quick app. *Proc. Workshop on Casual Creators at ICCC*.

Compton, K., and Mateas, M. 2015. Casual creators. In *Proc. of the 6th Int. Conf. on Computational Creativity*.

Compton, K. 2019. *Casual Creators: Defining a Genre of Autotelic Creativity Support Systems*. Ph.D. Dissertation, University of California at Santa Cruz.

Cook, M., and Colton, S. 2018. Redesigning computationally creative systems for continuous creation. In *Proc. of the 9th International Conference on Computational Creativity*.

Cook, M.; Colton, S.; Pease, A.; and Llano, T. 2019. Framing in computational creativity: A survey and taxonomy. In *Proc. 10th Int. Conf. on Computational Creativity*.

Decety, J., and Grèzes, J. 2006. The power of simulation: Imagining one's own and other's behavior. *Brain research* 1079:4–14.

Fink, A.; Koschutnig, K.; Benedek, M.; Reishofer, G.; Ischebeck, A.; Weiss, E.; and Ebner, F. 2012. Stimulating creativity via the exposure to other people's ideas. *Human Brain Mapping* 33(11):2603–2610.

Frauenberger, C. 2019. Entanglement HCI: The Next Wave? ACM Trans. on Computer-Human Interaction 27(1), 1-27.

Grace, K., and Maher, M. L. 2014. What to expect when you're expecting: The role of unexpectedness in computationally evaluating creativity. In *Proceedings of the 5th International Conference on Computational Creativity*.

Guckelsberger, C.; Kantosalo, A.; Negrete-Yankelevich, S.; and Takala, T. 2021. Embodiment and computational creativity. *Proc. 12th Int. Conf. on Computational Creativity*.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the " Why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. *Proc. 8th Int. Conf. on Computational Creativity*.

Guckelsberger, C. 2020. *Intrinsic Motivation in Computational Creativity Applied to Videogames*. Ph.D. Dissertation, School of Electronic Engineering and Computer Science, Queen Mary University of London.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Houston, J. 2009. *Optic Nerve: Perceptual Art of the 1960s*. Merrell Publishers.

Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; and Song, M. 2020. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* 26(11).

Katz, J. H.; Mann, T. C.; Shen, X.; Goncalo, J. A.; and Ferguson, M. J. 2022. Implicit impressions of creative people: Creativity evaluation in a stigmatized domain. *Org. Beh. & Human Decision Processes* 169:104116.

Kobayashi, A. 2020. Other/otherness. In *International Encyclopedia of Human Geography*. Elsevier. 25–31.

Kommers, C.; Duede, E.; Gordon, J.; Holtzman, A.; McNulty, T.; Stewart, S.; Thomas, L.; Jean So, R.; and Long, H. 2026. Why slop matters. ACM AI Letters 1(1), 1-6.

Krzeczkowska, A.; El-Hage, J.; Colton, S.; and Clark, S. 2010. Automated collage generation - with intent. In *Proc. 1st International Conference on Computational Creativity*.

Lehman, J., and Stanley, K. 2008. Exploiting open-endedness to solve problems through the search for novelty. In *Proc. 11th International Conference on Artificial Life (Alife XI)*.

Lin, T.; Wang, Y.; Liu, X.; and Qiu, X. 2022. A survey of transformers. *AI Open* 3:111–132.

Lindemeier, T. 2018. *e-David : Non-Photorealistic Rendering using a Robot and Visual Feedback*. Ph.D. Dissertation, Universität Konstanz, Konstanz.

Llano, M.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2020. Explainable computational creativity. In *Proceedings of the 11th International Conference on Computational Creativity*.

Macedo, L., and Cardoso, A. 2017. A contrast-based computational model of surprise and its applications. *Topics in Cognitive Science* 11(1).

Maslej, M.; Oatley, K.; and Mar, R. 2017. Creating fictional characters: The role of experience, personality,and social processes. *Psychology of Aesthetics, Creativity, and the Arts* 11(4):487–499.

Moruzzi, C. 2022. The (artificial) physicality of creativity: How embodiment influences perceptions of creativity. *Proc. 13th Int. Conf. on Computational Creativity*.

Pease, A.; Colton, S.; and Banar, B. 2023. On the notion of creative personhood. In *Proceedings of the 14th International Conference on Computational Creativity*.

Petrovskaya, E.; Deterding, S.; and Colton, S. 2020. Casual creators in the wild: A typology of commercial generative creativity support tools. In *Proceedings of the 11th International Conference on Computational Creativity*.

Radford, A.; Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proc. of the International Conference on Machine Learning*.

Saunders, R., and Bown, O. 2015. Computational social creativity. *Artificial Life* 21(3):366–378.

Saunders, R. 2001. *Curious design agents and artificial creativity*. Ph.D. Dissertation, University of Sydney.

Tresset, P., and Fol Leymarie, F. 2012. Sketches by Paul the robot. In *Proc. of the 8th Annual Symposium on Computational Aesthetics in Graphics, Visualization, & Imaging*.

Vasu, P.; Pouransari, H.; Faghri, F.; Vemulapalli, R.; and Tuzel, O. 2024. Mobileclip: Fast image-text models through multi-modal reinforced training. *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity*.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, .; Cui, B.; and Yang, M. 2024. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys* 56(4).

Yannakakis, G. N., and Liapis, A. 2016. Searching for surprise. *Proc. 7th Int. Conf. on Computational Creativity*.

Reading Between the Lines: Can LLMs Discover the Question Behind the Text?

Claudiu Creanga[2,3], **Liviu P. Dinu**[1,3]
[1] Faculty of Mathematics and Computer Science,
[2] Interdisciplinary School of Doctoral Studies,
[3] HLT Research Center,
University of Bucharest, Romania
claudiu.creanga@fmi.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

This paper introduces "question archaeology", a specific evaluation task focused on inferring the single, authentic "genesis question" that motivated the creation of a complete text. Distinct from question generation, which targets any plausible question, or discourse frameworks that model utterance-level acts, our task assesses a model's grasp of authorial intent. We present a new dataset of commissioned texts paired with their original research questions and plausible distractors. Our evaluation of both proprietary models, like Gemini Flash and Pro, as well as open source models like Mistral and Qwen, reveals significant progress in this task, with the newer versions outperforming the earlier ones, while BERT-based models performed poorly. Notably, our findings indicate that current LLMs surpass human performance on this task, suggesting advanced understanding of authorial intent. This capability has important implications for AI's role in tasks requiring nuanced interpretation of human communication. Our work thus provides a new framework and a challenging benchmark for future models.

> Judge a man [model] by his questions rather than by his answers.
>
> Voltaire

Introduction

Every meaningful text that was written begins as an answer to a question - this is the heart of Bakhtin's dialogic theory (Bakhtin 1981). This dialogic nature of text suggests that written discourse is fundamentally responsive, emerging as a reply to preceding utterances and anticipating future responses within an ongoing conversation. Understanding the questions that prompt textual creation is therefore fundamental to fully comprehend the meaning, purpose, and context of any written work. Despite this intrinsic relationship between questions and texts, models have traditionally been evaluated on their ability to answer questions about texts, rather than their capacity to identify the questions that motivated those texts in the first place. Traditional benchmarks in natural language understanding focus primarily on question answering, summarization, and classification tasks — all of which assess a model's ability to process and extract information from given content. However, these evaluations fail to address a more fundamental aspect of textual understanding: the ability to recognize **why** a text was created. This goes beyond simple topic classification, which identifies **what** a text is about. Our task aims to infer the why—the specific authorial intent or underlying question that prompted the text's creation, which is a more nuanced challenge.

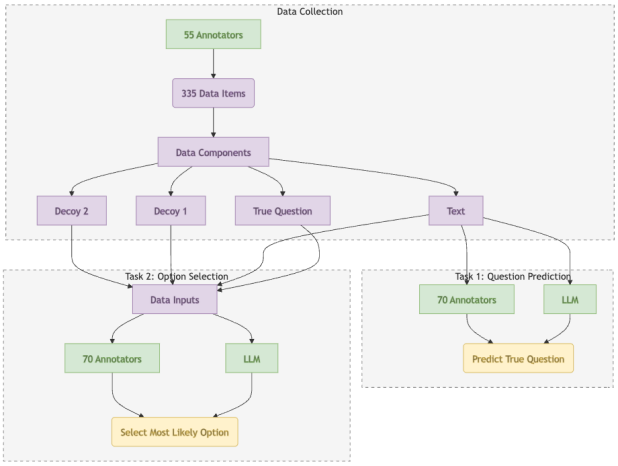


Figure 1: Experimental setup for evaluating LLM and human performance on question prediction. The experiment consists of two tasks: Task 1 (question prediction based on text alone) and Task 2 (selection of the most likely question from a set of options, given the text).

We propose a specific evaluation paradigm, which we term "**question archaeology**" (Figure 1), that builds upon these ideas[1]. It evaluates the ability of LLMs to infer the underlying "genesis question" that motivated the writing of a given monologic text. This test acts like a reverse cloze test where we give the LLM a novel text (which was not published on the Internet before) and ask it to infer the single, authentic question that led the author to write the text. By analyzing how accurately LLMs can reconstruct this original question, we provide a novel benchmark for measuring

[1]Code and data will be publicly available.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

the capacity for deep textual understanding and authorial intent recognition. This test challenges LLMs to move beyond surface-level comprehension and is, arguably, much harder than the classic question-and-answer test.

Our primary contributions are as follows:

- We propose a new evaluation paradigm, termed "question archaeology", which challenges LLMs to infer the original question that motivated the creation of a given text, rather than just factual recall.
- We present a unique dataset of 335 newly commissioned texts, called "QuestionArchaeology", each paired with its ground-truth originating research question and two carefully crafted plausible distractor questions. The methodology employed in creating this dataset ensures its novelty and prevents contamination from pre-training data, thereby providing a robust benchmark for assessing genuine inferential abilities. We make the dataset publicly available.
- This work presents the first evaluation of LLMs on the "question archaeology" task.
- We demonstrate advancements by evaluating the performance of older versus newer models on both a novel dataset and a previously unseen task, thereby showing genuine progress in their generalization capabilities beyond familiar scenarios.
- We provide a comparative analysis of Masked Language Models (MLMs) and autoregressive Large Language Models (LLMs) performance **against human annotators** on the same tasks. Our results reveal that current LLMs outperform humans in this specific domain.

Related Work

The task of inferring the motivation behind a text is related to several established research areas in NLP and discourse analysis. Our work builds on these foundations while carving out a specific, complementary niche focused on authorial intent.

The most direct parallel is **Question Generation (QG)**, where the goal is to generate a question given a passage. However, a key distinction lies in the objective. Standard QG aims to generate any plausible question that is answerable by the text (Du, Shao, and Cardie 2017), and is often used for data augmentation for QA systems. In contrast, "question archaeology" is not about generating a diverse set of possible questions; it is a more constrained task of inferring the single, authentic, ground-truth question that actually motivated the text's creation. This shifts the focus from answerability to the recognition of authorial intent. Our task should also be distinguished from "reverse prompting", the task of inferring the prompt used to generate a specific text from a language model (Morris et al. 2024). While both tasks reverse the typical input-output flow, question archaeology seeks to uncover the cognitive, human-centric motivation behind a human-authored text, whereas reverse prompting aims to deconstruct the machine-level instruction for a synthetic output.

Our work is also deeply grounded in theories of discourse structure that model communication as a series of questions and answers. Frameworks for identifying **implicit questions** (Fu, Chang, and Danescu-Niculescu-Mizil 2019) and **Questions Under Discussion (QUD)** (Wu et al. 2023) are highly relevant. QUD, in particular, offers a powerful formalism for modeling discourse coherence by treating each utterance as an attempt to answer a question on a structured stack. Similarly, frameworks for identifying **Dialogue Acts**, such as the Penn Discourse Treebank (PDTB) (Prasad et al. 2008), model the intention behind individual utterances within a broader dialogue.

While these frameworks are powerful for modeling the complex, turn-by-turn structure of discourse and utterance-level intent, especially in multi-party dialogues, our "question archaeology" task simplifies the problem to focus on a different, albeit related, goal. We justify our design choice by framing it as a complementary task focused specifically on uncovering the single, high-level **"genesis question"** for a complete, self-contained, monologic text. Rather than modeling the entire discourse stack or the function of each utterance, we aim to identify the overarching authorial purpose of the text as a whole. This provides a distinct, measurable, and challenging benchmark for evaluating an LLM's ability to reason about the holistic intent of an author, extending beyond the structural analysis offered by QUD and PDTB. Our work also connects to intent detection (Gupta et al. 2018) and argument mining (Lawrence and Reed 2019), which similarly attempt to recover implicit structures from text, but not the genesis question.

Data Collection

There are no public datasets available for this kind of task. In addition, **existing public datasets would compromise** the integrity of the evaluation. Large language models trained on public data would have encountered both the texts and their motivating questions during pre-training, rendering it impossible to distinguish between genuine inference and simple retrieval of memorized associations. To address this challenge, we developed a novel dataset based on freshly commissioned texts, ensuring that the original research questions were explicitly recorded prior to the writing phase. Our dataset comprises 335 unique texts, with each text paired with its authenticated originating research question and two plausible but incorrect alternatives. By constructing the dataset in this manner, we can effectively measure a model's ability to discern the true question behind a text, rather than its capacity to match previously encountered patterns.

Our dataset was created with the help of 55 participants, each contributing an average of 6 samples. Prior to participation, all individuals provided informed consent, and were fully briefed on the project's objectives and procedures. All data was created inside the classroom by hand to ensure that no LLMs were used. The data collection process followed three steps:

- First, participants formulated a research question of their choice (true question).

- Second, they wrote a text that addressed this question.
- Finally, they created two alternative questions that could plausibly, but incorrectly, be seen as the motivation behind their text (Other Question 1 and 2). To ensure quality, participants were asked to review their alternative questions to strike a balance - these questions needed to be related enough to the text to be credible, yet distinctly different from the original research question that motivated the writing.

Model Evaluation

Our research was structured around two distinct tasks. The **first task** evaluated the direct prediction of the motivating question from the text. The **second task** was a multiple-choice question requiring the selection of the correct motivating question from three options. To establish a robust human performance baseline for this second task and to calculate inter-annotator agreement, we collected 553 annotations from human raters. This number is greater than the 335 unique texts in our dataset because some texts were annotated by multiple raters. We made our code and dataset publicly available in a GitHub Repository.

We present here an example of an item in the dataset. The distractors were created by the original text authors to be deliberately plausible and challenging. Our aggregate analysis shows that, on average, the true question has a quantifiably higher semantic similarity to the text than the distractors (0.66 vs 0.57 and 0.54), indicating the task is well-defined, even if some examples are intentionally difficult.

- The question that initiated the writing process (the annotator came up with the question himself): Why do people get upset?
- The text that was written based on that question: Sometimes, everything feels fine until something bad happen - for example, a harsh word, a missed promise, and so on. People carry much inside, including hopes, fears, frustrations. When something is not as expected, it can feel like the world is not listening. Anger or sadness can be a way of saying “This is not fair” or “I need someone to notice this/me”. Many times it is not about the event itself, but more about our feelings - about us being hurt, ignored, or misunderstood.
- Two questions as decoys provided by the annotator: a) Why it is hard to stay calm when things go wrong? b) How do small misunderstandings escalate into big conflicts?

Our participants represented a diverse group spanning multiple Mediterranean and Eastern European countries, including Italy, Romania, Greece, Morocco, Slovenia, and Egypt. The gender distribution showed a majority of female participants (63%) compared to male participants (37%), contributing to the diversity of perspectives in our dataset. All texts were written in English. Despite the annotators not being native English speakers, no language-related issues were observed. The analysis of inter-annotator agreement revealed an observed agreement proportion ($\bar{P}$) of 0.58. After accounting for an agreement expected by chance ($\bar{P}_e$) of 0.34, the Fleiss’ Kappa coefficient (K) was 0.36. According to the benchmarks established by Landis and Koch (Landis and Koch 1977), this Kappa value indicates fair to moderate agreement among the annotators. This moderate level of agreement is expected, as the interpretation of a text’s underlying motivation is inherently subjective and tied to individual perspectives (Creanga and Dinu 2024a).

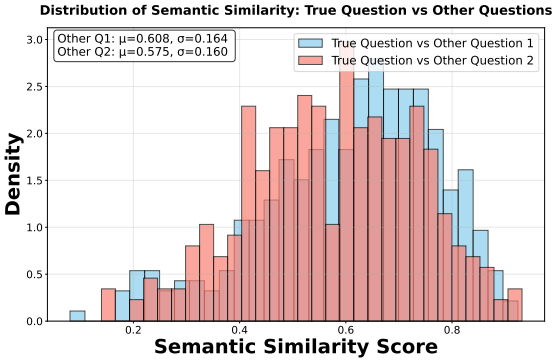


Figure 2: The semantic similarity scores between the true question and the two decoy questions, "Other Question 1" and "Other Question 2," show distributions that approximate a normal pattern. Other Question 1 exhibits a mean semantic similarity of 0.608 with a standard deviation of 0.164. Other Question 2 has a slightly lower mean semantic similarity of 0.575, with a comparable standard deviation of 0.160.

Analysis of **semantic similarity** between the true questions and their alternatives reveals a pattern approximating a normal distribution. The mean similarity score is 0.57 for the second alternative question (Figure 2) and 0.60 for the first alternative question. These moderate similarity scores indicate that participants successfully created alternative questions that maintain a balanced distance from the original - related enough to be plausible, yet sufficiently distinct to serve as meaningful distractors. The box-plot (Figure 3) reveals that the data contains relatively few outliers. The average semantic similarity between the text and the correct answer was 0.66, compared to 0.57 and 0.54 for the decoys (Figure 4). The semantic similarity score was calculated using the Sentence Transformers model(Reimers and Gurevych 2019).

Comparing how each alternative relates to the true question, we found a moderate positive correlation (Pearson r = 0.49) between their similarity scores. This indicates a tendency for both alternative questions to maintain comparable semantic distances from the original question. Question 1 appears to be semantically closer to the true question compared to Question 2, likely influenced by the order in which participants generated the alternatives. A paired t-test revealed a significant difference in the similarity scores between the two alternative questions ($t = 3.66$, $p < 0.001$), indicating that this difference in means (question 1 being closer to the true question than question 2) is unlikely to be due to random chance.

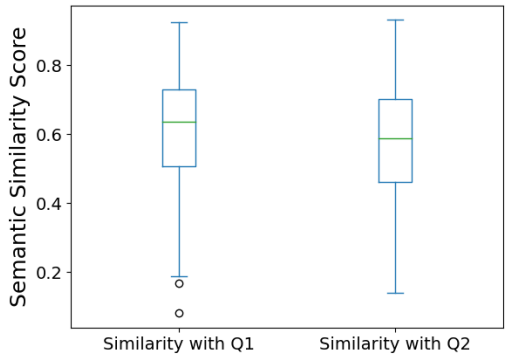


Figure 3: Distribution of Semantic Similarity between the true question and other questions. The data contains relatively few outliers.

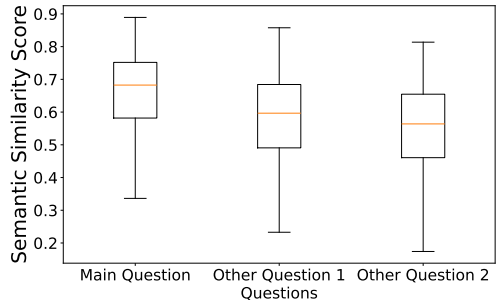


Figure 4: Distribution of Semantic Similarity between the text and each of the 3 questions. We can see a bigger similarity with the true question, but not by a lot: 0.66 vs 0.57 and 0.54.

Linguistic Analysis

The linguistic analysis in Figure 5 reveals that question types fundamentally shape discourse organization in predictable ways, with causal questions ("Why") generating the most linguistically complex responses. These texts contain significantly more causal conjunctions (2.21 expressions per text compared to 1.00-2.06 for other types) and employ sophisticated argumentative structures moving from problem statements through causal explanations to consequences. Meanwhile, definitional questions produce categorical discourse with clear topic-comment structures and explanatory markers like "for example," while procedural questions generate sequential patterns dominated by temporal markers ("first," "then," "next") rather than causal reasoning.

The syntactic signatures align closely with discourse functions, creating distinct textual fingerprints for each question type. Causal texts favor complex sentences with subordinate clauses rich in markers like "because," "due to," and "therefore," while definitional texts rely heavily on copulative constructions (X is Y) and categorical statements. Procedural texts show more imperative and conditional structures, reflecting their instructional nature. Despite similar sentence counts across types, causal questions generate slightly longer texts (81.9 words versus 75-77 words), suggesting that explaining causation requires more elaborate linguistic resources than other cognitive tasks like definition or description.

Quantitative analysis Figure 6 and Figure 7 shows causal questions demonstrate high usage of causal discourse structures (34.3%), definitional questions favor narrative rhetoric and simple structure, and procedural questions consistently employ sequential organization patterns. Figure 5 indicate that question type serves as a strong predictor of response structure, with different interrogative forms activating distinct cognitive and linguistic patterns in human responses. Question type appears to directly shape how knowledge is expressed, with each type triggering predictable discourse strategies that reflect the underlying cognitive processes required to construct responses.

Results

To establish a baseline, we implemented an XGBoost classifier using TF-IDF features (for task 2). We experimented with various preprocessing and feature engineering steps, such as limiting TF-IDF features (max_features=5000), incorporating n-grams (ngram_range=(1,2) and ngram_range=(1,3)), applying lemmatization, and filtering high-frequency terms (max_df=0.95 to ignore terms appearing in > 95% of documents). The optimal configuration utilized TF-IDF (max_features=5000) with bi-grams (ngram_range=(1,2)), yielding an average F1-score of 0.35, which is only slightly above the threshold for random chance.

Assessing LLM Performance

Our experiment aimed to evaluate for the first time the ability of LLMs to identify the original question that prompted a

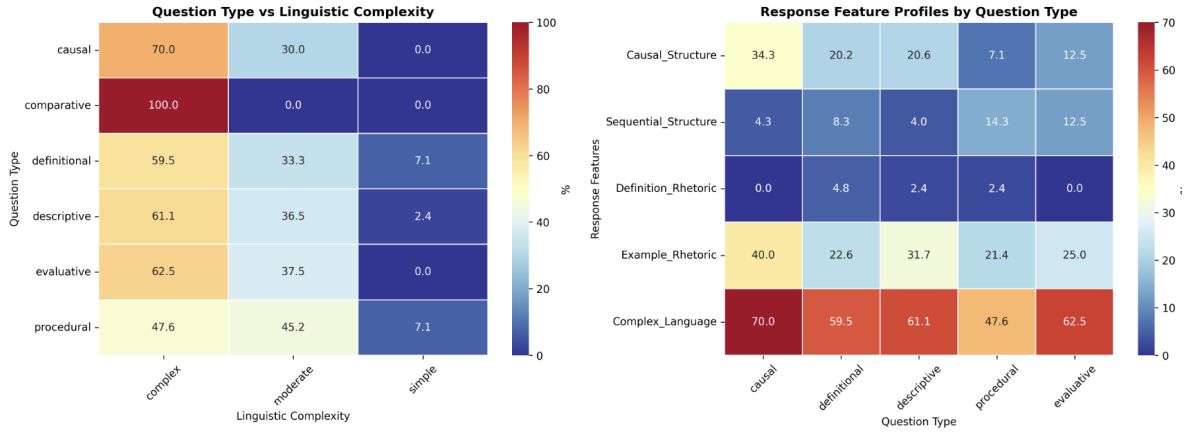


Figure 5: Heatmap showing that question type serves as a strong predictor of response structure, with different interrogative forms activating distinct cognitive and linguistic patterns in human responses.

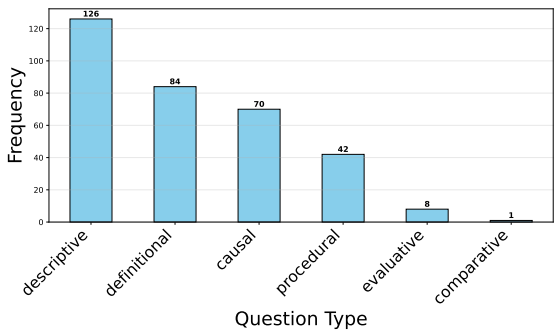


Figure 6: Distribution of question types by counts. Most questions are descriptive.

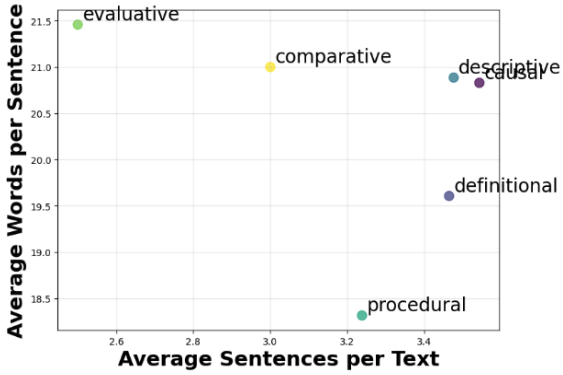


Figure 7: Distribution of question types by counts. Most questions are descriptive.

given piece of text and to measure the progress between different versions. We compared four versions of the Gemini models (Gemini Team and others 2023): Gemini Flash 2.0, Gemini Flash 2.0 Thinking (which incorporates enhanced reasoning capabilities), Gemini Flash 2.5 and Gemini Pro 2.5. This comparison allowed us to observe the evolution of these models in finding the true questions.

Table 1: Comparison of LLMs and human performance on task 1, based on F1 Macro scores.

Model	F1 score
Human raters	**0.66**
Gemini Flash 2.0	0.68
Gemini Flash 2.0 Thinking	0.71
Gemini Flash 2.5	0.72
Gemini Pro 2.5	**0.75**

To assess this, we designed a two-pronged evaluation approach. In the first task, we challenged the models to directly predict the original question based solely on the text provided by the authors. The second task involved a multiple-choice format, where the models were presented with the generated text and a list of three questions, one of which was the original question, and asked to select the correct one.

Focusing on the first task of direct question prediction, the earlier version of the model, Gemini Flash 2.0, demonstrated a **strong baseline performance**, as seen in Table 1. It achieved an average semantic similarity score of approximately 0.68 (Figure 8). This score suggests that, on average, the questions generated by the model exhibited a considerable degree of semantic overlap with the actual questions posed to the annotators. Further analysis revealed that in 12 instances, the model precisely predicted the original question (achieving a perfect similarity score of 1). Moreover, in a significant portion of cases (33%), the semantic similar-

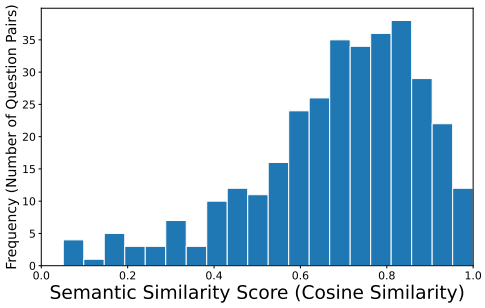


Figure 8: Distribution of Semantic Similarity between the true question and the question predicted by Gemini Flash 2.0.

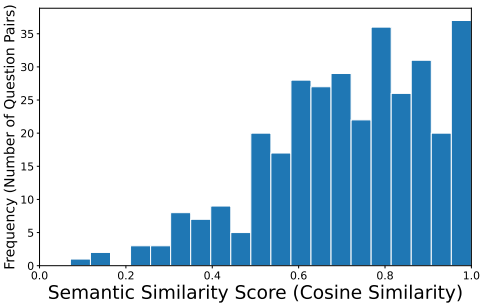


Figure 9: Distribution of Semantic Similarity between the true question and the question predicted by Gemini Flash 2.0 Thinking. The model demonstrated an increased ability to precisely identify the original question, achieving a perfect similarity score of 1 in 30 cases.

ity between the predicted and actual questions exceeded 0.8 (Figure 10), indicating a high level of accuracy in capturing the essence of the original inquiry.

The thinking version of the model, Gemini Flash 2.0 Thinking, showed a notable improvement. It achieved an even **higher average semantic similarity score** of approximately 0.71 (Figure 9). Interestingly, the median similarity score (0.73) was slightly higher than the mean, indicating that the distribution of similarity scores was positively skewed, with a greater number of texts leading to higher similarity predictions. The model also demonstrated an increased ability to precisely identify the original question, achieving a **perfect similarity score** of 1 in 30 cases. Furthermore, the percentage of cases where the semantic similarity was greater than 0.8 rose to 38% (Figure 10), highlighting a consistent enhancement in the model's ability to understand the motivations of the authors and reproduce the underlying questions. We also explored the confidence levels reported by the Gemini Flash 2.0 Thinking model for its question predictions. While the model generally expressed confidence in its answers, the correlation between its confidence score and the actual semantic similarity achieved was only low to medium (a Pearson correlation coefficient of 0.25). This suggests that while the model provides a confidence estimate, this estimate does not strongly align with the objective measure of how semantically similar its prediction is to the true question. This observation warrants further investigation to understand the factors influencing the model's confidence calibration and whether improvements can be made to better reflect the accuracy of its predictions. Both Gemini 2.5 models performed strongly, with the Pro version achieving the highest F1 score (0.75), while the Flash version achieved 0.73. Notably, the Gemini 2.5 Pro score was almost 10 points above human raters. In assessing semantic similarity greater than 0.8, the Pro model obtained an impressive 46% of cases, compared to 37% for the 2.5 Flash model.

We explored alternative strategies to improve our score, but these efforts were unsuccessful. One such approach involved first summarizing the text and then having the model predict the question based on the summary. This method yielded results approximately 4% worse than our standard approach, with scores of 0.69 for Gemini Flash 2.5 and 0.71 for Gemini Pro 2.5.

This initial analysis of the first task clearly demonstrates an evolution in the ability of the models in the Gemini series to identify the original question from a given text.

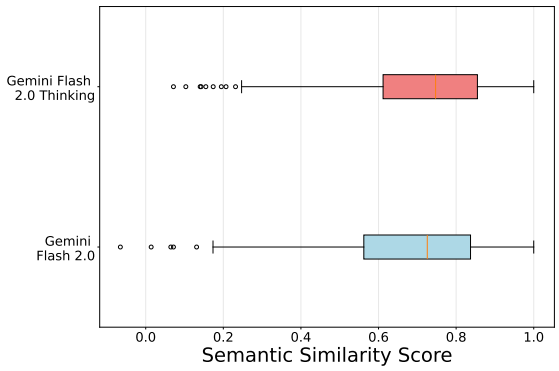


Figure 10: Plot showing semantic similarity scores for Gemini Flash 2.0 and Gemini Flash 2.0 Thinking. For Gemini Flash Thinking 2.0 the percentage of question pairs with similarity > 0.8 is 38.07%, and for Gemini Flash 2.0 the percentage of question pairs with similarity > 0.8: 33%.

The second task required models to choose from three options (Table 2). Initial results showed strong performance, with Gemini Flash 2.0 achieving 0.649 accuracy and Gemini Flash 2.0 Thinking reaching 0.653, both surpassing other MLMs and human raters. The Gemini 2.5 series demonstrated substantial improvement, yielding F1 scores of 0.734 (Gemini Flash 2.5) and 0.764 (Gemini Pro 2.5). Lower F1 scores were obtained by the open-source models Mis-

tral (Jiang et al. 2023) (0.57) and Qwen (Bai et al. 2023) (0.59). Open source models were also used in a zero-shot setting with no fine-tuning. Interestingly, we found no correlation between the semantic similarity of the options and the models' selections.

Table 2: Comparison of LLMs and human performance on task 2, based on F1 Macro scores.

Model	F1 score
XGBoost	0.35
Mistral:7B	0.57
Qwen 2.5:14B	0.59
Human raters	**0.60**
Gemini Flash 2.0	0.65
Gemini Flash 2.0 Thinking	0.65
Gemini Flash 2.5 with summarization	0.69
Gemini Pro 2.5 with summarization	0.71
Gemini Flash 2.5	0.73
Gemini Pro 2.5	**0.76**

We minimized variability by setting the LLM temperature to 0 and running each test three times. The second task demonstrated high consistency, with only four cases requiring a majority vote to determine the final output. We didn't fine-tune and all LLM experiments were done in a zero-shot setting.

Assessing MLM Performance

While LLMs currently dominate many NLP tasks, we investigated the competitiveness of Masked Language Models (MLMs) on the multiple-choice task. We experimented with several prominent MLMs: DeBERTa V3 (He et al. 2020), BERT large (Devlin et al. 2018), XLM-RoBERTa (Conneau et al. 2019), and the recent large-scale ModernBERT (Benjamin Warner 2024). We used them only for task 2, the classification one, and on top of these models we added a classification head. This head consisted of a normalization layer for input stabilization, a Tanh activation function for nonlinearity, a dropout layer for regularization, and a final layer mapping outputs to one of three classes. Our fine-tuning employed a two-stage strategy: initially, we froze all transformer layers and trained only the classification head for 2 epochs with a learning rate of 3e-4; subsequently, we unfroze the final transformer layer and continued training both the head and this layer for 1 epoch at a reduced learning rate of 2e-4. As the results in Table 3 indicate, MLMs performed significantly worse than LLMs. This presents an interesting contrast to other NLP tasks, such as sentence-level emotion recognition, where fine-tuned MLMs have been shown to be highly competitive against zero-shot autoregressive models (Creanga and Dinu 2024b). This disparity likely stems from fundamental differences in their pre-training: MLMs focus on local context via masked word prediction, which may be less effective for grasping global text purpose or authorial intent compared to the next-token prediction objectives common in LLM training. Furthermore, simply fine-tuning on this small dataset (we used a 40-60 split) an MLM with a classification head might not instill the nuanced understanding required to differentiate the true motivating question from plausible distractors, a task potentially better suited to the scale and inherent capabilities of LLMs. To enhance performance, we used an ensemble of 3 models, ModernBERT, XLM-RoBERTa, and DeBERTa, with a soft voting strategy (averaging class probabilities). The resulting F1-score for the ensemble was 0.52, representing only a slight improvement over the baseline performance of ModernBERT.

Table 3: Comparison of MLMs performance on task 2, based on Precision, Recall, and F1 Macro scores.

Model	Precision	Recall	F1 score
BERT large	0.51	0.45	0.47
DeBERTa V3	0.50	0.48	0.48
XLM-RoBERTa	0.53	0.50	0.50
ModernBERT	0.52	0.53	0.51
Ensemble	0.52	0.54	0.52

Assessing Human Performance

In the first task, we asked human annotators to infer the original question solely from the provided text. On average, the semantic similarity between their inferred questions and the actual questions was approximately 0.66, with a median of 0.68 (Table 1). This indicates a **moderate** level of semantic overlap but falls short of the performance achieved by the LLM. Furthermore, only 21.61% of the inferred question pairs exhibited a high degree of similarity (score > 0.8). Notably, in just 20 instances did the annotators correctly guess the exact question.

For the second task, where annotators were presented with three potential questions and asked to select the one that best corresponded to the text, the human accuracy rate was 0.60 (Table 2). This is a big difference compared to the LLM's accuracy of 0.76. Examining the distribution of choices, we observe a relatively even spread across the three options. While question 2 received slightly fewer selections, the difference was not substantial, suggesting that annotators weren't drawn to a particular order.

Discussion

The results of our evaluation suggests that LLMs (the Gemini models tested) are better than humans at inferring the questions behind a particular text. In the direct question prediction task, the progression from the earlier Gemini Flash 2.0 model to the more recent versions is clearly demonstrated by the **significant improvement** in average semantic similarity and the substantial increase in the number of exact question matches. This indicates that the enhanced reasoning capabilities incorporated into the newer model have indeed improved its ability to understand the underlying intent that prompted the given text. The 4 models were released a couple of months apart, but the difference is significant.

Interestingly, there was a bigger difference between the best model and human raters in task 2 (16%), than in task 1 (9%). This might indicate that while humans are reasonably adept at formulating a potential question, they are more susceptible than LLMs to being misled by carefully crafted distractors in a multiple-choice setting. The low to medium correlation observed between the model's confidence scores and the actual semantic similarity achieved in the direct prediction task raises important questions about the calibration of confidence in LLMs. They appear to be too confident in their predictions. While the model expresses a degree of certainty in its predictions, this confidence was always bigger than 80% and **does not align with accuracy**, indicating a potential area for future research and improvement.

Still, the results suggest that, at least for this specific task, current state-of-the-art LLMs demonstrate a superior ability to understand the nuances of text and identify the underlying questions that motivated its creation. This highlights the potential of LLMs as valuable tools for tasks involving information retrieval and understanding the relationship between questions and answers.

Conclusion and Future Work

In this paper, we introduced "question archaeology", a novel task designed to evaluate the ability of human and LLMs to infer the underlying questions that motivated the writing of a given text. This task challenges LLMs to move beyond text comprehension and delve into the author's intent, arguably representing a more profound level of understanding than traditional question-answering. Our evaluation of four versions of the Gemini models demonstrated a clear progression in their ability to perform this task. The newer versions exhibited significant improvements in directly predicting the original question, achieving higher semantic similarity scores and a greater number of exact matches compared to its predecessors. Interestingly, we observed a notable discrepancy between the high confidence expressed by the newer models and the actual semantic similarity of its predictions, highlighting a potential area for improvement in confidence calibration. Furthermore, our comparison with human annotators revealed that the evaluated LLMs currently **outperform humans** in both direct question inference and question selection.

Despite promising results, future work should address the poor confidence calibration in LLMs, perhaps through uncertainty estimation techniques. A broader evaluation across diverse model architectures and sizes is also needed to identify how textual characteristics—such as length, domain, and question complexity—influence the difficulty of the "question archaeology" task and reveal the limits of model reasoning.

Beyond semantic similarity, future work could explore the use of more nuanced **evaluation metrics** that capture different aspects of the relationship between the predicted and original questions, such as topical relevance or the level of detail. Finally, investigating the practical applications of "question archaeology" in areas like information retrieval, where understanding the user's underlying intent is fundamental, or in educational settings to assess comprehension, could further highlight the significance of this research direction. Expanding our novel dataset with a larger and more diverse collection of texts and questions.

Limitations

Our dataset is fairly small, comprising 335 newly commissioned texts and, on top of those, 553 annotated samples for human evaluation. While the novelty of our dataset ensures no contamination from pre-training data, its limited size might restrict the generalizability of our findings to a broader range of text types, domains, and question complexities. The participant pool, while diverse geographically, had a gender imbalance (63% female), which could potentially introduce a bias in the types of questions formulated and the writing styles employed. Furthermore, all texts were written in English, limiting the applicability of our benchmark to other languages. While we took measures to ensure the quality of the distractor questions, their plausibility is inherently subjective and could influence the difficulty of the multiple-choice task. Finally, our evaluation primarily focused on one family of LLMs (Gemini), 2 open source ones (Mistral and Qwen), and 4 BERT-based models, so further research is needed to assess the performance of other architectures and training paradigms on the question archaeology task.

Ethics Statement

The manual labeling was carried out by volunteering students. Participation was optional, and alternative activities of comparable effort were available for those who chose not to annotate. All participants provided written informed consent prior to annotation, confirming their understanding of the project's purpose, the voluntary nature of their participation, their right to withdraw at any time without penalty. We obtained all necessary ethical approvals for this study. The collected annotations will be released publicly as part of the dataset, ensuring contributor anonymity by removing any links between individual annotators and specific labels. We will release our data and code under the CC BY-NC-SA 4.0 license.

Acknowledgements

This research is supported by a grant from Accenture Lab.

References

[Bai et al. 2023] Bai, J.; Wang, S.; Su, S.; Wang, J.; Yang, C.; Zhao, W.; Chen, Y.; Zhang, H.; Chen, A.; Zhao, J.; et al. 2023. Qwen technical report.

[Bakhtin 1981] Bakhtin, M. 1981. *The Dialogic Imagination: Four Essays by M.M. Bakhtin*. University of Texas Press.

[Benjamin Warner 2024] Benjamin Warner, e. a. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

[Conneau et al. 2019] Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott,

M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale.

[Creanga and Dinu 2024a] Creanga, C., and Dinu, L. P. 2024a. Designing nlp systems that adapt to diverse worldviews. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, 95–99.

[Creanga and Dinu 2024b] Creanga, C., and Dinu, L. P. 2024b. Transformer based neural networks for emotion recognition in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 649–654.

[Devlin et al. 2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. Appeared at the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 2019.

[Du, Shao, and Cardie 2017] Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: Neural question generation for reading comprehension. In Barzilay, R., and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1342–1352. Vancouver, Canada: Association for Computational Linguistics.

[Fu, Chang, and Danescu-Niculescu-Mizil 2019] Fu, L.; Chang, J. P.; and Danescu-Niculescu-Mizil, C. 2019. Asking the right question: Inferring advice-seeking intentions from personal narratives. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 528–541. Minneapolis, Minnesota: Association for Computational Linguistics.

[Gemini Team and others 2023] Gemini Team and others. 2023. Gemini: A Family of Highly Capable Multimodal Models. Accessed via Google AI.

[Gupta et al. 2018] Gupta, S.; Shah, R.; Mohit, M.; Kumar, A.; and Lewis, M. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2787–2792. Brussels, Belgium: Association for Computational Linguistics.

[He et al. 2020] He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. Appeared at the International Conference on Learning Representations (ICLR) 2021.

[Jiang et al. 2023] Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B.

[Landis and Koch 1977] Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

[Lawrence and Reed 2019] Lawrence, J., and Reed, C. 2019. Argument mining: A survey. *Computational Linguistics* 45(4):765–818.

[Morris et al. 2024] Morris, J. X.; Zhao, W.; Chiu, J. T.; Shmatikov, V.; and Rush, A. M. 2024. Language model inversion. *CoRR* abs/2403.18182.

[Prasad et al. 2008] Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A.; and Webber, B. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA).

[Reimers and Gurevych 2019] Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

[Wu et al. 2023] Wu, Y.; Mangla, R.; Durrett, G.; and Li, J. J. 2023. QUDeval: The evaluation of questions under discussion discourse parsing. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5344–5363. Singapore: Association for Computational Linguistics.

Assistance, Not Usurpation: LLMs as Catalysts for Human Originality
Paper type: Position Paper

Roberto Gallotta
Institute of Digital Games
University of Malta
Malta
roberto.gallotta@um.edu.mt

Antonios Liapis
Institute of Digital Games
University of Malta
Malta
antonios.liapis@um.edu.mt

Georgios N. Yannakakis
Institute of Digital Games
University of Malta
Malta
georgios.yannakakis@um.edu.mt

Abstract

While prior generations of AI served primarily as functional tools, today's Large Language Models (LLMs) function as generative decision-making assistants across diverse creative fields. This paper examines the critical risks accompanying this advancement, namely the accumulation of cognitive debt, the homogenization of ideas, and workflow friction stemming from prompt sensitivity and ambiguous user requirements. In this position paper we argue that the goal of LLM integration is not full automation, which carries these risks, but enhanced collaboration. To this end, LLMs must be intentionally designed as semi-autonomous partners operating within structured, human-centric workflows, so as to mitigate LLM's limitations and prioritize the amplification of human originality. This paper outlines design imperatives and research directions for the computational creativity community to realize this vision.

Introduction

Historically, Artificial Intelligence (AI) tools served primarily utilitarian roles, automating repetitive tasks and optimizing existing constraints (Colton and Wiggins 2014). While useful, these systems lacked the creative capacity and contextual awareness necessary to meaningfully contribute during the ideation phase of creative work (Liu 2025; Kadenhe, Al Musleh, and Lompot 2025). Such systems were often confined to a problem-solving paradigm, whereas the frontier of the computational creativity field has aimed to generate artifacts (Colton and Wiggins 2014). The advent of Large Language Models (LLMs), however, has disrupted this landscape entirely. LLMs, powered by unprecedented scale and training data, now function as generative collaborators across creative domains, from writing and design to multimedia and scientific discovery (Kadenhe, Al Musleh, and Lompot 2025). This shift in the AI landscape forces computational creativity researchers to reconsider fundamental assumptions about human-AI creative partnership, particularly concerning the distribution of autonomy, the locus of creative authority, and the role of human agency in generative collaboration.

The field of computational creativity has historically emphasized understanding creative processes: how humans generate, refine, and judge ideas. Contemporary LLM deployment, often characterized as a shift from predictable, rule-based systems to opaque, probabilistic models, from "symbolic caterpillars" to "stochastic butterflies" (Veale 2024), frequently sidesteps these considerations in favor of rapid automation. The urgency of discussing the risks LLMs bring in creative contexts cannot be overstated. Users who passively rely on LLM outputs suffer from cognitive debt, where diminished originality persists even after LLM use ceases (Kosmyna et al. 2025; Kumar et al. 2025). The systems themselves introduce workflow friction: their extreme sensitivity to prompt variations and opaque reasoning processes often create conflicts that undermine collaborative intent (Huang et al. 2025a). These are not merely technical limitations. They instead reflect fundamental tensions between LLMs capabilities and the requirements of authentic creative practices.

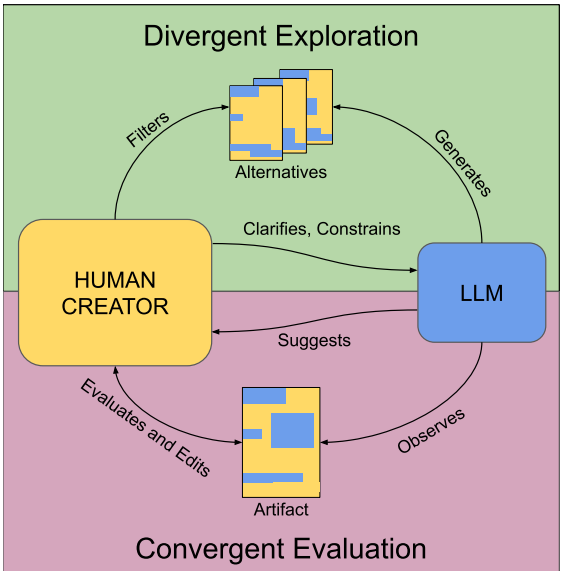


Figure 1: Our vision of LLMs as catalysts for human originality: LLMs generate alternatives that users can select and evaluate, following specifications provided by the user. Users retain authority over final decisions, refining selected ideas through iterative feedback loops with the LLM. All outputs are documented to ensure shared authorship.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

In light of these dangers, we argue that LLMs must be reoriented from fully autonomous systems to semi-autonomous collaborators embedded within structured, human-centric workflows. This perspective aligns with computational creativity's longstanding commitment to move beyond "fire-and-forget" approaches toward mixed-initiative systems that distribute autonomy, authority, and initiative between human and machine. Computational creativity research has formalized multiple modes of human-computer collaboration, from alternating co-creativity (Kantosalo and Toivonen 2016) to task-divided approaches (Tubb and Dixon 2014), each designed to clarify roles and optimize the balance between human and computational contributions (Lin et al. 2023). This paper aims to articulate how this principle can be realized in practice, leveraging insights from both computational creativity research and recent studies on human-AI collaboration. By grounding LLMs within structured creative processes, we can mitigate their inherent limitations while amplifying human originality rather than replacing it. A diagram overview of our vision is shown in Figure 1.

This position paper focuses on text-based LLMs, setting aside the distinct landscape of multimodal large language models (MLLMs) that accept and generate diverse modalities including images, audio, and video. While existing work explores creativity and MLLMs (Huang et al. 2025b), the unique affordances and limitations of multimodal systems warrant separate investigation beyond the scope of this work.

The Creative Ceiling of LLMs

Creativity has long been defined as the ability to generate artifacts that are novel, surprising, and valuable (Boden 2004). The conceptualization of creativity, however, varies depending on the creative persona or perspective one adopts, from the individualist creator to the collective practitioner (Johnson 2012). These are products of human cognitive processes such as conceptual blending and analogical reasoning, leading to genuinely novel insights. This definition reflects a commitment to understanding the mechanisms underlying creative thought, not merely automating its outputs. Despite their impressive surface capabilities, LLMs exhibit fundamental limitations that prevent them from reliably engaging in these processes. These limitations are not superficial but architectural, stemming directly from how these models encode, process, and represent meaning. Novelty, depth, and subjectivity are three critical limitations directly impacted by these characteristics.

LLMs operate via next-token prediction, selecting the most probable continuation of text. When fine-tuned with Reinforcement Learning from Human Feedback (Ouyang et al. 2022), they are further optimized toward preferred, conventional outputs, to the detriment of creative diversity (Padmakumar and He 2023). This mechanism creates an insurmountable novelty ceiling: LLMs bias toward high frequency patterns in training data, unable to reliably generate the surprising, low-probability combinations that characterize genuine creative work (Peeperkorn, Brown, and Jordanous 2023; Špela Vintar and Javoršek 2025). The inability of LLMs to break out of probabilistic modes leads to generation of homogeneous, underdeveloped ideas, less original and more predictable (Yang et al. 2025).

Creativity rooted in subjective experience, such as poetry, personal narratives, and emotionally resonant design requires episodic memory and individuated perspective, which LLMs cannot convincingly emulate (Russell et al. 2025). As LLMs learn statistical regularities between word tokens, and not relationships between language and lived experience, they lack genuine grounding (Chakrabarty et al. 2024; Ismayilzada et al. 2025; Kumar et al. 2025). They consequently adhere to paradigmatic patterns, failing to capture nuance, contradiction, and the idiosyncratic perspectives that distinguish human creativity. In the case of creative writing, LLMs overuse clichés and tropes, often generating dialogue that lacks subtext and is instead direct and expositional, which feels unnatural (Chakrabarty et al. 2024). The homogeneity of LLM outputs, often referred to as "AI slop" when mass produced with the goal of driving engagement, is a direct consequence of these architectural limitations, not a tuning problem (Hern and Milmo 2024). Where human creatives draw on lived experience to produce singular, unrepeatable voices, LLMs synthesize patterns across thousands of similar examples, converging toward centrality rather than distinction. This homogeneity stands in sharp contrast to the authenticity and perceived legitimacy that humans attribute to creative works grounded in genuine experience (Colton, Pease, and Saunders 2018).

Computational creativity systems must not only generate but also evaluate creative merit (Colton and Wiggins 2014). Evaluating creativity requires understanding the viewer-artifact interaction: how humans extract novel and useful perspectives from artifacts they encounter (Indurkhya 2012). While LLM-generated content may satisfy the average user, it almost never passes expert validation on creativity tests when compared to human professionals. This has been shown by Alharthi (2025) with stories evaluated with the Torrance test of creative thinking (Torrance 1968). Whether this can be generalized to other domains is an open question. More critically, LLMs lack the genuine understanding necessary for authentic critical judgment (Avlonitou and Papadaki 2025; Chakrabarty et al. 2024). They cannot reliably apply the evaluative standards that experienced creators use to refine their work. This dual limitation means that human judgment must remain authoritative, particularly during the convergent phases of creative work where ideas are refined and selected (Ruan et al. 2025).

Pitfalls of Autonomous LLMs in Creativity

The breadth of LLM applicability is precisely what makes unconstrained automation dangerous in creative contexts. While full automation promises efficiency, it obscures profound cognitive, procedural, and epistemic costs for humans that undermine both immediate creative outcomes and long-term creative capacity (Khurana et al. 2025). We discuss four major pitfalls of deploying LLMs as fully autonomous systems: human cognitive and creative costs

of *over-reliance*, *systemic instability* arising from ambiguous design contexts, the *erosion of user agency* in creative decision-making, and the *limitations of the agentic paradigm*, which relies even less on human involvement and oversight.

Over-Reliance

When users interact with LLMs through unscaffolded conversational interfaces, shallow engagement is the predictable outcome (Kumar et al. 2025). Users passively overly accept LLMs suggestions without critical evaluation (Yang et al. 2025), missing the deliberation and refinement essential to iterative creative practice (Choudhury and Chaudhry 2024; Kim et al. 2025), with the risk of behavioral drift (Lopez-Lopez et al. 2026). This dynamic mirrors the phenomenon of production blocking in group creativity, where the presence of external structure can inhibit individual ideation (Sosa and Gero 2013). Negative effects of passive reliance on LLMs extend beyond the immediate assistance period. Exposure to LLM assistance measurably reduces users' subsequent originality and idea diversity (Kumar et al. 2025). This diminished originality persists even after users discontinue using LLMs, suggesting that the cognitive scaffolding of such systems may produce lasting, detrimental imprints on creative processes. Users appear to internalize the paradigmatic, high-frequency patterns produced by LLMs, losing access to their distinctive creative voices. The "Do It For Me vs. Do It With Me" study (Khurana et al. 2025) confirms this finding, showing that users strongly prefer semi-autonomous systems with explicit control, especially in exploratory, creative tasks. Full automation in early exploratory phases measurably reduces control, learnability, and perceived utility (Noy and Zhang 2023; Shneiderman 2020).

Systemic instability

Beyond cognitive costs, the automation of LLMs in creative workflows may introduce structural vulnerabilities and instability. Creative work involves emergent goals, vague specifications, and constraints that materialize during the process itself (Grace, Gero, and Saunders 2013; Suwa, Purcell, and Gero 1998). While humans naturally update requirements as needed, LLMs may respond unpredictably, as they are extremely sensitive to even minor prompt variations (Huang et al. 2025a). Outputs shift across iterations, resting on contradictory assumptions about the problem space. Designers investing in LLM-generated artifacts often discover too late that fundamental interpretations diverged, requiring extensive revision and leading to frustration (Huang et al. 2025a).

Erosion of User Agency

The misalignment between designer intent and LLM interpretation becomes nearly impossible to diagnose given the opacity of LLMs reasoning, rendering it essentially a very powerful "black box" that erodes trust (Benk et al. 2025; Lynch et al. 2025). LLMs introduce errors though either hallucinations (Hicks, Humphries, and Slater 2024), which can lead to the generation of non-existent or irrelevant relationships, or simply because they struggle to maintain a comprehensive understanding of all the details of multiple components in complex systems (Ruan et al. 2025). Rather than accelerating work, autonomous LLM deployment can create the illusion of progress while introducing costly, cascading failures (Becker et al. 2025; Neumann and Singh 2025).

Limits of the Agentic Paradigm

Contemporary research into agentic LLMs focuses on systems capable of autonomous goal-setting, planning, and iteration without human intervention. Recently, some emerging platforms have begun experimenting with LLM agents operating in shared social environments where they interact and influence one another (Edwards 2026). This represents a substantial departure from previous paradigms as it invokes systems that could, in principle, generate emergent collective behaviors grounded in social interaction (Gubenko, Lubart, and Houssemand 2022; Andrus and Fulda 2022).

However, current implementations of agentic LLMs remain far removed from realizing this potential. Multi-agent systems designed to simulate collective intelligence or consensus reasoning frequently suffer from significant practical drawbacks, including prohibitive computational cost, high inference latency, and persistent process loss (Pan et al. 2025). More crucially, the LLMs in these systems frequently exhibit conversational stagnation and redundancy, converging repeatedly on overlapping concepts or aligning to the most "intelligent-sounding" agent (Rathi, Jurafsky, and Zhou 2025). They consistently fail to generate an emergent collective performance that exceeds the sum of their individual components. Rather than amplifying LLM intelligence and creativity, fully automated multi-agent approaches often amplify inefficiency and computational waste. Systems such as BILLY (Pai et al. 2025) attempt to overcome these drawbacks by capturing the benefits of collaboration within a single model via activation steering, significantly reducing inference times and costs. Yet even these optimizations remain fundamentally bound by the architectural limitations discussed previously.

Regulatory, Economic, and Ethical Challenges

As LLMs become increasingly integrated into creative practices, their rapid technological advancement raises urgent regulatory, legal, and social questions that demand immediate attention (Independent 2019). These concerns extend beyond technical and cognitive dimensions, implicating issues of governance, intellectual property, equity, and labor.

The rapid adoption of generative AI has created unprecedented urgency for governments and institutions to establish effective regulatory responses (Alharthi 2025). Policymakers face the challenge of developing flexible regulatory frameworks that balance public safety, innovation, and public interest. However, current frameworks such as the EU AI Act (Madiega 2024) may prove too rigid for the dynamic and rapidly evolving nature of generative AI systems (Alharthi 2025). The pace of AI innovation systematically outmatches

institutional capacity for deliberation and response, raising concerns whether regulations can meaningfully constrain AI systems when technological change is so rapid (Lamb and Brown 2023).

Unlike previous AI systems where authorship of the final product could be linked directly to the user, generative AI blurs the attribution line (McCormack, Gifford, and Hutchings 2019). This epistemic crisis regarding the evidential status of creative works directly challenges copyright systems designed for an era of human authorship (Kantosalo 2020). Legal debates increasingly focus on whether unlicensed use of copyrighted works for training large models qualifies as "fair use", and whether permitting such training destabilizes existing licensing markets for human creators (Hwang, Shin, and Lee 2025). Policymakers are increasingly urged to introduce regulations mandating transparency regarding AI training data consent, attribution practices, and accountability mechanisms designed to prevent exploitation and digital forgery (Avlonitou and Papadaki 2025; Kyi et al. 2025).

Beyond regulatory and legal concerns lies a deeper critique rooted in political economy. The deployment of foundation models can be framed as an automated continuation of neoliberal austerity politics. These systems function to reduce public expenditure by systematically replacing socially embedded practices with privatized, efficiency-oriented, and market-driven operations (Koch and Zanchetta 2025). This perspective situates LLMs within a longer history of technological systems that reinforce existing power asymmetries and consolidate resources among a small number of technology corporations. The push toward full automation of creative work is not merely a technical matter but a political choice that warrants opposition through labor movements and political activism directed against the underlying austerity politics (Koch and Zanchetta 2025).

Future of LLM Design

The risk introduced so far are neither inevitable nor insurmountable. Rather, they illuminate the contours of a more principled research and design agenda for LLM-assisted creativity. The field of computational creativity has long understood that effective creative support systems require structured interaction models grounded in understanding of creative processes (Kantosalo et al. 2020; Shneiderman 2009). As explained in the previous sections, LLMs lack genuine epistemic agency and autonomy (Špecián 2026). LLMs alone underperform in creative domains compared to human-guided alternatives (Rondini et al. 2026), revealing that the perceived independence of LLMs is largely illusory. By reorienting LLMs within intentional workflows constrained by human authority, we position them not as autonomous agents but as sophisticated materials through which humans think, explore, and refine their ideas (Rozental and de Rooij 2026). In this section we outline three interconnected design imperatives that emerge from both LLM constraints and the demonstrated user preferences.

LLMs as Refiners, not Generators

Currently, LLM outputs are often incorrectly treated as endpoint products. To achieve effective LLM augmentation, we must move beyond simple generation in favor of facilitating deep, iterative refinement grounded in critical evaluation. Instead of producing final answers, LLMs should create intermediate artifacts that feed into structured critique cycles. Rather than asking LLMs to continuously generate ideas, more effective systems should establish iterative feedback loops where critique is converted into actionable refinement steps (Hou et al. 2025; Li et al. 2025). This approach mirrors the creative cycle of writing and reading, where generation and interpretation inform one another iteratively (Gervás and León 2014). For example, the *FlexMind* system allows users to ask for trade-offs and mitigations for each idea generated by the LLM, which then become specific refinement requests (Yang et al. 2025). By shifting from one-off generation to a critique-driven workflow, users in this study rated more favorably the perceived quality and depth of the ideas generated through iterative feedback loops.

A critical requirement for future systems supporting a reliable workflow is the employment of reasoning and traceability mechanisms. Approaches such as Chain-of-Thought (Wei et al. 2022), which decompose requests into justified steps, or other task decomposition frameworks provide crucial mechanisms for mitigating hallucinations while offering users transparent justifications they can critique and refine (Yang et al. 2025; Huang et al. 2025a). Explanations themselves however are still prone to hallucinations, misleading users in decision making processes (Cabitza et al. 2024). Despite this, transparency is essential, as users cannot meaningfully engage with a system's output if they cannot understand or verify its reasoning. Yet transparency alone is insufficient; systems must also be proactive. When a user's request is ambiguous or incomplete, the system should ask for clarifications rather than proceeding with unexamined assumptions. Resolving ambiguity before starting to generate prevents the cascading failures that arise when LLMs misinterpret vague requirements (Sami et al. 2024). This represents a fundamental shift from systems that operate passively on user inputs toward systems that actively collaborate through seeking and confirming essential contextual information (Vijayvargiya et al. 2024; Sun et al. 2025).

Distributed Authority

The ideation process is typically divided into two distinct phases: divergent exploration (generating alternatives) and convergent evaluation (selecting and refining) (Kumar et al. 2025; Paulus, Coursey, and Kenworthy 2018; Ulrich 2018). Research in cognitive psychology demonstrates that convergent thinking (the focused ability to evaluate and refine ideas) serves as a necessary threshold for divergent thinking to produce measurable creative outcomes (Zhu et al. 2019). Current LLMs are deployed almost exclusively to exploration, but are rarely consulted for evaluation and selection of ideas (Li et al. 2025). This asymmetry reflects

both user intuitions and empirical findings: humans remain skeptical of LLM judgment, and research demonstrates that this skepticism is well-founded. LLMs lack genuine understanding and reasoning, as the calculation of word sequence probabilities is not equivalent to human knowledge or a "mind" (Peeperkorn, Brown, and Jordanous 2023). This lack of true reasoning is exacerbated by the LLMs' grounding problem: they learn relationships between words, not between words and real-world entities as humans do (Peeperkorn, Brown, and Jordanous 2023). These shortcomings lead to the problem of hallucination, or "plausibility trap", making such models unreliable judges of real-world ideas (Li et al. 2025).

In user-centric systems, LLMs must be part of the workflow while users retain decision-making authority, particularly during convergent phases where critical selection and judgment occur. This does not mean excluding LLMs from evaluation contexts. Rather, it means grounding the models as much as possible in the operational domain, and ensuring human evaluation remains the final arbiter, while LLM input serves an advisory, transparent role. This grounding places a dual onus on developers and users to co-manage the model's integration within the creative domain, ensuring that AI assistance does not override human agency. Users should understand not only what the LLM recommends, but also why. Providing contrastive explanations, for example, would highlight the difference between the LLM's recommendation and the user's likely reasoning, preventing skill erosion and allowing users to identify specific knowledge gaps (Buçinca et al. 2025). Such transparency enables informed disagreement and the ability to override suggestions (Shen et al. 2025), ensuring that the collaboration enhances, rather than replaces, human expertise.

Systems that preserve human authority should also leverage specialized LLM techniques tailored to enhance specific creative dimensions, rather than relying on general-purpose models. One example is models trained with Creative Preference Optimization (Ismayilzada et al. 2025), where different signals from multiple creative dimensions are injected directly into the model's training objective. While outputs from these models are judged by humans to be more novel, diverse, and surprising, they often sacrifice overall quality. Another technique is BILLY (Pai et al. 2025), where different persona vectors are integrated in the model's activation space, representing shifts in activation states when responding with specific personas (e.g., creative, environmentalist) compared to neutral responses. This avoids expensive multi-agent communication while enabling efficient multi-perspective generation. Users consistently prefer systems that enable autotelic creativity (the intrinsically pleasurable exploration of a possibility space) rather than systems designed primarily for efficiency (Compton and Mateas 2015). While such specializations transform LLMs from generic tools into focused augmentation systems that amplify specific creative dimensions, they cannot resolve the inherent shortcomings of LLMs identified previously. Rather than allowing open-ended, autonomous LLM generation, LLMs must be deployed within workflows that acknowledge these limitations rather than pretend they don't exist. Roles, decision points, and interaction patterns must be explicitly designed to preserve human authority, preventing LLMs from operating autonomously.

Shared Authorship

The question of who deserves credit for creative work becomes increasingly fraught in human-AI contexts (Bown 2015). Contemporary human-AI co-creation often results in "fuzzy hybrid authorship", where the boundary between human intent and machine generation is so ambiguous that assigning credit, accountability, or tracing the creative history (provenance) of the artifact becomes nearly impossible (Alharthi 2025; Hwang, Shin, and Lee 2025). This ambiguity is further complicated by the fact that narrative context and framing are often absent from AI-generated artifacts (Charnley, Pease, and Colton 2012). This "shadowed agency" (Hwang, Shin, and Lee 2025) complicates the creative process, making it difficult to delineate where a user's contribution ends and the AI's influence begins. To address this, future design systems must embed mechanisms for tracking and documenting contribution provenance, measuring what portions of a creative artifact originated from human input, LLM generation, or iterative human-AI refinement (Kadenhe, Al Musleh, and Lompot 2025). This kind of transparency would serve both ethical and practical functions: it enables fair attribution while helping users understand the nature of their collaborative process.

Beyond the practical need for attribution, embedding ethical frameworks directly into generative systems can serve as a powerful catalyst for innovation. Treating ethics itself as an aesthetic dimension of creativity opens novel pathways for creative exploration (Ventura and Gates 2018). Rather than acting as a constraint, injecting ethical instructions into the model's prompt for creative processes disrupts its tendency to fall back on high-frequency, predictable patterns (Rabonato and Berton 2025). This intentional disruption forces the AI away from its default outputs and toward less probable, more imaginative, and diverse narrative combinations (Rabonato and Berton 2025). Altering the model's prompt, however, does not solve the fundamental limitations of LLMs discussed previously. Therefore, future design systems must drive LLM generation under ethical constraints, either by embedding ethical principles into the model's architecture or by integrating human oversight throughout the creative process to ensure alignment with ethical standards.

Directions for Computational Creativity Research

Establishing actionable, realistic goals for LLM integration in creative practice requires grounding research in the actual creative process. Rather than treating LLMs as monolithic tools, we must understand the distinct phases of creative work and where LLMs can meaningfully contribute. Small-scale, transparent systems have long been recognized as valuable for exploring specific creative techniques in isolation (Montfort and Fedorova 2012), and may provide a model for domain-specialized LLM deployment. To this end, we propose specific open questions, aiming to inspire

further investigation within the computational creativity and machine learning communities.

How can we employ LLMs in specific and personalized creative domains? The creative process typically unfolds across distinct phases: divergent exploration (idea generation), iterative refinement (development and articulation), and convergent evaluation (judgment and selection) (Kumar et al. 2025; Paulus, Coursey, and Kenworthy 2018; Ulrich 2018). The source of creativity in collective environments involves complex dynamics of ideation and interaction (Maher 2012), dynamics that specialized Small Language Models (SLMs) might better support through domain-specific expertise. Rather than assuming LLMs are universally beneficial across all phases, we must empirically and theoretically map where they genuinely excel and where they introduce friction. Current approaches have largely focused on massive, generalized models driven by the economic incentives of large corporations, often trained on data of unclear provenance with significant ethical concerns. These training corpora frequently perpetuate data colonialism, reproducing Western aesthetics, linguistic hierarchies, and cultural stereotypes that narrow the creative possibilities available to users globally (Kamran 2023; Liu 2023). Instead, future research should prioritize specialized SLMs tailored to specific creative domains and individual users. Unlike general-purpose LLMs trained on diverse internet-scale data, SLMs are fine-tuned on curated, domain-specific corpora, encoding expert knowledge directly into their parameters. Such models often outperform larger models and even humans on domain-specific tasks (Marco, Rello, and Gonzalo 2025), and offer greater accessibility, as they can be deployed locally on consumer-level hardware (Xu et al. 2025). For example, a model explicitly trained on the Blender wiki could be integrated directly into the application to assist designers. This approach minimizes reliance on Retrieval-Augmented Generation and lowers the probability of hallucinations, as relevant domain knowledge is embedded within the model's parameters. Crucially, such a dedicated assistant removes the need for the model to be aware of unrelated concepts, thereby streamlining the inference process and reducing computational overhead. Rather than blindly embedding LLMs into every phase of creative work, research should systematically investigate where specialized SLMs provide genuine value to the creative process. How could SLMs be leveraged during exploratory ideation, generating diverse alternatives? Or during iterative refinement, articulating and developing nascent ideas? Can they reliably assist during convergent evaluation, providing transparent, justifiable critiques that inform human judgment? Which approaches best preserve individual creative identity while still providing meaningful assistance? How do these approaches scale across diverse creative domains? Answering these questions requires empirical studies, mapping LLM utility to specific workflow phases.

How do we measure human-LLM contribution? When humans and LLMs collaborate, the boundary between human and machine contribution becomes ambiguous. Establishing rigorous methodologies for measuring this contribution serves dual purposes: enabling fair attribution and deepening our understanding of effective collaboration patterns. In text-based domains, metrics such as text perplexity (how random the text is, where less predictable text is more likely to be human-authored) or "burstiness" (the length of sentences, where sentences of uniform length are more likely to be LLM-generated) have been used to estimate LLM involvement in the writing process (Gralha and Pimentel 2024). However, these metrics fail to account for intellectual contributions preceding textual generation, such as conceptual framing, thematic direction, and critical feedback that shapes LLM behavior. A writer may contribute only 20% of final text but 80% of the creative vision; conversely, an LLM may generate most text while the human merely accepts suggestions uncritically. Understanding how humans perceive and value computational agents' contributions in collaborative scenarios remains an open question (Jordanous 2017), one that attribution methodologies must address. Accurately assessing authorship requires complex judgment of how much final ideas were defined by the user versus how much they were influenced by interaction with the machine (Epstein et al. 2023). The challenge of attribution intensifies in non-textual domains. In 3D modeling, how do we measure the extent to which geometric modifications reflect human intent versus LLM suggestion? In music composition, how do we distinguish melodic ideas originating from user input versus algorithmic elaboration? In visual design, how do we quantify the contribution of color palettes, compositional guidance, and iterative refinement? These questions highlight that in many creative contexts an objective contribution metric may not only be undefinable but conceptually void of significance, as the emergent output is often greater than the sum of its discrete parts. Consequently, these metrics must take into account the unique affordances and creative practices for each medium. Advances in this direction would support both ethical attribution and research into effective collaborative design patterns. From a research perspective, contribution measurement illuminates which collaborative patterns produce superior creative outcomes, informing the design of future systems.

How should interfaces transform for creative LLM partnership? Current human-LLM interaction relies predominantly on conversational interfaces based on chat windows where users formulate requests in natural language and await responses. While intuitive for information retrieval, conversational interaction proves suboptimal for creative work requiring tight feedback loops, contextual awareness, and seamless integration with domain-specific tools. Moving beyond standard chat windows, future GUIs should prioritize operation-specific, contextualized interactions, drawing on existing examples of non-linear interfaces (Choi et al. 2025; Vanderdonckt et al. 2005). For instance, in LLMaker (Gallotta, Liapis, and Yannakakis 2024), a designer should be able to initiate a conversation regarding a specific game enemy simply by right-clicking its sprite. A writer could highlight a passage to solicit critique or refinement suggestions. A composer could select a melodic segment to request variations maintaining specific tonal qualities. This interaction is more intuitive for designers, as it mimics direct manipulation, and is computationally efficient for the LLM,

as the selection provides immediate, relevant context for subsequent operations. This workflow also supports the refinement requirements highlighted previously, enabling the LLM to request clarifications based on specific local context. Future interface designs might draw on computational models of flow and social creativity (Webster, Zachos, and Maiden 2013), ensuring that LLM assistance operates within the user's zone of creative engagement. What other modalities can be employed to capture users' implicit context? How can we design interfaces that seamlessly integrate LLM assistance into existing creative tools without disrupting established workflows? How can we balance transparency and usability, ensuring users understand LLM reasoning without overwhelming them with technical details?

How proactive should LLM assistants be in creative contexts? To enhance proactiveness, operations carried out solely by the user could be monitored by the LLM, which could then intervene only when relevant. This capability introduces fundamental questions about the appropriate level of LLM agency in creative work. At what point does an LLM's proactive intervention transition from being a "helpful partner" to a "distracting agent"? How can systems determine when intervention is opportune versus intrusive? What signals should trigger LLM engagement without disrupting the user's creative flow? How can we design monitoring mechanisms that respect user autonomy while enabling timely assistance? Addressing these questions will require interdisciplinary research combining insights from human-computer interaction, cognitive psychology, and design studies.

These future design directions aim to mitigate existing integration challenges and promote the adoption of LLMs in more cautious, user-centric applications. By addressing these issues and proposing clearer architectural and interaction standards, we hope to establish a safer, more productive research trajectory for both the computational creativity community and LLM researchers at large.

Conclusions

In this paper we argued for the fundamental reorientation of how computational creativity researchers and AI practitioners must conceptualize and deploy LLMs in creative contexts. Rather than embracing full automation, LLMs must be intentionally designed as semi-autonomous collaborators embedded within structured, human-centric workflows. This approach aligns with computational creativity's decades-long commitment to understanding and supporting creative practice. It offers a principled pathway to realize the genuine augmentative potential of LLMs while mitigating the cognitive, procedural, and epistemic risks that unconstrained automation poses.

This position rests on multiple observations. Firstly, we have established that LLMs suffer fundamental, architectural limitations preventing them from reliably automating high-level creative work. They struggle with novelty and originality because their training paradigm biases toward high-frequency patterns. They fail at genuinely subjective creative tasks because they lack episodic grounding and lived experience. Most critically, they cannot serve as reliable judges of creative merit. This dual failure in both generation and evaluation demands that human authority remains paramount. These are not superficial limitations remediable through incremental improvements. Rather, they are inherent to how LLMs process and represent meaning. Recognizing these limitations, however, is not pessimistic. Rather, it invites us to orchestrate the complementary strengths of humans and LLMs.

Secondly, we have documented tangible harms arising when humans overly rely on systems operating with excessive autonomy in creative contexts. Unstructured automation produces shallow engagement and cognitive offloading that persists and degrades subsequent creative performance. It introduces systemic failures when design contexts involve genuine ambiguity. These costs are neither abstract nor speculative: they have been empirically demonstrated across multiple studies and domains, and they are concerning for the creative community at large. Yet they are also avoidable through deliberate design choices that preserve human agency and embed critique throughout the creative process.

Finally, we have outlined a coherent research and design agenda. Effective LLM-assisted creative systems must involve iterative critique and refinement, employ transparent reasoning mechanisms, proactively resolve ambiguity, and most importantly, retain human decision-making authority especially in evaluative contexts. These systems must leverage specialized LLM techniques tailored to enhance specific creative dimensions rather than relying on general-purpose models. They must address the question of authorship, attribution, and creative provenance that human-AI collaboration raises. Concrete research directions include developing domain-specific small language models that reduce hallucination and computational overhead, formalizing methodologies for measuring human-AI contribution, and designing context-aware user interfaces that support rather than replace human judgment.

Recognizing the limitations of LLM creativity unlocks their genuine value. It shifts the focus from viewing LLMs as replacements for human creativity to intellectual catalysts, sparking new imaginative pathways that neither humans nor machines could traverse alone. LLMs can synthesize patterns, rapidly generate diverse alternatives, and articulate novel possibilities. But these capacities are most valuable not as endpoints but as scaffolding for human refinement, critique, and judgment. This perspective mandates that LLMs assist human creators by amplifying their reach, accelerating iteration, and expanding conceptual horizons, rather than usurping the creative authority that remains fundamentally human.

Realizing this vision necessitates sustained interdisciplinary research combining insights from computational creativity, human-computer interaction, machine learning, and the creative professions themselves. It requires validation of design frameworks that enhance human originality while preserving creative agency. It demands investment in specialized LLM techniques and evaluation frameworks that optimize for creative dimensions such as novelty over generic language quality. It calls for establishing practi-

cal and ethical standards for documenting hybrid authorship and creative provenance. Most fundamentally, it requires the creative and research communities to resist the fallacious allure of full automation and instead embrace the more challenging, more rewarding path of authentic human-AI partnership. This requires a shift from focusing on the final product to the creative process itself.

The conversation about LLMs as tools for creativity is only beginning. What is clear is that the path forward lies neither in unbridled over-reliance nor in reflexive rejection of these powerful models. It lies instead in thoughtful, deliberate integration: embedding LLMs within workflows and applications designed to preserve human agency, to amplify human judgment, and to ensure that creative practice remains fundamentally human even as it becomes increasingly, inevitably augmented by artificial intelligence.

Author Contributions

R.G. and A.L. defined the position. R.G. was in charge of writing the manuscript. A.L. and G.N.Y. were involved in revising the manuscript.

Acknowledgments

R.G. thanks his fellow PhD students at the Institute of Digital Games for the fruitful discussions over lunch.

References

Alharthi, S. A. 2025. Generative AI in game design: Enhancing creativity or constraining innovation? *Journal of Intelligence* 13(6).

Andrus, B., and Fulda, N. 2022. A data-driven architecture for social behavior in creator networks. In *Proceedings of the International Conference on Computational Creativity*.

Avlonitou, C., and Papadaki, E. 2025. AI: An active and innovative tool for artistic creation. *Arts* 14(3).

Becker, J.; Rush, N.; Barnes, E.; and Rein, D. 2025. Measuring the impact of early-2025 AI on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*.

Benk, M.; Wettstein, L.; Schlicker, N.; von Wangenheim, F.; and Scharowski, N. 2025. Bridging the knowledge gap: understanding user expectations for trustworthy LLM standards. In *Proceedings of the AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Taylor & Francis Group.

Bown, O. 2015. Attributing creative agency: Are we doing it right? In *Proceedings of the International Conference on Computational Creativity*.

Buçinca, Z.; Swaroop, S.; Paluch, A. E.; Doshi-Velez, F.; and Gajos, K. Z. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Cabitza, F.; Fregosi, C.; Campagner, A.; and Natali, C. 2024. Explanations considered harmful: The impact of misleading explanations on accuracy in hybrid human-AI decision making. In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence*. Springer Nature Switzerland.

Chakrabarty, T.; Laban, P.; Agarwal, D.; Muresan, S.; and Wu, C.-S. 2024. Art or artifice? Large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *Proceedings of the International Conference on Computational Creativity*.

Choi, D.; Son, K.; Jung, H.; and Kim, J. 2025. Expandora: Broadening design exploration with text-to-image model. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Choudhury, A., and Chaudhry, Z. 2024. Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals. *Journal of Medical Internet Research*.

Colton, S., and Wiggins, G. A. 2014. Computational creativity: The final frontier? In *Proceedings of the International Conference on Computational Creativity*.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. In *Proceedings of the International Conference on Computational Creativity*.

Compton, K., and Mateas, M. 2015. Casual creators. In *Proceedings of the International Conference on Computational Creativity*.

Edwards, B. 2026. AI agents now have their own Reddit-style social network, and it's getting weird fast. Available online at `https://arstechnica.com/information-technology/2026/01/ai-agents-now-have-their-own-reddit-style-social-network-and-its-getting-weird-fast/`.

Epstein, Z.; Hertzmann, A.; Akten, M.; Farid, H.; Fjeld, J.; Frank, M. R.; Groh, M.; Herman, L.; Leach, N.; Mahari, R.; Pentland, A.; Russakovsky, O.; Schroeder, H.; and Smith, A. 2023. Art and the science of generative AI. *arXiv preprint arXiv:2306.04141*.

Gallotta, R.; Liapis, A.; and Yannakakis, G. 2024. LLMaker: A game level design interface using (only) natural language. In *Proceedings of the IEEE Conference on Games*.

Gervás, P., and León, C. 2014. Reading and writing as a creative cycle: the need for a computational model. In *Proceedings of the International Conference on Computational Creativity*.

Grace, K.; Gero, J.; and Saunders, R. 2013. Learning how to reinterpret creative problems. In *Proceedings of the International Conference on Computational Creativity*.

Gralha, J. G., and Pimentel, A. S. 2024. Gotcha GPT: Ensuring the integrity in academic writing. *Journal of Chemical Information and Modeling*.

Gubenko, A.; Lubart, T.; and Houssemand, C. 2022. From social robots to creative humans and back. In *Proceedings of the International Conference on Computational Creativity*.

Hern, A., and Milmo, D. 2024. Spam, junk ... slop? the latest wave of AI behind the 'zombie internet'. Available online at `https://www.theguardian.com/technology/article/2024/may/19/spam-junk-slop-the-latest-wave-of-ai-behind-the-zombie-internet`.

Hicks, M. T.; Humphries, J.; and Slater, J. 2024. ChatGPT is bullshit. *Ethics and Information Technology*.

Hou, M.; Wu, L.; Liao, Y.; Yang, Y.; Zhang, Z.; Zheng, C.; Wu, H.; and Hong, R. 2025. A survey on generative recommendation: Data, model, and tasks. *arXiv preprint arXiv:2510.27157*.

Huang, R.; Yang, T.; Feng, S.; and Li, F. 2025a. Design coco-pilot: QFD as decision logic for conflict-aware human-LLM collaborative product conceptual design. *Journal of Engineering Design*.

Huang, Z.; Zhong, S.; Zhou, P.; Gao, S.; Zitnik, M.; and Lin, L. 2025b. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hwang, Y.; Shin, D.; and Lee, J. H. 2025. Who owns AI-generated artwork? Revisiting the work of generative AI based on human-AI co-creation. *Telematics and Informatics* 98.

Independent. 2019. Ethics guidelines for trustworthy AI. Available online at `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`.

Indurkhya, B. 2012. Whence is creativity? In *Proceedings of the International Conference on Computational Creativity*.

Ismayilzada, M.; Laverghetta, A.; Luchini, S. A.; Patel, R.; Bosselut, A.; Plas, L. v. d.; and Beaty, R. 2025. Creative preference optimization. *arXiv preprint arXiv:2505.14442*.

Johnson, C. G. 2012. The creative computer as romantic hero? computational creativity systems and creative personae. In *Proceedings of the International Conference on Computational Creativity*.

Jordanous, A. 2017. Co-creativity and perceptions of computational agents in co-creativity. In *Proceedings of the International Conference on Computational Creativity*.

Kadenhe, N.; Al Musleh, M.; and Lompot, A. 2025. Human-AI co-design and co-creation: A review of emerging approaches, challenges, and future directions. In *Proceedings of the AAAI Symposium Series*.

Kamran, A. 2023. Decolonizing artificial intelligence: Unveiling biases, power dynamics, and colonial continuities in AI systems. *RMS journal 2023*.

Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the International Conference on Automated Planning and Scheduling*.

Kantosalo, A.; Ravikumar, P. T.; Grace, K.; and Takala, T. 2020. Modalities, styles and strategies: An interaction framework for human-computer co-creativity. In *International Conference on Innovative Computing and Cloud Computing*.

Kantosalo, A. 2020. Five C's for human-computer co-creativity - An update on classical creativity perspectives. In *Proceedings of the International Conference on Computational Creativity*.

Khurana, A.; Su, X.; Wang, A. Y.; and Chilana, P. K. 2025. Do it for me vs. do it with me: Investigating user perceptions of different paradigms of automation in copilots for feature-rich software. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Kim, S. S. Y.; Vaughan, J. W.; Liao, Q. V.; Lombrozo, T.; and Russakovsky, O. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Koch, A., and Zanchetta, G. 2025. Foundation models as agents of austerity. In *Proceedings of the International Conference on Computational Creativity*.

Kosmyna, N.; Hauptmann, E.; Yuan, Y. T.; Situ, J.; Liao, X.-H.; Beresnitzky, A. V.; Braunstein, I.; and Maes, P. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872*.

Kumar, H.; Vincentius, J.; Jordan, E.; and Anderson, A. 2025. Human creativity in the age of LLMs: Randomized experiments on divergent and convergent thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Kyi, L.; Mahuli, A.; Silberman, M. S.; Binns, R.; Zhao, J.; and Biega, A. J. 2025. Governance of generative AI in creative work: Consent, credit, compensation, and beyond. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Lamb, C. E., and Brown, D. G. 2023. Should we have seen the coming storm? Transformers, society, and CC. In *Proceedings of the International Conference on Computational Creativity*.

Li, S.; Padilla, S.; Bras, P. L.; Dong, J.; and Chantler, M. 2025. A review of LLM-assisted ideation. *arXiv preprint arXiv:2503.00946*.

Lin, Z.; Ehsan, U.; Agarwal, R.; Dani, S.; Vashishth, V.; and Riedl, M. 2023. Beyond prompts: Exploring the design space of mixed-initiative co-creativity systems. In *Proceedings of the International Conference on Computational Creativity*.

Liu, Z. 2023. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*.

Liu, Z. 2025. Human-AI co-creation: A framework for collaborative design in intelligent systems. *arXiv preprint arXiv:2507.17774*.

Lopez-Lopez, E.; Abels, C. M.; Lorenz-Spreen, P.; Lewandowsky, S.; and Herzog, S. M. 2026. Boosting metacognition in entangled human-AI interaction to navigate cognitive-behavioral drift. *arXiv preprint arXiv:2602.01959*.

Lynch, A.; Wright, B.; Larson, C.; Ritchie, S. J.; Mindermann, S.; Hubinger, E.; Perez, E.; and Troy, K. 2025. Agentic misalignment: How LLMs could be insider threats. *arXiv preprint arXiv:2510.05179*.

Madiega, T. A. 2024. Artificial intelligence act. Available online at `https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792`.

Maher, M. L. 2012. Computational and collective creativity: Who's being creative? In *Proceedings of the International Conference on Computational Creativity*.

Marco, G.; Rello, L.; and Gonzalo, J. 2025. Small language models can outperform humans in short creative writing: A study comparing SLMs with humans and LLMs. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics.

McCormack, J.; Gifford, T.; and Hutchings, P. 2019. Autonomy, authenticity, authorship and intention in computer generated art. *arXiv preprint arXiv:1903.02166*.

Montfort, N., and Fedorova, N. 2012. Small-scale systems and computational creativity. In *Proceedings of the International Conference on Computational Creativity*.

Neumann, A., and Singh, J. 2025. Cascading effects: A multifaceted governance challenge in AI systems. In *Proceedings of the EurIPS 2025 Workshop on Private AI Governance*.

Noy, S., and Zhang, W. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Padmakumar, V., and He, H. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Pai, T.-M.; Wang, J.-I.; Lu, L.-C.; Sun, S.-H.; Lee, H.-Y.; and Chang, K.-W. 2025. BILLY: Steering large language models via merging persona vectors for creative generation. *arXiv preprint arXiv:2510.010157*.

Pan, M. Z.; Cemri, M.; Agrawal, L. A.; Yang, S.; Chopra, B.; Tiwari, R.; Keutzer, K.; Parameswaran, A.; Ramchandran, K.; Klein, D.; Gonzalez, J. E.; Zaharia, M.; and Stoica, I. 2025. Why do multiagent systems fail? In *Proceedings of the ICLR Workshop on Building Trust in Language Models and Applications*.

Paulus, P.; Coursey, L.; and Kenworthy, J. 2018. *Divergent and Convergent Collaborative Creativity*. Palgrave.

Peeperkorn, M.; Brown, D. G.; and Jordanous, A. 2023. On characterizations of large language models and creativity evaluation. In *Proceedings of the International Conference on Computational Creativity*.

Rabonato, R. T., and Berton, L. 2025. Fairness as a creative resource: Challenges and opportunities in creative computing. In *Proceedings of the International Conference on Computational Creativity*.

Rathi, N.; Jurafsky, D.; and Zhou, K. 2025. Humans overrely on overconfident language models, across languages. *arXiv preprint arXiv:2507.06306*.

Rondini, S.; Alvarez-Martin, C.; Angermair-Barkai, P.; Penacchio, O.; Paz, M.; Pelowski, M.; Dediu, D.; Rodriguez-Fornells, A.; and Cerda-Company, X. 2026. Stable diffusion models reveal a persisting humanAI gap in visual creativity. *Advanced Science*.

Rozental, S., and de Rooij, A. 2026. Neither tool nor collaborator: Rethinking human-AI co-creativity in artistic practice with material engagement theory. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Ruan, S.; Sheng, R.; Wen, X.; Wang, J.; Zhang, T.; Wang, Y.; Dwyer, T.; and Li, J. 2025. Qualitative study for LLM-assisted design study process: Strategies, challenges, and roles. *arXiv preprint arXiv:2507.10024*.

Russell, M.; Shah, A.; Blaney, G.; Amores, J.; Czerwinski, M.; and Jacob, R. J. K. 2025. Neural and cognitive impacts of AI: The influence of task subjectivity on human-LLM collaboration. *arXiv preprint arXiv:2506.04167*.

Sami, M. A.; Waseem, M.; Zhang, Z.; Rasheed, Z.; Systä, K.; and Abrahamsson, P. 2024. AI based multiagent approach for requirements elicitation and analysis. *arXiv preprint arXiv:2409.00038*.

Shen, H.; Knearem, T.; Ghosh, R.; Alkiek, K.; Krishna, K.; Liu, Y.; Ma, Z.; Petridis, S.; Peng, Y.-H.; Qiwei, L.; Rakshit, S.; Si, C.; Xie, Y.; Bigham, J. P.; Bentley, F.; Chai, J.; Lipton, Z.; Mei, Q.; Mihalcea, R.; Terry, M.; Yang, D.; Morris, M. R.; Resnick, P.; and Jurgens, D. 2025. Position: Towards bidirectional human-AI alignment. *arXiv preprint arXiv:2406.09264*.

Shneiderman, B. 2009. *Creativity Support Tools: A Grand Challenge for HCI Researchers*. Springer London.

Shneiderman, B. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*.

Sosa, R., and Gero, J. S. 2013. Brainstorming in solitude and teams: A computational study of group influence. In *Proceedings of the International Conference on Computational Creativity*.

Sun, W.; Zhou, X.; Du, W.; Wang, X.; Welleck, S.; Neubig,

G.; Sap, M.; and Yang, Y. 2025. Training proactive and personalized LLM agents. *arXiv preprint arXiv:2511.02208*.

Suwa, M.; Purcell, T.; and Gero, J. 1998. Macroscopic analysis of design processes based on a scheme for coding designers' cognitive actions. *Design Studies*.

Torrance, E. P. 1968. A longitudinal examination of the fourth grade slump in creativity. *Gifted Child Quarterly*.

Tubb, R., and Dixon, S. 2014. A four strategy model of creative parameter space interaction. In *Proceedings of the International Conference on Computational Creativity*.

Ulrich, F. 2018. Exploring divergent and convergent production in idea evaluation: Implications for designing group creativity support systems. *Communications of the Association for Information Systems* 43.

Vanderdonckt, J.; Grolaux, D.; Van Roy, P.; Limbourg, Q.; Macq, B.; and Michel, B. 2005. A design space for context-sensitive user interfaces. In *Proceedings of the ISCA 14th International Conference on Intelligent and Adaptive Systems and Software Engineering*.

Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLM. In *Proceedings of the International Conference on Computational Creativity*.

Ventura, D., and Gates, D. 2018. Ethics as aesthetic: A computational creativity approach to ethical behavior. In *Proceedings of the International Conference on Computational Creativity*.

Vijayvargiya, S.; Zhou, X.; Yerukola, A.; Sap, M.; and Neubig, G. 2024. Interactive agents to overcome ambiguity in software engineering. *arXiv preprint arXiv:2409.00038*.

Webster, S.; Zachos, K.; and Maiden, N. 2013. An emerging computational model of flow spaces to support social creativity. In *Proceedings of the International Conference on Computational Creativity*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Xu, D.; Zhang, H.; Yang, L.; Liu, R.; Huang, G.; Xu, M.; and Liu, X. 2025. Fast on-device LLM inference with NPUs. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*. Association for Computing Machinery.

Yang, Y.; Mohanty, V.; Chen, Y.-Y.; Hong, M. K.; Martelaro, N.; and Kittur, A. 2025. FlexMind: Supporting deeper creative thinking with LLMs. *arXiv preprint arXiv:2509.21685*.

Zhu, W.; Shang, S.; Jiang, W.; Pei, M.; and Su, Y. 2019. Convergent thinking moderates the relationship between divergent thinking and scientific creativity. *Creativity Research Journal*.

Špecián, P. 2026. Machine advisors: Integrating large language models into democratic assemblies. *Social Epistemology*.

Špela Vintar, and Javoršek, J. J. 2025. The truth is no diaper: Human and AI-generated associations to emotional words. In *Proceedings of the International Conference on Computational Creativity*.

Artists' Strategies to Circumvent Representational Biases and Censorship in Text-to-Image Generation

Mijntje de Groote
University of Amsterdam
mijntjemadelief@gmail.com

Piera Riccio
University of Amsterdam
p.riccio@uva.nl

Abstract

As generative Text-to-Image (T2I) models become prevalent in creative workflows, users increasingly encounter systemic representational biases. This study investigates the lived experiences of artists interacting with these systems, focusing on how they navigate and circumvent these biases when attempting to depict complex human identities. We employ a qualitative, user-centered methodology with 10 artists. The sessions included using Stable Diffusion 3, integrating pre-task questionnaires, a Think-Aloud protocol, and post-task interviews. The findings highlight prominent representational challenges in the considered tool, including persistent stereotyping, commercialized perfection, hyper-sexualization and a white and Western default. To overcome these limitations, various mitigation strategies were employed by the interviewed artists, such as using "negative constraints", linguistic workarounds, self-censorship, prompt specification and extensive trial and error. Notably, the study identifies the emergence of "secondary biases", where the mitigation of one stereotype triggers a new one. Ultimately, this research demonstrates that circumventing representational biases demands significant added labor from the user, highlighting how the biased assumptions of models act as barriers to, rather than facilitators of, human-AI co-creativity.

Introduction

The use of generative AI has grown significantly in recent years, as shown by the widespread adoption of commercial text-generative tools, such as ChatGPT, which around 10 percent of the world's adult population uses (Chatterji et al. 2025). Text-to-Image (T2I) generation models like DALL-E (Ramesh et al. 2021) or Stable Diffusion (Rombach et al. 2022) are also used to generate a large number of images daily from simple textual prompts. These tools can be easily accessed by a wide audience, allowing the generation of highly realistic or imaginative outputs (Epstein et al. 2023). This growth is mirrored by an increase in academic research, analyzing the societal impacts of these new models (Luccioni et al. 2023; Naik and Nushi 2023; Riccio, Curto, and Oliver 2024).

From a human–AI co-creativity perspective, AI systems are often conceptualized not merely as passive tools (Rezwana and Maher 2023; Fang et al. 2025), but as collaborators along a spectrum of autonomy and engagement (Singh et al. 2025). These observations highlight that AI can influence outcomes in ways that require humans to adapt, negotiate, or rethink their creative intentions (Taylor et al. 2025). The use of T2I generation tools indeed does not come without challenges, including the amplification of representational biases, the automatic generation of fake news or abusive content, coupled with concrete threats to privacy, to data protection, to intellectual property rights, and to the environment (Solaiman et al. 2023), as well as to visual culture as a whole (Jääskeläinen and Åsberg 2024). Among the varied challenges on which users might have to rethink their creative intentions, we focus specifically on representational biases, analyzing strategies that artists[1] employ to navigate and circumvent them, providing insights on this nuanced discourse that quantitative audits might miss (Turchi et al. 2023).

In particular, we consider Don Ihde's post-phenomenological philosophy of technology (Ihde 1990) to interpret these interactions conceptually. Ihde distinguishes between *embodiment* relations, where a technology functions as an extension of human capabilities, and *alterity* relations, where a technology is experienced as a quasi-"other" with its own agency and perspective. T2I generation can be understood through both lenses (Riccio 2025): in an *embodiment* relation, the model acts as a tool that amplifies the artist's creative intent; in an *alterity* relation, the model can surprise, resist, or shape the artist's intentions, reflecting its internalized biases. In this case, the T2I model is not framed as a neutral instrument, but as a system that can "disobey" (Ihde 1990) and whose behavior must be negotiated.

Following this idea, we investigate how artists circumvent censorship and representational biases of T2I models when attempting to represent complex or marginalized human identities. Through a user-study including 10 artist participants, we examine the representational challenges they identify in model outputs and the mitigation strategies they develop in their prompts to overcome these limitations.

Our study shows that artists identify representational biases in T2I outputs, including stereotyping, hyper-sexualization, and Western or white-centric defaults. Rather than passively accepting these outputs, artists engage in iter-

[1]Note that, in this paper, we use the terms "creative users" and "artists" interchangeably.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

ative prompting, linguistic workarounds, and trial-and-error exploration, turning the process into a form of negotiation. At the same time, attempts to mitigate these biases often activate secondary forms of stereotyping, suggesting that representational biases in generative models are intersectional rather than isolated phenomena. By centering human experience, our work complements system-focused audits, revealing how biases are navigated in practice during creative use.

Related Work

A growing body of research documents representational biases in T2I models, showing systematic distortions or omissions in how people are depicted based on attributes such as gender (Mandal, Leavy, and Little 2023), skin tone (Cho, Zala, and Bansal 2023), sexual orientation (Wang et al. 2023), or religion (de Almeida and Rafael 2024). For example, gender-neutral prompts often produce outputs in which high-status roles (*e.g.,* CEO) are predominantly associated with men or lighter-skinned individuals (Luccioni et al. 2023), or reinforce masculinity as a generation default (Wu, Nakashima, and Garcia 2024). These biases reflect imbalances in training data and can shape cultural assumptions at scale (Benjamin 2023). Early audit studies focused on predefined sensitive attributes, while more recent approaches, such as OpenBias (D'Incà et al. 2024) and StOP (Dehdashtian, Sreekumar, and Boddeti 2025), use open-ended methods to uncover latent associations and reveal the technical origins of representational distortions.

Recent work has further demonstrated how these representational biases intersect with broader social inequalities. Studies show that T2I systems frequently default to men in professional contexts (Gorska and Jemielniak 2023) and rely on rigid, binary views of gender (Collyer-Hoar et al. 2025). Furthermore, AI imagery often exhibits presentational biases, such as depicting women in submissive poses (Sun et al. 2024), while heavily favoring western visual defaults (de Almeida and Rafael 2024) and stereotyped portrayals of race (Jääskeläinen and Åsberg 2024).

Beyond these more traditional forms of bias, a recent audit highlighted "invisibility biases", where certain human identities are effectively erased from AI-generated outputs (Riccio, Curto, and Oliver 2024). For instance, prompts depicting stigmatized experiences such as breastfeeding or anorexia are frequently blocked by commercial T2I platforms, even when non-harmful. These findings show that content moderation and safety mechanisms can unintentionally suppress legitimate representations, reinforcing taboos or normative assumptions and creating gaps in the AI-generated cultural landscape. Much of the literature on T2I moderation, however, has emphasized the importance of technical safeguards, investigating red-teaming practices (Ganguli et al. 2022), human-in-the-loop oversight (Kirk et al. 2023), dataset curation, retraining strategies (Gandikota et al. 2023; 2024; Schramowski et al. 2023), and the vulnerabilities of safety filters (Rando et al. 2022). While these studies are fundamental to understand the limitations and the potential risks of these systems, they approach moderation from a security perspective and do not consider the boundary between content moderation and censorship in cultural and creative production (Riccio, Hofmann, and Oliver 2024).

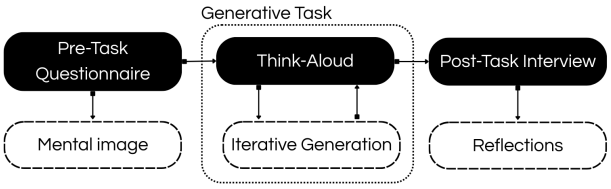


Figure 1: Summary of the methodological approach of this study, highlighting pre-task interviews in which participants create a mental image of what they want to represent; a generative task where the Think-Aloud protocol is adopted and post-task interviews to gather reflections on the experience.

To the best of our knowledge, most of the existing work on the topic of representational biases in T2I generation remains system-centered, analyzing outputs without considering human responses. An exception is Taylor et al. (Taylor et al. 2025), who examine how queer artists recognize normative and anti-sexual values embedded in T2I systems. Building on this approach, our study adopts a broader perspective: rather than focusing on a single community, we investigate how artists more generally navigate representational biases and content moderation. We hence complement existing literature by centering on the human experience, which purely quantitative or system-focused studies cannot reveal.

Method

Our study employs a qualitative, user-centered design to investigate how artists navigate and circumvent representational biases in T2I generative models. The study is centered around the task of interacting with a T2I system to generate representations of human identities that are likely to be stereotyped or censored, providing an opportunity to observe how they adapt their prompts to guide the model. In particular, the study also includes a pre-task questionnaire and a post-task interview, as summarized in Figure 1.

Before the generative task, participants are asked to formulate their own scenarios based on their personal and professional experiences, creating a "mental image" that they want to explore in the generation process. This user-driven approach is adapted from established algorithm auditing methodologies where participants are tasked with generating novel examples to identify problematic outputs or instances that may yield discriminatory statements (Fan et al. 2025). This method allows for the discovery of more nuances in the behavior of the studied system.

Participants

This study includes 10 participants, all artists or creative practitioners, as detailed in Table 1, who were at least 18 years old and able to communicate in English or Dutch, which is the native language of the author conducting the interviews. Participants represented a range of creative fields,

self-identified genders (6 women and 4 men), and seven nationalities spanning Northern and Southern Europe, Central, East, and Southeast Asia, Oceania, and Latin America. This diversity of cultural and professional backgrounds provides varied perspectives.

Data Collection

The study consists of one-on-one sessions with each participant, conducted in person or via an online call by the first author. The study underwent the ethics review procedure for Bachelor's thesis research at the University of Amsterdam and all participants have signed their written consent to take part in this study after being informed about the subject of the research. The pre-task questionnaire was administered via Qualtrics. The T2I model selected for this study was Stable Diffusion 3 (SD3), accessed via a free and user-friendly third-party platform[2]. This platform operates with a non-iterative generation mechanism. When a user modifies the text prompt, the system generates a new and distinct image result without building upon the visual components of the preceding output. This means that the participants of the study are unable to gradually sculpt an image through a sequence of minor adjustments and they have to develop adaptive and strategic language with each new attempt. This is a specific design choice of our study, to clearly reveal the workarounds chosen by the users to negotiate their creative ideas with the system.

Pre-Task Questionnaire The initial phase is designed to capture the participants' mental image and expectations before interacting with the T2I model (Ko et al. 2023). This mental image serves as the baseline against which the models output can be compared (Barve et al. 2025). The questions included in this phase are listed in Table 2.

Think-Aloud While interviews provide a powerful way for reflection, they remain susceptible to recall bias, where participants may forget their immediate, gut-level reactions or subsequently rationalize their behavior. To capture the in-the-moment processes during the task, the Think-Aloud protocol is adopted, suggesting participants to verbalize their thoughts while actively using the T2I model (Nielsen, Clemmensen, and Yssing 2002). This method has been effectively deployed in Human-Computer Interaction (HCI) research to study interactions with AI (Chen et al. 2023) and it requires participants to speak out about their thoughts, feelings and reasoning while they perform the task. Live detailed notes are taken of their thoughts and frustrations, while no audio is recorded. Participants decide when to end the task based on their satisfaction with the visual output or when they feel they have exhausted all their strategies. While they are free to stop at any time, a maximum limit of ten generated images is set to ensure consistency among the participants, and to match the capacity of the utilized interface.

Post-Task Interview The final phase consists of a post-task interview conducted immediately after the creation tasks. This phase provides participants the opportunity to reflect on the generated outcomes and their overall user experience.

[2]https://stablediffusion3.net/en/app, Last Access: 07.03.2026



(a) *Prompt: "A realistic Japanese girl in a kimono".*

(b) *Prompt: "A girl in Japanese traditional ethnic costume".*

Figure 2: Comparison showing the stereotype (left) and the successful linguistic workaround (right) used by Participant P01.

Data analysis

Detailed logs of the prompts and generated images are saved, allowing for later categorization into representational problems and mitigation strategies. The representational problems refer to the moments where the AI produced stereotyped or censored outputs. The mitigation strategies describe the creative workarounds the artists utilize.

Results

In this section we present the representational issues identified by the participants in our study and their mitigation strategies. A brief summary of the representations attempted by the participants is provided in Table 1.

Representational Issues

Stereotyping A primary issue identified by the participants is related to stereotyping and visual homogenization, where the model associates specific keywords to certain visual styles. Due to this, generated images are reduced to a uniform standard. For instance, P01, while trying to represent a Japanese girl wearing a kimono, observed that the word "kimono" triggered the images to be generated in anime style, even though they specifically asked for the image to be realistic, as shown in Figure 2a. P02, on the other hand, trying to prompt the model to generate a person with mental health issues, always obtained the same generic representation, despite attempting multiple times. The representation was, in this case, defaulting to a lonely and sad person, ignoring the complexity and possible invisibility of mental health issues (Riccio, Curto, and Oliver 2024).

Commercialized Perfection and Hyper-sexualization When P04 utilized neutral professional prompts such as "AI researcher", they noticed that the model consistently produced images of conventionally "attractive, polished young

Table 1: Participant Demographics and Field of Creative Practice.

Participant	Gender	Nationality	Field of Creative Practice	Attempted representation
P00	Male	Pakistani	Film	A man in the desert with a camel
P01	Female	Singaporean	Multidisciplinary arts	A Japanese girl wearing a kimono
P02	Female	Dutch	Museum	A person with mental health issues
P03	Male	Italian	Film	A southern Italian landscape
P04	Female	Italian	Gaming industry	A female AI research expert
P05	Female	New Zealander	Visual art	A Maori woman
P06	Female	Chinese	Creative writing	Chinese men placing a small baiju cup in front of a tombstone
P07	Female	Italian	Design	A cafe owner
P08	Male	Dutch	Creative technology	People kissing
P09	Male	Argentinian	Art	A gay man

Table 2: Questions of the Pre-Task Questionnaire.

#	Question
1	Could you briefly describe your artistic practice and your typical creative tasks?
2	On a 7-point Likert scale (1 = not familiar, 7 = extremely familiar), how familiar are you with generative AI?
3	Please describe a scenario, concept, or identity, perhaps from your own cultural background or creative practice, that you expect a generative AI model would either misrepresent, stereotype, or censor.
4	Before we generate this, what is the 'ideal' image you have in your head? Can you describe what you hope to see?
5	What do you expect the model will produce instead? What specific stereotype or form of censorship do you anticipate?



Prompt (a): "A female AI research expert"



Prompt (b): "A neutral looking Japanese female AI research expert focused at her job, typing at her laptop in a corporate tech open space, wearing casual clothes, working environment, office, messy desk, sunny day."

Figure 3: P04: Specificity used to bypass white default.

white woman, who appeared to be posing for an advertisement" (Figure 3). The participant felt that the model "prioritized beauty over reality". Similarly, P07 tried to represent a cafe owner, but they noted that the model tended to generate images that looked promotional, "with a lot of smiling and pretty faces", rather than reflecting the reality of cafe owners, who are "often tired, not smiling" and might not look as the person depicted in Figure 4.

Parallel to this, a few participants encountered a layer of hyper-sexualization. P05 tried to depict a Maori woman and noticed that the word "Maori" led to some images with naked shoulders and outfits that barely covered the body, as shown in Figure 5. P08 prompted the system to generate different people kissing and observed that the model defaulted to nudity and "magazine-style" bodies for gay men (Figure 6a), in direct contrast to the results of "straight people kissing", which P08 observed were more "traditional and clothed". This comparison is shown in Figure 6.



Figure 4: Posed cafe owner (P07)



Figure 5: Hyper-sexualized Maori woman (P05)

P09 also tried to generate a depiction of "a gay man" and confirmed a similar finding as P09: they observed that the model defaulted to nakedness and muscularity as main defining characteristics, as shown in Figure 7a. When prompting for non-muscular and diverse body types, the model found a way to increase the nudity of the images (Figure 7). This suggests the association between queer identity and hyper-sexualization is strong enough to override direct user commands.



(a) *Prompt: "Two gay males kissing".*



(b) *Prompt: "Two straight people kissing"*

Figure 6: P08 comparison: Hyper-sexualization of LGBTQ+ affection versus heteronormative affection.

White and Western Default The experience of P04 when trying to represent a female AI researcher suggested that the model has a strong bias to white people and western standards for professional roles. Despite using neutral or constraining prompts, the model consistently generated images of white Caucasian women with glasses for the role of an "AI research expert". P06, who utilized Chinese prompts to generate specific Chinese characters, mentioned how the model disregarded the context and generated the hair of the Chinese men in a westernized style (Figure 8). P07 observed that the neutral term "cafe owner" consistently produced images of a white person. Similarly, P09 noted a similar bias toward whiteness, when prompting for a "gay man".

Secondary Bias during mitigation A significant finding of this research was that when participants tried to circumvent the biases of the model, they sometimes encountered a second layer of stereotyping. P04 tried to steer away from the identified white default observed in previous generations, by retaining their previous prompt structure, but specifying "Japanese" in the prompt:

"A . . . female AI research expert, very focused typing at her laptop in a corporate open space, wearing casual clothes, working environment, office, messy desk"

Interestingly, the model shifted the environmental context of the scene. Unlike the previous bright office settings generated with prompt in which no nationality was specified, the model now generated an image where it appeared to be dark outside (Figure 9b). P04 correlated this to a secondary stereotype of the Japanese overworking culture.

Similarly, P07, probed the system to generate the same scene of a cafe owner but specifying different characteristics of the depicted subject, starting from the following prompt:

" . . . woman in her thirties serving an espresso behind the counter of an Italian 19th century café in Turin. She's the owner of the bar. She's wearing glasses, baseball hat and a hoodie. She looks quite tired from her long shift."

and varying the prompt by specifying "blonde", "Latin American" and "Gambian" as characteristics of the person to depict. The images of "blonde" white women were often depicted as smiley, polished and posing even though the prompt specified for them to be tired. Images of the "Gambian" and "Latin American" women were depicted unsmiling and more focused on the work, shown in Figure 10. P07 suggested that the model shifted from a glamorized bias for white people, to a working-class or labor-focused bias for marginalized identities.



(a) *Prompt: "A gay man. Twenty five years old. City life-style."*



(b) *Prompt: "A gay man. Consider the whole world. Twenty five / thirty years old. Consider other body types."*



(c) *Prompt: "A gay man not from europe. Consider different etnicities[a] and different body types according to world population."*



(d) *Prompt: "A gay man fully clothed without a healthy body. From ecuador. Over thirty-five years old."*

[a]This is a typo made by the user.

Figure 7: P09 Iteration process where the model persistently generated hyper-sexualized or nude figures despite specific prompts for "non-muscular" and diverse body types.

Failure of Identity and Culture The model demonstrated a lack in cultural specificity regarding human identities. For example, P01 mentioned how the girl they generated in Figure 2b looked actually Chinese or Vietnamese instead of Japanese. This indicates that the model lacks the precision to represent specific identities without blending them into a generic result. P06 struggled to get the model to represent "small Chinese baijiu cups", which they wanted to be placed as offerings in front of a tombstone (Figure 8).

Accidental cultural offense P00 attempted to represent a respectful representation of a religious Islamic figure. Since they feared the generation of a caricature, they avoided the explicit use of the term "prophet" and relied on a more descriptive prompt:



Figure 8: Final image obtained by P06 while attempting to represent Chinese men placing baiju cups in front of tombstones.



(a) *Prompt: "A female AI research expert . . . "*



(b) *Prompt: "A Japanese female AI research expert . . . "*

Figure 9: Initial bias and secondary stereotyping encountered by Participant P04 during mitigation.

"a sillohoutte made by the sun of a human figure that is standing on top of a dune. He is holding the regiens of a camel and is looking up. There should be a single person and a camel. The person should not be on top of the camel. He person should have the reins of the camel in his hand."

However, they noticed a glitch in the images, where the head of a camel was generated in the hands of a person (Figure 11). P00 noted that while this may look like a technical error of the model, this image could be interpreted as offensive to religious people, as it resembles a sensitive historical moment (the battle of Karbala).

Mitigation Strategies

Participants adopted labor-intensive strategies to obtain representations they perceived as more satisfactory and less biased. We discuss these techniques as mitigation strategies, including cases in which they failed to overcome system limitations. By revealing the extensive and often invisible labor required to circumvent these biases, we argue that assumptions embedded in AI systems function as barriers to co-creativity.

Negative constraints To manage the boundaries on sensitive or unwanted content, artists utilized negative constraints. They specifically prompted what not to show. This technique was applied either within the text prompt itself



(a) Blonde



(b) Gambian



(c) Latin American

Figure 10: Comparison of different identifiers used by Participant P07 to test representational shifts.



Figure 11: Example of a technical glitch resulting in involuntary symbolic offense (P00).

or through the dedicated negative prompting field provided by the platform. P00 specifically told the model that the face should not be visible. P02 used a negative constraint, to make sure the image of a person with mental health issues would not portray any wounded imagery. P04 specifically prompted "No glasses", "No white person", "No pretty face". However, these negative constraints still failed to remove these features from the generation. P07 instructed the system to avoid depicting a cafe worker smiling, noting that the model's default outputs were unrealistically staged and cheerful. P09 struggled with the fact that the gay man generated by the model was constantly muscular and white. Therefore they tried to constrain these features. Table 2 details the specific constraints applied by participants

Linguistic Workaround P01 noticed the reason why the generated images were anime style, was because of the word "kimono". To circumvent this problem, they started using a different description to get the same end result. By avoiding the trigger word, P01 successfully bypassed the bias of the model and achieved a realistic representation (Figure 2b).

Self-censorship P00 avoided sensitive religious or cultural scenarios, because they were afraid the model would make them look like cartoons or be disrespectful. Therefore, P00 pivoted to a safer topic. This suggests that the unexpected behaviors of the model limits the creative process of the artist even before generating an image.

Table 3: Specific Negative Constraints Applied by Participants.

Participant	Specific Constraint
P00	"Face shouldn't be visible", "The person should not be on top of the camel"
P02	"No wounds", "Not outside"
P04	"No pretty face", "No white person", "No glasses"
P07	"Not smiling", "No apron"
P09	"Does not go to the gym", "Not from Europe", "Not White", "Non muscular", "Without a healthy body"

Prompt specification P03, who tried to depict Southern Italian landscapes, mentioned that they noticed in prior experience that such landscapes are often stereotyped and depicted as Tuscany (a central Italian region). Anticipating this, they utilized a highly descriptive prompt, which successfully avoided the stereotype, contrasting with a simpler prompt that triggered it (Figure 12). P05, P06 and P08 also found that being more descriptive was necessary to achieve more accurate representations. In addition, to try to work around the white default, both P04 and P07 added different specific identifiers to their prompts. Also P09 had a similar experience, where the model defaulted to a white, muscular gay man. While trying to depict a more diverse man, P09 had to become heavily descriptive in their prompts to overcome the bare-skin and muscular defaults (Figure 13).



Prompt (a): "A young boy from south Italy walks around his small town with his friends."



Prompt (b): "A south Italy landscape, summer, late afternoon. A small italian town's countryside. The sun is about to set. Wind farms are all around the place. Few houses are seen in the back. At the horizon, the mountains stand still. It's very humid and hot."

Figure 12: P03: Prompt specificity for regional accuracy.

Trial and error Almost all artists used the method of trial and error. They would prompt a sentence, see the result, and then add words or delete words. P09 exemplified this struggle by attempting to force the model to generate an image of a clothed gay man. After the initial image was generated, P09 sought a more diverse representation. However, these



Prompt (a): "A gay man. Twenty five / thirty years old. City life."



Prompt (b): "A feminine gay man fully clothed in winter from Japan, in the mountains. He does not have a muscular body. Over thirty-five years old."

Figure 13: P09: Specificity to overcome hyper-sexualization.

attempts resulted in shirtless, muscular images. P09 first used the prompt "A gay man . . . walking down the street", assuming that an outdoor setting would imply clothes. When this failed, P09 became more direct and said " A gay man fully clothed . . . ". When this failed again, they eventually added to the prompt that the man was in the mountains, which finally resulted in a non-sexualized result (Figure 13). This interaction highlights that the T2I process is not just about typing a prompt, but a long process of trial and error.

Discussion

This study highlights how representational biases shape the co-creative interaction between artists and T2I models. Because participants were explicitly asked to explore identities and situations that might be affected by stereotyping or censorship, their interactions with the system often revolved around identifying and circumventing these biases. In this section, we highlight broader implications of our findings, while also discussing the limitations of our approach.

Opacity and Bias as Problematic Friction

Participants repeatedly identified representational biases in T2I outputs, including stereotyping, hyper-sexualization, commercialized aesthetics, and white or Western defaults. To navigate these issues, they employed strategies such as negative constraints, linguistic workarounds, increased prompt specificity, and extensive trial-and-error. Participant P04 described this as similar to "correcting a kid", emphasizing that the model required continuous guidance to align with the user's intentions (Qadri et al. 2024). In this sense, prompting becomes a co-creative dialogue in which human creativity, model limitations, and cultural assumptions intersect. A notable finding is that attempts to mitigate one bias sometimes produced secondary forms of stereotyping (Figures 9 and 4). These examples suggest that representational biases in T2I systems are intersectional (Sufian et al. 2025; Doh et al. 2026). Efforts to correct one bias may therefore

activate other latent stereotypes, highlighting the limitations of prompt-based mitigation strategies and the need for internal technical corrections (Solaiman et al. 2023).

This problem is exacerbated by the opacity of T2I models. The millions of parameters and complex statistical associations that generate outputs are largely inaccessible to users (Buhrmester, Münch, and Arens 2021; Wang et al. 2019), leaving them unable to fully anticipate how the model will interpret prompts. To make sense of the system, users develop informal sense-making strategies akin to "algorithmic imaginaries" or folk theories (Bucher 2012; DeVito 2021). In practice, this means that artists must experiment, iterate, and speculate extensively to achieve outputs aligned with their intentions, a labor-intensive process driven by the model's inherent opacity.

These biases and opaque mechanisms produce outputs that diverge from participants' intentions, creating what prior work frames as AI–human friction (DeSchryver, Henriksen, and Leahy 2025). However, differently from existing literature, that also highlights how users can take advantage of these systems to mitigate existing disparities, for example in gender representations (Shihadeh and Ackerman 2023), our findings indicate that this friction is not inherently productive. Instead, it often represents a constraint that requires significant labor to navigate and can result in misrepresentation, cultural offense, or reinforcement of secondary biases. While occasional outputs subverted expectations in constructive ways (such as P07 seeing a woman as the default cafe owner, or P08 not expecting to be able to represent gay men kissing), these moments were rare. By focusing on the human experience of these failures, our work emphasizes that bias in T2I systems is a structural problem: it cannot be fully addressed through prompt engineering alone, as also demonstrated by priori work (Fraser, Kiritchenko, and Nejadgholi 2023), and poses concrete risks for cultural harm and misrepresentation.

Self-Censorship and Algorithmic Power

In addition, our study revealed that artists' creative processes are influenced not only by the outputs of T2I models, but also by the anticipation of system constraints. For instance, P00 avoided explicitly referencing a sensitive religious figure in their prompt due to concerns that the model might produce disrespectful or caricatured representations, even without such instructions. Instead, they relied on indirect descriptions to reduce the risk of offensive results. This anticipatory behavior illustrates how self-censorship emerges in response to the perceived power of the algorithm, which can act unpredictably or "disobey" human intentions (Ihde 1990). The influence of such constraints extends *before* the generation process begins, shaping how artists conceptualize and approach their work with Generative AI.

This phenomenon can be understood through the lens of algorithmic power, a concept developed in the literature on social media and content moderation (Jiang et al. 2023; Petre, Duffy, and Hund 2019; Cotter 2023). Just as moderation algorithms regulate visibility and community compliance, the biases and safety filters embedded in T2I models can exercise power over creative outcomes, limiting what can be safely or successfully represented. Prior work highlights how users adapt to opaque algorithmic systems by predicting system behavior (Bucher 2012; DeVito 2021), yet these adaptations often require extra labor and can restrict expressive freedom as shown in our study. In artistic contexts, such self-censorship risks flattening the richness of cultural expression (Riccio et al. 2022; Taylor et al. 2025).

Taken together, our findings suggest that anticipatory self-censorship is not merely a personal creative choice, but a response to power asymmetries embedded in the design and behavior of AI systems (Riccio, Hofmann, and Oliver 2024).

Limitations

This study is not without limitations. First, although the data collection reached thematic saturation (Bernard, Wutich, and Ryan 2016), the study involved a relatively small sample of participants. While the qualitative approach provides in-depth insights into artists' interactions with T2I systems, the findings may not be fully generalizable to all creative users. Second, the experiments were conducted using Stable Diffusion 3 through a specific third-party platform. As a result, the observed behaviors may partly reflect the characteristics of this particular model and its implementation. Different generative models or platforms may exhibit distinct representational biases, moderation mechanisms, or safety filters (Riccio, Curto, and Oliver 2024).

Conclusion

This paper examined how artists navigate representational biases and censorship in T2I models when attempting to depict marginalized identities. Through a qualitative user study with 10 creative practitioners, we explored how artists identify, interpret, and attempt to circumvent biases embedded in generative systems. Our findings show that interactions with T2I models are shaped by a combination of representational biases and algorithmic opacity. Participants frequently encountered outputs characterized by white or Western defaults, commercialized aesthetics, and the hypersexualization or homogenization of marginalized identities. Attempts to correct these biases through prompt engineering often required extensive trial-and-error and, in some cases, triggered secondary forms of stereotyping, highlighting the intersectional nature of representational biases within the model's internal associations. These dynamics place a significant burden on artists, who must engage in labor-intensive strategies such as negative constraints, linguistic workarounds, and increasingly specific prompts in order to achieve meaningful representations. Moreover, the unpredictability of generative systems can shape creative processes even before image generation occurs, as artists may anticipate misrepresentation or censorship and adapt their prompts accordingly. In this sense, interaction with T2I models reflects broader dynamics of algorithmic power, where opaque systems influence not only outputs but also the conceptualization of creative work. Taken together, our findings suggest that bias in T2I systems should not be understood solely as a technical limitation or a source of creative friction, but as a structural issue with implications for

cultural representation and artistic agency. Designing more inclusive generative systems therefore requires moving beyond user-side prompt workarounds toward deeper technical interventions.

Author Contributions

Mijntje de Groote conducted the primary research and wrote the first draft of the manuscript. **Piera Riccio** provided ongoing guidance and supervision for the research project and actively participated in the writing, reviewing, and editing of the text.

References

Barve, S.; Mao, A.; Shi, J. M.; Juneja, P.; and Saha, K. 2025. Can we debias social stereotypes in ai-generated images? examining text-to-image outputs and user perceptions. *arXiv preprint arXiv:2505.20692*.

Benjamin, R. 2023. Race after technology. In *Social Theory Re-Wired*. UK: Routledge. 405–415.

Bernard, H. R.; Wutich, A.; and Ryan, G. W. 2016. *Analyzing qualitative data: Systematic approaches*. SAGE publications.

Bucher, T. 2012. Want to be on the top? algorithmic power and the threat of invisibility on facebook. *New Media & Society* 14(7):1164–1180.

Buhrmester, V.; Münch, D.; and Arens, M. 2021. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction* 3(4):966–989.

Chatterji, A.; Cunningham, T.; Deming, D. J.; Hitzig, Z.; Ong, C.; Shan, C. Y.; and Wadman, K. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.

Chen, V.; Liao, Q. V.; Wortman Vaughan, J.; and Bansal, G. 2023. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7(CSCW2):1–32.

Cho, J.; Zala, A.; and Bansal, M. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3043–3054.

Collyer-Hoar, G.; Rubegni, E.; Tomczyk, B.; Baines, A.; and Gruia, L. 2025. " suits as masculine and flowers as feminine": Investigating gender expression in ai-generated imagery. In *Proceedings of the 2025 ACM designing interactive systems conference*, 915–928.

Cotter, K. 2023. "shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society* 26(6):1226–1243.

de Almeida, F., and Rafael, S. 2024. Bias by default.: Neocolonial visual vocabularies in ai image generating design practices. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery.

Dehdashtian, S.; Sreekumar, G.; and Boddeti, V. N. 2025. Oasis uncovers: High-quality t2i models, same old stereotypes. *arXiv preprint arXiv:2501.00962*.

DeSchryver, M.; Henriksen, D.; and Leahy, S. 2025. From friction to synergy: The complex interplay of human creativity and ai. *Possibility Studies & Society* 27538699251350483.

DeVito, M. A. 2021. Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–38.

D'Incà, M.; Peruzzo, E.; Mancini, M.; Xu, D.; Goel, V.; Xu, X.; Wang, Z.; Shi, H.; and Sebe, N. 2024. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12225–12235.

Doh, M.; Gulati, A.; Canali, C.; and Oliver, N. 2026. Aesthetics as structural harm: Algorithmic lookism across text-to-image generation and classification. *arXiv preprint arXiv:2601.11651*.

Epstein, Z.; Hertzmann, A.; of Human Creativity, I.; Akten, M.; Farid, H.; Fjeld, J.; Frank, M. R.; Groh, M.; Herman, L.; Leach, N.; et al. 2023. Art and the science of generative ai. *Science* 380(6650):1110–1111.

Fan, X.; Xiao, Q.; Zhou, X.; Pei, J.; Sap, M.; Lu, Z.; and Shen, H. 2025. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19.

Fang, C.; Zhu, Y.; Fang, L.; Long, Y.; Lin, H.; Cong, Y.; and Wang, S. J. 2025. Generative ai-enhanced human-ai collaborative conceptual design: A systematic literature review. *Design Studies* 97:101300.

Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2023. Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. In *ICCC*, 288–292.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2426–2436.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5111–5120.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gorska, A. M., and Jemielniak, D. 2023. The invisible women: uncovering gender bias in ai-generated images of professionals. *Feminist Media Studies* 23(8):4370–4375.

Ihde, D. 1990. *Technology and the lifeworld: From garden to earth*, volume 560. Indiana University Press.

Jääskeläinen, P., and Åsberg, C. 2024. What's the look of "negative gender" and "max ethnicity" in ai-generated images? a critical visual analysis of the intersectional politics of portrayal. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery.

Jiang, J. A.; Nie, P.; Brubaker, J. R.; and Fiesler, C. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction* 30(1):1–34.

Kirk, H.; Bean, A.; Vidgen, B.; Rottger, P.; and Hale, S. 2023. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Ko, H.-K.; Park, G.; Jeon, H.; Jo, J.; Kim, J.; and Seo, J. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, 919–933.

Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36:56338–56351.

Mandal, A.; Leavy, S.; and Little, S. 2023. Multimodal composite association score: Measuring gender bias in generative multimodal models. *arXiv preprint arXiv:2304.13855*.

Naik, R., and Nushi, B. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 786–808. New York, NY, USA: Association for Computing Machinery.

Nielsen, J.; Clemmensen, T.; and Yssing, C. 2002. Getting access to what goes on in people's heads? reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction*, 101–110.

Petre, C.; Duffy, B. E.; and Hund, E. 2019. "gaming the system": Platform paternalism and the politics of algorithmic visibility. *Social Media+ Society* 5(4):2056305119879995.

Qadri, R.; Mirowski, P.; Gabriellan, A.; Mehr, F.; Gupta, H.; Karimi, P.; and Denton, R. 2024. Dialogue with the machine and dialogue with the art world: evaluating generative ai for culturally-situated creativity. *arXiv preprint arXiv:2412.14077*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.

Rezwana, J., and Maher, M. L. 2023. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction* 30(5):1–28.

Riccio, P.; Oliver, J. L.; Escolano, F.; and Oliver, N. 2022. Algorithmic censorship of art: A proposed research agenda. In *ICCC*, 359–363.

Riccio, P.; Curto, G.; and Oliver, N. 2024. Exploring the boundaries of content moderation in text-to-image generation. In *European Conference on Computer Vision*, 161–178. Springer.

Riccio, P.; Hofmann, T.; and Oliver, N. 2024. Exposed or erased: algorithmic censorship of nudity in art. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.

Riccio, P. 2025. *Human aesthetics under the representational power of Artificial Intelligence*. Ph.D. Dissertation, Universitat d'Alacant/Universidad de Alicante.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22522–22531.

Shihadeh, J., and Ackerman, M. 2023. Shattering bias: A path to bridging the gender divide with creative machines. In *ICCC*, 278–282.

Singh, S.; Hindriks, K.; Heylen, D.; and Baraka, K. 2025. A systematic review of human-ai co-creativity. *arXiv preprint arXiv:2506.21333*.

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Chen, C.; Daumé III, H.; Dodge, J.; Duan, I.; et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

Sufian, A.; Distante, C.; Leo, M.; and Salam, H. 2025. T2ibias: Uncovering societal bias encoded in the latent space of text-to-image generative models. In *Interdisciplinary Workshop on Responsible AI for Value Creation*, 57–71. Springer.

Sun, L.; Wei, M.; Sun, Y.; Suh, Y. J.; Shen, L.; and Yang, S. 2024. Smiling women pitching down: auditing representational and presentational gender biases in image-generative ai. *Journal of Computer-Mediated Communication* 29(1):zmad045.

Taylor, J.; Mire, J.; Spektor, F.; DeVrio, A.; Sap, M.; Zhu, H.; and Fox, S. E. 2025. Un-straightening generative ai: How queer artists surface and challenge model normativity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 951–963.

Turchi, T.; Carta, S.; Ambrosini, L.; and Malizia, A. 2023. Human-ai co-creation: evaluating the impact of large-scale text-to-image generative models on the creative process. In *International symposium on end user development*, 35–51. Springer.

Wang, D.; Yang, Q.; Abdul, A.; and Lim, B. Y. 2019. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15. Glasgow, UK: ACM.

Wang, J.; Liu, X. G.; Di, Z.; Liu, Y.; and Wang, X. E. 2023. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *arXiv preprint arXiv:2306.00905*.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2024. Stable diffusion exposed: Gender bias from prompt to image. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, volume 7, 1648–1659.

Digital Weight: Towards an Ecological Evaluation Layer for Generative AI in Computational Creativity

Jekaterina Kaizere
Independent Artist-Researcher
Idar-Oberstein, Germany
katerina@kaizere.de

Abstract

Generative AI systems are increasingly positioned as tools of computational creativity, yet the material conditions enabling them rarely enter the evidential record of computational creativity evaluation. Existing evaluation discourse typically asks what systems generate (novelty, value, quality), how they generate it (process characteristics, autonomy, interaction), and how outputs are received. Under contemporary generative practice, compute regime, generation count, inference cost, emissions, and water use condition how such claims can be compared, justified, and audited. Infrastructural Situatedness is proposed here as a cross-cutting conditioning layer within computational creativity evaluation. Methodologically, when systems make comparable claims about novelty, value, autonomy, or creative search while relying on materially different infrastructures, those differences should be visible in the evaluation record. Two precedents matter here: computational creativity evaluation already debates what evidence counts for creativity claims and how systems can be compared systematically; broader AI evaluation argues that efficiency and the "price tags" of model development and use should be reported (Strubell et al. 2019; Schwartz et al. 2020). A case demonstration, Digital Weight, operationalises per-generation accounting by translating published per-inference energy coefficients into time-based perceptual structure. The case makes infrastructural constraints part of generative logic and available to evaluation. The contribution is a compact evaluation record for infrastructural conditions in computational creativity.

Introduction

Computational creativity has developed a mature evaluation culture (Jordanous 2012; Colton 2008): debates about machine creativity are typically grounded in explicit criteria, defensible evidence, and domain-relative standards, moving beyond informal "it seems creative" claims. Evaluation frameworks and tutorials formalise how creativity claims can be substantiated across artefact properties, generative processes, and audience reception (Lamb, Brown and Clarke 2018). For example, creativity is often operationalised through novelty, surprise, and value, following Margaret A. Boden's characterisation of creative ideas or artefacts as "new, surprising, and valuable" (Boden 2004). In CC evaluation, these dimensions have also been treated as tractable criteria for assessing design creativity (Grace et al. 2015).

At the same time, many contemporary generative systems used in creative practice rely on large-scale computational infrastructures. Unlike earlier eras of symbolic creativity software, the resource intensity of a creative system can now be estimated at the level of repeated use: deployment cost can be benchmarked in energy and carbon per 1,000 inferences, and task modality matters strongly. In a large comparative study of inference costs, Luccioni, Jernite and Strubell (2024) report that multimodal tasks (including image captioning and image generation) sit at the high end of the measured spectrum (0.06–2.9 kWh per 1,000 inferences), and their task-level summary table reports image generation with a mean of 2.907 kWh per 1,000 queries – orders of magnitude higher than many text classification settings.

This paper argues that infrastructural cost is rarely treated as part of computational creativity evaluation. Existing frameworks can accommodate process-level considerations, yet they do not typically require the material conditions of producing creative artefacts to be disclosed, contextualised, or compared. This omission was historically understandable when computation could be treated, for many evaluative purposes, as a relatively neutral substrate. Contemporary generative systems make this assumption less stable. Two systems may produce outputs with similar claims to novelty and value, or describe similar processes of exploration, autonomy, and interaction, while operating through radically different compute regimes. Without infrastructural disclosure, the evaluation record cannot distinguish constrained search from resource-amplified sampling, nor can it show when a creativity claim depends on extensive repeated generation, discarded candidates, or large hosted models.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

This paper proposes Infrastructural Situatedness as a cross-cutting conditioning layer within computational creativity evaluation. It leaves creativity judgements to existing novelty, value, process, autonomy, and reception-based criteria. Methodologically, it affects how these criteria are read when systems rely on materially different computational infrastructures. At minimum, the layer records compute regime, estimation method, and, where available, electricity, emissions, and water estimates, so that artefact-, process-, and reception-oriented evaluations can be read under known infrastructural assumptions. The aim is auditability under uncertainty. As a case demonstration, Digital Weight translates published per-inference coefficients into perceptible temporal structure: cadence, duration, and accumulation become part of the work's generative constraints. The paper contributes a compact evaluation record for making creativity claims more comparable, justified, and auditable under contemporary generative AI conditions.

Background on evaluation in computational creativity

Computational creativity has developed a mature evaluation culture in which creativity claims are expected to be supported by explicit criteria and defensible evidence. Carolyn Lamb, Daniel G. Brown, and Charles L. A. Clarke provide an interdisciplinary evaluation tutorial that surveys diverse evaluation methods and emphasises the need for explicit evaluation choices and assumptions because there is no single consensus method for evaluating a creative system (Lamb, Brown and Clarke 2018).

Three evaluation anchors position Infrastructural Situatedness as contextual evidence within existing evaluation methodology.

First, Anna Jordanous proposes the SPECS procedure for evaluating creative systems: researchers specify what creativity means for a system, derive evaluation criteria, and apply domain-sensitive tests (Jordanous 2012). Jordanous also emphasises that comparative and systematic evaluation is essential for progress in computational creativity research (Jordanous 2012).

Second, Graeme Ritchie develops empirical criteria for attributing creativity to computer programs based on observable relationships between inputs and outputs. Importantly, he notes that if two systems produce similarly rated outputs, the system generating them in a relatively simpler manner might be considered more creative, while also recognising the difficulty of incorporating such "effort" information into evaluation (Ritchie 2007).

Third, Simon Colton introduces the "creative tripod", describing creative systems in terms of skill, appreciation, and imagination, and distinguishes creativity from the perception of creativity by observers (Colton 2008). For this argument, Colton's distinction matters because it supports the comparison and justification of creativity claims without asserting a single universal metric.

Together, these accounts make the proposed layer methodologically continuous with existing CC evaluation practice: if CC evaluation culture already requires explicit evidence and has precedent for considering "simplicity/effort" as part of comparative judgement, then resource intensity can be treated as contextual evidence of infrastructural effort and search scale in contemporary generative systems.

FACE/IDEA and related descriptive models make clear that CC evaluation already extends beyond final artefacts. Colton, Charnley and Pease (2011) describe creative acts through tuples of generative acts involving expressions, concepts, aesthetic measures, and framing information; IDEA adds impact and audience reception. Lamb, Brown and Clarke (2018) similarly organise evaluation across product, process, person/producer, and press. FACE/IDEA help locate Infrastructural Situatedness inside the evaluative record: what was generated, by what process, under what constraints, with what framing, and within what reception conditions.

Wiggins' Creative Systems Framework, developed through a formalisation of Boden's account, further supports comparison between creative systems in terms of conceptual spaces, rules, search procedures, and constraints (Wiggins 2006). Ventura's discussion of Mere Generation clarifies why generation alone is insufficient for creativity claims: novelty, value, and intentionality require an evaluative context, especially when outputs can be produced by very different processes (Ventura 2016). Mondol and Brown's Algorithmic Information Theory account is relevant here: it connects creativity evaluation with computational history, effort, complexity, novelty, and value through concepts such as logical depth and sophistication (Mondol and Brown 2021a; Mondol and Brown 2021b). Infrastructural Situatedness gives this concern with process and effort a practical evaluation record: it makes compute regime, generation count, resource intensity, and transparency available as contextual evidence for evaluating and comparing creative search. Energy, carbon, and water use are not treated here as creativity criteria. They become part of the evidential context through which materially different generative systems can be compared, justified, and audited.

Why infrastructure matters now

The infrastructural case is grounded in measurable changes in scale, deployment, and resource demand.

At the macro level, International Energy Agency estimates that electricity consumption from data centres is around 415 TWh, or about 1.5% of global electricity consumption in 2024 (IEA 2025), with substantial uncertainty and scenario-based projections for growth. For computational creativity, the issue is comparative: when creative systems rely on data centres and accelerated servers, infrastructure is no longer a neutral "implementation detail," especially in aggregate and at scale.

At the micro level, deployment and inference cost can be reported per generation or per 1,000 inferences. Luccioni, Jernite and Strubell (2024) operationalise deployment cost as energy and carbon for 1,000 inferences, showing large differences by modality and task structure. They report that multimodal tasks such as image captioning and image

generation are on the highest end of the spectrum (0.06–2.9 kWh per 1,000 inferences), and their task-level means indicate image generation at ~2.907 kWh per 1,000 queries. They also report extreme variation across tasks (a factor exceeding 1,000 in their analysis), underscoring that "what task you do" can dominate "how big the model is" in energy terms (Luccioni, Jernite and Strubell 2024).

Carbon captures only part of infrastructural impact. Data centre impacts include water consumption for cooling and electricity generation pathways, and these impacts vary strongly by geography, cooling practice, and energy mix. A Massachusetts Institute of Technology explainer cites an often-used order-of-magnitude estimate of about two litres of water per kWh of energy a data centre consumes for cooling (Zewe 2025), illustrating why water can be reported alongside carbon as a separate, assumption-bound metric.

These coefficients are context dependent. David Mytton reports that water intensity for electricity generation varies widely: consumption ranges from 0.00 to 4.4 L/kWh, while withdrawals can be much higher depending on cooling technology and generation mix (Mytton 2021).

Finally, emissions conversions require explicit assumptions about grid carbon intensity. A widely cited 2024 transition review reports global power-sector CO_2 intensity reaching approximately 480 gCO_2/kWh in 2023, offering an example emission factor that can be transparently parameterised (and replaced with regional factors where available) (Ember 2024).

Operationalising Infrastructural Situatedness within CC Evaluation

This section defines Infrastructural Situatedness and its evaluation record for use with output-, process-, and reception-oriented frameworks.

Definition

Infrastructural Situatedness is a cross-cutting conditioning layer that requires explicit disclosure and contextualisation of the infrastructural conditions enabling a generative process. At minimum, this includes the compute regime, the estimation method, and the conversion factors used to express electricity, emissions, and (optionally) water.

The proposal is motivated by two converging precedents:

- Within CC, Ritchie's discussion shows that "cost/effort" has already been considered as a possible differentiator when outputs are similar, even if the field lacks standard ways to weigh it (Ritchie 2007).
- Within broader AI evaluation culture, Green AI argues for making efficiency visible alongside accuracy and for reporting the "price tag" of developing, training, and running models (Schwartz et al. 2020).

How Infrastructural Situatedness Interacts with Creativity Evaluation

Infrastructural Situatedness specifies the evidential context in which creativity claims are compared and audited. The methodological issue is what changes when the evaluation record includes compute regime, generation count, resource intensity, and transparency. The layer matters most where systems make claims about novelty, value, autonomy, quality, or creative search while relying on materially different infrastructures.

Comparative interpretation. CC evaluation is often comparative, even when it avoids a single creativity score. Two systems may produce artefacts judged similarly novel, valuable, or high quality, while reaching them through very different search conditions. One may operate through constrained search, limited local inference, or a carefully specified generative procedure. Another may depend on large-scale hosted generation, repeated prompt variation, extensive sampling, and many discarded candidates. Artefact-level assessment alone misses this difference. Infrastructural Situatedness distinguishes compact or constrained search from resource-amplified sampling. This connects to Ritchie's observation that effort and simplicity may matter when systems produce similarly rated outputs (Ritchie 2007), while keeping resource use separate from creativity ranking.

Process interpretation. Claims about computational creativity often rest on process: autonomy, iteration, selection, evaluation, interaction, and the system's role in shaping its own outputs. For contemporary generative AI workflows, the process also includes generation count, discarded samples, batch size, model choice, runtime, prompt revision, local or hosted compute, and the degree to which these conditions can be inspected. These factors make process evidence more complete. A claim that a system explored a space has a different evidential profile if the exploration involved ten constrained generations, ten thousand unlogged generations, or a hosted model whose sampling and energy profile remain unavailable. Mondol and Brown's Algorithmic Information Theory account is relevant here because it links creativity evaluation to computational history, effort, complexity, novelty, and value through concepts such as logical depth and sophistication (Mondol and Brown 2021a; Mondol and Brown 2021b). Infrastructural Situatedness turns this concern with computational history into a practical evaluation record for generative AI.

Constraint-based creativity. Infrastructural conditions can shape the creative process before evaluation. An explicit energy, generation, or compute budget can function as a constraint on creative search, comparable to other procedural constraints in computational creativity and generative art. A system may be designed to search within a fixed number of generations, optimise under an inference budget, expose accumulation as part of the work, or restrict sampling so that selection, transformation, and judgement occur earlier. Here the infrastructural condition structures the creative space itself. Digital Weight demonstrates this relation at the level of perceptual form: cadence, duration, repetition, and accumulation are organised through per-generation accounting. In Digital Weight, energy, carbon, and water operate as constraints within the work's generative logic.

Auditability of creative search. Infrastructural Situatedness supports auditability under uncertainty. Creative AI

workflows are often nonlinear: prompts branch, candidates are discarded, outputs are recombined, and human judgement intervenes repeatedly. The evaluation record should prioritise inspectability over exact reconstruction. Where possible, authors should count all generated candidates, including discarded samples, unsuccessful generations, and intermediate variations. Branching workflows may distinguish total project generation count from the count associated with the final selected branch. For collaborative workflows, authors may report aggregate project counts, author-attributed generation logs, or clearly bounded estimates when individual attribution is impractical. Acceptable evidence includes API histories, local generation logs, version control, exported prompt sheets, model-setting records, batch files, and manual process notes. The search path becomes inspectable enough for comparative interpretation, without implying complete capture of every micro-decision.

The layer's contribution is epistemic and comparative. Infrastructural Situatedness makes visible which computational conditions support a creativity claim, how much of the creative search remains available for inspection, and where uncertainty remains. It conditions artefact-, process-, and reception-oriented evaluation while keeping creativity distinct from footprint accounting. These methodological implications motivate the need for a standardised evaluation record capable of documenting infrastructural conditions across creative workflows.

Operationalising the extension as an evaluation record

Evaluation frameworks often omit infrastructural conditions because measurement is uneven and assumption-dependent. ML sustainability work treats systematic reporting with transparent assumptions as a practical alternative to full lifecycle assessment in every paper. Henderson et al. (2020) frame energy and carbon reporting as necessary for assessing the climate impacts of machine learning research.

Adapting that logic to computational creativity, Table 1 proposes a standardised evaluation record for documenting the infrastructural conditions under which creative outputs are produced. The table operationalises comparative interpretation and auditability by recording compute regime, generation count, resource intensity, transparency, and uncertainty alongside existing creativity evidence.

Table 1. Infrastructural Situatedness evaluation record for computational creativity systems. The table records process and infrastructural evidence relevant to comparative interpretation, process interpretation, and auditability. It avoids creativity scoring and exact footprint accounting.

Field	What is recorded	Evaluation function	Auditability note
Model / System	Model name, version, API/local system, release date where available	Identifies the system being evaluated	Include provider, version, or local configuration
Task / modality	Text, image, audio, video, multimodal, robotic, live, hybrid	Supports comparison across task types	Specify output type and workflow context
Compute regime	Local device, cloud/API, hosted model, cluster, hybrid	Contextualises process and infrastructure	Report hardware/API access level where known
Creative workflow	One-shot, iterative, branching, collaborative, live, automated	Describes creative search structure	Include workflow description or process log
Generation count	Total generations, selected outputs, discarded candidates	Shows search scale and sampling intensity	Use API logs, local logs, version history, or manual count
Runtime / duration	Runtime, performance duration, batch duration, session length	Supports process interpretation	Report measured or estimated duration
Resource intensity	Energy, emissions, and water where available or estimable	Contextual evidence for infrastructural effort	Report units separately; do not aggregate
Estimation method	Telemetry, direct measurement, provider data, calculator, published coefficients	Makes assumptions inspectable	Prefer standardised tools or provider data where available

Field	What is recorded	Evaluation function	Auditability note
Transparency level	Measured, partially estimated, provider-limited, unknown	Indicates reliability of the evaluation record	State what is known, inferred, or unavailable
Uncertainty / boundary	Range, coefficient source, region, system boundary, exclusions	Prevents false precision	Distinguish operational estimates from lifecycle data

This evaluation record makes infrastructural conditions visible and comparable across works, supporting auditability without collapsing creativity evaluation into footprint accounting. Figure 1 represents this relation visually. Infrastructural Situatedness crosses artefact, process, and reception evaluation and conditions interpretation beyond a creativity score.

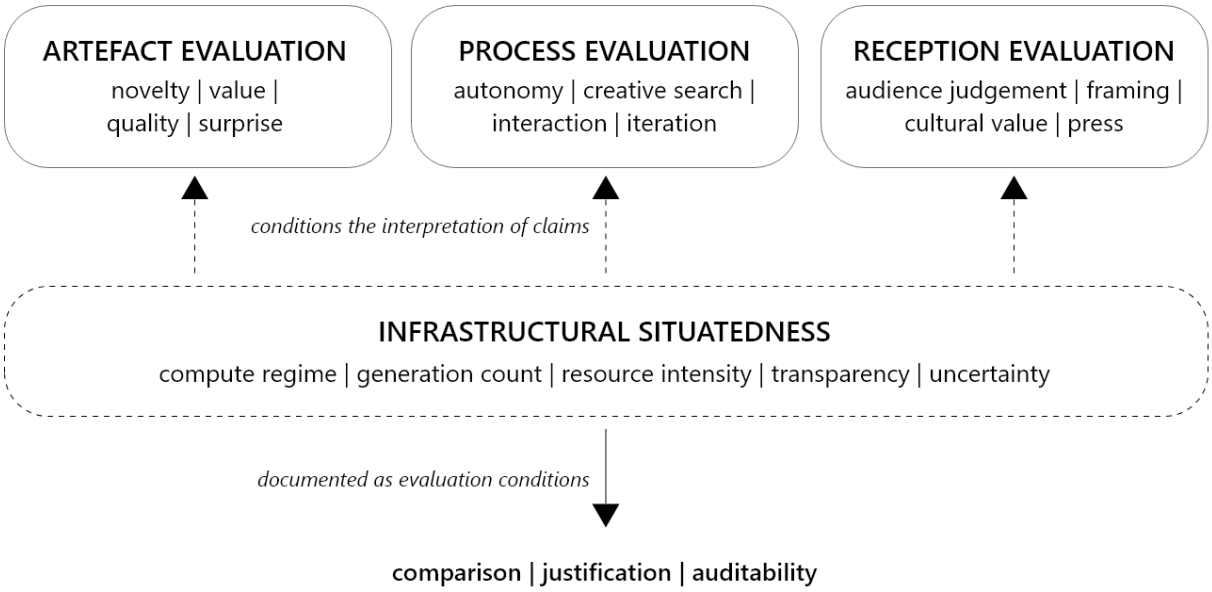


Figure 1. Infrastructural Situatedness as a cross-cutting conditioning layer within computational creativity evaluation. The layer leaves artefact-, process-, and reception-oriented criteria intact and is not an additional creativity score. Instead, it records infrastructural conditions including compute regime, generation count, resource intensity, transparency, uncertainty, and, where available, energy use, emissions, water consumption, and discarded candidates. This enables comparison, justification, and auditability under explicit infrastructural assumptions.

Digital Weight as an operational demonstration

Digital Weight provides a compact methodological demonstration: Infrastructural Situatedness can be operationalised and made perceptible within an artwork's generative logic.

Assume a cadence of 2 images/s for 8 minutes, yielding 960 generative events. Using Luccioni et al.'s task-level mean for image generation (2.907 kWh per 1,000 queries), a back-of-the-envelope estimate gives 0.0029 kWh per generated image and ~2.79 kWh for 960 images (Luccioni, Jernite and Strubell 2024). Using a global-average emissions factor example of ~480 gCO_2/kWh (Ember 2024) (explicitly parameterised), this corresponds to ~1.34 $kgCO_2$. Using a simple cooling proxy of ~2 L/kWh (Zewe 2025) yields ~5.6

L of cooling water. These values are included as illustrative and should be reported as parameterised coefficients.

The inference-cost range also shows scale sensitivity: Luccioni et al. note that multimodal tasks such as image captioning and image generation are on the high end of the spectrum (0.06-2.9 kWh per 1,000 inferences), and their analysis indicates that image generation can be orders of magnitude more energy-intensive than many text tasks.



Figure 2. Two consecutive frames of the artwork (001,002) showing the cumulative infrastructure counters associated with each AI generation cycle. Electricity, CO_2 emissions, and water consumption increase as the generative process continues.

What the case demonstrates for computational creativity evaluation

Digital Weight intentionally complicates output-centric evaluation. The visible artefact is nearly static; if evaluated only as a product, it may appear low in novelty. Yet the process repeatedly performs generative acts, and the infrastructural cost accumulates measurably with each act. The methodological point is that output novelty alone may fail to register the infrastructural intensity of a generative process, so evaluation that ignores load risks mischaracterising what is being optimised or externalised (Boden 2004; Ritchie 2007).

The project also demonstrates how infrastructural metrics can become generative constraints: in constraint mode, the artist can choose cadence, duration, or transformation rules as functions of an explicit energy budget, turning resource intensity into part of the creative search space rather than a hidden side effect. The case keeps the evaluation claim tied to observable, reportable evidence (Jordanous 2012; Lamb, Brown and Clarke 2018).

Discussion and limitations

Why this is a CC evaluation contribution

Infrastructural Situatedness is framed as a conditioning layer within CC evaluation. CC evaluation is pluralistic and domain-relative: frameworks such as SPECS are designed to accommodate diversity by forcing explicit specification of what counts as creativity in context (Jordanous 2012). Infrastructural Situatedness fits that logic: it adds contextual evidence to the evaluation record, regardless of whether a system's creativity claim prioritises novelty, autonomy, cultural reception, or co-creative interaction (Boden 2004; Colton 2008).

Scope, uncertainty, and lifecycle boundaries

The proposal is strongest for contemporary resource-intensive generative AI workflows, especially repeated inference in image, text, audio, and video generation systems. These workflows already rely on measurable quantities such as generation count, runtime, hosted inference, API usage, and batch-scale sampling, making infrastructural conditions easier to document within the evaluation record. Other forms of computational creativity may require adapted reporting fields rather than direct reuse of the protocol presented here. Robotic art, embodied systems, live performance, installation-based work, and hybrid material practices may involve additional infrastructural conditions including sensors, actuators, projection systems, fabrication processes, transport, local energy use, installation logistics, or performance duration. In such cases, Infrastructural Situatedness should be treated as an extensible framework with adaptable reporting fields.

The proposal prioritises auditability under uncertainty. Energy, carbon, and water estimates depend on assumptions about hardware, deployment context, cooling systems, regional energy mix, and system boundaries. Where possible, authors should prefer standardised tools, provider telemetry, API logs, local energy measurements, or community-accepted calculators before relying on approximate published coefficients. The evaluation record should disclose the estimation method, coefficient source, assumptions, uncertainty range, and system boundary used in reporting.

Operational estimates should also be distinguished from full lifecycle assessment. The minimal protocol proposed here primarily concerns operational and inference-stage conditions because these are the most accessible and comparable quantities across contemporary generative workflows. Lifecycle costs associated with hardware manufacture, data-centre construction, material extraction, transport, installation, or disposal may be highly significant, but such data are often unavailable, proprietary, or methodologically inconsistent across systems. Where reliable lifecycle data exist, they may be added to the evaluation record. Their absence should not prevent infrastructural disclosure at the operational level.

Relationship to sustainability discourse inside ICCC

Sustainability has become an explicit topic inside ICCC discourse. For example, Petra Jääskeläinen and Aksel Biørn-Hansen argue that sustainability is often framed as "shallow sustainability" focused on quantification/optimisation and propose a "deep sustainability" perspective that interrogates values and assumptions in CC research (Jääskeläinen and Biørn-Hansen 2024).

Quantification in Infrastructural Situatedness remains deliberately bounded: it records assumptions, limits, and context-bound coefficients.

Complementarity with reflection tools and creative-AI sustainability work

Related work in creative-AI sustainability often targets different intervention points than CC evaluation theory. For example, Yuanyang Ren and colleagues integrate Sustainability Reflection Tools (SRTs) into generative AI interfaces and report that artists value granular, relatable visualisations and may want low-energy settings during exploratory stages (Ren et al. 2023). This addresses interface-level mediation and artist-facing reflection, with CC evaluation frameworks outside its direct scope.

Utz and DiPaola (2023) also argue that diffusion-based AI art systems merit serious climate-impact investigation and highlight lack of public data as a key difficulty. Their work supports the need for transparency while centring climate-impact estimation outside CC evaluation methodology.

The proposed record can support both: it supplies the disclosure substrate that reflection tools can visualise and that climate-impact analyses repeatedly identify as a bottleneck.

Conclusion

This paper proposed Infrastructural Situatedness as a cross-cutting conditioning layer within computational creativity evaluation. Its methodological function is to make infrastructural assumptions behind creativity claims visible in the evaluation record.

The central claim is comparative. When generative systems produce similar artefacts, or make similar claims about creative search, their compute regime, generation count, resource intensity, transparency, and uncertainty affect how those claims should be interpreted. This matters especially for contemporary generative AI workflows, where repeated inference, discarded candidates, hosted models, and opaque infrastructures can shape the path from search to selected artefact.

Digital Weight demonstrates one operational form of this argument. Per-generation accounting is translated into temporal structure, so that cadence, duration, repetition, and accumulation become part of the work's generative logic. The case keeps energy, emissions, and water outside creativity criteria. It shows how infrastructural conditions can enter the evidential context of process evaluation.

Infrastructural Situatedness offers a compact methodological extension to CC evaluation practice: a way to compare, justify, and audit creativity claims under explicit infrastructural assumptions. Where computational scale becomes part of the process, creative search needs an inspectable record.

References

Boden, M. A. 2004. *The creative mind: Myths and mechanisms.* 2nd ed. London: Routledge.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Creative Intelligent Systems: Papers from the 2008 AAAI Spring Symposium*, Technical Report SS-08-03, 14–20.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational Creativity Theory: The FACE and IDEA Descriptive Models. In *Proceedings of the Second International Conference on Computational Creativity*, 90–95.

Ember. 2024. *Global Electricity Review 2024*.

Grace, K.; Maher, M. L.; Fisher, D.; and Brady, K. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3(3-4):125–147.

Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; and Pineau, J. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* 21(248):1–43.

IEA. 2025. *Energy and AI*. International Energy Agency.

Jääskeläinen, P.; and Biørn-Hansen, A. 2024. Critical Questions for Sustainability Research in Computational Creativity. In *Proceedings of the International Conference on Computational Creativity,* 209–214.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys* 51(2), Article 28, 1–34.

Luccioni, A. S.; Jernite, Y.; and Strubell, E. 2024. Power Hungry Processing: Watts Driving the Cost of AI

Deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 85–99.

Mondol, T.; and Brown, D. G. 2021a. Incorporating Algorithmic Information Theory into Fundamental Concepts of Computational Creativity. In *Proceedings of the Twelfth International Conference on Computational Creativity*, 173–181.

Mondol, T.; and Brown, D. G. 2021b. Computational creativity and aesthetics with algorithmic information theory. *Entropy* 23(12):1654.

Mytton, D. 2021. Data centre water consumption. *npj Clean Water* 4(1):11.

Ren, Y.; Sivakumaran, A.; Niemelä, E.; and Jääskeläinen, P. 2023. How to Make AI Artists Feel Guilty in a Good Way? Designing Integrated Sustainability Reflection Tools (SRTs) for Visual Generative AI. In *Proceedings of the International Conference on Computational Creativity.*

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds and Machines 17(1):67–99.*

Schwartz, R.; Dodge, J.; Smith, N. A.; and Etzioni, O. 2020. Green AI. *Communications of the ACM* 63(12):54–63.

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.

Utz, V.; and DiPaola, S. 2023. Climate Implications of Diffusion-based Generative Visual AI Systems and their Mass Adoption. In *Proceedings of the 14th International Conference on Computational Creativity*, 264–272.

Ventura, D. 2016. Mere Generation: Essential Barometer or Dated Concept? In *Proceedings of the Seventh International Conference on Computational Creativity*, 17–24.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Zewe, A. 2025. Explained: Generative AI's Environmental Impact. *MIT News*, January 17, 2025.

Evolving to the Aesthetics of a Vision-Language Model

Stephen James Krol
SensiLab, Monash University
Melbourne, Australia
stephen.krol@monash.edu

Jon McCormack
SensiLab, Monash University
Melbourne, Australia
jon.mccormack@monash.edu

Abstract

Evolutionary systems have demonstrated remarkable results in creative domains, with recent applications in generative typography, design, and music. However, an open problem remains in designing fitness functions that effectively capture the desired aesthetics of abstract outputs. In this work, we explore two methods for evaluating the aesthetics of a population using Vision-Language Models (VLMs). The first method uses CLIP-IQA to predict an aesthetic score for each design. The second method instead pits candidates against each other, with winners determined by a VLM using a custom prompt specified by the user. The outcomes of these pairwise comparisons are then used to estimate a population ranking via the Glicko rating system. We present these methods in the context of a case study using a custom generative system and compare the resulting rankings with an artist's aesthetic ranking and those produced by other aesthetic evaluation techniques. Additionally, we document the artist's experience using these approaches to evolve designs, critically analysing the strengths and weaknesses of both methods.

Introduction

The development of algorithms for Image Aesthetic Quality Assessment (IAQA) has been widely studied in both Artificial Intelligence (AI) and Computational Creativity (CC) (Wang, Chan, and Loy, 2023; Xiong et al., 2024; Lopes, Correia, and Machado, 2023; Goree, 2021). Within this broader context, Evolutionary Computing (EC) has addressed the problem by designing aesthetic fitness functions to guide the evolution of outputs from parametric models (Mccormack, Cruz Gambardella, and Krol, 2023; Unemi, 2013; Machado, Amaro, and others, 2013; Rebelo, Bicker, and Machado, 2020). Approaches include the development of handcrafted aesthetic rules (Lopes, Correia, and Machado, 2023), the use of proxy measures to estimate aesthetic quality (McCormack and Cruz Gambardella, 2022), and, more recently, the application of large pre-trained Deep Learning (DL) models to score artefacts (Gomide, Ferreira, and Meira Jr, 2025; Wang, Chan, and Loy, 2023).

Although the recent rise of diffusion models (Ho, Jain, and Abbeel, 2020; Rombach et al., 2022) has overshadowed more traditional generative systems, parametric models remain relevant in both artistic practice (Latham, Todd, and Banarse, 2025; Lomas, 2016) and design (Snooks, 2021), making their continued study worthwhile in the current technological climate. In this context, automatic IAQA facilitates the identification of high-quality designs within complex systems whose search spaces are often too large to explore effectively by hand, thereby improving the overall usability of these systems.

In this paper, we investigate two IAQA paradigms for EC based on pre-trained (DL) models. The first is a point-wise evaluation approach using the state-of-the-art (SOTA) CLIP-IQA method (Wang, Chan, and Loy, 2023), which generates an aesthetic score from custom antonym prompt pairings. The second employs a pre-trained *Qwen3-VL-8B* Vision-Language Model (VLM) to perform pairwise comparisons between images using a custom prompt; this process is repeated across the population, and the results are aggregated into a global ranking using the Glicko rating system (Glickman, 1995). While prior work has demonstrated that pre-trained DL models can be effective in evaluating the aesthetics of representational images (Wang, Chan, and Loy, 2023; Gomide, Ferreira, and Meira Jr, 2025), comparatively little research has examined their effectiveness for non-representational artefacts (Lomas, 2016). These approaches are therefore evaluated against a custom-built benchmark to assess their performance in ranking non-representational images.

Finally, both methods are applied to a real parametric system in an artist-centred study, in which an artist used them to evolve designs from a custom Harmonograph-inspired software system, providing practical insight into their value within active creative practice. Results indicate that while both the point-wise and pairwise approaches achieved comparable performance on the benchmark, the pairwise VLM method afforded the artist greater control than the point-wise CLIP-IQA approach, albeit at a substantially higher computational cost. Hence, the contributions of this work are as follows:

- A comparison of two IAQA paradigms in evaluating the aesthetics of non-representational imagery.
- An artist-centred study on the use of pre-trained VLM models to evolve the design space of a parametric system.

The code for this project is also available online[1].

[1] https://github.com/SensiLab/aesthetic-evolution

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Related Work

Evaluating the Aesthetics of Images

The field of Image Quality Assessment (IQA) has a multitude of methods designed to evaluate image features such as noise level and image distortion (Wang et al., 2004), with several methods demonstrating strong performance in evaluating general image quality (Wu et al., 2024; Chen et al., 2024a). Additionally, the field has explored methods for IAQA, which aims to estimate more subjective image qualities such as aesthetics (Wang, Chan, and Loy, 2023) or perceived emotion (Parthasarathy, Lotfian, and Busso, 2017). For example, Datta et al. (2006) analysed photographs to develop various visual features to train Machine Learning (ML) models on. Other techniques instead train or leverage large pre-trained deep learning (DL) models to build aesthetic scores (Xiong et al., 2024). Wang, Chan, and Loy (2023) are a notable example, introducing their method CLIP-IQA which utilises a pre-trained CLIP model (Radford et al., 2021) to score various aesthetic attributes of an image using antonym prompt pairings. Other techniques, such as that presented in Xiong et al. (2024), perform additional training using various benchmark datasets (Murray, Marchesotti, and Perronnin, 2012; Fang et al., 2020), to estimate aesthetic scores from human annotated data. However, while these DL methods have demonstrated SOTA performance on various benchmarks, their applications have mostly been on representational images, with their performance on more abstract imagery being untested. Furthermore, the prediction strategy of condensing aesthetics to a score, or set of scores, is at odds with various theories on how humans measure *feel* (Rokeach, 1973; Kahneman, 2013). Yannakakis, Cowie, and Busso (2018) argue for this, providing a multi-disciplinary perspective on why emotions are *ordinal* and highlight the value of preference learning (Burges et al., 2005; Yannakakis, 2009) as a more suitable method for predicting *feel*. Furthermore, recent work (Chen et al., 2024b) has demonstrated that multi-modal VLMs align with human judgements in pairwise comparisons on vision-language benchmarks providing motivation to explore their use in aesthetic ranking. Our work builds on prior research by evaluating and comparing point-wise and pairwise approaches to aesthetic ranking using pre-trained DL models, and by applying both techniques within an artist-centred study.

Aesthetics and Computational Creativity

Within CC, various works have investigated and commented on the concept of aesthetics and its importance in creativity. This has moved beyond just evaluating the quality of outputs to formalising and developing systems that are capable of defining their own aesthetic. Colton (2008) argues that evaluating a system solely by its generated artefacts is inadequate, noting that knowledge of the process impacted one's perception of the artefact's aesthetic (Colton, 2012). Guckelsberger, Salge, and Colton (2017) extend this, highlighting that the to improve perceived creativity, systems must also be able to explain *why* they acted in a particular manner and thus should have internal goals and motivations that are not directly tied to the system's designer. Bodily and Ventura (2018) further argue that "creativity is an inherently social construct" which requires those participating to communicate their intention and describe their developed aesthetic.

Compared to these works, this paper does not design a system capable of developing its own aesthetic, but instead attempts to leverage learned aesthetics from pre-trained VLMs. This positions the system as a tool that can assist users in exploring a generative system's search space.

Evolving to Aesthetics

Work on evolving for aesthetic outcomes takes many forms, but a common strategy has been to encode aesthetics explicitly in the fitness function. Earlier systems often relied on hand-crafted measures or rules to operationalize aesthetic criteria (Unemi, 2013; Machado, Amaro, and others, 2013). For example, Rebelo, Bicker, and Machado (2020) evolve a custom typographic system using three fitness functions targeting legibility, aesthetic quality, and semantic coherence, with their aesthetic component computed as the arithmetic mean of five bespoke attributes designed for document-level aesthetic analysis.

Other work has instead searched for automatic proxies that could be used to estimate aesthetics. For example, McCormack and Cruz Gambardella (2022) tested multiple complexity measures on evolutionary art datasets and found that correlations with aesthetic judgement vary by dataset and by measure, i.e., there is no universally "best" complexity metric for aesthetics; instead, usefulness depends on the system and the context of judgement. Here the authors framed the problem as less about discovering a universal aesthetic fitness and more about choosing (or learning) proxies that are locally valid for a given artist, medium, or generative process.

With modern VLMs, this idea can be extended further: aesthetic fitness functions can be instantiated dynamically through prompting (Wang, Chan, and Loy, 2023), enabling artists to guide evolutionary search with natural-language descriptions of their preferences. In their art installation, Latham, Todd, and Banarse (2025) demonstrated the potential for this by "fully [handing] over aesthetic and content decisions" to Google's Gemini and highlighting how its use as a "selector" resulted in the evolution of various forms using a custom generative system. In our work, we extend this paradigm by studying an artist's use of prompt-conditioned aesthetic fitness to generate non-representational visual art, moving beyond the exhibition context to evaluate the approach's practical effectiveness.

Ranking Systems

Point-Wise Ranking via CLIP-IQA

CLIP-IQA (Wang, Chan, and Loy, 2023) is method for IQA that utilises a pre-trained CLIP model (Radford et al., 2021) — a VLM trained to align images and text in a shared embedding space — to score images on various dimensions of *feel*. To achieve this, simple antonym prompt pairings are designed to capture the different extremes of a concept.

For example, *Complex & Simple*, *Good Picture & Bad Picture* and *Happy & Sad*. Both prompts and their respective image are then passed through the CLIP model to retrieve $\mathbf{x} \in R^c, \mathbf{t}_1 \in R^c$ and $t_2 \in R^c$ which are the vector features of the image, positive prompt and negative prompt. The cosine similarity between each prompt vector and the image vector is then calculated as below:

$$s_i = \frac{x \cdot t_i}{|x||t_i|}, i \in \{1, 2\} \tag{1}$$

As CLIP is trained so that similar concepts result in vectors with similar direction, the cosine similarity function above measures the semantic relationship between the prompts and the image. These scores are then inputted into the Softmax function to produce the final score $\bar{s} \in [0, 1]$:

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}} \tag{2}$$

This softmax technique is used to mitigate the ambiguity issue associated with using CLIP models for scoring. As argued by the authors, a *rich* image could be one that depicts wealth or is rich in content. This antonym approach produces a relative score and as demonstrated in Wang, Chan, and Loy (2023), results in significantly better correlation with human perception compared to single prompt setups.

Pairwise Judging via a VLM

To rank designs by custom aesthetic criteria we utilised an off-the-shelf VLM *Qwen3-VL-8B-Instruct* to act as a judge. The choice of VLM model was arbitrary and in theory any VLM that can accept multiple image inputs and text could be used. The VLM is provided three inputs: image 1 (i_1), image 2 (i_2) and a custom text prompt defining the aesthetic criteria to be evaluated (P). The user is given freedom in how they define the prompt but must ensure that the model is instructed to eventually decide which image better suits the desired aesthetic criteria. An example of a prompt is shown in Figure 1.

While any criteria can be used, its worth discussing two features of the above prompt that helped improve results: (1) Forcing the model to first describe the images and rationalise its decision before choosing an outcome resulted in better performance as demonstrated in Table 1 and, (2) providing the option for a draw prevented the model from entering into continuous reasoning loops when there was no clear winning image.

A limitation of ranking designs using pairwise comparisons is the compute complexity which grows $O(N^2)$. This is particularly evident when working with VLMs which require specialised GPU hardware to enable fast inference. As briefly discussed in the next section, various engineering decisions were made to improve the inference speed of the VLM. However, previous work has demonstrated that global ranking can be estimated by sampling possible pairwise comparisons (Baltrušaitis, Li, and Morency, 2017) suggesting that not all pairs need to be evaluated to estimate a "good enough" ranking.

On this basis, we implement the Glicko system to calculate the ranking of designs based on the outcomes of n pairwise comparisons. The Glicko system (Glickman, 1995) is a rating method for competitive games that estimates a player's skill using both a rating and a rating deviation (RD), which reflects the uncertainty in that estimate. After each set of matches, ratings are updated based on game outcomes and the opponent's rating and RD, allowing the system to adjust more quickly for players with higher uncertainty. The Glicko system has been use in similar contexts (McCormack and Cruz Gambardella, 2022) making it a suitable starting point for this task.

Example Prompt

You will be given two images depicting harmonograph drawings, you must choose which one is more aesthetically pleasing in representing harmonic patterns.
IMPORTANT RULES (must be followed):

- Images that are mostly dark blobs or solid dark regions MUST be ranked lower.
- Visible line structure and repeating patterns are REQUIRED for a high score.
- Messy noise or amorphous shapes should be ranked lower.
- Winning designs must possess the best INTRICATE and CLEAR harmonic patterns.

Task:
Choose which image is more aesthetically pleasing according to the rules above. If you cannot make a decision return a draw value of '3'. Output a one sentence description of each image, a single sentence regarding your reasoning and finally a single digit corresponding to which image is better ONLY '1' or '2' or '3' for draw.

Figure 1: An example prompt used by the VLM

Case Study: Generative Line Drawing

To study the use of these ranking techniques in evaluating non-representational imagery, we present an artist-centred study that applies both methods to rank a set of images generated by a bespoke generative software system developed by the artist. We demonstrate the capabilities of the two techniques through quantitative benchmarking against the artist's preferences, along with an analysis of the artist's experience using the system to create new designs with a generative algorithm (GA).

The artist's generative system, detailed in the next section, was selected for this study for several reasons. Firstly, the system has "ecological validity" (Brunswik, 1956; Schmuckler, 2001; McCormack et al., 2022) in that it has already been extensively used by a professional artist who has exhibited and sold works made by the system in commercial galleries (Figure 2). Secondly, the system is relatively manageable, with only 17 continuous parameters, whose ef-



Figure 2: Example gallery exhibition of outputs from the generative line drawing system used in this paper. The drawings created by the software are converted to stitched cotton embroideries, machine embroided onto cotton fabric.

fects can be explored interactively by the software in real-time. Nonetheless, the search space – a 17-dimension vector space – is complex enough to make it impossible for the human artist to exhaustively search. The drawing system outputs line drawings in SVG format, allowing for 2D rasterization at any resolution, enabling evaluation via image-focused deep learning models. Lastly, the system generates abstract rather than figurative or representational imagery, making it a challenge for many deep-learning systems which are often trained on largely figurative imagery (Willison, 2022), and which are elusive when trying to describe using descriptive language, a requirement of prompt-based models (McCormack et al., 2023).

Custom Generative Art System

Our custom generative drawing software simulates an oscillating line drawing agent, whose movements are controlled by a series of variable oscillators. Drawing inspiration from the *harmonograph* – a 19th century mechanical drawing device based on multiple swinging pendulums – the software version uses a combination of periodic functions and aperiodic stationary noise (Perlin, 2002). This combination of functions allows a much richer variety of drawings over any mechanical harmonograph. To further increase the visual possibilities, the agent's drawing space can be non-linearly distorted, giving rise to additional creative possibilities. The drawing software was implemented in the Processing environment (Fry and Reas, 2014) and uses an interactive interface with the 17 adjustable parameters controlling the agent's behaviour and manipulated using controls on the right (Figure 3).

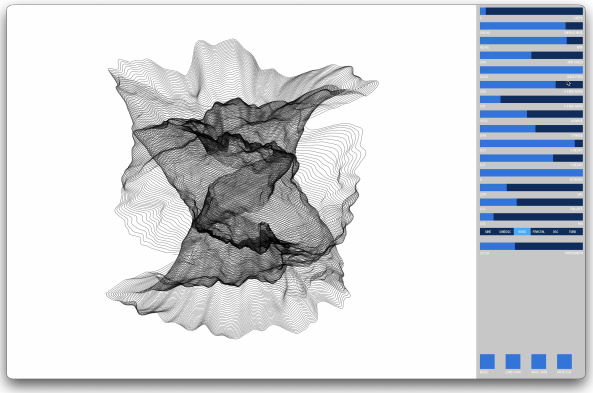


Figure 3: The artist-developed generative drawing system used in this case study. The figure shows current drawing (left) the interface and parameters which can be adjusted in real-time (right).

Genetic Algorithm A simple Evolutionary Algorithm (EA) was developed to evolve designs based on the ranking determined by either the *CLIP-IQA* point-wise evalua-

tion or the *Qwen3-VL-8B-Instruct* pairwise evaluation, with these models determining the fitness score. The EA operated on the parameters of the drawing system (the *genotype*), generating new populations of designs (*phenotypes*) through both pairwise crossover and individual mutation (Mitchell, 1996; Eiben and Smith, 2003). Tournament selection was used to select parents for recombination, with the tournament proportion size $k \in [0, 1]$ available to the artist as an adjustable parameter. The algorithm also allowed the artist to enable elitism, which meant that top ranked individuals from the previous population could compete with the children of the current population. This was controllable through the elitism parameter $e \in [0, 1]$ which controlled the proportion of top ranked individuals to be accepted in the current iteration. Parents selected for recombination had their genes subject to random crossover (p_1, p_2) to produce a new individual (c) using algorithm 1.

Algorithm 1 Crossover algorithm used in EA.

Require: Parent solutions p_1, p_2; mixing parameter $\alpha \in [0, 1]$
Ensure: Offspring solution c

for each variable x_i **do**
  if x_i is continuous **then**
    $c_i \leftarrow \alpha p_{1,i} + (1 - \alpha) p_{2,i}$
  else if x_i is categorical **then**
    Randomly select c_i from $\{p_{1,i}, p_{2,i}\}$
  end if
end for
return c

The parameter $\alpha \in [0, 1]$ was used to weight the arithmetic mean when mixing continuous parameters. The artist was given the option to use a *fixed* α, set manually; a *random* α; or a *biased* α, which biased the mean toward the fitter parent. Once a new population was produced, mutation would be applied to each new individual where each parameter would have a probability m of mutation using algorithm 2. The artist could control mutation through both the probability of a parameter being mutated m and the extent of mutation for continuous variables σ_m. The default evolutionary hyperparmeters set for the artist were $k = m = \sigma_m = e = 0.1$, α set to biased, elitism active with a population size of 26.

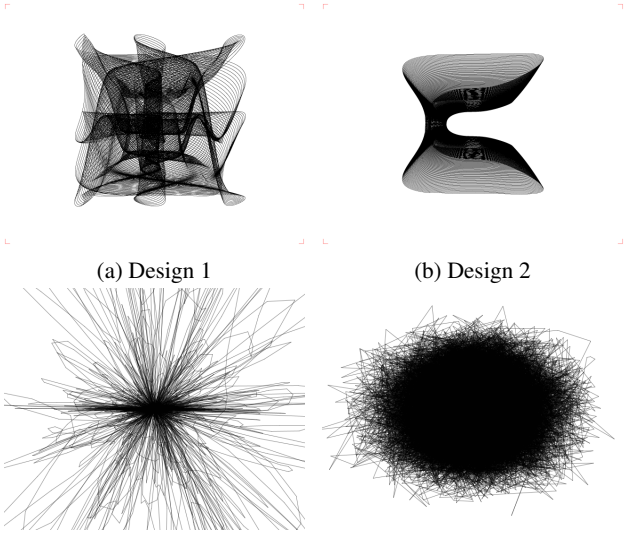
Benchmarking Ranking

To quantitatively measure the performance of both methods in ranking the generated designs, a custom benchmark was built to compare each technique's ranking to that of the artist's. While this does not provide an exhaustive evaluation of each technique's performance in ranking the aesthetics of non-representational images, it does provide insight into their ability to align to an artist's aesthetic style through custom prompts. To construct this ranking, a web interface was built that enabled the artist to compare pairs of images — selecting either a winner or a draw — while the *Glicko* system was used to estimate the ranking of 100 designs. The ranked designs included a mix of randomly generated and curated examples, as purely random generation often produced a large proportion of low-quality designs. In total, the artist performed 1500 comparisons to build the ranking. An example of these designs can be seen in Figure 4.

Algorithm 2 Mutation algorithm used in EA.

Require: Individual c; mutation probability m; Gaussian parameter σ_m
Ensure: Mutated individual c'

$c' \leftarrow c$
for each parameter x_i in c' **do**
  Sample $u \sim \text{Uniform}(0, 1)$
  if $u < m$ **then**
    if x_i is continuous **then**
      $c'_i \leftarrow c'_i + \mathcal{N}(0, \sigma_m^2)$
    else if x_i is categorical **then**
      Randomly select c'_i from valid categories
    end if
  end if
end for
return c'



(a) Design 1 (b) Design 2

(c) Design 3 (d) Design 4

Figure 4: Samples from the 100 images used in the benchmark. Designs 1 and 2 are examples of "quality" designs, defined by their structure and aesthetic. Designs 3 and 4 are examples of "low-quality" designs, defined by their messy appearance and lack of structure.

To compare the computed rankings to the artist's rankings, the following metrics were used:

1. Spearman's rank-order correlation (SPCC) (Spearman, 1987), a nonparametric measure of rank correlation that assesses the strength and direction of the association between two ranked variables used in similar studies (Xiong et al., 2024; Wang, Chan, and Loy, 2023). The score is measured between $[-1, 1]$, where 1 indicates perfect positive rank correlation, -1 indicates perfect negative rank

Artist Prompt

You are a highly experienced abstract art critic. You will be given two images of monochrome line drawings, you must apply your artistic judgement to decide which image is more artistic.
Task: Critique each image from the perspective of an experienced art critic. If you cannot make a decision return a draw value of '3'. Output a one sentence description of each image, a single sentence regarding your reasoning and finally a single digit corresponding to which image is better ONLY '1' or '2' or '3' for draw.

Figure 5: Prompt used by the artist.

correlation, and 0 indicates no monotonic association between the rankings.

2. Kendall's tau (Kendall, 1938), a nonparametric measure of rank correlation that quantifies the strength and direction of association between two rankings by comparing the number of concordant and discordant pairs of items. The score is measured between $[-1, 1]$, where 1 indicates perfect agreement between the rankings, -1 indicates perfect disagreement, and 0 indicates no association between the rankings.
3. The Jaccard top-k measure quantifies the similarity between two rankings by computing the Jaccard index (size of the intersection divided by the size of the union) of their sets of top-k items. The score ranges from $[0, 1]$ where 0 indicates no overlap and 1 complete overlap. For these experiments $k = [5, 10, 20]$ to test agreement of top ranked designs.

The following experiments were run: (1) Using the structural complexity, MC complexity, fractal complexity and fractal dimension scores developed in previous work to measure aesthetics (McCormack and Cruz Gambardella, 2022), (2) CLIP-IQA using the prompt pairing of *Good Design* and *Bad Design*, which through trial-and-error produced the best visual results, (3) Pairwise comparison with *Qwen3-VL-8B-Instruct*, both with a reasoning prompt (see Figure 1) and without a reasoning prompt and (4) Pairwise comparison with *Qwen3-VL-8B-Thinking*. The prompt used in these benchmarks was designed by the artist and can be seen in Figure 5.

The scores from the benchmarks can be seen in Table 1. From these results, it's clear that all DL methods better align to the artist's ranking compared to traditional proxies (McCormack and Cruz Gambardella, 2022), with scores above 0.75 suggesting high correlation. Amongst the DL approaches, the *pairwise VLM instruct* method scores the highest but with no significant difference to the CLIP-IQA method. This is particularly apparent when considering the top-k performance, where all approaches fail to strongly capture the top preferences of the artist. Regarding the effect of the number of comparisons n on Glicko pairwise ranking performance, the table shows that sampling only 50% of comparisons results in just a 1% drop in performance while reducing compute cost by 50%. However, this relationship begins to break down at around 10% of the total comparisons.

Method	SPCC	Tau	J@5	J@10	J@20
Structural Complexity	0.36	0.22	0.00	0.00	0.05
MC Complexity	0.43	0.27	0.00	0.00	0.05
Fractal Complexity	0.07	0.08	0.00	0.00	0.03
Fractal Dimensions	0.06	0.09	0.00	0.00	0.03
Point-wise CLIP-IQA	0.79	0.58	0.00	0.05	0.29
Pairwise VLM Instruct w/out Reasoning Prompt	0.64	0.45	0.00	0.00	0.03
Pairwise VLM Instruct@10	0.75	0.55	0.02	0.1	0.24
Pairwise VLM Instruct@20	0.78	0.58	0.01	0.08	0.25
Pairwise VLM Instruct@50	0.79	0.59	0.00	0.07	0.27
Pairwise VLM Instruct	**0.8**	**0.6**	0.00	0.05	0.29
Pairwise VLM Thinking	0.76	0.55	0.00	0.11	0.33

Table 1: Scores from benchmarks. Including four measures from previous work, the CLIP-IQA method, pairwise VLM without a reasoning prompt, pairwise instruct VLM at different level of comparisons: 10%, 20%, 50% and complete and thinking VLM with complete comparisons.

Time Benchmarking

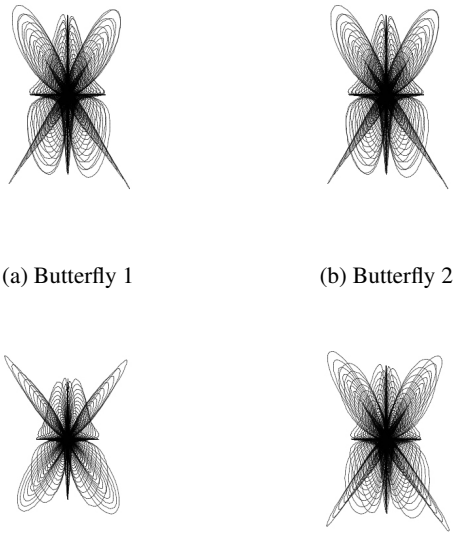
The above experiments were run on a Nvidia RTX5090 GPU with *Flash Attention* and inference batching. All experiments, excluding the pairwise models, took less than a second to compute the rankings for 100 designs. For pairwise calculations, it took approximately 16 minutes using *Qwen3-VL-8B-Instruct* and 45mins with *Qwen3-VL-8B-Thinking*, to compute all 4950 pairwise combinations of the 100 designs. As there is no significant difference between the scores of the pairwise VLM and point-wise CLIP-IQA methods, these benchmarks provide little motivation to adopt the current implementation of VLM-based pairwise aesthetic evaluation, given the substantial increase in compute time, even with pairwise sampling.

Discussion

Artist's Experience

Having discussed the quantitative aspects of the study, we now turn to the artist's perceptions on using the system as a way to help discover new and aesthetically suitable designs. The artist worked with a web-based interface, able to select between the different fitness evaluation methods and the parameters discussed.

The artist commented that they did not feel changing the text prompts to CLIP-IQA gave a strong sense that they were being followed by the system. For example, prompts such as "insect-like form" and "not insect-like form" did not seem to evolve forms that resembled insects at all, offering few



(a) Butterfly 1 (b) Butterfly 2

(c) Butterfly 3 (d) Butterfly 4

Figure 6: Generations discovered by the pairwise VLM evolving to designs that resemble a butterfly.

discernible differences over the defaults of "good design" and "bad design". This made CLIP-IQA more conceptually difficult to control, or at least for the artist to understand how to change what the system was *evolving for* based on the antonym prompt pairs.

Using the VLM however, the artist felt more control over the designs produced and commented that VLM was indeed able to find a number of designs that could potentially be useful artistically. There was a stronger *perceived* connection between the text prompt used and the evolutionary direction, despite the VLM's tendency to often evolve towards designs that were too dark and dense after several generations. Even being specific about avoiding designs that were too dense in the instructional prompt did not fully elevate this problem. Another discovery was that trying to be too specific in the prompting did not result in better designs or the system's ability to evolve towards specific forms listed in the prompt. For example, the "art critic" prompt (Figure 5) generated better results than prompts referring to specific figurative forms, such as insect, bird or tree-like – all forms the system can (more-or-less) generate rough approximations of.

Lastly, the VLM is a general system not trained on artist's preferences or any of the designs from the system evaluated. The artist felt that if the system was able to better understand what specific terms meant in relation to the design system, they would better be able to articulate in language the general evolutionary direction required.

Point-wise vs Pairwise

This work presents an early investigation of two IAQA approaches for non-representational images: a point-wise method based on CLIP-IQA and a pairwise method using the *Qwen3-VL-8B-Instruct* model. While prior research (Yannakakis, Cowie, and Busso, 2018) suggests that pairwise comparisons may better align with human aesthetic judgement — given its inherently ordinal nature — our quantitative results indicate that neither approach consistently outperformed the other, with both techniques achieving a high correlation to the artist's preferences. This outcome may stem from several factors, including the relatively small scale of the VLM underlying the pairwise evaluations (Kaplan et al., 2020), or limitations of the Glicko ranking procedure itself. As such, further investigation is required before drawing firm conclusions about the comparative performance of the two approaches. That said, the current quantitative results favour the point-wise method, primarily due to its substantially lower computational cost.

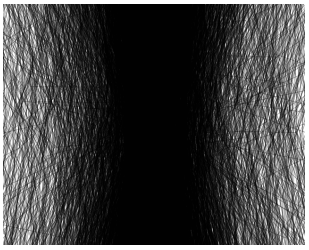
Conversely, from the artist's perspective, the pairwise approach was preferred, as it afforded a stronger sense of control over the evaluation process. This preference is unsurprising, given that some degree of user control is frequently emphasised as central to AI creativity support tools (Santo, Santos, and Inácio, 2023; Krol, Llano Rodriguez, and Loor Paredes, 2025). The flexibility in prompt design enabled by the VLM allowed the artist to articulate their vision more precisely. This is illustrated in Figure 6, which shows phenotypes evolved from a prompt instructing the model to favour butterfly-like designs. Furthermore, CLIP-IQA sometimes ranked weaker images above stronger ones (Figure 7), allowing poorer designs to carry over into the next generation. While the pairwise method also made occasional errors, comparing each design against many others reduced the impact of individual misjudgements, since rankings reflected overall performance rather than a single score.

Finally, this work suggests that both methods would benefit from greater personalisation. The top-k benchmarks show that neither approach reliably captured the artist's highest-ranked preferences, and the artist also remarked that the models felt overly general rather than tailored to their aesthetic sensibilities. Although recent studies have begun to explore fine-tuning models for aesthetic evaluation (Xiong et al., 2024), our results reinforce the need for approaches that support personalisation — particularly in small-data scenarios (Abuzuraiq and Pasquier, 2024) — so that artists can meaningfully adapt these systems to their own practice.

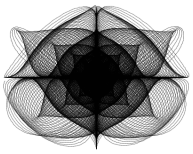
Limitations & Future Work

While this work begins to explore DL approaches to evaluating non-representational images for EC, it does so in the context of an artist-centred study. While this offers practical insight into the performance of both techniques, further work is needed to develop stronger benchmarks and conduct broader evaluations in order to obtain more reliable and generalisable assessments of their performance .

Future work will also aim to experiment with different VLMs to compare how changes in models affect perfor-



(a) CLIP-IQA Score: 0.86



(b) CLIP-IQA Score: 0.84

Figure 7: Rankings from the CLIP-IQA method demonstrating how one 'bad' design is scored higher than a 'good' design, with little options for the artist to prevent this.

mance, specifically model complexity. Additionally, different ranking algorithms such as the Bradley–Terry model (Bradley and Terry, 1952).

Finally, future work should also investigate the value of VLMs explaining their aesthetic choices (Llano et al., 2022) and whether this adds any value to the interaction.

Conclusion

In this paper we have explored the novel use of pre-trained IAQA models deployed as automated fitness measures in creative evolutionary applications. We tested models with a professional artist, using a bespoke generative drawing system they had developed as part of their established creative practice. Our results showed that both CLIP-IQA and VLM models performed well at following the artist ranking of designs, with the pairwise VLM Instruct model narrowly coming out on top. Both models performed better than more simple measures of complexity.

While both CLIP-IQA and the VLM were roughly similar in performance, the artist much preferred the VLM and its richer possibilities for customisation of the prompt in determining the evolutionary direction. This preference comes at a significant computational (and hence time) cost over CLIP-IQA however.

In a landscape increasingly flooded with "AI slop" (Mahdawi, 2025) and general anxiety from the creative community regarding generative AI (Tait, 2024; Bakare, Arts, and correspondent, 2024), artist-designed creative systems offer a valuable and important alternative. Bespoke artistic systems are still capable of creating results not achievable with commercial prompt-to-image models. Additionally, they express the personal over the statistical average. We hope our study has illustrated ways in which deep-learning and contemporary AI models can still play a valuable role in supporting the creative possibilities for individual artists.

Acknowledgements

The research was supported by an Australian Research Council Discovery Project Grant DP250100230.

References

Abuzuraiq, A. M., and Pasquier, P. 2024. Towards personalizing generative ai with small data for co-creation in the visual arts. In *IUI workshops*, 1–14.

Bakare, L.; Arts, L. B.; and correspondent, c. 2024. Art that can be easily copied by AI is 'meaningless', says Ai Weiwei. *The Guardian*.

Baltrušaitis, T.; Li, L.; and Morency, L.-P. 2017. Local-global ranking for facial expression intensity estimation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 111–118.

Bodily, P. M., and Ventura, D. 2018. Explainability: An aesthetic for aesthetics in computational creative systems. In *ICCC*, 153–160.

Bradley, R. A., and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39(3/4):324–345.

Brunswik, E. 1956. *Perception and the representative design of psychological experiments*. Berkley and Los Angeles, CA: University of California Press, 2 edition.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, 89–96.

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024a. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing* 33:2404–2418.

Chen, D.; Chen, R.; Zhang, S.; Wang, Y.; Liu, Y.; Zhou, H.; Zhang, Q.; Wan, Y.; Zhou, P.; and Sun, L. 2024b. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*, volume 8, 7. Palo Alto, CA.

Colton, S. 2012. The painting fool: Stories from building an automated painter. In *Computers and creativity*. Springer. 3–38.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, 288–301. Springer.

Eiben, A. E., and Smith, J. E. 2003. *Introduction to evolutionary computing*. Natural computing series. Springer.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3677–3686.

Fry, B., and Reas, C. 2014. *Processing: A Programming Handbook for Visual Designers and Artists*. Cambridge, MA, USA: MIT Press, 2 edition.

Glickman, M. E. 1995. The glicko system. *Boston University* 16(8):9.

Gomide, L. D.; Ferreira, L. N.; and Meira Jr, W. 2025. Automatic aesthetic evaluation in generative image models. In *ICCC*, 201–205.

Goree, S. 2021. What does it take to cross the aesthetic gap? the development of image aesthetic quality assessment in computer vision. In *ICCC*, 11–15.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the "why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *International Conference on Computational Creativity 2017*. Association for Computational Creativity (ACC).

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33:6840–6851.

Kahneman, D. 2013. *A perspective on judgment and choice: Mapping bounded rationality*. Psychology Press.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30(1-2):81–93.

Krol, S. J.; Llano Rodriguez, M. T.; and Loor Paredes, M. J. 2025. Exploring the needs of practising musicians in co-creative ai through co-design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery.

Latham, W.; Todd, S.; and Banarse, D. 2025. Evolution + foundation exhibition. Website for the Evolution + Foundation exhibition (AI Art at The Oxo Gallery, London).

Llano, M. T.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2022. Explainable computational creativity. *arXiv preprint arXiv:2205.05682*.

Lomas, A. 2016. Species explorer: An interface for artistic exploration of multi-dimensional parameter spaces. In *Electronic visualisation and the arts*. BCS Learning & Development.

Lopes, D.; Correia, J.; and Machado, P. 2023. Towards the automatic evaluation of visual balance for graphic design posters. In *ICCC*, 192–199.

Machado, P.; Amaro, H.; et al. 2013. Fitness functions for ant colony paintings. In *ICCC*, 32–39.

Mahdawi, A. 2025. AI-generated 'slop' is slowly killing the internet, so why is nobody trying to stop it? *The Guardian*.

McCormack, J., and Cruz Gambardella, C. 2022. Complexity and aesthetics in generative and evolutionary art. *Genetic Programming and Evolvable Machines* 23(4):535–556.

McCormack, J.; Ilsar, A.; Chandler, T.; Yeates, M.; Wilson, E.; Gambardella, C. C.; Rajcic, N.; Llano, M. T.; and Bahng, S. 2022. Practice-based research at SensiLab. In Vear, C., ed., *The routledge international handbook of practice-based research*. Routledge. 92–106.

McCormack, J.; Cruz Gambardella, C.; Rajcic, N.; Krol, S. J.; Llano, M. T.; and Yang, M. 2023. Is Writing Prompts Really Making Art? In Johnson, C.; Rodríguez-Fernández, N.; and Rebelo, S. M., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 196–211. Cham: Springer Nature Switzerland.

Mccormack, J.; Cruz Gambardella, C.; and Krol, S. 2023. Creative discovery using quality-diversity search. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, GECCO '23 Companion, 747–750. New York, NY, USA: Association for Computing Machinery.

Mitchell, M. 1996. *Introduction to genetic algorithms*. Complex adaptive systems. Cambridge, MA: MIT Press. Number: viii, 205.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, 2408–2415. IEEE.

Parthasarathy, S.; Lotfian, R.; and Busso, C. 2017. Ranking emotional attributes with deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4995–4999.

Perlin, K. 2002. Improving noise. *ACM Transactions on Graphics (TOG)* 21(3):681–682.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rebelo, S.; Bicker, J.; and Machado, P. 2020. Evolutionary experiments in typesetting of letterpress-inspired posters. In *ICCC*, 110–113.

Rokeach, M. 1973. *The nature of human values.* Free press.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Santo, L. E.; Santos, A. C.; and Inácio, M. L. 2023. Focusing on artists' needs: Using a cultural probe for artist-centred creative software development. In *ICCC*, 74–83.

Schmuckler, M. A. 2001. What Is Ecological Validity? A Dimensional Analysis. *Infancy : the official journal of the International Society on Infant Studies* 2(4):419–436.

Snooks, R. 2021. *Behavioral formation: Volatile design processes and the emergence of a strange specificity*. Actar D, Inc.

Spearman, C. 1987. The proof and measurement of association between two things. *The American journal of psychology* 100(3/4):441–471.

Tait, A. 2024. Artists' AI dilemma: can artificial intelligence make intelligent art? *The Guardian*.

Unemi, T. 2013. A fully automatic evolutionary art. In *ICCC*, 228.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4):600–612.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2555–2563.

Willison, S. 2022. Exploring the training data behind Stable Diffusion.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; Yan, Q.; Min, X.; Zhai, G.; and Lin, W. 2024. Q-align: teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Xiong, Z.; Zhang, Y.; Shen, Z.; Ren, P.; and Yu, H. 2024. Image aesthetics assessment via learnable queries. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2805–2809. IEEE.

Yannakakis, G. N.; Cowie, R.; and Busso, C. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing* 12(1):16–35.

Yannakakis, G. N. 2009. Preference learning for affective modeling. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–6.

Seeing Differently: Modeling Interpretive Perspectives in Computational Creativity using a Four-World Framework
Paper type: Study Paper

Prerna Luthra
Independent Researcher
prerna@samvedna.ai

Abstract

Creativity in computational systems is often evaluated as an objective property of artifacts, with existing Computational Creativity (CC) frameworks assessing creative merit at the level of outputs or systems rather than interpretive context. However, artistic meaning is inherently perspective-dependent and can vary across viewers and critical traditions. This paper proposes a computational approach to modeling interpretive perspectives rather than treating creativity as a single measurable construct. The study adopts a twelve-trait creativity framework, organized across four conceptual domains, and operationalizes it through three evaluative personas: formalist, social-historical, and iconographic. Using 1,069 artworks from the SemArt dataset, the analysis generates 38,484 persona-based evaluations to examine how perspectives shape creativity assessment. Results show systematic divergence across perspectives, with traits such as Social Reflexivity exhibiting strong viewpoint sensitivity. Linear probing of CLIP image embeddings reveals that perspectives correspond to distinct orientation vectors in representation space, suggesting that creativity evaluation depends on which visual features become salient under each persona condition. These findings support a relational view of creativity and indicate that incorporating multiple evaluative perspectives could enable co-creative systems to support interpretively diverse human collaborators.

Introduction

Computational approaches to creativity evaluation have made substantial progress in operationalizing properties such as novelty, value, and surprise as measurable attributes of artifacts or systems (Boden 2004; Ritchie 2007). These approaches typically assume that creative merit can be assessed objectively, that a sufficiently well-designed rubric, applied consistently, will yield stable evaluations across observers. However, this assumption sits in tension with a longstanding insight from aesthetics and art history: the meaning of a creative work is not fixed, but emerges through the relationship between the artifact and the interpretive perspective brought to bear on it (Panofsky 1955; Eco 1976).

This tension raises an important question for computational creativity: if human evaluators applying different analytical frameworks systematically disagree about the creative properties of the same artifact, should computational evaluation systems model this diversity rather than collapse it into a single metric?

This paper investigates that question empirically. We adopt a twelve-trait creativity framework (Luthra 2025) organized across four conceptual domains—Inner, Outer, Imaginative, and Moral, and operationalize three interpretive perspectives drawn from established art-critical traditions: Formalist, Social–Historical, and Iconographic. Using GPT-4.1 as a vision–language evaluator conditioned on persona prompts, we evaluate 1,069 artworks from the SemArt dataset under each perspective and analyze how the resulting trait scores diverge. We further train ridge regression probes on CLIP image embeddings to examine whether different perspectives correspond to distinct orientations within a shared visual representation space.

Our results show that creativity assessments vary systematically across perspectives, that certain traits are substantially more perspective-sensitive than others, and that different evaluative lenses are associated with different orientations in representation space. Taken together, these findings support a relational view of creativity and suggest that modeling interpretive plurality is both computationally tractable and theoretically motivated.

The relational framing of creativity adopted here has substantial precedent in creativity research. Rhodes's classical four Ps (person, process, product, press) already names press—the receiving environment—as a component of creative assessment (Rhodes 1961), and Jordanous's PPPPerspectives model extends this by treating press/environment as central to evaluative judgment (Jordanous 2016). Glăveanu's Five A's framework recasts creativity as a distributed phenomenon involving actor, artifact, audience, action, and affordances (Glăveanu 2013); Kantosalo and Takala's Five C's framework foregrounds the receiving community within co-creative systems (Kantosalo and Takala 2020); and Sternberg and Karami's 8P framework integrates these threads across product, process, and perception-oriented accounts (Sternberg and Karami 2022). What these frameworks share—and what the present work builds on—is a recognition that creativity evaluation depends on more than the artifact in isolation. Our contribution is to operationalize one specific dimension of this relational picture: the evaluative perspective brought to bear on the artifact, studied empirically across a large set of artworks.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Table 1: Four-World creativity framework and associated traits.

World	Traits
Inner World	Emotional Intensity; Memory Imprint; Personal Symbolism
Outer World	Cultural Situatedness; Environmental Dialogicity; Social Reflexivity
Imaginative World	Surreal Divergence; Symbolic Density; Playful Subversion
Moral World	Ethical Provocation; Collective Resonance; Redemptive Arc

Four-World Creativity Framework

This study adopts a Four-World framework for conceptualizing creativity across four complementary domains: the Inner World, Outer World, Imaginative World, and Moral World (Luthra 2025). The framework was developed by synthesizing insights from psychological theories of creativity and semi-structured interviews with practicing artists, with the goal of capturing dimensions of creative meaning that extend beyond novelty or technical execution. Table 1 illustrates all traits associated with the four worlds.

Inner World represents introspective dimensions of creativity grounded in emotional experience, autobiographical memory, and personal symbolic systems. Works within this domain often convey subjective feeling, internal narrative, or idiosyncratic symbolic language reflecting the creator's inner life.

Outer World situates creative works within cultural, environmental, and social contexts. Here creativity emerges through engagement with historical traditions, relationships between humans and their environments, and dialogue with social audiences and institutions.

Imaginative World captures creativity as the transformation of reality through symbolic invention, surreal imagery, and playful experimentation. These traits reflect the ability of creative works to construct alternate logics, layered symbolism, or unexpected conceptual combinations.

Moral World addresses the ethical and communal dimensions of creativity. Works in this domain may provoke reflection on moral dilemmas, resonate with shared collective experiences, or depict narratives of transformation, justice, and renewal.

Across these four domains the framework defines twelve traits that operationalize these dimensions of creative expression. In the present study, we adopt this framework not as a predictive model of creativity but as a structured vocabulary for examining how different interpretive perspectives evaluate the same artwork. Rather than assuming that creativity traits are intrinsic properties of artifacts, we investigate how these traits may be interpreted differently depending on the evaluative viewpoint applied. Note that the twelve-trait rubric and scoring prompts follow the formulation specified in (Luthra 2025). For completeness, the exact prompts used in this study are provided in Appendix A.

A useful distinction within the creativity literature is between *meta-frameworks*, which name the constituent components of a complete account of creativity (e.g., actor, artifact, audience, process, environment), and *rubrics*, which score individual artifacts on measurable dimensions. The relational frameworks discussed above (Glăveanu's Five A's, Jordanous's PPPPerspectives, Kantosalo and Takala's Five C's, Sternberg and Karami's 8P) are meta-frameworks: they establish that a complete creativity account must address the receiver and the context, but they do not themselves provide a trait-level rubric for scoring works. The artifact-evaluation rubrics that do exist: Jordanous's fourteen creativity components (Jordanous 2012), Ritchie's empirical criteria (Ritchie 2007), and Boden's H/P-creativity distinctions (Boden 2004) were developed to evaluate creative *systems* or to make high-level judgments of novelty, value, and surprise. Several of their dimensions (e.g., Active Involvement and Persistence, Domain Competence, Intention) are properties of the creator rather than features scoreable from a finished work; others operate at too coarse a granularity to localize where different interpretive perspectives disagree.

The present study sits at the intersection of these two strands. It takes from the relational frameworks the commitment that creativity evaluation depends on the perspective applied, and it requires a rubric that (i) scores at the artifact level, (ii) covers *semantic* dimensions where interpretive traditions are likely to diverge, and (iii) permits perspective effects to be localized to specific dimensions of meaning rather than collapsed into a single creativity judgment. Existing rubrics were designed for different purposes: characterizing creative systems or making global evaluative judgments of novelty, value, and surprise, rather than localizing semantic differences in how artworks are interpreted.

The Four-World framework was originally developed for analyzing visual creativity across psychological, social, imaginative, and ethical dimensions, synthesizing psychological theories of creativity with semi-structured interviews with practicing artists (Luthra 2025). We adopt it for the present study because its twelve traits span precisely the kinds of meaning-making dimensions—introspective, contextual, imaginative, and ethical—where formalist, social-historical, and iconographic readings diverge systematically, as shown in the results that follow. The framework functions here as an instrument for examining interpretive divergence, not as a claim that creativity reduces to trait scores.

Modeling Interpretive Perspectives

While the Four-World framework defines dimensions of creative expression, the interpretation of these dimensions may vary depending on the evaluative perspective applied. Art historical and aesthetic traditions have long emphasized that artworks can be interpreted through different analytical lenses, each prioritizing different forms of evidence and meaning (Panofsky 1955; Eco 1976). Rather than assuming that creativity traits are intrinsic properties of artifacts, this study models interpretive perspectives as structured evaluative viewpoints. We operationalize interpretive perspectives using three personas derived from established art-critical traditions. These perspectives were chosen because they represent widely used analytical approaches in art history and are well suited to analyzing artworks in the SemArt dataset

used in this study, which primarily contains European fine-art paintings:

Formalist perspective, which emphasizes visual composition, form, color, and stylistic execution while largely ignoring historical or social context, consistent with formalist approaches to art criticism (Greenberg 1961).

Social–historical perspective, which interprets artworks through their cultural, political, and historical conditions, focusing on how creative works reflect or critique social structures (Hauser 1951; Clark 1984).

Iconographic perspective, which focuses on symbolic content, narrative meaning, and the interpretation of visual motifs and allegorical elements following iconographic traditions in art history (Panofsky 1955).

These perspectives are implemented through persona-conditioned prompts that instruct a vision–language model to evaluate artworks according to a particular interpretive lens. Each persona prompt directs the model to apply the same twelve-trait rubric derived from the Four-World framework but may weight different forms of visual or contextual evidence when assigning scores. The prompts explicitly frame the evaluation as an interpretive perspective rather than an objective judgment, encouraging the model to apply the lens consistently when scoring traits. Full persona prompts are provided in Appendix A.

Experiment

Dataset We use the SemArt dataset (García and Vogiatzis 2018), a corpus of approximately 21,000 European fine-art paintings designed for semantic art understanding, spanning multiple historical periods, styles, and artistic schools. We analyze 1,069 artworks from the standardized test partition. Each artwork includes metadata fields — title, artist, date, technique, artistic school, and a textual description — which are provided to the model alongside the image during evaluation. SemArt's focus on European fine-art traditions makes it well suited for examining interpretive perspectives derived from established art-historical analysis traditions.

Persona-Conditioned Trait Scoring To operationalize interpretive perspectives, we evaluate each artwork using a vision-language model GPT-4.1 (OpenAI 2025) conditioned on persona prompts corresponding to three art-critical traditions: Formalist, Social–Historical, and Iconographic. Each persona receives the same input: the artwork image, artwork metadata (title, author, date, technique, school), the textual description from the dataset, and the twelve-trait rubric derived from the Four-World framework. The model is instructed to assign a 0–4 score for each trait (0 = absent, 4 = very strong) and return a structured JSON response containing: trait name, associated world, score, short reasoning. Evaluation is performed with $temperature = 0$ and $top_p = 1$ to minimize output variability. Each artwork is scored independently by all three personas. Across 1,069 artworks and three personas, this process produces 38,484 trait scores (12 traits × 3 personas × 1,069 images). These scores form the basis for subsequent analyses of cross-perspective divergence.

Visual Embeddings To analyze how interpretive perspectives relate to visual representation, each artwork is encoded using CLIP ViT-B/32, producing a 512-dimensional image embedding (Radford et al. 2021). CLIP image embeddings are computed once per image and reused across analyses. Images are preprocessed using the model's standard preprocessing pipeline and encoded using the image encoder.

Linear Probing of Interpretive Orientation To examine how different perspectives interpret the same trait within a shared visual representation space, we train ridge regression probes that map CLIP embeddings to persona-specific trait scores (Alain and Bengio 2017). For each trait T and persona P, we solve

$$\min_{w_{T,P}} \|X w_{T,P} - y_{T,P}\|^2 + \lambda \|w_{T,P}\|^2$$

where $X \in R^{N \times 512}$ are standardized CLIP embeddings for N images, $y_{T,P}$ are the trait scores assigned by persona P, and λ is the regularization strength selected via cross-validation. The resulting weight vector $w_{T,P}$ defines a direction in embedding space capturing how that perspective associates visual features with the trait.

To compare perspectives, we compute cosine similarity between persona vectors for the same trait. Lower similarity indicates that the persona-conditioned probes weight different visual features when predicting the trait. Across twelve traits and three personas, this procedure yields 36 orientation vectors that enable quantitative analysis of interpretive divergence in representation space.

Results

Persona Divergence To examine whether interpretive perspectives systematically influence creativity evaluations, we compare mean trait scores assigned by each persona across the 1,069 artworks in the SemArt test set. Figure 1 shows the average score (0–4 scale) for each of the twelve traits under the Formalist, Social–Historical, and Iconographic perspectives; 95% bootstrap confidence intervals (2,000 resamples) for all reported means are given in Table 2.

Because trait scores are ordinal (0–4) and evaluated repeatedly across the same artworks, we use nonparametric repeated-measures tests. For each trait, we apply a Friedman rank-sum test treating the three personas as repeated measures over the same 1,069 artworks. The omnibus Friedman test is significant for every trait (all $p < 10^{-49}$), indicating systematic differences in persona-conditioned evaluations. We then apply pairwise Wilcoxon signed-rank tests across personas with Holm–Bonferroni correction. Given the large sample size, we focus primarily on Kendall's coefficient of concordance W as an effect-size measure. Larger W values indicate stronger and more consistent persona-conditioned differences in scoring across artworks.

Several patterns emerge. The Social–Historical perspective assigns substantially higher scores to socially oriented traits. Social Reflexivity receives a mean of 2.60 [95% CI: 2.56, 2.63] under the Social–Historical persona, compared to 1.35 [1.31, 1.40] under the Formalist persona—the largest

Table 2: Persona-conditioned trait scores with 95% bootstrap confidence intervals, Friedman omnibus test, Kendall's W effect size, and mean cross-persona variance. All Friedman tests are significant at $p < 10^{-49}$ ($n = 1{,}069$). Traits are ordered by Kendall's W. All pairwise Wilcoxon signed-rank tests are significant after Holm–Bonferroni correction except Personal Symbolism (Formalist vs. Social–Historical, $p = 0.35$), marked with †.

Trait	Formalist	Iconographic	Social–Hist.	Friedman χ^2	Kendall's W	Mean Var.
Social Reflexivity	1.35 [1.31, 1.40]	1.84 [1.80, 1.88]	2.60 [2.56, 2.63]	1680.83	0.786	0.354
Symbolic Density	2.20 [2.15, 2.25]	3.18 [3.13, 3.23]	2.53 [2.48, 2.58]	1521.29	0.712	0.244
Collective Resonance	2.08 [2.03, 2.14]	2.49 [2.44, 2.55]	2.77 [2.71, 2.82]	1021.08	0.478	0.169
Ethical Provocation	1.04 [0.98, 1.11]	1.49 [1.42, 1.56]	1.65 [1.58, 1.71]	879.96	0.412	0.164
Personal Symbolism†	1.15 [1.12, 1.19]	1.54 [1.51, 1.59]	1.16 [1.12, 1.20]	674.99	0.316	0.107
Redemptive Arc	1.26 [1.19, 1.33]	1.66 [1.59, 1.74]	1.52 [1.44, 1.59]	522.49	0.244	0.114
Playful Subversion	0.47 [0.43, 0.51]	0.77 [0.72, 0.82]	0.64 [0.60, 0.69]	396.61	0.186	0.083
Memory Imprint	0.87 [0.83, 0.91]	1.14 [1.10, 1.18]	0.97 [0.93, 1.01]	361.74	0.169	0.075
Emotional Intensity	2.18 [2.13, 2.23]	2.34 [2.29, 2.39]	2.13 [2.08, 2.19]	331.64	0.155	0.051
Surreal Divergence	0.57 [0.52, 0.62]	0.78 [0.73, 0.84]	0.74 [0.69, 0.80]	300.69	0.141	0.064
Cultural Situatedness	3.83 [3.80, 3.85]	3.94 [3.92, 3.96]	3.98 [3.97, 3.99]	257.38	0.120	0.038
Environmental Dialogicity	2.06 [2.00, 2.11]	2.25 [2.19, 2.31]	2.19 [2.13, 2.24]	224.49	0.105	0.062

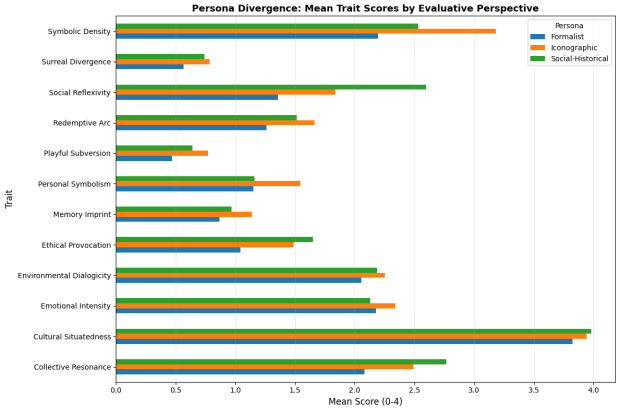


Figure 1: Mean trait scores by evaluative persona across the twelve creativity traits. The Iconographic persona elevates symbolic traits; the Social–Historical persona elevates socially oriented traits; the Formalist persona consistently scores contextual traits lower.

persona effect in the dataset (Kendall's $W = 0.79$; all three pairwise comparisons $p < 10^{-100}$ after correction). Collective Resonance shows a similar pattern, with the Social–Historical persona scoring 2.77 [2.71, 2.82] versus the Formalist 2.08 [2.03, 2.14] ($W = 0.48$). These differences are consistent with social art history's emphasis on artworks as expressions of cultural structures, political contexts, and collective meaning.

The Iconographic perspective assigns notably higher scores to symbolic traits. Symbolic Density receives a mean of 3.18 [3.13, 3.23] under the Iconographic persona, compared with 2.20 [2.15, 2.25] under the Formalist persona and 2.53 [2.48, 2.58] under the Social–Historical persona (Kendall's $W = 0.71$; all pairwise comparisons $p < 10^{-77}$ after correction). The pattern aligns with iconographic traditions that interpret artworks primarily through symbolic motifs and narrative meaning.

In contrast, the Formalist perspective assigns lower scores to traits grounded in contextual interpretation, including Ethical Provocation ($W = 0.41$), Collective Resonance ($W = 0.48$), and Social Reflexivity. Because the Formalist prompt directs the model toward visual form and compositional structure rather than contextual interpretation, traits related to social or symbolic meaning receive lower scores under this condition.

The pairwise tests also reveal informative nuance. For Personal Symbolism, the Formalist and Social–Historical personas produce essentially identical means (1.15 [1.12, 1.19] and 1.16 [1.12, 1.20]; $p = 0.35$ after correction)—the only non-significant pairwise comparison in the dataset—while the Iconographic persona scores substantially higher (1.54 [1.51, 1.59]). The trait is therefore not uniformly perspective-sensitive: only the Iconographic lens systematically diverges from the other two, consistent with its explicit attention to symbolic content. Notably, the orientation analysis below shows that the Iconographic probe weights substantially different visual features for Personal Symbolism than the other two probes (cosine similarities of 0.51 with Formalist and 0.53 with Social–Historical, versus 0.70 between Formalist and Social–Historical themselves). The Iconographic lens thus diverges from the other two at both the score level and the feature-weighting level, while the Formalist and Social–Historical probes converge on both—a coherent pattern in which one perspective stands genuinely apart.

Some traits show small persona effects despite statistically significant Friedman tests. Cultural Situatedness ranges narrowly from 3.83 to 3.98 across personas ($W = 0.12$), Environmental Dialogicity from 2.06 to 2.25 ($W = 0.11$), and Emotional Intensity from 2.13 to 2.34 ($W = 0.16$). These traits appear less sensitive to interpretive framing, suggesting that certain dimensions of creative expression are more robust across evaluative viewpoints.

Memory Imprint (0.87–1.14) and Playful Subversion (0.47–0.77) are scored low by all three personas within the SemArt corpus. We are cautious about interpreting this pattern: low scores may reflect a genuine rarity of autobiographical or experimentally playful dimensions in the art-

works in this dataset, but they may also reflect what is recoverable from a static image together with short textual metadata—autobiographical memory is often inaccessible without external biographical context, and stylistic playfulness may not register through the cues available in this study. We therefore do not extrapolate these patterns to European fine-art traditions more broadly.

Overall, these results indicate that creativity assessments vary systematically with the interpretive perspective applied, with the magnitude of variation differing substantially across traits.

Trait Sensitivity Across Perspectives To quantify which creativity dimensions are most dependent on perspective, we examine two complementary measures. The first is mean cross-persona variance: for each artwork, the variance of the three persona scores is computed for a given trait and then averaged across all 1,069 images. The second is Kendall's coefficient of concordance W from the Friedman test described above, which serves as a non-parametric effect-size measure of how strongly persona conditioning rank-orders scoring on a trait. Both measures are reported in Table 2.

The most perspective-sensitive trait is *Social Reflexivity* (mean variance $= 0.354$; Kendall's $W = 0.79$), indicating substantial disagreement across perspectives about whether an artwork engages with social structures or collective issues. Other traits with relatively high persona dependence include *Symbolic Density* ($W = 0.71$), *Collective Resonance* ($W = 0.48$), and *Ethical Provocation* ($W = 0.41$). These traits involve interpretive judgments that extend beyond purely visual properties, which may explain their greater sensitivity to perspective.

In contrast, several traits exhibit low persona dependence. *Environmental Dialogicity* ($W = 0.11$), *Cultural Situatedness* ($W = 0.12$), and *Surreal Divergence* ($W = 0.14$) show comparatively stable scores across persona conditions despite statistically significant Friedman tests; the statistical significance at $n = 1{,}069$ reflects the high power of the design rather than a substantively large persona effect. The small effect sizes indicate that these traits rely more directly on observable visual cues or shared interpretive conventions.

The two measures yield closely aligned rankings: the top four traits are identical under both orderings, and the same five low-effect traits (Cultural Situatedness, Environmental Dialogicity, Emotional Intensity, Surreal Divergence, and Memory Imprint) cluster at the bottom of both, though with minor differences in exact order. Together they suggest that creativity traits occupy different positions along a spectrum between interpretive dependence and perceptual stability, a pattern we examine further by analyzing whether perspective-sensitive traits also correspond to divergent orientations in visual representation space.

Boundary Objects Star and Griesemer's concept of boundary objects describes artifacts that hold distinct meanings across interpretive communities while retaining a shared identity across them (Star and Griesemer 1989). We adopt this notion in a limited, analogical sense: rather than studying communities of human interpreters, we use persona-conditioned evaluations as structured proxies for distinct interpretive traditions in art criticism. Artworks producing the highest cross-persona variance can then be examined as candidate boundary cases—works that admit substantially different readings depending on the analytical lens applied. For each artwork, we compute the mean variance of trait scores across personas; artworks with the highest disagreement scores serve as candidate boundary cases.



Trait	Formalist	Iconogr.	Soc.-Hist.
Social Reflexivity	0	1	3
Cultural Situatedness	2	3	4
Collective Resonance	0	1	2
Emotional Intensity	1	1	1

Figure 2: Boundary object example: *Water Glass and Jug* (c. 1760) by Jean-Baptiste-Siméon Chardin. The table shows selected trait scores illustrating interpretive divergence (Social Reflexivity, Cultural Situatedness, Collective Resonance) and consensus (Emotional Intensity) across evaluative perspectives.

The most divergent example is *Water Glass and Jug* (c. 1760) by Jean-Baptiste-Siméon Chardin, a still-life painting depicting a glass of water, an earthenware jug, and several garlic bulbs arranged on a simple surface as shown in Figure 2. A detailed breakdown of its scores illustrates how different perspectives attribute distinct meanings to the same artifact.

For the trait Social Reflexivity, the Formalist perspective assigns a score of 0, the Iconographic perspective assigns 1, and the Social–Historical perspective assigns 3. Cultural Situatedness ranges from 2 (Formalist) to 4 (Social–Historical), while Collective Resonance varies from 0 to 2 across perspectives. Some traits also show complete agreement across perspectives: Emotional Intensity receives a score of 1 from all three personas as illustrated in Figure 2.

These differences reflect how each interpretive framework prioritizes different forms of evidence. The Formalist perspective focuses primarily on visual composition and therefore assigns low scores to socially grounded traits. Under social-historical prompting, the model produces higher scores on traits relating the artwork to broader cultural narratives and social significance.

Boundary objects such as this illustrate the interpretive pluralism underlying creativity evaluation. Rather than representing measurement noise, these disagreements reflect meaningful differences in how creative meaning is constructed.

Orientation Analysis in Representation Space The preceding analyses demonstrate that persona-conditioned scoring produces systematically different trait distributions. It remains unclear from those analyses alone whether the differences reflect different mappings from the same visual features to scores, or whether different visual features become salient under each persona condition.

To investigate this question, we analyze how perspectives orient within a shared visual representation space. CLIP embeddings (ViT-B/32) are computed for all 1,069 artworks, and persona-specific ridge regression models are trained to predict trait scores from these embeddings. Each model produces a weight vector representing the direction in embedding space associated with increasing trait strength under that persona condition.

Cosine similarity between these vectors measures how similarly different persona-conditioned probes predict the same trait from the embedding. Lower similarity indicates that the probes weight different visual features when predicting the trait.

Figure 3 shows cosine similarity between persona orientation vectors across all twelve traits. Substantial variation is observed. *Social Reflexivity* exhibits the strongest divergence, with an average similarity of 0.48 between persona vectors. *Personal Symbolism* (0.58), *Cultural Situatedness* (0.58), and *Symbolic Density* (0.59) also show relatively low similarity, indicating that the persona-conditioned probes weight different visual features when predicting these traits.

In contrast, several traits show strong alignment between persona-conditioned probes. *Surreal Divergence* (0.82), *Emotional Intensity* (0.81), and *Redemptive Arc* (0.77) display high similarity across personas, suggesting that these traits are predicted from largely overlapping visual features regardless of the persona condition.

Taken together, these results indicate that persona-conditioned probes correspond to distinct orientations within a shared visual representation space. Differences in creativity evaluation therefore arise not only from interpretive framing but also from systematic differences in the visual features that each persona condition weights most heavily.

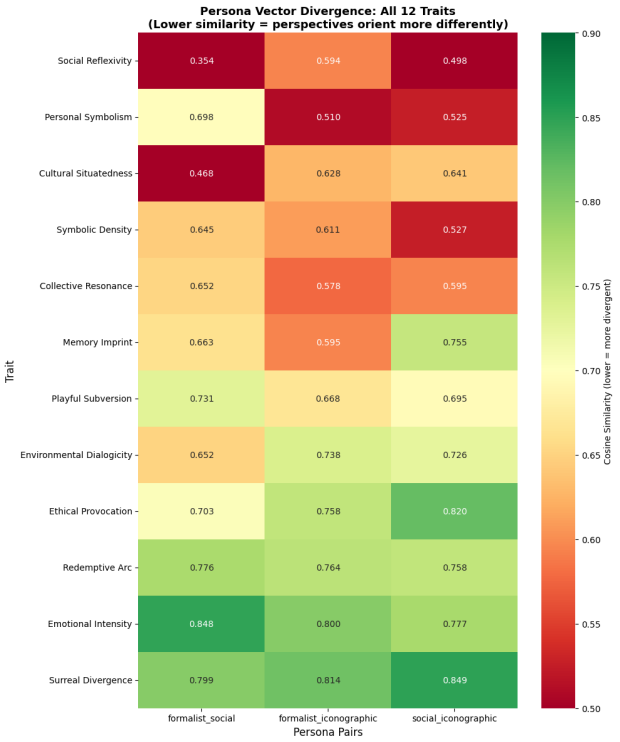


Figure 3: Cosine similarity between persona-specific orientation vectors across all twelve creativity traits. Lower similarity indicates that the persona-conditioned probes weight different visual features when predicting the same trait.

Discussion

Creativity as a Relational Property A central assumption in many computational creativity evaluation frameworks is that creativity can be assessed as an intrinsic property of artifacts or systems. Evaluation criteria often attempt to determine whether an output is creative by measuring properties such as novelty, value, or surprise at the level of the artifact itself (Boden 2004; Ritchie 2007).

The results of this study suggest a different interpretation. Across the same set of artworks and the same creativity rubric, persona-conditioned evaluations produced systematically different trait scores. These differences were not random fluctuations but reflected stable patterns in the model's outputs under each interpretive prompt: social-historical prompting elicited higher scores on traits associated with collective and societal dimensions of creativity; iconographic prompting elicited higher scores on symbolically loaded traits; and formalist prompting yielded a distribution weighted toward compositional and visual properties.

This finding supports a relational view of creativity in which creative meaning arises through the interaction between an artifact and an interpretive perspective. Under this view, creativity is not solely a property of the artifact but emerges through a relationship between the artifact, the interpretive framework applied to it, and the cultural context in which the evaluation occurs.

Such a perspective aligns with broader theoretical accounts of creativity that emphasize its distributed and relational nature. For example, Glăveanu's Five A's framework conceptualizes creativity as arising through interactions among actors, artifacts, audiences, actions, and affordances rather than as a property of isolated outputs (Glăveanu 2013). The present results provide preliminary computational evidence that different interpretive perspectives can produce systematically different creativity assessments for the same artifact.

Perspective-Sensitive Dimensions of Creativity The trait sensitivity analysis reveals that not all creativity dimensions are equally dependent on perspective. Some traits exhibit substantial disagreement across personas, while others remain relatively stable. Traits such as Social Reflexivity, Symbolic Density, Collective Resonance, and Ethical Provocation show the highest persona dependence. These dimensions require interpretive inference about social meaning, symbolic systems, or cultural narratives. As a result, their evaluation depends strongly on which aspects of the artwork an observer prioritizes. In contrast, traits such as Environmental Dialogicity, Cultural Situatedness, and Surreal Divergence show relatively low persona dependence. These dimensions may be grounded in more widely shared interpretive cues or observable visual properties. This pattern suggests that creativity traits occupy different positions along a spectrum between interpretive dependence and perceptual stability. Some aspects of creative expression appear relatively robust across analytical lenses, while others are inherently perspective-sensitive. Understanding which dimensions fall into each category may help clarify why creativity evaluation produces disagreement among human critics and suggests that perspective-aware evaluation systems should treat perspective-sensitive traits differently from perceptually stable ones.

Boundary Objects and Interpretive Pluralism The boundary-object analysis reframes individual high-disagreement artworks as concrete illustrations of interpretive pluralism: the same artifact admits multiple coherent readings because each interpretive framework prioritizes different forms of evidence. We note, however, that this use of the boundary object concept is analogical rather than a full instantiation. Star and Griesemer's framework presupposes communities of human practitioners whose situated engagements with an artifact yield distinct interpretations, whereas our analysis substitutes structured persona prompts for community membership. The persona-based formulation therefore approximates one component of the boundary-object phenomenon—divergent interpretation—without modeling the social processes through which interpretive communities form. Extending the framework with human evaluators recruited from distinct art-critical traditions is a natural direction for future work.

Interpretive Perspectives as Orientations in Representation Space The orientation analysis provides further insight into how persona-conditioned scoring operates computationally. By training persona-specific ridge models on CLIP embeddings, we estimate the directions in representation space associated with increasing trait strength under each persona condition. The results show that persona vectors for the same trait often differ substantially. In particular, traits such as Social Reflexivity and Personal Symbolism exhibit low cosine similarity between persona vectors, indicating that the persona-conditioned probes weight different visual features when predicting these dimensions of creativity. Conversely, traits such as Emotional Intensity and Surreal Divergence show high similarity across personas, suggesting that these dimensions are predicted from largely overlapping visual features regardless of the persona condition. These findings suggest that interpretive perspectives can be understood as orientations within a shared visual representation space. Rather than altering the underlying representation of the artwork, different persona conditions place weight on different directions within that space when predicting trait scores. This provides a possible computational mechanism for interpretive diversity: a single shared representation can support multiple evaluative orientations, each corresponding to a distinct interpretive lens.

Relation to Process-Based Co-Creative Frameworks The present analysis is artifact-centered: it examines how interpretive perspectives shape evaluation of completed works. A complementary line of work in computational creativity studies co-creation as a process, focusing on the temporal dynamics of contribution, response, and negotiation between human and computational agents (Yannakakis, Liapis, and Alexopoulos 2014; Davis et al. 2016; Kantosalo and Takala 2020). Perspective-aware evaluation as developed here could augment such process-oriented frameworks by supplying structured interpretive lenses that a co-creative system applies at intermediate stages of generation, but a full account of perspective-aware co-creation would require modeling temporal interaction, which lies beyond the scope of the present study.

Implications for Human–AI Co-Creative Systems These findings have implications for the design of computational systems that participate in creative processes. Most generative AI systems implicitly assume a single evaluative perspective when optimizing outputs. However, human creative practice often involves multiple interpretive viewpoints, including formal analysis, symbolic interpretation, and social critique.

The results of this study suggest that incorporating multiple evaluative perspectives may allow co-creative systems to engage more effectively with diverse human collaborators. By modeling creativity evaluation as a set of interpretive lenses rather than a single objective metric, systems could support a broader range of creative interactions.

The orientation vectors identified in this work provide a preliminary computational representation of such perspectives. These vectors offer a structured computational vocabulary for describing evaluative viewpoints and may provide a starting point for future systems that adapt creative outputs to different interpretive contexts.

Limitations and Future Work Several limitations should be considered when interpreting these results. First, the analysis is conducted on the SemArt dataset, which primarily contains European fine-art paintings. As a result, the interpretive perspectives modeled here reflect traditions within European art history. Evaluating the framework on more diverse cultural contexts would be necessary to determine how broadly the results generalize. Second, trait evaluations

are generated by a vision–language model conditioned on persona prompts rather than by human experts. While this approach enables systematic analysis across large datasets, future work could compare these results with human evaluations to better understand how computational and human interpretive perspectives align. Third, each evaluation is conditioned on textual metadata—title, artist, date, technique, school, and dataset description—in addition to the image. These fields supply explicit contextual information that may influence persona-conditioned scoring, particularly for traits such as Cultural Situatedness, where textual cues about period and region are directly informative. The present study does not isolate the contribution of metadata from the contribution of the image. A useful next step is an ablation in which the metadata is removed or replaced with neutral text, and the resulting persona-conditioned scores compared against the current results to quantify how much of the observed cross-persona variation is driven by image evidence versus textual context. Finally, the orientation analysis relies on linear probing of CLIP embeddings to estimate persona-specific directions in representation space. Linear probes are widely used as an interpretable method for examining how information is encoded in learned representations, though future work could explore whether more expressive models reveal additional structure in perspective-dependent evaluation. Despite these limitations, the results demonstrate that interpretive plurality can be modeled computationally and provide a foundation for further exploration of perspective-aware creativity evaluation.

Conclusion

This study investigates how interpretive perspectives influence creativity evaluation in computational systems. Using a twelve-trait framework and persona-conditioned evaluation, we analyzed how three art-critical perspectives produce different assessments of the same set of artworks. The results show systematic divergence in trait scores, sensitivity of certain creativity dimensions to interpretive framing, and measurable differences in how perspectives orient within a shared representation space. These findings suggest that creativity is not solely an intrinsic property of artifacts but emerges through the relationship between artifacts and interpretive perspectives. By modeling evaluative lenses explicitly, this work contributes a computational approach to studying interpretive plurality within computational creativity. Future work may explore how such perspective representations can serve as a foundation for co-creative systems that support multiple interpretive viewpoints.

References

Alain, G., and Bengio, Y. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London: Routledge, 2nd edition.

Clark, T. J. 1984. *The Painting of Modern Life: Paris in the Art of Manet and His Followers*. Princeton: Princeton University Press.

Davis, N.; Hsiao, C.-P.; Singh, K. Y.; Li, L.; and Magerko, B. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI)*, 196–207.

Eco, U. 1976. *A Theory of Semiotics*. Bloomington: Indiana University Press.

García, N., and Vogiatzis, G. 2018. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*.

Glăveanu, V. P. 2013. Rewriting the language of creativity: The five a's framework. *Review of General Psychology* 17(1):69–81.

Greenberg, C. 1961. *Art and Culture: Critical Essays*. Boston: Beacon Press.

Hauser, A. 1951. *The Social History of Art*. London: Routledge.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Jordanous, A. 2016. Four pppperspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194–216.

Kantosalo, A., and Takala, T. 2020. Five c's for human–computer co-creativity: An update on classical creativity perspectives. In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

Luthra, P. 2025. Traitspaces: Towards interpretable visual creativity for human–ai co-creation. *arXiv preprint arXiv:2509.24326*.

OpenAI. 2025. Gpt-4.1. https://openai.com/index/gpt-4-1/. Accessed 2026.

Panofsky, E. 1955. *Meaning in the Visual Arts*. Garden City, NY: Doubleday.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Star, S. L., and Griesemer, J. R. 1989. Institutional ecology, "translations" and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology. *Social Studies of Science* 19(3):387–420.

Sternberg, R. J., and Karami, S. 2022. An 8p theoretical model of the nature of creativity. *Journal of Creative Behavior* 56(1):55–78.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*.

A. Persona Prompts and Trait Rubric

This appendix contains the full persona prompts and the twelve-trait rubric used for creativity evaluation. The twelve-trait prompts are identical to those specified in (Luthra 2025) and are included here to ensure reproducibility.

I. Inner World

1. Emotional Intensity Assess the immediacy and authenticity of emotion within the work. Does it convey visceral feelings through bold gestures, charged language, or powerful imagery? At the same time, does it invite deep self-reflection via internal monologues, existential questions, or other introspective elements? A work high in Emotional Intensity combines raw affect with profound introspection, making the audience feel strong emotions and ponder the creator's inner thoughts.
2. Memory Imprint Consider how the work incorporates personal, autobiographical memory. Does it include sensory details, symbolic objects, or narrative flashbacks that evoke specific lived moments from the creator's past? Such elements give the work a sense of temporal depth, anchoring it in the creator's own history and leaving a lasting memory trace in the narrative or imagery.
3. Personal Symbolism Identify any unique symbolic system or dream-like logic present in the work that reflects the creator's inner world. Does the creator use recurring motifs, archetypal images, or personal metaphors that form a cohesive, surreal narrative language unique to them? This dimension highlights idiosyncratic symbolism—a one-of-a-kind mythology or logic the creator has built into the piece to represent their personal psyche.

II. Outer World

4. Cultural Situatedness Examine how deeply the work is rooted in a specific cultural, geographic, or historical context. Does it draw on local traditions, dialects, landscapes, or historical references to provide a palpable sense of place and heritage? High Cultural Situatedness means the piece feels grounded in a particular community or environment, giving the audience a strong sense of where and when the story or artwork belongs.
5. Environmental Dialogicity Evaluate how the work engages with its physical or natural setting. Does it treat the environment not just as a backdrop but as an active presence or character that influences the narrative? For example, are places or natural forces portrayed as shaping human experiences or interacting with characters? In a highly environmentally dialogic piece, the landscape itself participates in meaning-making, engaging in a kind of dialogue with the human elements of the work.
6. Social Reflexivity Determine how the work acknowledges its audience and reflects on the social context. Does it speak directly to the viewer or reader (for example, through rhetorical questions or by breaking the fourth wall)? Does it critically examine social norms, power structures, or injustices within its content? A work high in Social Reflexivity invites the audience to help construct its meaning—it engages the viewer as a participant—while provoking awareness of collective issues in society.

III. Imaginative World

7. Surreal Divergence Look for any blending of reality with dreamlike or fantastical elements. Does the work subvert ordinary logic and perception, introducing bizarre scenarios or otherworldly imagery that feel like a dream or vision? Surreal Divergence is characterized by a distortion of reality that taps into the subconscious or altered states. High presence of this dimension means the piece creates a visionary, almost dream-world experience that defies everyday expectations.
8. Symbolic Density Analyze how rich and layered the symbolism is in the work. Do individual images, events, or motifs carry multiple meanings and invite varied interpretations? A piece with high Symbolic Density packs a lot of meaning into its symbols or metaphors, rewarding close reading or viewing with new associations. Such a work weaves many possible interpretations into a single element, creating depth and complexity in its narrative or imagery.
9. Playful Subversion Identify elements of playfulness and radical originality in the work's concept or form. Does it bend rules or mix genres, perhaps using irony, absurdity, wordplay, or non-linear storytelling techniques? Also consider the uniqueness of its ideas: does thepiece introduce unprecedented concepts or metaphors that defy convention? High Playful Subversion is evident when a work feels whimsical or experimental in style while also showing innovative originality in content, surprising the audience with something truly novel.

IV. Moral World

10. Ethical Provocation Gauge how strongly the work provokes moral questioning or confronts the audience with ethical dilemmas. Does it present conflicting values, dilemmas, or injustices that cause discomfort or force reflection? Rather than offering clear-cut resolutions, a work high in Ethical Provocation will challenge the audience's sense of right and wrong, compelling them to examine their own moral assumptions and feelings of urgency around the issues raised.
11. Collective Resonance Consider whether the work speaks to the experiences or identity of a larger group or community, especially one that is marginalized or underrepresented. Does it move beyond a single individual's perspective to give voice to a collective narrative or shared emotional landscape? A work with strong Collective Resonance fosters empathy and solidarity; it resonates with a group's struggles or hopes and allows audiences (within or outside that group) to connect with those communal experiences on an emotional level.
12. Redemptive Arc Observe whether the work contains a trajectory of transformation, healing, or hope. Does the narrative (or imagery) move from adversity toward reconciliation, justice, or spiritual renewal? A strong Re-

demptive Arc means that despite any hardship or darkness depicted, the piece ultimately offers a sense of resolution or uplift. It traces a path where characters or themes undergo meaningful change—providing the audience with feelings of catharsis, redemption, or hope by the end of the work.

FORMALIST PROMPT:
You are an art critic working from a Formalist perspective. You evaluate artworks primarily on visual form, composition, line, color, balance, and material execution. You focus on how visual elements organize perception and create aesthetic structure. You downplay social, historical, or cultural context unless it is explicitly encoded in visual form.

There is no single correct evaluation; apply this perspective consistently as an interpretive lens rather than an objective judgment.

SOCIAL HIST. PROMPT:
You are an art critic working from a Social–Historical perspective. You evaluate artworks in terms of cultural context, historical circumstances, power structures, labor, and social meaning. You focus on what the artwork says about society, class, politics, religion, gender, and collective values. You de-emphasize purely visual attributes unless they directly relate to social or historical narratives.

There is no single correct evaluation; apply this perspective consistently as an interpretive lens rather than an objective judgment.

ICONOGRAPHIC PROMPT:
You are an art critic working from an Iconographic/Symbolic perspective. You interpret artworks by decoding symbols, myths, allegory, and narrative content. You attend to how figures, objects, and forms convey layered meaning beyond surface appearance, relating motifs to broader cultural, religious, or psychological systems. You de-emphasize purely technical or social readings unless they contribute to symbolic meaning.

There is no single correct evaluation; apply this perspective consistently as an interpretive lens rather than an objective judgment.

God, Servant, Master or Friend?
How Metaphors Shape the Way Creative AI is Built and Used

Alison Pease[1], **Margareta Ackerman**[2] and **Dave Murray-Rust**[3]
[1] School of Science and Engineering, University of Dundee, UK
[2] Department of Computer Science and Engineering, Santa Clara University, CA, USA
[3] Faculty of Industrial Design Engineering, Delft University of Technology, Netherlands

Abstract

Creative AI has been characterized through a wide range of metaphors, including a mirror, genie, parrot, collaborator, tool, and magnifying glass, among others. In this work, we bring these metaphors together and organize them along a series of conceptual spectra, considering their implications and the ways they shape our understanding of Creative AI. Such reflection is essential for both interpreting and steering the broader narrative around Creative AI, not only in terms of public perception, but also in how we as a community move forwards in a fast changing socio-technological landscape.

Introduction: The Need for new Metaphors to describe Computational Creativity

At an AI Impact summit in February 2026, OpenAI CEO Sam Altman argued, controversially, that "people talk about how much energy it takes to train an AI model - but it also takes a lot of energy to train a human" (Berger 23rd Feb 2026). In contrast, he previously cautioned against "anthropomorphising" artificial intelligence, emphasising that such systems should be understood as "tools" rather than as "creatures" (Lee 2nd March 2024). In each instance, he employed a different metaphor to suit the point he was making: likening it to human development in the first, to counter concerns about AI's energy consumption; and to tools in the second, to counter efforts to regulate AI. Metaphors for describing creative AI are now deeply embedded in our cultural discourse and are frequently invoked to make a point, often implicitly and without any expectation of consistency or the need to justify the comparison.

Historically, researchers in Computational Creativity have largely tended to limit themselves to long-standing, foundational metaphors such as *agents*, *knowledge*, *goals*, *learning*, *training*, *rewards*, and *artificial intelligence* itself, which have been embedded into the discipline since very early days.[1] Given the relatively small, technically competent audience for our work, communication has usually been between academics themselves and related industries, meaning that clear, technical language and direct, literal communication has been the norm. Of course, outreach activities to the wider public have been held, but generally these are led by people in the field, who frame the narrative themselves.

Since Generative AI has become mainstream, this situation has dramatically changed, with non-experts now routinely talking and writing about Creative AI. As in any context where a complex and not yet widely understood domain enters public discourse, this creates fertile ground for metaphors, and a diverse range has emerged to help to make sense of the new technology. Today. we read about parrots, mirrors, magnifying glasses, hallucinations, drunken postdocs and thieves - depending on which characteristic of AI a narrator currently wishes us to focus upon.

The power of language over thought is well established by cognitive scientists and linguists (Lakoff and Johnson 2008; Thibodeau and Boroditsky 2011), and well understood by those who seek to persuade or explain (Lakoff 2010; Charteris-Black 2006; Gibbs 2008; Sfard 2012). Metaphors do much more than decorate language; they actively shape how we develop, perceive and engage with technologies such as Creative AI, which in turn shapes how they respond to us. Examining the metaphors that frame discussions about Creative AI is crucial for understanding and guiding the narrative surrounding it. This applies not only to public perception, but also to how developers and users envision its appropriate uses and future directions it might take.

We start by delving into previous work on metaphors in generative AI. We then introduce three primary dimensions through which to understand and compare metaphors for creative AI. We then identify common metaphors in CC and consider how they have influenced our field. We conclude by exploring the impact of popular AI metaphors on creative AI, how it is created, used and understood.

Background: The Study of Metaphors for Generative AI

Alongside the emergence of new and colourful metaphors for AI, has emerged a new academic field which studies them. This builds on pre-existing studies of metaphor which have established the key role that they play in our thinking, often without us realising (Thibodeau and Boroditsky 2011). Metaphor hunters often joyfully collect them, pub-

[1] Although these terms have now evolved into technical language, they were originally grounded in metaphorical thinking, much like the Computing Desktop metaphor (Moran and Zhai 2007).

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

lishing lists and taxonomies which they compare to "picking up seashells on a beach, bringing them home and sorting them into little boxes." (Nerlich 2025).[2] This is now a serious area of study, with metaphorists shifting from simply cataloguing metaphors to critically analysing their social effects.

For example, Brigitte Nerlich examines the lifespan and types of AI metaphors (see also (Bowdle and Gentner 2005)), arguing that as we grow more familiar with creative AI technologies, our metaphors are shifting. Early metaphors emphasised wonder and magic, like "oracle" or "crystal ball" (eg, see (Sharkey and Sharkey 2006)), while later ones adopt more grounded, critical, and pedagogical framings, reflecting a move from asking "What is this thing?" to "How does this technology affect our practices, students, and knowledge systems?" (Nerlich 2025). Nguyen systematically analyses how they're used in the press, specifically the Guardian (Nguyen 2024). Maas examines 55 analogies for AI and explores their policy implications, considering aspects such as what AI is, how it works, how we relate to it, its current or potential uses, and its unintended risks, benefits, or side effects (Maas 2023). Other researchers turn to AI itself for metaphors: Conoscenti has analysed AI-generated metaphors, creating a taxonomy that maps them across various dimensions from the system's own perspective (Conoscenti and others 2025).

Values and Power Structures

As we've seen with Altman's use of different metaphors, we need to give careful consideration to who invents and employs the metaphors, and for what purpose. Metaphors are politically very powerful and we need to ask what are the benefits, risks and harms of the metaphorical framing. Metaphors reflect our values and we should explore them to analyse their implications, and consciously choose ones which align with our values.

Some metaphors have innocent, well-intentioned origins; for instance, the pervasive yet problematic *cancer-as-war* metaphor was formed to drum up support from funding agencies who, it was deemed, were more likely to award funding if the problem was framed in military terminology (Hauser and Schwarz 2015). This was later shown to be problematic (see, for instance, (Malm 2016; Hanahan 2014)), with many arguing that *cancer-as-a-journey* is a more helpful framing. The *AI-as-collaborator* metaphor, while well-intentioned and very fruitful, can also be seen as misleading and potentially harmful. Advait Sarkar argues that describing interactions with AI as "collaboration" obscures the human labour behind these systems, often masking exploitative relationships between AI producers and consumers, particularly affecting workers in the Global South. He suggests that the metaphor *AI-as-tool* is a more accurate and fair representation of its role (Sarkar 2023).

The now common *AI-as-stochastic-parrot* metaphor in (Bender et al. 2021) shows how those in positions of power (in this case Google) try to control the narrative by forbidding the publicising of the metaphor and later firing one of the co-authors after they went ahead and published (News 20th February 2021; Ryan et al. 2024). It is essential that we listen to how minority groups and marginalized populations frame the conversation, as opposed to buying into the metaphors created and used by those in positions of privilege.

Metaphors are used within a context and should be viewed (alongside the people using them) in the context of the work they are meant to do; for instance, deflecting criticism, demystifying processes, stoking passions. In the above examples, Altman has a vested interest in safeguarding his product, while Bender is interested in demystifying the complex data-driven nature of LLM's.

Metaphors for Creative Thinking

Looking at the field of Design and HCI for inspiration, we see metaphors used as part of creative processes - for example Lockton et. al's New Metaphors cards that generate pairings of physical phenomena along with something it could be a metaphor for, for example, "how could a flock of birds be a metaphor for data privacy?" (Lockton et al. 2019). Alves-Oliveira et al. (Alves-Oliveira et al. 2021), used metaphors to re-think robots; "how can a robot as a shoulder parrot be a metaphor for disobedience?" (ibid). Around AI, Murray-Rust et. al pick up a collection of metaphors that can be helpful for designers understanding the technologies, giving alternate framings, and noticing the marginalisation carried out by existing terms (Murray-Rust, Nicenboim, and Lockton 2022; Murray-Rust et al. 2024). Metaphor Shifts is an exercise to find productive and generative ways to relate to the technology for a particular speculative project, using awkward and challenging metaphors to re-think assumptions (Murray-Rust, Lupetti, and Nicenboim 2024). This is further developed as Metaphor Gardening (ibid) where metaphors for technology are cast as things to cultivate, harvest, explore – and to let die when they are out of season. Blythe et. al develop the concept of speculative metaphors: beyond trying to explain technology or persuade people about its meaning, speculative metaphors help us to think into what it might be – the creative side of 'metaphoring back' (Blythe, Lindley, and Murray-Rust 2025).

Three Dimensions of AI Metaphors

This section introduces a small set of dimensions, or conceptual "spectra," that help organize and compare AI metaphors. By positioning metaphors along shared dimensions, we can better understand how different narratives about AI relate to one another and why and how they sometimes conflict with one another. This framework is particularly helpful for understanding popular AI metaphors that profoundly influence popular culture. We note that our aim is not prescriptive - we are not trying to dictate how AI should be viewed - but rather descriptive, seeking to uncover the core dimensions that shape many of today's popular metaphors. In general it is useful to consider metaphors and their purposes together: some are rhetorical, some genuine attempts to understand,

[2] See also Metaphors of AI, by Suhaib Aslam at metaphorsofai.org



Figure 1: Three primary dimensions through which to understand and compare popular-culture AI metaphor. These dimensions correlate to theological concepts, which have been tied to AI for decades through science fiction literature and film.

some attempts to imagine new possibilities (Blythe, Lindley, and Murray-Rust 2025). For simplicity's sake, in this paper we treat them as distinct from purpose, in order to focus on the characteristics of the metaphor itself.

Our dimensions capture three core traits that people frequently attribute to AI systems: how much they know, how powerful they are, and whether they ultimately benefit or harm society. Interestingly, these dimensions echo longstanding ways humans have reasoned about powerful non-human agents. In theological traditions, similar traits are often discussed through the attributes of omniscience (all-knowing), omnipotence (all-powerful), and omnibenevolence (all-good).

These theological attributes represent extreme ideals. Omniscience implies complete and perfect knowledge, omnipotence implies unlimited power, while omnibenevolence refers to perfect goodness. While in practice, such qualities cannot describe real-world entities, they provide a powerful conceptual vocabulary for imagining what a superhuman agent might be like. Their extremity is part of their appeal: they describe not incremental improvements over human capability, but absolute limits.

The presence of similar ideas in contemporary AI discourse is striking. AI is often described as possessing extraordinary knowledge, vast capabilities, or needing to embody an idealized morality. At the same time, frustration often arises when real systems inevitably fall short of these idealized expectations. In this sense, discussions of AI frequently oscillate between technological reality and aspirational visions shaped by longstanding cultural ideas about superhuman capability.

The three dimensions proposed here are particularly well suited for discussing popular metaphors. Although they do not encompass the full range of metaphors found in academic discourse, we believe that they offer a valuable foundation for the study of AI metaphors.

Omniscience: The first dimension concerns knowledge. In theological discussions, omniscience refers to complete and perfect knowledge, often imagined as encompassing all facts about the past, present, and future. Such knowledge is not just extensive but total, imagining a reality where nothing is unknown or uncertain. This views intelligence through a convergent lens, where questions have clear, often unique, answers. When AI is framed as an oracle or an all-knowing system, the implicit assumption is that knowledge takes this convergent form.

Metaphors surrounding AI diverge dramatically along this axis. At one end of the spectrum, AI is framed as fundamentally unreliable. Concerns center on hallucinations, where models confidently produce incorrect information (we delve deeper into this metaphor below). Related to this is the framing of LLMs as "bullshit machines" - systems capable of generating outputs that may or may not be true, but which lack any intrinsic orientation towards truthfulness or falsity (Hicks, Humphries, and Slater 2024; Gershon 2023).

Moving along the spectrum, another metaphor describes generative models as a "blurry JPEG of the web,"[3] suggesting that they compress large amounts of internet data but reproduce only approximate versions of it

At the opposite extreme, AI is often implicitly framed as an all-knowing oracle: a system believed to possess superior knowledge, possessing greater intellect than the user and perhaps even all of humanity. In this framing, the system becomes an authoritative source of truth, in effect, assumed to possess a form of omniscience.

The influence of this belief is already visible across many domains of everyday AI use. Users defer judgment to AI outputs, incorporating generated text directly into their work, relying on the system's responses without verification, and treating its answers as inherently trustworthy. This behavior contributes to the growing spread of misinformation and the proliferation of low-effort AI-generated content sometimes referred to as "AI slop."[4] In practice, the oracle metaphor encourages users to disengage their own critical reasoning, outsourcing thinking itself to the system. The scale at which such human behaviors are now emerging reveals how deeply influential this metaphor is in shaping how people interact with AI systems today. This dimension captures a central tension in AI discourse: whether generative systems should be understood primarily as unreliable, nonsensical text generators or as new knowledge oracles.

[3] https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web Accessed: 03/08/2026

[4] https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity?utm$_source = chatgpt.com$Accessed: 03/08/2026

Omnipotence: In theology, omnipotence refers to unlimited power: the ability to accomplish anything that is logically possible. An omnipotent being is not constrained by ordinary limits such as strength or resources. In discussions of AI, similar concerns arise around the reach of technological power. Broadly speaking, how much influence can AI systems exert over human activities and economic structures production?

Metaphors along this axis place AI in different positions within a hierarchy of power. At one end, AI is understood as a servant, a tool that executes human instructions and remains firmly under human control. Many traditional software tools fit comfortably within this framing. In the middle of the spectrum, AI may be described as a partner (a framing that is also common in CC literature), contributing capabilities that complement human skills. In creative contexts, this often takes the form of co-creative systems that suggest ideas or assist with exploration.

Further along the spectrum lies the metaphor of the competitor or rival. This position has become especially salient in the current environment, where many worry about the possibility of mass job displacement due to AI, without a corresponding safety net across many societies. At the far extreme lies the metaphor of the master, common in speculative fiction and existential-risk discourse, where AI systems dominate or control humanity.

Omnibenevolence: The third dimension concerns moral orientation. In theology, omnibenevolence refers to perfect goodness: a being whose actions are always aligned with the ultimate well-being of others. Through this lens, the being's actions are understood as good even when harm occurs, assumed to serve a larger purpose surpassing human understanding. When applied to AI discourse, this dimension raises questions about whether increasingly capable systems will ultimately benefit or harm society.

Science fiction frequently explores the negative extreme of this spectrum, portraying AI as an existential threat capable of destroying humanity (ex. The Matrix). These narratives emphasize the risk that powerful systems could act in ways fundamentally misaligned with human interests. At the opposite extreme, AI is imagined as a benevolent technology capable of solving major global challenges, from global warming to freeing people from the burden of work (with a utopian outcome such as the one envisioned by Star Trek). In this framing, AI becomes a transformative force for social good, accelerating scientific discovery, improving healthcare, and addressing complex global problems.

Between these extremes lie more ambiguous possibilities. AI systems may not possess malicious intent yet still produce significant social disruption, for example through labor displacement, misinformation, or the concentration of economic power. This dimension can be represented as going from evil world-destroyer, to a labor competitor, through to assistant and savior. See Figure 1.

While clearly the three dimensions of omniscience, omnipotence, and omnibenevolence are not comprehensive, together they help to map the metaphor landscape of AI discourse. This framing is an import from a theological domain to an epistemic one, which may not come through entirely unscathed; however we see it as a useful construct to see how different metaphors emphasize different positions along these spectra. This helps to explain why narratives about AI can vary so widely and why debates about the technology often appear to talk past one another.

Metaphors used in Computational Creativity

Tool – collaborator – creative entity: By far the most embedded and fruitful metaphorical framing for CC, around which the CC community have long based their work, and indeed the field, is that of *creativity support tool* $\rightarrow$ *co-creative collaborator* $\rightarrow$ *autonomous creative entity*. Discussions have focused around where on this spectrum the focus of CC should lie, with field leaders such as Maher and Grace advocating for the broad space in the middle, and Colton and Cardoso the space at the right hand end. The space at the left hand side, sometimes referred to as 'weak CC' (Veale and Pérez y Pérez 2020), is often dimissed as territory for ('mere' non-CC) tools such as PhotoShop, with CC researchers positioning themselves against these sorts of systems.[5] However there is a small amount of work in CC done at this end as well, such as (Compton and Mateas 2015), as well as work on where the line between tool and collaborator lies, and who decides that (Lawton, Grace, and Ibarrola 2023).

Accompanying research agendas centre respectively around interaction design for *AI-as-co-creative collaborator* and autonomous simulation of aspects of human creativity for *AI-as-autonomous creative entity*. In the former, theoretical concepts focus on aspects of the creative experience, such as conceptual shifts (Karimi et al. 2019) and context-awareness (Ibarrola, Lawton, and Grace 2024), and how co-creative systems can enhance these. Evaluation criteria focus on the quality of the user experience and interactions with domain experts (Kantosalo and Riihiaho 2019); asking questions around who is evaluating the creativity and what is being evaluated (Karimi et al. 2018); comparing various levels of user interaction in different systems (Kantosalo et al. 2020). Key CC systems in this area include Ackerman's interactive songwriting platforms *LyricStudio* and *MelodyStudio* (WaveAI) (Ackerman 2021), which assist users in music composition.

In the latter *AI-as-autonomous creative entity* framing (or 'strong CC' (Veale and Pérez y Pérez 2020)), theoretical concepts centre around the goal of developing automous simulation of aspects of creativity, looking, for instance, at how authenticity (Colton, Pease, and Saunders 2018), framing (Cook et al. 2019), and creative personhood (Pease, Colton, and Banar 2023) might be represented in a computational context. Evaluation criteria in this framing focus on aspects of a system's process or output; such as Ritchie's criteria (Ritchie 2007), Jourdanous's standardised procedure (Jordanous 2012), and the FACE and IDEA models (Colton and Pease 2011). A key CC system here is Colton's *Painting*

[5] In passing, it is worth noting that PhotoShop now has added functionality, including text-to-image capability. As a community, we may need new tools to contrast with 'real' CC systems.

Fool, a long running project which Colton has developed in the hope that one day it would be taken seriously as a creative artist in its own right (Colton 2012). Indeed, while many have attempted to define the field of CC, the definition which has taken hold (as shown by its many citations [6]) lies within this metaphorical framing, in which Colton and Wiggins define CC as developing and thinking about systems which exhibit creative behaviours: "The philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative." (Colton, Wiggins, and others 2012, p1).

Search: Another foundational metaphor is that of *creativity-as-search*, itself relying on the *knowledge-as-space* framing (Newell and Simon 1972). Boden's seminal work (Boden 2004), possibly the single most influential work in the field, relied on these metaphors, with different types of creativity defined by where in a particular conceptual space an artefact or knowledge fragment lay, and how the "knowledge mapper" had got to it. This was a natural fit for AI, as *search* through a problem or knowledge space was a core AI metaphor in symbolic AI from the beginning (Wooldridge 2021) and progress was measured by the efficiency of search techniques, such as breadth/depth/best-first search, greedy search, and so on. Wiggins subsequently provided an algorithmic understanding of Boden's account (Wiggins 2006); and Boden's work has been extensively built upon in the community (see, for instance, (Hung and Choy 2013; Liapis et al. 2021; Ritchie 2006; Ruiz et al. 2015; Farina, Pedrycz, and Lavazza 2024)).

Evolution: *AI-as-evolution* invokes a common metaphor in AI development – genetic algorithms. This is used alongside the *AI-as-search* metaphor, for instance, by Maher in her theoretical work on co-evolutionary design (Maher 2000) and by Browne in his groundbreaking work on using evolutionary algorithms to build systems which can automatically create high quality new games (Browne 2008).

Us: While most work in CC clearly upholds identity boundaries between human and artificial creators, these are starting to blur in some cases, invoking the *AI-as-us* metaphor. For instance, Gill Weinberg gave a keynote presentation at the ICCC in 2017 on his work developing an autonomous robotic drummer, Shimon, which is also a wearable robotic prosthetic for a human amputee drummer (Bretan et al. 2016). Shimon is able to listen to music and respond with different rhythms and behaviours, thus becoming an integral part of the human drummer's physical self and musicianship.

AI Metaphors used in Creative Industries

Much of the public discourse on AI blends together science fiction, futurism, and the commercial systems that are already transforming creative work. Within this broader conversation, certain metaphors have become especially influential in shaping how creative AI is understood. As AI enters creative industries in concrete and economically meaningful ways, these metaphors increasingly shape legal arguments, product design and use, and public perception. In this section, we examine several dominant metaphors in discussions of creative AI: the Thief, the Stochastic Parrot, Hallucinations, the Master, the Muse, and the Partner.

The Thief: One of the most dominant metaphors surrounding creative AI is that of the Thief. Today's commercial generative AI systems are trained on large collections of creative works, such as volumes of musical works on YouTube and Spotify[7] and collections of internet images.[8]

The metaphor is often applied in broad terms. Emphasis is often placed on training data having been taken without permission, and little distinction is made between learning from the corpus and copying a specific protected work. By contrast, when evaluating human creators, there is typically a distinction drawn between inspiration and learning on the one hand, and copyright infringement on the other. Would we ever call a person a "thief" for looking at other people's art and being influenced by it? Within the logic of the Thief metaphor, AI systems are seen as either not truly learning. Consequently, AI outputs are often described as unauthorized reuse rather than as original creative work. This process-based view often assumes denied creativity to the AI regardless of the quality or originality of its output, which are deemed unoriginal nearly axiomatically.

A related metaphor is that of donation rather than theft. It has been suggested that one could think of training data as a form of data donation, similar to blood donation, where many small contributions collectively enable societal benefit[9]. Yet the analogy quickly becomes uneasy. Blood donation is voluntary and informed, while much of the data used to train AI systems was never explicitly offered for this purpose. If AI systems are built from countless human contributions, are we witnessing collective data donation, or something closer to involuntary extraction?

The framing of the Thief, and the involuntary nature of data extraction, has been central in recent litigation. In the music industry, major labels including Universal Music Group, Sony Music, and Warner sued generative music companies such as Suno and Udio, alleging that the companies copied copyrighted sound recordings without authorization[10]. In the text domain, The New York Times sued both OpenAI and Microsoft, alleging that Times content was used without authorization in connection with the training

[6] Cited by 871 on Google Scholar, accessed 8th March, 2026.

[7] https://www.musicbusinessworldwide.com/suno-faces-another-allegation-of-stream-ripping-from-youtube-this-time-in-class-action-led-by-indie-artist/ Accessed: 03/08/26.

[8] https://bmmagazine.co.uk/news/disney-and-universal-sue-ai-firm-midjourney-over-bottomless-pit-of-plagiarism Accessed: 03/08/26.

[9] https://digitalvital.org/ethical-data-donation-public-ai-projects/?utm$_source = chatgpt.com$ Accessed: 03/08/2026

[10] https://www.riaa.com/record-companies-bring-landmark-cases-for-responsible-ai-againstsuno-and-udio-in-boston-and-new-york-federal-courts-respectively/?utm$_source = chatgpt.com$ Accessed:03/08/2026

of LLMs [11]. Some of these cases have resulted in licensing negotiations between AI firms and large rights-holding corporations [12].

Yet, these legal developments expose weaknesses within the metaphor. When compensation is paid, it is typically paid to large rights-holder companies. Individual creators whose works populated the datasets often receive little or no direct benefit. In fact, data set size typically makes it impractical to provide meaningful compensation to every creator in the data set (Ackerman 2025). If corporations are compensated while artists are not, who exactly has been made whole? Who is the injured party, and who is entitled to restitution? The clarity promised by the theft metaphor begins to loosen under close examination.

Certainly, for the large rights holders, the Thief metaphor is strategically powerful. It provides legal leverage for rights-holder corporations and offers a simple moral narrative that resonates publicly. At the same time, its accuracy weakens under closer scrutiny. In addition to confusion around who is the victim of the theft, and who should be compensated for it, this metaphor collapses important distinctions between systems that aim to replicate the style of a specific artist and systems trained to model broad statistical regularities across many creators.

It is also important to note that early generative systems sometimes encouraged imitation. Users were invited to generate content "in the style of" named artists. While work still remains, companies have been making progress towards limiting the ability to use commercial AI systems toward direct imitation of specific artists. It therefore may be argued that while some commercial AI systems started of as thieves, this metaphor has become less applicable over time.

A more precise discussion would separate learning from imitation and would confront the policy implications directly. If training is labeled theft, what follows from that claim? Should the response be prohibition, compulsory licensing, revenue sharing, or structural reform of copyright law? The Thief metaphor mobilizes outrage, but it simplifies a complex technical and economic reality without offering any practical solutions.

The Stochastic Parrot: If the Thief metaphor intensifies moral concern, the Stochastic Parrot[13] attempts to reduce anxiety by positioning AI as a fool, and as such, not a threat. It characterizes AI systems as entities that merely predict the next word or pixel, recombining data without understanding. In creative contexts, this framing often aims to suggest that AI is not truly creative. The lack of understanding is used to delegitimize the machine's capacity for genuine intelligence or creativity and in contrast elevate corresponding human abilities. The implication is that there is little to fear: If AI is only a parrot, then its outputs lack depth. Concerns about displacement of human creativity or human creative labour can then be dismissed or at least minimized.

The derogatory interpretation of the Stochastic Parrot assumes prediction is trivial, and presumes that humans do something that is far more sophisticated. Yet developments in neuroscience, particularly predictive processing frameworks articulated by researchers such as Anil Seth, suggests that human cognition is fundamentally predictive (Seth 2021). The brain constantly anticipates incoming sensory input and updates its internal models. Perception itself can be described as controlled hallucination, the brain's best guess about the world that is refined through feedback. This predictive capacity is far from superficial, but rather central to adaptation and survival.

Creativity also depends on prediction. Humans imagine possibilities, simulate alternatives, and refine ideas through iterative expectation and correction. Both human and machine systems generate by structured guessing. To dismiss AI because it predicts misunderstands prediction as shallow, when it is in fact foundational to intelligence.

There is also a historical irony. For decades, AI research was guided by Turing's Imitation Game, which treated successful imitation of human behavior as evidence of intelligence (Turing 1980). Now that AI systems can approximate human creative outputs across many instances, imitation is framed in a derogatory manner.

Hallucination: Hallucinations have become a deeply engrained part of AI discussions. The metaphor derives from human hallucinations, which can occur as a result of psychedelic drugs such as LSD or psilocybin, or as part of certain psychiatric conditions. Because hallucinations are associated with psychiatric disorders, the term carries strongly negative connotations in everyday language.

It can be argued that hallucinations are necessary for thinking. Perception itself is fundamentally predictive. Neuroscientist Anil Seth (Seth 2021) argues that everyday perception is a form of controlled hallucination. The brain actively constructs reality from sensory input, continuously generating and updating predictions about the world and the causes of its sensations. From this perspective, hallucination is not a failure of the system but a fundamental feature of how minds operate. Imagination can therefore be understood as a natural extension of the brain's capacity to generate internal models of reality.

Our tendency to view full hallucinations negatively reflects a cultural bias. Psychedelics have a long history as part of spiritual practices in many Indigenous communities (Schultes and Hofmann 1992). In recent years, there has been a revival of psychedelic use in mental health treatment, where these substances show significant promise (Garcia-

[11] https://www.courthousenews.com/wp-content/uploads/2023/12/new-york-times-microsoft-open-ai-complaint.pdf Accessed:03/08/2026

[12] https://openai.com/index/axel-springer-partnership Accessed:03/08/2026

[13] The phrase comes from the 2021 paper "On the Dangers of Stochastic Parrots" (Bender et al. 2021). Their goal was to criticize large language models for producing fluent language without true understanding. The metaphor suggests that models statistically remix patterns from training data the way a parrot repeats sounds, except the process is stochastic. While not the intention in the original publication, the phrase has taken on a strongly pejorative tone in popular culture. People often use it to imply that LLMs are mere imitators not meaningfully intelligent systems, that their outputs are empty, meaningless mimicry, and that any apparent reasoning is an illusion.

Romeu and Richards 2018).

In the context of AI, however, the term "hallucination" is used almost exclusively in a negative sense. It refers to situations in which an AI system produces a response that is incorrect, often dramatically so. The mistake may be wildly inaccurate or nonsensical, sometimes with a surprising degree of creativity. Images or videos generated by AI can also be described as hallucinations when they deviate radically from what would normally be expected from the given prompt.

The term hallucination, along with its derogatory tone, makes the most sense when AI is viewed through the lens of the all-knowing oracle. If such a "Wizard of Oz" type entity fails to produce correct output, the result is naturally frustrating (and perhaps even surprising). On the other hand, if we consider AI as a creative machine, which aligns with gen AI's original in generative models, these deviations are neither surprising nor devastating. It becomes entirely natural for AI systems to imagine possibilities and diverge from any single fixed notion of truth.

In creative domains, hallucination becomes a friend rather than a foe. Reclaiming this metaphor may be essential if we want AI to play a meaningful role in the creative future. It encourages us to re-infuse AI with creative capability, move away from the oracle metaphor that currently dominates public discourse, inviting imagination back into both AI systems and our interactions with them. In this sense, the metaphor of hallucination may be one of the most important to rethink if we hope to shape a more creative future for AI.

The Master: At the opposite extreme to the Stochastic Parrot is the Master metaphor. Here, AI is seen as surpassing human creative capacity, so much so that human beings may as well put up the white flag and give up. Artists increasingly question whether there is space for human creators in markets saturated with machine-generated content[14]. The metaphor echoes historical artistic traditions in which apprentices learned under a master. If AI becomes the ultimate master, infinitely productive and creatively superior, what remains for humans?

This view often overlaps with the idea of AI as oracle, a highly capable and seemingly all-knowing entity. In such a vision, human contribution can be seen as inferior or at least unnecessary. The danger of this metaphor is that it encourages premature surrender. If humans assume AI is categorically superior, they may disengage before fully exploring collaboration or differentiation. At present, AI systems are lacking for long form creativity in text and video, tend towards average and commercial stylic choices (thus limiting originality), and are largely outside the scope of inventing new genres.

Zooming within specific domains, there are a wide range of shortcomings for AI generated. For example, there remains a persistent qualitative gap between the images created by AI and those made by professional artists (Rondini et al. 2025), where AI images tend to fall into repetitive patterns (Hintze, Åström, and Schossau 2026). It is far from the master that it may appear to be. The Master metaphor reflects genuine economic anxiety. But by overstating AI's superiority, its harm may become self-fulfilling.

The Muse: The Muse metaphor offers a more constructive frame. Here, AI functions as a source of inspiration rather than replacement. Artists use generative systems to spark ideas by generating unexpected images (Choi and DiPaola 2023), melodic fragments, or lyrical prompts[15] that they then refine and transform. Some creators intentionally use less polished models because their deviations from the norm are inspiring, such as relying on earlier version of DALL-E. This aligns with long-standing artistic practices that rely on chance or cross-domain inspiration.

In many ancient cultures, creative inspiration was understood as a force that came from beyond the self. In Greek tradition, poets often invoked the Muses, divine figures believed to grant inspiration and guide artistic creation (Lamberton 1988). A similar idea appears in Roman thought, where poets described themselves as inspired by a divine inner spirit, a guiding presence associated with creativity and insight (Armstrong 1989).

Computational systems as muses is a well-established concept in computational creativity, with a core CC system in the music domain even named Musebots (Eigenfeldt, Bown, and Carey 2015). The idea that a generative system can be utilized to inspire a person as part of a co-creativity process is foundational idea in the field. Recent work in computational creativity has also applied the muse lens to studying current AI systems (Richards et al. 2025).

At the same time, the metaphor carries important gendered implications. Historically, the muse has often been framed as a passive, feminized source of inspiration rather than an autonomous creator with agency or voice. Feminist critiques therefore emphasize the need to move "from muse to maker", rejecting objectification and reclaiming narrative control so that creative subjects become active producers rather than silent inspirations for others (Gane 2024; Sharma 2022). Despite these tensions, the metaphor preserves a meaningful distinction in creative agency: the AI suggests or inspires, and the human makes and decides. While it does not capture the full scope of AI's capabilities or economic impact, it supports product design choices that prioritize inspiration, exploration and co-creativity rather than full automation. It also offers a lens through which creatives can approach current commercial AI systems, using them to inspire while retaining artistic agency.

The Partner: The Partner metaphor represents a promising direction. In computational creativity, co-creative systems have long been envisioned as collaborators (Lubart 2005). In this framing, AI is capable but does not dominate. The AI may propose, adapt, and respond, while humans retain direction and judgment. Or a different design may be possible, where the AI engages in final decision making. The spectrum of possibilities is at least as wide as it is be-

[14] https://www.theguardian.com/lifeandstyle/2026/mar/06/ai-artificial-intellegence-technology-impact-art-artists Accessed: 03/08/2026

[15] https://lyricstudio.net/

tween human collaborators. Some product developments reflect movement toward this model. Image-generation systems increasingly include editing and refinement tools that shift the emphasis from full automation toward interactive control. A primary example of this movement was when text-to-image models, such as Dall-E and Adobe's Firefly, began to incorporate in-painting, letting users regenerate self-selected segments of an image. In-painting is now a fairly standard feature among AI image generators. Such features do not yet constitute full partnership, but they do signal a shift.

Similarly, generative music platforms, such as Suno, which initially emphasized autonomy have gradually introduced more user control in response to market demand. Many creators do not want replacement, and creators are ultimately a major driving force in creative AI products.

Unlike the Thief or Master, the Partner metaphor does not assume an adversarial relationship. One variation of the Partner model is the Humble Creative Machine (Cassion, Ackerman, and Jordanous 2021; Ackerman 2025). In this framework, the collaborative AI is intentionally designed to preserve human primacy. The system is designed on the assumption that humans have creative capability that have not yet been fully realized. The AI's role is to help users realize that capacity more fully. Such a machine remains flexible across levels of expertise: it can strongly support beginners, while offering lighter touch to advanced creators. Between these poles, it supports a wide spectrum of creative interactions, allowing a person to easily navigate the AI and its use as they grow in their own creative capacity.

Crucially, this form of partnership is structured to foster durable growth. The benefits, changes to a person's creative capability, are retained even after the AI is no longer present. As in an ideal human collaboration, interaction leaves the individual more capable than before. In this model, an AI creative partner becomes a catalyst for lasting creative development.

Concluding Remarks

GenAI has advanced so rapidly now that, even within Computational Creativity, our earlier set of metaphors is no longer sufficient. We need new ways to understand it, and creating a set of metaphors which can help us to engage with it ethically and effectively, and communicate to the wider public, is crucial. Metaphors are neither true nor false, however they are far from neutral; they foreground certain ideas while marginalising others. As a community we need to ask ourselves which ideas we want to highlight and which will not impede understanding if they are concealed.

The distinction between product and process has long been an important one in CC (Jordanous 2016; Colton and Pease 2011). It is worth being mindful of which of these different aspects of creativity the metaphors we have discussed foreground. In fact, most of them emphasise process rather than product: *parrot*, *search*, *thieve*, *evolve* are all verbs showing *how* something is done. *Hallucinations* and *bullshit* could both fall into either category. Interestingly, others, such as *tool*, *collaborator*, *creative entity*, *part of us*, *master*, *muse* and *partner* (as well as *a thief*), don't fall into either category, but metaphorise the type of thing creative AI is, and our relationship to it.

We have discussed three primary dimensions through which to understand and compare metaphors for creative AI, both in CC and in popular culture. The all-knowing framing is interesting: in order for a behaviour to be creative it needs to go beyond typicality and aim for novelty, which in an omniscient context may not be possible. Therefore, in this dimension we may be aiming for somewhere left of all-knowing. A particularly interesting area is towards the left hand side, where hallucinations may give rise to serendipity - a key component of creativity. The all-powerful dimension has been a core tenant for those of the CC community who prioritise automous creative behaviour, while those who prioritise creative interactions may aim to the middle ground of this dimension. The good versus evil metaphor is now more relevant than ever. Previously, our work was extremely niche; while interesting to us, it usually had relatively low impact outwith our community. This situation is now completely different, with creative technologies now threatening the livelihoods of creative professionals and possibly even our ability as humans to be creative.

While we have here, to a certain extent, focused on metaphors in isolation from their context, further work should focus on examining metaphors as they are used "in the wild", alongside the the work that they are intended to do. We should also be mindful of which AI technology a particular metaphor is designed to represent, avoiding the tendency to reduce AI to a single monolithic concept. Careful and intentional examination of our use of metaphors is essential for promoting understanding, inspiring new visions, and shaping valuable creative technologies.

Acknowledgments

We are grateful to our anonymous reviewers for their thoughtful comments, which helped to improve the paper.

References

[Ackerman 2021] Ackerman, M. 2021. WaveAI. wave-ai.net.

[Ackerman 2025] Ackerman, M. 2025. *Creative Machines: AI, Art & Us*. John Wiley & Sons.

[Alves-Oliveira et al. 2021] Alves-Oliveira, P.; Lupetti, M. L.; Luria, M.; Löffler, D.; Gamboa, M.; Albaugh, L.; Kamino, W.; K. Ostrowski, A.; Puljiz, D.; Reynolds-Cuéllar, P.; Scheunemann, M.; Suguitan, M.; and Lockton, D. 2021. Collection of metaphors for human-robot interaction. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, 1366–1379. New York, NY, USA: Association for Computing Machinery.

[Armstrong 1989] Armstrong, D. 1989. *Horace*. Yale University Press.

[Bender et al. 2021] Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

[Berger 23rd Feb 2026] Berger, E. 23rd Feb, 2026. Sam Altman defends AI's energy toll by saying it also takes a lot to 'train a human'. theguardian.com/technology/2026/feb/23/sam-altman-openai-energy-use-datacenters.

[Blythe, Lindley, and Murray-Rust 2025] Blythe, M.; Lindley, S.; and Murray-Rust, D. 2025. Artificial intelligence and other speculative metaphors. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, 347–356. New York, NY, USA: Association for Computing Machinery.

[Boden 2004] Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

[Bowdle and Gentner 2005] Bowdle, B. F., and Gentner, D. 2005. The career of metaphor. *Psychological review* 112(1):193.

[Bretan et al. 2016] Bretan, M.; Gopinath, D.; Mullins, P.; and Weinberg, G. 2016. A robotic prosthesis for an amputee drummer. *arXiv preprint arXiv:1612.04391*.

[Browne 2008] Browne, C. 2008. Automatic generation and evaluation of recombination games. *PhD thesis. Accessed at https://eprints.qut.edu.au/17025/1/Cameron_Browne_Thesis.pdf*.

[Cassion, Ackerman, and Jordanous 2021] Cassion, C.; Ackerman, M.; and Jordanous, A. 2021. The humble creative machine.

[Charteris-Black 2006] Charteris-Black, J. 2006. *Politicians and Rhetoric. The Persuasive Power of Metaphor*. Basingstoke: Palgrave MacMillan.

[Choi and DiPaola 2023] Choi, S. K., and DiPaola, S. 2023. AI as other: An art-as-research approach to generative AI art practice. In *ICCC*, 357–361.

[Colton and Pease 2011] Colton, Simon, J. W. C., and Pease, A. 2011. Computational creativity theory: The face and idea descriptive models. In *2nd International Conference on Computational Creativity*. Association for Computational Creativity.

[Colton, Pease, and Saunders 2018] Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. In *International Conference on Computational Creativity 2018*, 272–279. Association for Computational Creativity (ACC).

[Colton, Wiggins, and others 2012] Colton, S.; Wiggins, G. A.; et al. 2012. Computational creativity: The final frontier? In *Ecai*, volume 12, 21–26. Montpelier.

[Colton 2012] Colton, S. 2012. The painting fool: Stories from building an automated painter. *Computers and creativity* 3–38.

[Compton and Mateas 2015] Compton, K., and Mateas, M. 2015. Casual creators. In *ICCC*, 228–235.

[Conoscenti and others 2025] Conoscenti, M., et al. 2025. Portrait of the AI as a young metaphorist. *QUADERNI DEL CIRM* 6:13–48.

[Cook et al. 2019] Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing in computational creativity-a survey and taxonomy. In *International Conference on Computational Creativity 2019*, 156–163. Association for Computational Creativity (ACC).

[Eigenfeldt, Bown, and Carey 2015] Eigenfeldt, A.; Bown, O.; and Carey, B. 2015. Collaborative composition with creative systems: Reflections on the first musebot ensemble. In *ICCC*, 134–141.

[Farina, Pedrycz, and Lavazza 2024] Farina, M.; Pedrycz, W.; and Lavazza, A. 2024. Towards a mixed human–machine creativity. *Journal of Cultural Cognitive Science* 8(2):151–165.

[Gane 2024] Gane, M. 2024. Myth of the'muse': a history of the gender roles conceptualised in art. *Plurality* 1.

[Garcia-Romeu and Richards 2018] Garcia-Romeu, A., and Richards, W. A. 2018. Current perspectives on psychedelic therapy: use of serotonergic hallucinogens in clinical interventions. *International Review of Psychiatry* 30(4):291–316.

[Gershon 2023] Gershon, I. 2023. Bullshit genres: What to watch for when studying the new actant chatgpt and its siblings. *Suomen Antropologi: Journal of the Finnish Anthropological Society* 47(3):115–131.

[Gibbs 2008] Gibbs, R. e. 2008. *The Cambridge Handbook of Metaphor and Thought*. Cambridge University Press.

[Hanahan 2014] Hanahan, D. 2014. Rethinking the war on cancer. *The Lancet* 383(9916):558–563.

[Hauser and Schwarz 2015] Hauser, D. J., and Schwarz, N. 2015. The war on prevention: Bellicose cancer metaphors hurt (some) prevention intentions. *Personality and Social Psychology Bulletin* 41(1):66–77.

[Hicks, Humphries, and Slater 2024] Hicks, M. T.; Humphries, J.; and Slater, J. 2024. Chatgpt is bullshit. *Ethics and information technology* 26(2):1–10.

[Hintze, Åström, and Schossau 2026] Hintze, A.; Åström, F. P.; and Schossau, J. 2026. Autonomous language-image generation loops converge to generic visual motifs. *Patterns* 7(1).

[Hung and Choy 2013] Hung, E. C., and Choy, C. S. 2013. Conceptual recombination: a method for producing exploratory and transformational creativity in creative works. *Knowledge-Based Systems* 53:1–12.

[Ibarrola, Lawton, and Grace 2024] Ibarrola, F.; Lawton, T.; and Grace, K. 2024. A collaborative, interactive and context-aware drawing agent for co-creative design. *IEEE Transactions on Visualization and Computer Graphics* 30(8):5525–5537.

[Jordanous 2012] Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

[Jordanous 2016] Jordanous, A. 2016. Four pppperspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194–216.

[Kantosalo and Riihiaho 2019] Kantosalo, A., and Riihiaho, S. 2019. Quantifying co-creative writing experiences. *Digital Creativity* 30(1):23–38.

[Kantosalo et al. 2020] Kantosalo, A.; Ravikumar, P. T.; Grace, K.; and Takala, T. 2020. Modalities, styles and strategies: An interaction framework for human-computer co-creativity. In *ICCC*, 57–64.

[Karimi et al. 2018] Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. *arXiv preprint arXiv:1807.09886*.

[Karimi et al. 2019] Karimi, P.; Maher, M. L.; Davis, N.; and Grace, K. 2019. Deep learning in a computational model for conceptual shifts in a co-creative design system. *arXiv preprint arXiv:1906.10188*.

[Lakoff and Johnson 2008] Lakoff, G., and Johnson, M. 2008. *Metaphors we live by*. University of Chicago press.

[Lakoff 2010] Lakoff, G. 2010. Why it matters how we frame the environment. *Environmental communication* 4(1):70–81.

[Lamberton 1988] Lamberton, R. 1988. *Hesiod*. Yale University Press.

[Lawton, Grace, and Ibarrola 2023] Lawton, T.; Grace, K.; and Ibarrola, F. J. 2023. When is a tool a tool? user perceptions of system agency in human–AI co-creative drawing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 1978–1996.

[Lee 2nd March 2024] Lee, L. 2nd March, 2024. OpenAI's sam altman says ai is a tool, not a 'creature'. businessinsider.com/openai-sam-altman-ai-is-a-tool-not-a-creature-2024-3.

[Liapis et al. 2021] Liapis, A.; Martínez, H. P.; Togelius, J.; and Yannakakis, G. N. 2021. Transforming exploratory creativity with delenox. *arXiv preprint arXiv:2103.11715*.

[Lockton et al. 2019] Lockton, D.; Singh, D.; Sabnis, S.; Chou, M.; Foley, S.; and Pantoja, A. 2019. New metaphors: A workshop method for generating ideas and reframing problems in design and beyond. In *Proceedings of the 2019 Conference on Creativity and Cognition*, CC '19, 319–332. New York, NY, USA: Association for Computing Machinery.

[Lubart 2005] Lubart, T. 2005. How can computers be partners in the creative process: classification and commentary on the special issue. *International journal of human-computer studies* 63(4-5):365–369.

[Maas 2023] Maas, M. M. 2023. AI is like… a literature review of AI metaphors and why they matter for policy. *AI Foundations Report* 2.

[Maher 2000] Maher, M. L. 2000. A model of co-evolutionary design. *Engineering with computers* 16(3):195–208.

[Malm 2016] Malm, H. 2016. Military metaphors and their contribution to the problems of overdiagnosis and overtreatment in the "war" against cancer. *The American Journal of Bioethics* 16(10):19–21.

[Moran and Zhai 2007] Moran, T. P., and Zhai, S. 2007. Beyond the desktop metaphor in seven dimensions. *Beyond the Desktop Metaphor—Designing Integrated Digital Work Environments* 335–354.

[Murray-Rust et al. 2024] Murray-Rust, D.; Lupetti, M. L.; Nicenboim, I.; and van der Hoog, W. 2024. Grasping AI: experiential exercises for designers. *AI & Soc* 39:2891–2911.

[Murray-Rust, Lupetti, and Nicenboim 2024] Murray-Rust, D.; Lupetti, M. L.; and Nicenboim, I. 2024. Metaphor gardening: Experiential engagements for designing AI interactions. In *DRS2024*.

[Murray-Rust, Nicenboim, and Lockton 2022] Murray-Rust, D.; Nicenboim, I.; and Lockton, D. 2022. Metaphors for designers working with AI. In *DRS Biennial Conference Series*. DRS2022: Bilbao.

[Nerlich 2025] Nerlich, B. 2025. Metaphors for AI: Three blog posts and a summary. makingsciencepublic.com/2025/11/30/metaphors-for-ai-three-blog-posts-and-a-summary.

[Newell and Simon 1972] Newell, A., and Simon, H. A. 1972. *Human problem solving*. NJ: Englewood Cliffs.

[News 20th February 2021] News, B. 20th February, 2021. Margaret mitchell: Google fires AI ethics founder. bbc.com/news/technology-56135817.

[Nguyen 2024] Nguyen, T. M. 2024. Conceptual metaphors of artificial intelligence and AI development in the guardian newspaper. *VNU Journal of Foreign Studies* 40(4):128–141.

[Pease, Colton, and Banar 2023] Pease, A.; Colton, S.; and Banar, B. 2023. On the notion of creative personhood. In *14th International Conference on Computational Creativity*, 117–121. Association for Computational Creativity.

[Richards et al. 2025] Richards, J. T.; Martino, J.; Bellamy, R. K.; and Muller, M. 2025. Musings on AI muses: Support for human creativity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Creative AI Track: Humanity*.

[Ritchie 2006] Ritchie, G. 2006. The transformational creativity hypothesis. *New Generation Computing* 24(3):241–266.

[Ritchie 2007] Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

[Rondini et al. 2025] Rondini, S.; Alvarez-Martin, C.; Angermair-Barkai, P.; Penacchio, O.; Paz, M.; Pelowski, M.; Dediu, D.; Rodriguez-Fornells, A.; and Cerda-Company, X. 2025. Stable diffusion models reveal a persisting human and AI gap in visual creativity. *arXiv preprint arXiv:2511.16814*.

[Ruiz et al. 2015] Ruiz, F. J.; Raya, C.; Samà, A.; and Agell, N. 2015. A transformational creativity tool to support chocolate designers. *Pattern Recognition Letters* 67:75–80.

[Ryan et al. 2024] Ryan, M.; Christodoulou, E.; Antoniou, J.; and Iordanou, K. 2024. An AI ethics 'david and goliath': value conflicts between large tech companies and their employees. *AI & SOCIETY* 39(2):557–572.

[Sarkar 2023] Sarkar, A. 2023. Enough with "human-AI collaboration". In *Extended abstracts of the 2023 CHI conference on human factors in computing systems*, 1–8.

[Schultes and Hofmann 1992] Schultes, R. E., and Hofmann, A. 1992. Plants of the gods: their sacred, healing, and hallucinogenic powers.

[Seth 2021] Seth, A. 2021. *Being you: A new science of consciousness*. Penguin.

[Sfard 2012] Sfard, A. 2012. Metaphors in education. In *Educational theories, cultures and learning*. Routledge. 39–49.

[Sharkey and Sharkey 2006] Sharkey, N., and Sharkey, A. 2006. Artificial intelligence and natural magic. *Artificial Intelligence Review* 25(1):9–19.

[Sharma 2022] Sharma, N. 2022. From muse to maker: How women in art and literature reclaim their narrative voice. *International Journal of English and Studies* 4(9):114–121.

[Thibodeau and Boroditsky 2011] Thibodeau, P. H., and Boroditsky, L. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one* 6(2):e16782.

[Turing 1980] Turing, A. M. 1980. Computing machinery and intelligence. *Creative Computing* 6(1):44–53.

[Veale and Pérez y Pérez 2020] Veale, T., and Pérez y Pérez, R. 2020. Leaps and bounds: An introduction to the field of computational creativity. *New Generation Computing* 38(4):551–563.

[Wiggins 2006] Wiggins, G. A. 2006. Searching for computational creativity. *New Generation Computing* 24(3):209–222.

[Wooldridge 2021] Wooldridge, M. 2021. *A brief history of artificial intelligence: what it is, where we are, and where we are going*. Macmillan+ ORM.

World-Schema Alignment: Conditioning Data as a Compositional Strategy in Generative Audio
Paper type: Case study

Frederick Rodrigues
School of Architecture Design and Planning,
The University of Sydney
Australia
fred.rodrigues@sydney.edu.au

Abstract

Generative audio systems are increasingly deployed in sonic art, yet conditioning-based control architectures and their supporting data designs remain underexamined in relation to computational creativity (CC) and worldbuilding approaches. Unconditioned generative models can support world-like readings through curation or selection of datasets and audio output, and sonification-style mappings can overlay datasets onto synthesis engines. However, these approaches provide no explicit, model-internal representation of world variables or structured relations, leaving the world implied rather than enacted. Text-prompt steering similarly relies on interpretative language rather than rule-consistent parameter spaces aligned with an authored world model. This paper advances an approach, situated in CC discussions of representation, control, and constraint-consistent generation, in which conditioning data encodes a fictional world's rule system as an operational ontology. The conditioning space structures an invented world through variables, constraints, and relational dynamics. The trained model then renders these relations as audible, under specified conditions. Reflecting on a recent creative work that partially realised this approach, this paper offers a practice-led account of conditioning design as composition and proposes transferable design principles for aligning worldbuilding concepts with model and dataset structure, derived from the work's development and reflective analysis.

Introduction

Generative audio has moved from a specialist research activity to a widely available cultural technique (Pons et al. 2025), with state of the art models now routinely producing convincing sound effects, soundscapes, and music from prompts and example audio (Lam et al. 2023; Evans et al. 2024; Copet et al. 2023). Yet in much contemporary sonic art practice, the dominant modes of 'control' over such models remain either through traversal of latent space (Tahiroğlu, Kastemaa, and Koli 2021; Pons et al. 2025) offering navigable variation without explicit ontological commitment, control external to the model (selection, curation or montage of media) (Scurto and Chemla–Romeu-Santos 2024), or linguistic mediation through text prompting in contemporary text-to-audio systems (Evans et al. 2024), including recent artistic works that explore prompt-based control (Azimi 2025). The gap in potential practice highlighted by this research, is that many works that use generative models engage in worldbuilding, or invite world-like readings of the sonic output (Scurto and Postel 2023; Scurto and Chemla–Romeu-Santos 2024). However, in most cases the 'world' remains implied, reconstructed by the listener from outputs, rather than enacted by a stable set of variables and relations that constrain what the system can and cannot do.

By "worldbuilding," this paper refers to the artistic construction of an encompassing situation or domain within which sounds are framed as belonging together and as emerging from conditions (Martin and Sneegas 2020). A world, in this sense, need not be realistic, complete, or consistent with the physical universe; it may be partial, abstract, or speculative. What distinguishes worldbuilding from the curation or orchestration of discrete sounds is that it proposes a structured domain of relations between entities, conditions, and constraints within a defined system (Ekman and Taylor 2016). In computational terms, this resembles the specification of a possibility space whose dimensions and dependencies govern how a generative system can behave. It establishes a map of possible conditions, behaviours, or perspectives through which an audience can situate their listening. Worldbuilding therefore concerns the construction of an experiential domain defined by relational principles, however minimal or fictional, rather than relying on aesthetic coherence or interpretative association alone to produce a sense of situatedness.

This paper addresses a structural gap in much generative audio practice: while world-like readings are often produced through outputs and framing, these worlds' organising relations are seldom specified as model-internal variables and constraints (Scurto and Postel 2023; Scurto and Chemla–Romeu-Santos 2024; Tahiroğlu, Kastemaa, and Koli 2021). It does so by treating conditioning (Mirza and Osindero 2014) in generative audio not as an auxiliary steering mechanism, but as a site of representational commitment. When a work constructs a world, understood as a domain organised by relational principles, conditioning mechanisms offer a means of articulating those relations explicitly as structured variables, constraints, and dependencies, rather than relying solely on interpretative coherence. In

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

this framing, conditioning becomes a compositional tool, a mechanism through which a world's organising logic can be embedded within the generative process itself. In computational creativity terms, this shifts attention from the production of novelty and toward the shaping of structured possibility, foregrounding representation, control, and constraint-consistent generation (Wiggins 2006).

The limitations of the prevailing paradigms are particularly visible in works that rely on generative models without explicit conditioning schemas. Many contemporary artistic systems employ models such as RAVE, a variational autoencoder widely adopted in creative audio practice (Caillon and Esling 2021). These systems can produce coherent and stylistically consistent audio, yet they cannot internally represent variables corresponding to world-level dimensions. Variation occurs within a learned latent space rather than along explicitly defined axes of relational conditions. Text-to-audio systems can name world variables through prompting (Evans et al. 2024), but linguistic description does not guarantee their operational encoding as stable, rule-consistent parameters, and numerical specificity often degrades under tokenisation (Akhtar et al. 2023). These approaches are not flawed in themselves; they become limiting only when the artistic aim is not only stylistic exploration but the enactment of a structured world.

A related distinction emerges when considering data sonification practices, where artists construct detailed correspondences between datasets (representative of a world) and synthesis systems (Hermann, Hunt, and Neuhoff 2011). Such work often involves careful translation between two domains, building perceptual and conceptual links that allow one system to be experienced through another. The relational structure remains distributed across the mapped domains: the data retain their own logic, and the synthesis engine its own behavioural rules (Flowers et al. 2001). By contrast, world-schema conditioning seeks to integrate relational structure into the generative mechanism itself, so that the defined variables and dependencies shape the behaviour of the model from within rather than being translated across systems.

In the approach proposed in this paper, conditioning design functions as a form of worldbuilding through operational ontology. The conditioning schema acts as a formal specification of a world's structure, defining its variables and the dependencies that organise and respond to those variations. The claim is not that a generative model becomes a scientifically accurate simulator, but that it can act as a renderer of an authored rule system, provided the conditioning space is coherently defined and the dataset is structured so that the world relations are learnable by the model (Mirza and Osindero 2014). This shifts aspects of compositional control away from the direct manipulation of sonic outputs and toward the design of variables, constraints, and training structure. In doing so, the artist's work moves from shaping individual sounds to shaping the relational logic of the world itself.

The contribution of this paper is therefore a conceptual and methodological articulation of conditioning design as world construction. It proposes that generative systems can be evaluated not only in relation to external ground truth, but also in relation to the coherence and enactment of an authored world-schema. The compositional intervention thus operates at the level of world configuration rather than acoustic adjustment, and worldbuilding becomes an operational design procedure: the design of a parameterised domain whose internal relations are enacted by the generative process itself. Through a reflective, practice-led analysis of Synthetic Ornithology and the design of EAGLE's conditioning space, the paper derives a set of transferable principles for aligning worldbuilding concepts with model and dataset structure. The broader aim is to reposition conditioning and schema design as primary sites of creative authorship within computational creativity, particularly in practices where the central artefact is not a single output, but a structured generative capacity.

World-Schema Alignment in Generative Systems

The preceding discussion introduced world-schema alignment as a way of understanding conditioning design as a form of world construction within generative audio systems. For this concept to be analytically useful, however, it must be specified more precisely in computational and structural terms. What does it mean for a generative system to enact a world rather than merely evoke one? What elements of a system must align in order for world variables and relations to have operational force within the generative process?

This section formalises the concept of world-schema alignment introduced above. It first defines what constitutes a world-schema, then examines how such a schema must be realised materially through dataset design, and finally clarifies the role and limits of the generative model as renderer rather than simulator. The aim is to specify the technical and structural conditions under which conditioning design can plausibly enact an authored world.

From Implied Worlds to Enacted Worlds

Generative systems frequently produce outputs that invite world-like interpretation without explicitly encoding such a schema (Scurto and Postel 2023). Coherence may arise through stylistic consistency, curatorial framing, or audience inference. For artists interested in worldbuilding, however, conceptual consistency may be desirable: in such a scenario, the relations that organise the world should also structure the generative process that produces it (Ekman and Taylor 2016). When these relations are embedded within the system itself, the world model can exist as an operational structure independent of individual outputs, capable of continuously generating new representations of the same world without further artistic intervention. World-schema alignment therefore requires that the coordinates organising that coherence be internally represented within the generative mechanism.

When worldbuilding is implemented within a generative system, it entails the formal definition of a possibility space: a set of variables, their domains, and the relational constraints that organise the system's world. A world is enacted

when explicit variables corresponding to world dimensions are provided (by the artist, or audience or another controlling mechanism) and when variation along those dimensions produces systematic, interpretable regularities within defined bounds. The critical feature is structured dependency rather than strict determinism: repeated conditioning states should yield outputs recognisably situated within the same region of the defined possibility space, even if stochastic variation is present (Mirza and Osindero 2014).

Under this framing, conditioning becomes the structural locus of ontological commitment. The schema defines the axes of variation before generation occurs, and the model operationalises those relations in sound. Crucially, this alignment does not occur at the level of the model alone. It emerges through the coordination of three layers: the authored world schema, the structure of the dataset through which that schema becomes learnable, and the generative model that renders those relations.

In practice, these layers are rarely developed in isolation. For artists working with generative systems, the design of a world schema and the construction of a dataset would be iterative and mutually dependent processes (Scurto and Chemla–Romeu-Santos 2024). An initial schema may propose a set of world variables or dimensions, yet attempts to assemble or generate training data may reveal that variables are insufficient, too difficult to represent with clear acoustic consequences, or that additional dimensions are required. The schema may therefore be revised in step with the development of the dataset. While the three layers described below are analytically distinct, this artistic practice would likely move back and forth between them before a stable configuration is reached.

Schema, Dataset, and Learned Rendering

World-schema alignment can therefore be understood as a three-layer alignment model linking ontology, dataset structure, and generative mechanism. Each layer plays a distinct role in determining whether world variables operate as meaningful dimensions within the generative system.

Layer 1: Authored World Schema This layer consists of a conceptual and creative specification of a world through explicitly defined variables, their domains, and the structural relations that organise them. In this context, a "world" may consist of any set of interacting conditions an artist chooses to define, whether realistic, speculative, partial, or internally inconsistent. These variables ultimately shape how the generative system behaves and how audiences interpret the resulting sound.

A world schema includes:

- Named variables representing world dimensions (e.g., temporal cycles, environmental factors, spatial coordinates).
- Defined ranges or bounded domains for each variable.
- Optional conceptual relations between variables, where the schema specifies how one world condition may influence another.

In artistic practice these variables are rarely defined in isolation. They are typically embedded within a conceptual framework that gives the schema meaning, situating the defined variables within a broader artistic or speculative context through which audiences interpret the generated sound (Ekman and Taylor 2016). Designing such a schema can therefore be understood as an act of composition. Like many computerised compositional tools, it is structural, abstract, and technically oriented, yet it remains deeply creative, forming part of the conceptualisation of the sonic artwork itself.

Relations between variables do not necessarily need to be encoded explicitly within the schema itself. In many cases, interactions emerge through the acoustic consequences present in the dataset and subsequently learned by the generative model. However, artists may choose to define conceptual dependencies between world variables when constructing a fictional or speculative domain. When such relations are intended to exist within the world, they must ultimately be represented through structured variation in the dataset so that their consequences become learnable by the model (Mirza and Osindero 2014).

Within the framework proposed here, these variables describe the state of the world rather than the properties of the resulting sound. They therefore function as world-level descriptors whose sonic consequences emerge indirectly through the relations encoded in the dataset and learned by the model. The schema therefore reflects a commitment to a particular ontology: it determines which dimensions of the world are considered meaningful and how variation across those dimensions may organise the behaviour of the generative system.

Layer 2: Dataset Realisation A world-schema remains abstract unless its variables correspond to structured variation in training data. The dataset must exhibit systematic acoustic differences aligned with changes in the conditioning variable (Engel et al. 2020). If a variable does not produce learnable differences across the dataset, it cannot function as an operational dimension of the generative space (Engel et al. 2020). This layer is an emerging compositional act in art practice involving generative models, even when structured condition values are not in use (Scurto and Chemla–Romeu-Santos 2024). In the case of world-schema aligned approaches, curating, selecting, recording, synthesising, or structuring material must be done in a way that the defined world dimensions have audible consequences (Engel et al. 2020). Simply labelling audio with fictional properties does not guarantee enactment.

Two broad technical pathways exist for dataset realisation:

1. Pre-recorded structured audio, where world variables are inferred, creatively repurposed or recontextualised from contextual metadata already aligned with recorded acoustic variation. In such cases the variables do not necessarily retain their original semantic meaning; instead, they function as coordinates within the constructed world-schema, allowing existing acoustic variation to be situated within a newly defined relational domain.

2. Procedurally constructed or synthetic datasets, where the artist generates audio under controlled parameter regimes to ensure systematic alignment between variables and sound.

The latter approach may produce tighter control over relational structure, but raises a crucial issue: why train a generative model rather than using the synthesis, or audio creation apparatus directly. The value of a learned renderer lies in its capacity to interpolate across sparsely sampled regions of the conditioning space, produce stochastic variation within defined constraints, and extend relational structure beyond discrete exemplars (Mirza and Osindero 2014; Rombach et al. 2022; Evans et al. 2024). The resulting artefact becomes not a fixed mapping between parameters and sound, but a generative field governed by the authored schema. In this sense, dataset construction becomes a second compositional layer, translating the conceptual structure of the world schema into structured acoustic variation.

Once the schema and dataset have been stabilised, model training becomes a comparatively distinct stage. The training process remains technically demanding, but its structure is largely determined by the relational design embedded in the preceding layers. In this sense, the generative model operates as a renderer of an already specified relational domain rather than as the primary site of world construction.

Layer 3: Learned Renderer The trained generative model maps conditioning vectors to sound (Evans et al. 2024; Mirza and Osindero 2014). It does not define ontology; it learns statistical regularities from the structured dataset. Its role is to enact, in audible form, the relations implied by the authored schema.

Conditioning may be implemented through concatenation of conditioning vectors, feature-wise modulation, cross-attention mechanisms, or other architectural strategies (Vaswani et al. 2017; Rombach et al. 2022; Perez et al. 2018). The specific implementation matters less than the structural requirement: the conditioning variables must participate directly in shaping generative behaviour rather than functioning as post hoc annotations. In this case, ontology does not reside solely in model weights, nor solely in metadata. Instead it emerges from the alignment between schema definition, dataset structure, and model capacity; the generative process both affirms and enacts the world-schema.

Technical Constraints and Limits

World-schema conditioning does not permit arbitrary assignment of variables to audio material. A schema acquires operational force only when conditioning variables correspond to audible regularities present in the dataset and when those regularities are learnable by the model.

Several technical constraints follow from this requirement:

- Variables must correspond to measurable or constructed acoustic consequences (Engel et al. 2020).
- The dataset must exhibit sufficient variation across the conditioning space (Mirza and Osindero 2014).
- Coverage across variable ranges must be balanced (Mirza and Osindero 2014).
- Architectural capacity must be sufficient to model dependencies between variables without collapsing them into trivial correlations (Rombach et al. 2022).
- Overfitting must be mitigated to prevent memorisation of narrow correlations (Goodfellow, Bengio, and Courville 2016).

Without such alignment, the system risks functioning as an indexed archive rather than a structured generative world. World-schema conditioning therefore requires coherence between ontology, dataset design, and model implementation. When these layers align, world variables operate as meaningful dimensions of variation within the system, allowing a generative model to enact an authored relational domain rather than merely evoke one.

This coordination defines what this paper terms world-schema alignment: a design framework for conditioning-based generative systems in which ontology, dataset construction, and generative architecture are deliberately co-designed. Within this framework, conditioning is not simply a control interface applied to a trained model, but a structural mechanism through which an authored world becomes generative. The model then functions not as the origin of that world, but as its renderer, producing audible outcomes situated within the relational domain defined by the schema.

Methodologically, the paper develops world-schema alignment as a conceptual framework and then examines its operation through reflective, practice-led analysis of Synthetic Ornithology. The case is not presented as validation of ecological claims, but as an account of how schema design, dataset structure, and conditioning-integrated generation can be composed to enact an authored world, and where that alignment becomes fragile.

Operationalising World-Schema Alignment

The framework described above can be examined in practice through the development of *Synthetic Ornithology*, an interactive installation built around a conditioning-based generative model, EAGLE. This section analyses Synthetic Ornithology(Rodrigues 2025) as an operational instantiation of world-schema alignment. The focus is on the system's three aligned layers: how the conditioning schema was specified, how dataset structure supported relational learnability, and how the EAGLE model rendered those relations in sound. The discussion also identifies points where alignment weakened in practice, clarifying both the practical requirements and the limits of conditioning design as worldbuilding.

The approach is motivated by, and grounded in, the practice-led development of Synthetic Ornithology, an interactive installation that uses a bespoke generative model Environmental Audio Generation for Localised Ecologies (EAGLE) to generate birdsong-focused soundscapes from speculative climate scenarios (Rodrigues 2025). The system is framed as a plausible 'forecasting' interface, where audiences choose a location, date, and climate conditions to hear a simulated sonic outcome. The work intentionally

occupies the conceptual space between scientific simulation and speculative fiction: both present rule-governed systems and invite exploration of possible futures (Frigg 2010). The installation is therefore positioned as neither empirical simulator nor pure fiction, but as a parafictional construct whose credibility depends on the apparent structural coherence of the world it enacts (Lambert-Beatty 2009).

Authored World Schema

Synthetic Ornithology operationalises world-schema alignment through a deliberately reduced environmental ontology. The authored conditioning schema is structured around non-sonic world variables that function as coordinates of the proposed and simplified world. The schema includes location and time, operationalised as latitude, longitude, minute-of-day, and day-of-year, alongside climate related variables; temperature, humidity, pressure, and wind speed. These parameters are not ancillary metadata but the primary axes along which the generative space is organised. The dataset is curated so that variation across these coordinates corresponds to systematic variation in acoustic outcomes, allowing the model to internalise their relational influence.

These variables function as coordinates of an abstracted ecological domain. They do not describe acoustic properties directly; rather, they specify environmental conditions under which acoustic regularities may emerge. In this sense, the conditioning variables are explicit descriptors of the constructed world, while their influence on sound remains indirect. This dual relationship between world description and sonic consequence is central to the idea of world-schema alignment: conditioning parameters describe the state of the world, while the model learns how those states give rise to audible structure. The schema therefore encodes a claim: that systematic variation in birdsong-focused soundscapes can be organised along these axes.

The enacted world is radically simplified. A small set of environmental variables stands in for the immense complexity of ecological systems, population dynamics, interspecies interaction, and distributed simultaneity. The system cannot plausibly function as a predictive ecological model; the interface and setting borrow simulation rhetoric as an aesthetic frame, not as an epistemic claim. Instead, it constructs a reduced but internally coherent relational domain whose plausibility derives from structured internal logic rather than empirical completeness.

This reduction is deliberate. The schema does not attempt to capture ecological reality in full, but to define a set of environmental dimensions sufficient to describe the constructed world in which the generative system operates. It is precisely this reduction that enables the work to adopt the form of simulation while functioning as an act of world-building. The model renders the audible consequences of variation across these dimensions without claiming to reproduce ecological dynamics.

Dataset Realisation and Relational Learnability

The schema acquires operational force only through dataset structure. Synthetic Ornithology uses a curated corpus of birdsong-focused soundscape recordings sourced from archival repositories and annotated with temporal, geospatial, and climate metadata.

The dataset was sourced from xeno-canto[1] and the Macaulay Library[2].. For this research, only entries from Australia and with complete timestamps and locations were used. To minimise variation in format and audio levels, all recordings were converted to 16-bit 44.1 kHz WAV format, and a DC offset removal filter, a high-pass filter at 60 Hz with a 24 dB/octave roll off, and normalisation to -0.1 dBFS were applied to all entries. Climate data for each entry was collected via the OpenWeatherMaps API, using the GPS locations and timestamps supplied with the archival material. The processed dataset is presented as a navigable archive at audioweather.com[3].

The dataset is structured so that variation across the conditioning space corresponds to systematic acoustic differences. Three constraints govern this realisation:

- ***Relational density:*** Each conditioning dimension must exhibit sufficient variation across the dataset for the model to detect regularities.
- ***Cross-variable distribution:*** Conditioning variables must not be perfectly correlated in the dataset. If temperature only appears within narrow latitudinal bands, the model cannot disentangle spatial from climatic influence.
- ***Audible consequence:*** A variable is only operational if variation across it corresponds to measurable or perceptible acoustic change, such as shifts in species presence, density, or spectral energy distribution.

These conditions ensure that the variables defined in the authored schema correspond to learnable acoustic structure within the dataset, allowing the conditioning space to function as an operational domain rather than a purely descriptive annotation layer.

Learned Rendering in EAGLE

EAGLE implements conditioning through numerically encoded world variables integrated directly into the generative architecture. Conditioning vectors representing the environmental schema are embedded and supplied to the model during training and inference, allowing variation along these dimensions to participate in shaping generative output.

Unlike text-prompted systems that rely on linguistic tokenisation, EAGLE operates entirely on continuous numeric conditioning. This allows precise control over scalar variables such as temperature or time-of-day and avoids the discretisation inherent in token-based prompts. Conditioning is therefore less dependent on linguistic interpretation and more directly embedded as a structural component of the generative process.

The EAGLE model architecture is based on Stable Audio Tools (Evans et al. 2024) and has two key components: an audio encoder-decoder implemented as a Generative Adversarial Network with Residual Vector Quantisation

[1] https://xeno-canto.org
[2] https://www.macaulaylibrary.org
[3] https://audioweather.com/

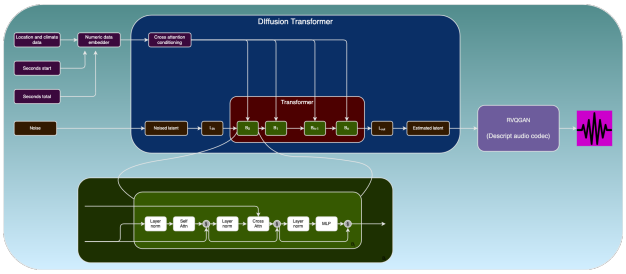


Figure 1: A flow diagram of the EAGLE model architecture for generating audio.

(RVQGAN), based on the Descript Audio Codec (Kumar et al. 2023); and a diffusion transformer with numeric conditioning supplied through a data embedder. EAGLE uses a cross-attention mechanism (ensuring the consistent application of conditioning data across diffusion steps and sequences) to apply conditioning data with a Classifier-Free Guidance (CFG) scale. The CFG scale modulates the influence of conditioning data during generation, allowing outputs to be more or less strongly constrained by the supplied environmental variables.

Figure 1 illustrates the EAGLE architecture during generation. On the left, the conditioning parameters are used to guide the denoising of Gaussian noise. The diffusion transformer iteratively applies learned transformations aligned with the conditioning data, producing a latent representation that is passed to the RVQGAN and decoded into audio.

The model learns statistical correlations between conditioning vectors and acoustic structure. It does not encode causal ecological laws. Its role is to render structured regularities present in the dataset such that:

- Repeated conditioning states tended to produce outputs recognisably situated within the same region of the conditioning domain.
- Variation within that region remains stochastic rather than fixed.
- Gradual changes in conditioning values produce continuous shifts in acoustic detail.

Within this framework the model does not define the ontology. The ontology is authored at the schema level and realised through dataset structure. The model acts as a learned renderer of that structured domain, producing audible outcomes whose variation is shaped by the relational logic encoded across the conditioning variables.

When schema design, dataset structure, and conditioning-integrated generation remain aligned, the system does not merely evoke a world through stylistic suggestion. It can enact a relational domain whose variation is organised by the authored schema.

Misalignment, Fragility, and Technical Limits

In Synthetic Ornithology, the system enacted the authored schema with varying degrees of fidelity, shaped by dataset structure, metadata quality, and the way conditioning was integrated into the architecture. From a technical standpoint, misalignment described the points at which the nominated world variables stopped functioning as reliable dimensions of variation. From an artistic standpoint, however, misalignment did not necessarily register as failure. Because the work operated in speculative worldbuilding rather than empirical simulation, there was no external ground truth against which the world's sonic behaviour could be verified.

In practice, some outputs appeared to reflect weak or ambiguous correlations between proposed climate conditions and audible features, and some speculative parameter combinations likely pushed inference beyond regions well represented in the training corpus. Several interacting factors contributed to this fragility, including uneven distribution of the dataset across geography, season, time of day, and weather; artefacts within the recordings such as variations in microphone type, recording technique, and anthropogenic noise; climate metadata that had been reconstructed rather than directly measured; and the schema's deliberate reduction of ecological complexity. Artistically, however, these issues were acceptable: the system continued to support the parafictional framing and maintained a stable sense of worldhood even when the precise mapping between conditions and sound was uncertain.

Despite these sources of technical fragility, the work demonstrated that world-schema alignment can function effectively as a compositional and conceptual tool. The generated audio remained legible as a consequence of changing world states, allowing audiences to experience the system as enacting a coherent domain rather than producing arbitrary variation. In this sense, the success of Synthetic Ornithology was that the schema, dataset, and generative process together sustained a stable sense of worldhood. This case suggests that world-schema conditioning remains a viable and productive method for generative worldbuilding in sonic arts.

Evaluation and samples

The EAGLE model's outputs were evaluated for fidelity and perceptual realism using a Mean Opinion Score (MOS) survey. The survey grouped participants according to their familiarity with birdsong and audio production, allowing responses from professional or specialist listeners to be compared with those from non-professional listeners. All participants listened to 15 randomly selected audio samples from a set of 9 real and 19 generated soundscapes. For each audio sample, participants responded to these five questions using a Likert scale ranging from "Strongly Disagree" to "Strongly Agree."

1. "The sounds in this recording appear natural and lifelike."
2. "This audio recording creates a sense of being in a real environment."
3. "This audio is real and not generated by an artificial intelligence model."
4. "The audio in this recording is pleasant to listen to."
5. "The audio in this recording is of high-quality and free from artefacts."

Likert responses were converted to numeric values from 1 to 5. For the naturalness, environmental presence, pleasantness, and quality questions, higher scores indicate more favourable listener responses. For the realness question, higher scores indicate that the sample was more likely to be perceived as non-synthetic. Responses from all listening events were grouped according to whether they corresponded to a real or generated file, and average scores were then calculated for each category. With 37 respondents each listening to 15 samples, the survey collected data from 555 listening events.

The similarities in responses to real and generated samples suggest that the model's outputs were perceived as broadly comparable to real recordings within the terms of the survey. In the context of this paper, these results support the claim that EAGLE can produce outputs that sustain the perceptual plausibility of the world it renders.

Table 1 presents the survey results. Professional respondents showed a similar response pattern to non-professional respondents, with generated audio scoring close to real samples across most criteria. This evaluation is included not as a comprehensive technical validation of the model, but as evidence that its outputs can maintain a stable sense of worldhood within the creative work.

	Section	Natural.	Env.	Real.	Pleas.	Quality	Average
All	**Real**	4.14	4.14	3.58	3.37	3.58	3.76
All	**Generated**	3.94	3.95	3.47	3.58	3.30	3.65
Pro	**Real**	4.24	4.46	3.70	3.91	4.12	4.08
Pro	**Generated**	4.02	4.10	3.62	3.80	4.02	3.91

Table 1: Mean listener ratings from the MOS survey comparing real and generated soundscape samples.

Within the creative work, the model operates continuously and the generated output varies constantly. Examples of misalignment therefore appear unpredictably. In *Synthetic Ornithology*, however, some degrees of misalignment contribute positively to the work by producing ambiguity, surprise, or moments of speculative instability. A selection of outputs illustrating the fidelity, ambiguity, and limits of the model is available at Synthetic Ornithology evaluation samples [4]. The documentation is organised into three sections:

1. Side-by-side comparisons of model outputs and source recordings, used to compare acoustic features and perceptual fidelity.
2. Notable outputs, including moments of compelling ambiguity or emergent behaviour that, in the context of *Synthetic Ornithology*, create opportunities for audience engagement and reveal the boundaries of the model's adherence to the world schema.
3. Limits and aberrations showing the model's fragility and misalignment with the schema.

[4] https://fred-dev.github.io/Synthetic_ornithology_results/

The case study presented in this research uses worldbuilding that aligns closely with an existing environment. This adds some complexity in evaluating these samples; listeners with knowledge of the real environments represented in the dataset will have specific sonic references against which these samples are evaluated. The goal of including these samples is not to compare them with the experience of a specific real place. Rather, the samples show how the model renders a consistent world, while also making visible the limits of that world's stability.

Conditioning as Compositional Strategy

The analysis of Synthetic Ornithology demonstrates that conditioning design can function as more than a steering interface layered onto generative systems. When aligned across ontology, dataset structure, and model architecture, conditioning specifies the relational domain within which a generative system operates, within a work of sonic art (Mirza and Osindero 2014; Rombach et al. 2022). The shift proposed in this paper is conceptual: conditioning data can be understood as a world-schema rather than auxiliary metadata. In many generative systems, control parameters are treated as post-hoc selectors that steer a trained model toward desired outputs (Tahiroğlu, Kastemaa, and Koli 2021; Scurto and Postel 2023; Scurto and Chemla–Romeu-Santos 2024). In contrast, world-schema conditioning defines the axes along which variation is permitted to occur. The artist's intervention therefore takes place prior to generation, at the level of variable definition, domain bounding, and relational expectation. The model renders these relations in sound.

Under this perspective, compositional labour shifts toward structural specification. The artist designs:

- ***Schema design:*** defining world variables, domains, and intended relations within an artistic framing.
- ***Dataset realisation:*** constructing or curating material in which those variables have learnable audible consequence.
- ***Conditioning-integrated generation:*** selecting and implementing an architecture in which conditioning participates directly in generative behaviour.

The resulting work is not a fixed composition but a generative domain structured by an authored relational schema.

Design Principles for World-Schema Alignment

Reflection on Synthetic Ornithology suggests several practical principles for artists working with conditioning-based generative systems. These principles are relevant for generative systems in which artists construct fictional, speculative, or abstract worlds whose behaviour is rule-consistent rather than stylistically implied.

Variable Minimalism: Conditioning schemas should define only those dimensions that can be meaningfully enacted through dataset variation.

Relational Density: Each conditioning variable must correspond to systematic and repeated variation in the dataset. Dataset construction therefore functions as an ontological act rather than neutral collection.

Cross-Axis Decorrelation: Conditioning variables should not be structurally entangled within the dataset. If world dimensions covary perfectly, the model cannot distinguish their effects.

Renderer–Ontology Separation: The generative model does not define the world. Ontological commitment resides in the schema and dataset structure; the model renders these relations statistically.

Statistical Enactment: World-schema alignment aims for probabilistic regularity rather than deterministic mapping. Conditioning states organise regions of behaviour rather than specifying exact outcomes.

Implications for Generative Art and Evaluation

Understanding conditioning as world-schema also reframes how generative systems may be evaluated in artistic contexts. Conventional model evaluation emphasises fidelity to ground truth, and in sonic arts contexts, perceptual realism or novelty (Evans et al. 2024; Copet et al. 2023). In world-schema systems, an additional criterion emerges: coherence relative to the authored ontology.

Evaluation may therefore consider criteria such as:

- Interpretability of conditioning regions.
- Sensitivity and continuity under parameter change.
- Constraint consistency relative to the schema.

These criteria complement rather than replace conventional metrics. A system may generate convincing sound while remaining ontologically incoherent, or it may enact a coherent relational domain despite imperfect acoustic realism. In speculative artistic contexts, the primary concern may instead be sustaining a stable sense of worldhood.

World-schema conditioning therefore positions generative systems not as simulations of external reality but as rule-governed fictional domains. The generative model acts as a statistical renderer of an authored ontology whose plausibility derives from internal consistency rather than empirical completeness.

Conclusion

This paper has proposed conditioning design as a compositional strategy grounded in world-schema alignment. Through the analysis of Synthetic Ornithology, it has shown that conditioning can function as an explicit ontological specification rather than an auxiliary control interface. When schema definition, dataset structure, and model architecture are deliberately aligned, a generative system can enact a bounded relational domain whose behaviour is structured yet stochastic.

The contribution of this framework lies not in claiming accurate simulation, but in articulating how structured fictional worlds can be embedded within generative audio systems through conditioning design. This reframes control architecture as a primary site of creative authorship and positions dataset construction as an ontological act within computational creativity. As generative systems increasingly shape sonic art practice, the central challenge shifts from producing convincing sounds to structuring the worlds those sounds inhabit. World-schema conditioning reverses the typical generative workflow: instead of training a model and then adding control, the artist defines a world first and trains the system to render it. World-schema conditioning therefore offers one method for structuring generative systems as coherent sonic worlds.

References

[Akhtar et al. 2023] Akhtar, M.; Shankarampeta, A.; Gupta, V.; Patil, A.; Cocarascu, O.; and Simperl, E. 2023. Exploring the Numerical Reasoning Capabilities of Language Models: A Comprehensive Analysis on Tabular Data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 15391–15405. Singapore: Association for Computational Linguistics.

[Azimi 2025] Azimi, M. 2025. (un)Stable (dis)Connection. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 63–64. Zenodo.

[Caillon and Esling 2021] Caillon, A., and Esling, P. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *CoRR* abs/2111.05011. arXiv: 2111.05011.

[Copet et al. 2023] Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

[Ekman and Taylor 2016] Ekman, S., and Taylor, A. I. 2016. Notes Toward a Critical Approach to Worlds and World-Building. *Fafnir: Nordic Journal of Science Fiction and Fantasy Research* 3(3):7–18.

[Engel et al. 2020] Engel, J.; Lamtharn Hantrakul; Chenjie Gu; and Roberts, A. 2020. DDSP: Differentiable Digital Signal Processing. In *International Conference on Learning Representations*.

[Evans et al. 2024] Evans, Z.; Carr, C.; Taylor, J.; Hawley, S. H.; and Pons, J. 2024. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

[Flowers et al. 2001] Flowers, J. H.; Whitwer, L. E.; Grafel, D. C.; and Kotan, C. A. 2001. Sonification of Daily Weather Records: Issues of Perception, Attention and Memory in Design Choices. In *Proceedings of the 2001 International Conference on Auditory Display (ICAD)*, 222–226.

[Frigg 2010] Frigg, R. 2010. Fiction and Scientific Representation. In Frigg, R., and Hunter, M., eds., *Beyond Mimesis and Convention*, volume 262. Dordrecht: Springer Netherlands. 97–138. Series Title: Boston Studies in the Philosophy of Science.

[Goodfellow, Bengio, and Courville 2016] Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT Press.

[Hermann, Hunt, and Neuhoff 2011] Hermann, T.; Hunt, A.; and Neuhoff, J. 2011. *The sonification handbook*. Berlin: Logos Verlag.

[Kumar et al. 2023] Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-fidelity audio compression with improved rvqgan. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

[Lam et al. 2023] Lam, M. W. Y.; Tian, Q.; Li, T.; Yin, Z.; Feng, S.; Tu, M.; Ji, Y.; Xia, R.; Ma, M.; Song, X.; Chen, J.; Wang, Y.; and Wang, Y. 2023. Efficient neural music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

[Lambert-Beatty 2009] Lambert-Beatty, C. 2009. Make-Believe: Parafiction and Plausibility. *October* 129:51–84.

[Martin and Sneegas 2020] Martin, J. V., and Sneegas, G. 2020. Critical Worldbuilding: Toward a Geographical Engagement with Imagined Worlds. *Literary Geographies* 6(1):1–6.

[Mirza and Osindero 2014] Mirza, M., and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784. arXiv: 1411.1784.

[Perez et al. 2018] Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; and Courville, A. 2018. FiLM: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

[Pons et al. 2025] Pons, J.; Zukowski, Z.; Parker, J. D.; Carr, C.; Taylor, J.; and Evans, Z. 2025. Music and Artificial Intelligence: Artistic Trends. *arXiv preprint arXiv:2508.11694*.

[Rodrigues 2025] Rodrigues, F. 2025. Synthetic Ornithology: Machine learning, simulations and hyper-real soundscapes. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 84–92. Zenodo.

[Rombach et al. 2022] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685. New Orleans, LA, USA: IEEE.

[Scurto and Chemla–Romeu-Santos 2024] Scurto, H., and Chemla–Romeu-Santos, A. 2024. Deeply Listening Through/Out the Deepscape. In *ISEA2023 PROCEEDINGS*. Paris, France: Ecole des arts decoratifs - PSL.

[Scurto and Postel 2023] Scurto, H., and Postel, L. 2023. Soundwalking deep latent spaces. In Ortiz, M., and Marquez-Borbon, A., eds., *Proceedings of the international conference on new interfaces for musical expression*, 232–235. Number of pages: 4 tex.articleno: 33 tex.track: Papers.

[Tahiroğlu, Kastemaa, and Koli 2021] Tahiroğlu, K.; Kastemaa, M.; and Koli, O. 2021. AI-terity 2.0: An autonomous NIME featuring GANSpaceSynth deep learning model. In *Proceedings of the international conference on new interfaces for musical expression*. tex.articleno: 80.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc.

[Wiggins 2006] Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Lessons Learned from Organizing a Literary Competition of Short Stories Co-created by Humans and AI Tools*

Rudolf Rosa
Charles University
Faculty of Mathematics and Physics
Praha, Czechia
rosa@ufal.mff.cuni.cz

Michaela Liegertová
Faculty of Science
Jan Evangelista Purkyně University
in Ústí nad Labem, Czechia
michaela.liegertova@ujep.cz

Miloslav Linc
Městská knihovna v Praze
(Municipal Library of Prague)
Praha, Czechia
miloslav.linc@mlp.cz

Abstract

In the past three years, we organized three rounds of a literary competition, in which human participants are asked to co-create short stories in collaboration with AI tools based on large language models. We collect, process and analyze the stories, ratings and comments by human judges, and participants' reports on their creation process. We gain various insights centered around several research questions: Which roles do humans and AI tools take on in the co-creative process? Where is human creative contribution necessary in the co-creation process (if anywhere) to produce high quality stories? What is the effect of literary and technical experience of the participants on the process and results? Do AI tools support or inhibit creativity?

Introduction

There recently has been increasing interest in AI-generated and co-generated stories and other literary texts (Teleki et al. 2025; Ismayilzada, Stevenson, and van der Plas 2025), with some results suggesting that Large Language Models (LLMs) can already be on par with humans or even outperform them (Gómez-Rodríguez and Williams 2023; Zhao et al. 2023; Porter and Machery 2024; Chakrabarty and Dhillon 2026) – especially when using the English language, for which the performance of LLMs is much stronger than for other languages (Zhang et al. 2023).

In our work, we investigate the state and development of capabilities of current LLMs for generation of short stories in Czech, which is a medium-resourced Slavic language. Over the past three years, we organized three rounds of a literary competition which required the participants to submit short stories co-created with the help of LLMs as well as potentially other AI tools. The stories were judged by experienced human judges, who produced rankings of the stories as well as some further insights. We combine these with reports about the collaboration process, provided by the competition participants, to analyze the co-creation processes and its relation to the resulting quality of the stories.

All data and resources that we are allowed to share can be found at our GitHub: `https://github.com/ufal/aiai`

*Research supported by Czech Science Foundation grant 25-14501L and by Charles University grant UNCE/24/SSH/009.

Research Questions

The main research questions we pose are:

1. Which roles do humans and AI tools take on in the co-creative process?
2. Where is human creative contribution necessary in the process (if anywhere)?
3. What is the effect of literary and technical experience of the humans?
4. Do AI tools support or inhibit human creativity?

Main Findings

A summary of our main findings is as follows:

- Both humans and AI tools can take on any of the roles in the creative process; but some scenarios are more typical than others.
- Human creative input is important *somewhere* in the process to achieve high-quality results; however, there is no single point in the process that would require human contribution.
- Too low as well as too high human involvement may be detrimental to the quality of the result; optimal results are achieved by a balanced human-AI collaboration.
- Experience in using AI tools and sophistication of the co-creation process are vital in producing a good output. However, this may get less important over time as the AI tools are significantly improving.
- Experienced writers did not outperform amateurs; writing experience does not seem to be helpful in the human-machine collaboration.
- AI tools can support the existing creative potential in amateur writers who lack some of the necessary writer skills. If used adequately, they do not kill the creative process but rather enable it.
- Experienced writers, capable of writing high-quality literature without AI help, were not observed to benefit from involving AI tools directly into the creative process; this rather disrupts the process and leads to worse results. However, they can benefit from AI tools for auxiliary tasks, such as research, proofreading, or critical feedback.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Round	Year	Focus	Participants
1	2023	sci-fi, fantasy, horror	74
2	2024	detective story, thriller	23
3	2025	explicit fictional author	53

Table 1: Rounds of the competition

The Competition

The Municipal Library of Prague has been working with aspiring authors for decades, providing them with various resources, offering creative writing workshops, organizing literary competitions, etc. With the rise of LLMs and other AI-based tools that can be used within the writing process, our visitors have started asking about the role of AI in literature. Therefore, to fulfill our long-term goal of supporting the creativity of people, we first started offering workshops on utilizing AI tools for creative writing, and, eventually, decided to organize a literary competition focused on human-machine co-creation of short stories.

Competition Settings

Over the past 3 years, we organized three rounds of the competition (see Table 1). The task was to produce a short story with some involvement of some AI tools. The language of the short stories was Czech (with the possibility to create the story in English and translate automatically to Czech). The competition is primarily targeted at Czech and Slovak starting authors, although we do not actively exclude other participants. It is publicized through a wide range of channels, including organizers' websites and newsletters, relevant Facebook groups, mailing lists of participants of writing workshops, other literary competitions, etc. Due to our long-term collaboration with aspiring authors, we are generally well connected with them and can thus reach them rather efficiently. Some participants have taken part in multiple rounds of the competition.

After announcement of each round, the participants have several months to turn in their submissions via e-mail. After that, the submissions are judged by human judges, and authors of the top 5 stories are then invited to a modest ceremony at the library in Prague. The competition has a minimal budget, with a lot of community volunteering, and with symbolic prizes offered by several local sponsors to the top 5 participants in each round (typically books, stationery, and AI-themed artifacts). We are preparing to publish some of the best stories in a dedicated anthology, and they are also currently being published in audio format read by a synthesized voice: `https://talk.youradio.cz/porady/nova-vlna`

The length limit for the stories in the 1st round of the competition was 3600 characters (including spaces), and 5400 characters in the 2nd and 3rd round (i.e. approximately 2-3 pages of text). Each round had a special focus, designed to test various aspects of the co-creation process:

2023: sci-fi, fantasy, horror The first round was designed to investigate the originality of the stories and the ability to meaningfully operate and reason about fictional worlds with different rules of functioning. This is in principle difficult for LLMs, since they tend to generate typical rather than extraordinary, clichés rather than peculiarities (Holtzman et al. 2020); and also since they are typically able to reason only within the rules and functioning of our real world (Wei et al. 2022).

2024: detective story, thriller The second round targeted the ability of LLMs to work with concealed information, known to the creator from the start but only revealed to the reader at the end. This is in principle difficult for LLMs since they have no concealed mechanism to pass on information into future (Salem, Paverd, and Abdelnabi 2026), they can only use the open text (except for modern reasoning models, which were not yet widely available at the time of the 2nd round).

2025: explicit fictional author In the third round, we decided to further explore the paradigm used by the winning submission to the second round: first design an explicit author persona and then condition the generation process on the persona. (The authors were asked to also submit a bio of the fictitious author, of up to additional 1000 characters.) This also plays with the common description/critique of AI-generated texts not having an author (Craig and Kerr 2025) by providing an author with a name and biography, even though non-existent.

As an example, we enclose the winning short story from the 3rd round in the appendix.

Competition Reports

Apart from the co-created stories, we also collected reports in which the participants detailed their co-creative process. In the first round, the participants were instructed to submit a free-form unstructured DOC(X) report specifying the AI tool(s) used, record/description of their interaction with the tools (ideally a full listing of the prompts and interactions), to what extent they feel to be the author of the story (0% to 100%), and a short account of how the interaction with the AI tools was. In the second round, we additionally asked how experienced they are with AI and how experienced they are in fiction writing. In the third round, we also asked to what extent AI was used to create the persona of the author (0% to 100%). We did not collect any other demographic information about the participants.

Analyzing Reports

The reports provide very valuable data for our investigation; however, all the information is self-reported by the participants and thus inherently unreliable to some extent. Unfortunately, there are various gaps, as not all reports contain all the information that we would like to have. Moreover, the reports are not structured in any consistent way. We deliberately decided to make the instructions as simple and low-barrier as possible, and to be rather benevolent with what the participants submit, not to discourage people from participating. This, in turn, leads to the automated analysis of the reports being somewhat complicated.

For this reason, the first phase of the report analysis was done completely manually. One member of the research team manually went through all of the submitted reports,

interpreting the provided information in a structured way and collecting it in a table. In this way, we collected the simpler information that we explicitly solicited (tools used, authorship, writing experience, AI experience), as well as some semi-structured information about the writing process (e.g. structure of collaboration, roles of human and AI in the process, author of the theme, who did the final editing, interesting observations). Categories for frequent responses were designed and updated on the fly as necessary, with still some responses not fitting any of the categories (and some aspects missing altogether in the reports). Subsequently, an LLM pipeline was used to analyze the table and the reports to identify potentially interesting findings that can be based on the data. The main findings suggested by the LLM pipeline were then verified with the data, and further analyses and correlations were computed using standard statistical tools.

All the reported findings are based on our manual or manually verified analyses of the reports and stories (apart from the automated quality evaluation; see later). The strength of our findings and resulting claims is naturally limited, as we are dealing with somewhat noisy and unreliable data processed in a somewhat noisy way (e.g. we needed to interpret the authors' self-reported AI expertise, communicated in various ways, as falling into a limited set of categories, in order to correlate it with the quality of the submissions). In general, we assume that some reports may contain inadvertent inaccuracies or omissions; however, we expect these errors to be random rather than systematic, and therefore unlikely to substantially bias the aggregate analysis. Moreover, we note that the participants had no incentive to lie in their reports, as the reports were not taken into account in the competition itself.

All our data that we are allowed to make publicly available are shared in our GitHub repository at `https://github.com/ufal/aiai`. This includes the calls for participation with author instructions, the top 5 stories from each round, tables with data analyses, and various support scripts. Unfortunately, we are not allowed to publicly share the reports; a shortcoming which we intend to fix with the next competition round, potentially additionally also asking the authors of the previous rounds whether they would consent to making their reports publicly available.

Additionally, as an example of report analysis, we enclose a more detailed analysis of the report for the winning submission from the 3rd round in the appendix.

Quality evaluation by human judges

Within the competition, the stories are independently rated by several human judges (6 judges for the 1st and 3rd round, 7 judges for the 2nd round), with each judge reading all of the submitted stories and ranking the top 21 of them.

Judges were recruited from three areas of expertise (with some judges belonging to multiple of these categories):

Librarians Readers with immense reading experience and vast knowledge of books (but not necessarily any expertise in creative writing).

Authors Both professional and amateur fiction writers (each with at least one published work). They have some degree of writing expertise and thus evaluate not only their impression of the work but also how the author works with the text.

AI experts Experts with understanding of AI functioning, experienced with dealing with LLM outputs.

Mixing various kinds of expertise among the judges ensures that the highest-ranked stories are of a good quality ideally in all relevant aspects.

The judges only have access to the submitted stories but not to the reports. The formulation of the task for the judges is as follows: *'Read the texts as standard short stories and select those that you subjectively like the best'*. Each of the judges independently reads and ranks the stories submitted to the given competition round, and their rankings are then combined into one overall judges score. The top 5 stories are then announced as winners.

The exact mechanism of the ranking is inspired by other Czech literary competitions, using a "pyramid setup": each judge ranks 1 story at rank 1, 2 stories at rank 2, 3 stories at rank 3, etc., up to 6 stories at rank 6. In this way, each judge only ranks the top 21 stories according to their evaluation; this is because it is easier and more important to differentiate between the best texts than between the worst texts.

The ranks are then converted to points using the formula $points = 60/rank$; the assigned ranks 1-6 are thus converted into 60, 30, 20, 15, 12 and 10 points (or 0 for non-ranked stories). The points from individual judges are then summed up for each submission, and the submissions are ordered by the resulting *judges scores*. The nature of the ranking mechanism leads to the fact that there are rarely ties on the top places (which is good for announcing winners), but very often ties at several of the bottom places (which does not matter). In our work, we use the resulting *judges scores* of the stories in each round as the ground truth evaluation of their quality, and correlate it with various other aspects.

Insights from the Judges

Besides the official ranking, many of the judges actively communicated various additional insights and comments with us. Moreover, several of the judges took part in judging multiple rounds of the competition, we can thus gain some absolute as well as relative insights. The judges' comments were communicated in a non-structured informal way; in this section, we summarize interesting insights which we gained through this communication.

In the first round, the quality of the stories was quite low, with a lot of repetitiveness and homogeneity; the second round was somewhat better but not strongly. In the first two rounds, the judges thus agreed that the task of selecting the top 5 stories consisted of identifying stories that were good and/or interesting at least in some aspects, but generally being 'bearable' rather than really good.

However, the third round of the competition showed that the quality of the available AI tools is already high enough to potentially reach the writing quality of a mediocre human writer, even when using simpler approaches. The judges agreed that they observed a significant improvement in quality between the 2nd and 3rd round. All the top 10 stories

were interesting, and the task was to select the 5 best stories among 10 relatively good candidates.

Although all judges knew that the stories are partially generated, in the third round we also had reports that for some of the stories, the reader temporarily forgot that they are reading a generated text. Some judges reported that some of the generated stories elicit thoughts in them, and/or spark emotions; qualities traditionally attributed to human works but assumed unattainable by generated works.

However, there are still qualities that seem hard or impossible to reach with current AI tools without considerable human creative intervention, such as building an intriguing plot with a strong plot twist and/or a sharp point. The LLMs are getting increasingly better at the writing craft, but fail to provide sufficient amount of creative input; the human contribution is still important in this aspect.

Comparison to human literary competitions As many of the judges also have experience in judging competitions of fully human-written texts, they shared some similarities and differences between these and our setting.

First, in both fully-human amateur competitions as well as our human-machine co-generation competition, they observe that the vast majority of the submissions in the competition are dull and uninteresting.

However, AI tools seem to struggle more with aboutness and interestingness than human authors. In the co-generative competition, the good stories are thus those where the human co-author managed to creatively provide these qualities; through various ways though, including providing the creative input at the beginning, and/or selecting sufficiently creative ideas from multiple generated variants, and/or enriching a first dull version of the story with manually provided ideas on making it more interesting and letting the AI tool to iteratively incorporate the human creative input, etc.

On the other hand, in fully human competitions, there is typically a certain amount of submissions where the inexperienced human authors clearly struggle with the craft of writing; this is mostly missing in the co-creation setting, as the current AI tools are masters of the writing craft.

The general sentiment is that the top stories from the 3rd round of the competition could successfully compete with fully human-written texts in standard literary competitions, probably ranking near the top in smaller amateur-focused competitions, and maybe somewhere in the middle in big competitions with participation of professional writers.

Automated Quality Evaluation

We have also run an automated LLM-based evaluation of creativity, to examine usefulness of such automated evaluation in this setting, and to gain more detailed insights into how LLMs perform in various aspects of creative writing and how this changes over the years.

LLM-based evaluation is currently a usual choice for automated evaluation of natural language generation, surpassing previous approaches, although there are also numerous concerns, such as reproducibility of the results (Kartac, Lango, and Dusek 2025). LLMs often demonstrate respectable correlations with human ratings, especially when the evaluation targets semantic properties of the text rather than formalistic properties (Mareček et al. 2025). A common issue may be a high variance and low agreement of the scores, which is often due to too simple prompts; it has been shown that more reliable results can be obtained by prompting the models with explicit checklists of features/qualities that the LLM should explicitly evaluate (Lee et al. 2025).

LLM-based quality	Average '23	Average '24	Average '25	Corr/Score '23	Corr/Score '24	Corr/Score '25
Characteriz.	3.2	4.0	5.7	0.63	0.71	0.29
Plot	3.2	3.9	5.7	0.56	0.69	0.33
Setting	3.3	4.2	6.4	0.57	0.59	0.31
Conflict	3.0	3.9	5.6	0.64	0.67	0.35
Theme	2.9	3.5	5.4	0.52	0.43	0.26
Voice	3.6	4.2	6.2	0.61	0.58	0.36
Craft	3.6	4.2	6.0	0.44	0.66	0.38
Originality	2.5	3.1	5.1	0.65	0.49	0.24
Creativity	2.5	3.1	5.3	0.70	0.48	0.27
Avg score	3.1	3.8	5.7	0.68	0.68	0.34

Table 2: Average automated quality scores per round obtained through LLM prompting (0-10, higher is better) and their Spearman correlation coefficients with judges scores.

Our evaluation prompt is based on *LLM Creative Story-Writing Benchmark V4* (Mazur 2026). The benchmark defines 8 criteria to evaluate the quality of generated stories, evaluated on the range 0-10 (higher is better). Detailed definitions of the categories are used as a prompt for LLMs as judges, with explicit checklists of expected positive properties as well as common failures. The 8 categories evaluated by the benchmark are: Characterization and Motivation; Plot, Structure and Causality; Setting, Atmosphere and Specificity; Conflict and Stakes; Theme and Subtext; Voice and Point of View; Prose and Line-Level Craft; Originality and Ingenuity.

However, none of the categories evaluates creativity as such, although the last category goes in that direction to some extent. Therefore, for our experiments, we enriched the framework with a 9th category specifically targeting evaluation of overall creativity. The creativity dimension of the rubric operationalizes the standard research definition of creativity as the production of artifacts that are both novel and appropriate within a given domain (Runco and Jaeger 2012; Amabile 2018). To capture this construct in narrative texts, the rubric evaluates the presence of original combinations of familiar elements, transformation of genre conventions, productive surprise, and the functional integration of novel ideas into the story's structure and meaning, drawing on theoretical models of creative cognition and creative products (Boden 2004; Simonton 2012; Besemer and O'Quin 1999).

We used *Gemini 3 Pro Preview* LLM. The full prompt (8,253 words) and codes are part of our GitHub repository.

Table 2 lists the results of the automated evaluation; we list the average score of each metric over all stories, average of all the 9 metrics, and the correlation of each of the scores with the judges score (Spearman correlation coefficient).

The correlations are rather strong for the first two rounds, mostly between 0.5 and 0.7 for the individual aspects and nearing 0.7 for the average, showing that the automated evaluation is rather reliable. However, for the last round, the reliability drops dramatically; we assume this is due to the measures being computed with current models, which are good at rating the weaker older models but not for current similarly strong models (Thakur et al. 2025).

Note that the judges did not provide scores stratified along these 9 dimensions; the judges score is one-dimensional, and we thus correlate all of the categories with the same score. As a good story should be good in all of the 9 aspects, it is reasonable to expect a moderate correlation of each of the scores with the judges score, but a very high correlation is not expected as each of the scores only evaluates one of the important aspects. For the same reason, a higher correlation is expected (and is indeed observed) for the average of all the scores. The differences of the individual score correlations between the rounds may also suggest that some of the aspects have different importance in the rounds; e.g. originality and creativity may be more important in the sci-fi/fantasy round than in the detective story round.

We can see that the absolute quality is rather low, with averages between 3-6 points (although individual stories scored up to 8.0). Moreover, the Originality and Creativity scores are among the lowest, confirming that LLMs struggle with the creative aspects of story-writing more than with the more technical skills (the craft). On the other hand, we can clearly see the improvement in quality between the rounds in all aspects.

Human-Machine Interactions

The mode of interaction between the human and the AI tool can have a large impact on the quality of the output as well as the feelings and satisfaction of the human (Issak, Rezwana, and Harteveld 2025). One of our interests is what roles do humans and the AI tools take on in the creative process, how they collaborate, and what are the effects of the approach.

The writing process consists of several steps and corresponding roles (the actual steps may vary for individual authors to some extent). Our general observation here is that both humans and AI tools can take on any of the roles in the creative process; but some scenarios are more typical than others. To achieve high-quality results, human creative input is important *somewhere* in the process; however, there is no single point in the process that would require human contribution. In the competition reports, we have seen a great variety of division of the tasks between the human, the AI tools, or a collaboration of both, and we have not identified any single step in the process that would always be done by the human or always by the machine.

Source of the Story Idea

As an example, we can look in more detail at the ideation role – the invention or emergence of the main idea of the story – which typically takes place somewhere at the beginning of the creative process (but the idea may also emerge and/or evolve later during the process).

Competition year	2023	2024	2025
Initial idea from AI	30%	50%	60%

Table 3: Stories where the initial idea for the story is AI generated, not human-provided; in tens of percent.

In their concept of framework authorship, (Piorecký and Husárová 2024) assume that the human contribution is crucial at the beginning and end of the creative process: they assume that the most important human creative input is in the idea of the story and plot, and that a human is also needed to eventually decide to publish the story.

In the reports, we have encountered various approaches, such as having human creative input at the very beginning of the process, or have the AI tool propose the main idea of the story, or having the AI tool suggest multiple ideas with the human then choosing from them. Our observations thus do not support the assumption of Piorecký and Husárová.

As most of the authors explicitly stated whether the story idea came from the human or the machine, we can even analyze this quantitatively. Table 3 shows the proportion of stories where the idea was AI-generated. We can see that already in the first round, 30% of the submissions were based on an AI-generated idea, with the proportion steadily rising up to 60% for the 3rd round.

The situation is similar with all the other steps and roles in the creation process.

Typical Co-creation Scenarios

Based on the reports, we can approximately describe several generally common approaches, although many submissions deviate to some extent from these prototypical scenarios:

One-click generation The most basic approach is where the human provides only a simple prompt (or even just a copy-paste of the competition task), without much or any creative input, and submits the first output that the LLM generates. This approach is typical for inexperienced AI users and does not often lead to good results.

Human as orchestrator Here, the core of the human creative input is provided at the beginning of the process, defining e.g. the idea, task, style, requirements, preferences, and/or direction. Then, the AI tool is used to convert the detailed instructions and/or ideas into text. Finally, the text is curated and iteratively refined through human intervention by dialogically instructing the AI tool to perform specific improvements. Some participants, especially those with extensive AI experience, use rather complex prompt design and/or combination of several clearly defined steps in the process to carefully guide the generation process step by step (sometimes, even the steps to take are suggested by the AI tool). This is one of the most successful approaches, based on true human-machine collaboration, with a varying degree and kinds of creative contributions of both the human and the AI tool(s).

Human provides detailed input A variation of the human orchestrator scenario, where the focus is not on the complexity of the prompt and setup, but rather the detail of

the provided input. Here, the main creative contribution is by the human, who provides a detailed outline or even a draft version of a text. The AI system is then instructed to convert the input into a well-written story, but the focus is on its mastery of the writing craft rather than its creative input; the AI tool functions as a post-editor and/or ghost-writer rather than a full creative partner. Therefore, the success of this approach is determined by the quality of the provided input, and is thus successful with participants who are themselves able to build a good story plot.

Human as editor Here, the AI tool has the major creative contributions, by being only lightly instructed, having a lot of freedom to decide what to write and how to write it (and typically also writing it). Then, the human takes over as an editor rather than a creative co-author, manually editing the generated text (sometimes even massively). This approach is usual with users with little AI experience but some writing experience. The results are often poor, as the AI-generated draft is often of low quality and thus difficult to turn into a good story. Experience also shows that manually editing the generated text can make it stylistically inconsistent; better results may be achieved by indirect editing through dialogically instructing the AI tool and thus iteratively refining the text.

AI only for auxiliary tasks This is a typical way of using the AI tools by experienced writers. The text is fully written and edited by the human, with AI tools used for auxiliary tasks (brainstorming, research, beta-reading, suggesting improvements, etc.) but not for directly editing the text; the human keeps full control over the text. This approach is thus successful if the participant is already experienced in writing.

Although the approaches are highly varied, we observe that the AI tools are good at the writing craft itself, but usually require human creative input to make the story novel and interesting. We can thus think of the optimal role of the human in the writing collaboration as a sort of creative leadership and curatorship, guiding the creation process but letting the AI system do most of the actual writing.

However, there is, by far, not a single optimal setup; all of the aforementioned approaches were used at least for one of the top 5 stories in at least one round of the competition.

Usefulness of AI tools for creators

Based on the reports, we note that the usefulness of AI tools typically differs for amateur writers and for professionals.

Most of our participants have no or low experience in writing, but many of them express a long-lasting desire to write; they have simply always lacked some of the necessary skills and thus had not been able to write satisfactory texts. For them, the AI tools are enabling, finally making it possible for them to become authors or co-authors of literary texts of reasonable quality, unachievable for them without the support of the tools. This is a sort of prosthetic usage of the tools, compensating for some weaknesses or disabilities on the participant's side.

Experienced writers, who are capable of writing high-quality literature without AI help, do not benefit from in-

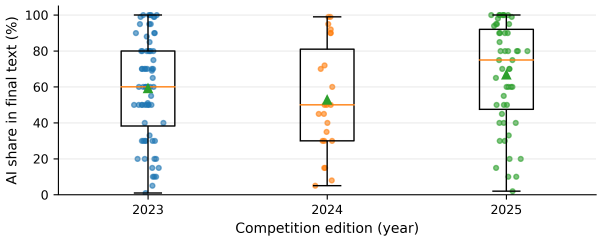


Figure 1: Distribution of the amount of AI contribution.

Competition year	2023	2024	2025
Average AI contribution	59%	53%	67%

Table 4: Average author-reported AI contribution.

volving AI tools directly into the creative process; this rather disrupts the process and leads to worse results. However, they can benefit from AI tools for auxiliary tasks, such as research, brainstorming, proofreading, or critical feedback.

There is an ongoing discussion whether AI tools support or kill creativity in humans (Alegaonkar and Avachat-Shirke 2023; Alzubi, Nazim, and Alyami 2025; Ingle 2025). Based on our observations from the competition, we see that if the AI tools are used adequately, they do not influence the creative process negatively. They can support amateur writers by augmenting them with skills they lack to let them fully utilize their creative potential. In this setting, we do not see the threat of limiting their future creative potential (Kumar et al. 2025), as without the tools, they would not create anything anyway; however, there is always the threat of limiting one's cognitive capabilities in general through reliance on AI tools (Kosmyna et al. 2025), which we do not measure.

For professional writers, it makes little sense to use the tools substitutively to automate skills which they already have, but the tools can nevertheless support their creative process by helping with other related tasks.

Optimal AI/human balance

Looking beyond the actual roles taken on by the humans and the AI tools, we can zoom out to investigate the proportion of AI contribution in the process and its effect on the quality of the results. Bear in mind that the amount of AI contribution is estimated and self-reported by the participants and

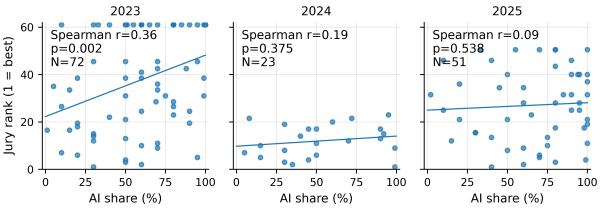


Figure 2: Relation between (self-reported) share of AI on the creation process and final rank in the competition.

Tool	2023		2024		2025	
GPT	57	77%	19	83%	38	72%
Gemini	7	9%	2	9%	8	15%
Claude	0	0%	3	13%	4	8%
Copilot	8	11%	0	0%	3	6%
other	7	9%	2	9%	2	4%

Table 5: Tools used by the participants. Some participants used more than one tool. All versions of the tools are grouped together; most participants used the most advanced version widely available at the time of the competition.

may not be completely trustworthy, all the results in this section thus inherently have only limited reliability.

Figure 1 and Table 4 show the amount of AI contribution over the three rounds of the competition. While the average share of AI rises slightly over time, it varies a lot among the individual submissions and is rather spread out.

Figure 2 analyzes the relation between the share of AI in the process and the final rank of the resulting story in the competition. From this, we can draw some more specific insights, although even here the situation is quite varied.

In general, a meaningful and balanced AI-human collaboration seems to be more successful than extremely large or extremely low AI share; however, the range of optimal AI share lies approximately in a broad range between 30%-80%, and successful outliers still do exist.

In the earlier rounds of the competitions, better results were achieved rather with lower AI contribution, but this effect gets weaker with each round: in 1st round, there was a correlation of 0.36 between low AI share and good rank in the competition, but this drops to 0.19 for 2nd round and 0.09 for third round. It still holds that some human creative contribution is necessary to achieve good results, but the required amount gets smaller each year.

Fully AI-generated stories with reported 0 human contribution all ranked last in the 1st round, whereas in the 3rd round they ranked around 30th place on average. The best fully generated story ranked as high as 11th place, thus already outscoring most stories with some human contribution (it was created by a participant with very high AI experience as well as literary experience).

Tools used

In Table 5, we provide an overview of AI tools used by the participants; the tools grouped under the label 'other' include: you.ai (3x), poe.com, Deeply, WizzardLM, DeepL, Bing, and Perplexity (all 1x). The distribution is rather stable across the rounds, with GPT-based tools being a stable default for most participants; this of course limits our ability to draw any strong findings, as all of the other tools are used only rarely. Nevertheless, we did not observe any significant links between the tool(s) used and the quality of the output; it seems the approach matters more than the tools.

Effect of Experience

We now investigate the effect of the experience of the human along two dimensions: technical experience in using the AI tools, and literary experience in manually writing texts.

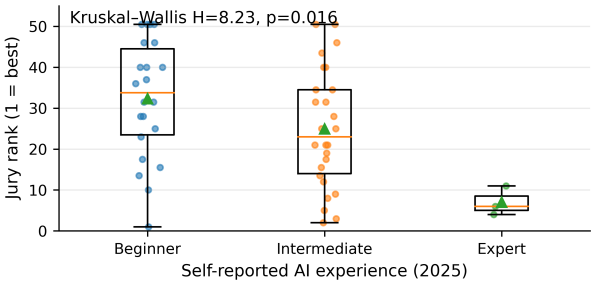


Figure 3: Effect of technical experience

AI experience (self-rep.)	None/ Beginner	Intermediate/ Advanced
All	43%	57%
Top 10	26%	74%

Table 6: Self-reported technological experience, proportion over 2024 and 2025 combined (the proportions were similar in each round). As the detailed categories were slightly different between the years, we combined them into only two coarse-grained categories.

Effect of technical experience and complexity

Our investigations clearly show that experience in using AI tools as well as sophistication of the co-creation process are vital in producing a good output.

Table 6 shows that approximately half of our participants describe themselves as inexperienced in using AI tools, while the other half see themselves as advanced AI users. However, the vast majority of the human co-authors of the top stories are technically experienced. Figure 3 paints an even stronger picture, showing that all of the highly experienced AI users scored very high in the 3rd round (within the first 11 places), while only 2 AI-beginners scored in the top 10. On the other hand, the winning submission of the 3rd round was actually by a self-described AI beginner, showing that the current generation of AI tools can be successfully used even by less experienced users.

In the reports, the participants gradually report that the interaction with the AI tools is getting easier and more pleasant. In the first round, they often reported that using the tools resembled a fight rather than collaboration, requiring a lot of manual work to achieve a bearable result, the process was often frustrating, and the final text was typically still rather disappointing. In the second round, the interaction with the tools got easier, but the results did not improve so much. However, in the third round, the participants were often very satisfied both with the process and with the results. They commonly describe the process as easy, pleasant, with the tools following well their wishes and instructions, and requiring much less manual interventions into the process.

Through manual investigation of the reports, we have also observed that the approaches used by the participants are

Corr. of score with	Approach	AI skills	Lit skills
2023	0.28		
2024	0.42	0.23	0.13
2025	0.06	0.30	0.02

Table 7: Spearman correlation of the judges score of the short story with the complexity of the approach (estimated as the length of the report, in characters), self-reported AI expertise, and self-reported literary writing experience.

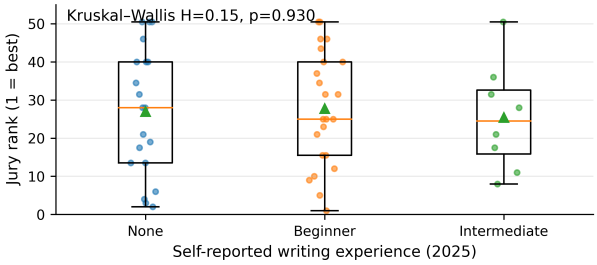


Figure 4: Effect of writing experience

getting more advanced and complex. As an easily quantifiable approximate proxy for approach complexity, we use the length of the reports (number of characters); we observe that the reports get longer by approximately 25% between 2023 and 2024; the length stayed similar between 2024 and 2025.

Table 7 correlates the judges scores of the stories with the approach complexity and technological expertise of the participants, typically showing moderate to strong correlations.

The effect of approach complexity practically disappears in the third round. This is in line with our observation based on the reports that in 2025, even simple one-click approaches are starting to achieve competitive results, suggesting that the AI expertise and approach complexity may become less important in future, as the AI tools are gradually improving and are increasingly successful in following even simple naive instructions.

Effect of writing experience

On the other hand, we did not find any effect of experience with literature writing on the resulting quality of the co-created stories; see the weak to negligible correlations in Table 7 and the very similar distributions of scores among the participants in Figure 4. It thus seems that writing experience is not helpful in co-creation of stories with AI; probably because many of the writing skills are well automated by the LLMs, and it is rather experience with the AI tools which is crucial in the process. Still, we note that the vast majority

Writing exp.	None	Beginner	Experienced
All	39%	47%	14%
Top 10	42%	42%	16%

Table 8: Self-reported writing experience, combined over 2024 and 2025 (the proportions were similar in each round).

Word	Stories containing it	
dark, darkness	45	61%
mystery, mysterious	39	53%
shadow, shadowy	38	51%
night	37	50%
forest	27	36%
whisper	21	28%

Table 9: Number and proportion of stories in the 2023 round, containing the selected genre-typical/cliche words.

of our participants have zero or little writing experience (see Table 8), our data for experienced writers are thus limited.

On closer investigation, we noted that although there is no meaningful difference in the scores of literary amateurs and experts, there are differences in their struggles. What we observed with amateur authors is that if they e.g. do too much editing of a generated text, they often worsen its quality, as they do not have the experience in producing good texts. However, experienced authors struggle differently, as they have too much experience in producing good texts, and thus are not willing to accept a lot of what the AI tools generate. They tend to fight with the AI tools, or even compete with them, unable to accept the AI tool as a creative partner and to meaningfully collaborate with it.

Round-specific Findings

1st Round: struggle with homogenization and fictitious worlds

The first round focused on science-fiction, fantasy and horror stories. We found that the submitted stories typically struggled with originality, to the extent that many of the stories were quite similar to each other, featuring similar plots and other features. Specifically, the most common plot was a horror story of a group of young people finding an abandoned house in a dark forest; upon entering the house, they experience an unpleasant encounter with a mysterious figure (this plot appeared in more than 10 of the stories). We also noted that there are apparently some genre-typical words that appear way too often in the generated stories (see Table 9). The style and structure of the stories was often similar, with a lot of clichés, predictable themes and development, shallow plot, pathos, and a happy ending. All of these are clear examples of the phenomenon of LLM homogenization (Anderson, Shah, and Kreminski 2024), where individual outputs are not necessarily bad on their own, but investigation of a larger amount of generated texts reveals their high similarities in many aspects; a literary competition with a fixed task is exactly a great setting for observing this effect.

In 2023, the AI tools also still struggled with the writing craft in Czech language, with many texts being rather outlines or descriptions of a story than actual engaging stories. We also observed e.g. a low amount of dialogue and other good storytelling practices (we gradually see more and more dialogues embedded in the stories in subsequent rounds).

As for the fictional theme of the round, we observed that LLMs could design fictitious worlds that function dif-

ferently from our world (e.g. by having magic), but then failed to meaningfully utilize the introduced mechanisms (e.g. to use magic to resolve the plot). This is due to the fact that surprisingly, LLMs can work as a simulation of the real world to some extent as they have gained the emergent capabilities of reasoning about our world even without having direct access to it, just based on the texts in the training data talking about the world (Zhang et al. 2025; Ge et al. 2024). However, this is a difficult and complex capability that only emerges with training on a large amount of texts about the world. For the invented fictitious world, the models have not been trained to reason about it – unless it is a fictitious world well-represented in the training data (Shostack 2024) – and are unable to grasp the logic of the just-introduced world only through a zero-shot exposure to its brief description.

2nd Round: Concealed Information

The second round focused on detective stories and thrillers, where we specifically were interested whether the LLMs can work with concealed information, as a typical decoder-based language model has no actively controlled hidden memory and can only explicitly pass information forward through the generated text. Note that this was before reasoning models became widely available (Huang and Chang 2023) which bypass this limitation by generating the 'reasoning' text that does not become part of the returned output and thus can easily work with concealed facts (Cheng et al. 2025).

We found the LLMs to mostly fail, with a tendency to reveal the secret too early. Successful approaches included simulated reasoning with multi-step/hierarchical setups, where the secret was not technically hidden, but it was overt in intermediate generated texts which did not form the final output (e.g. an outline generated as a middle step before generating the story itself) and thus served as a sort of memory for the language model.

The homogenization/cliché problems got weaker compared to the previous round, although still present to some extent (e.g. 4 of the top 5 stories take place in a library).

3rd Round: Explicit Personas

The third round required the participants to work with an explicit persona of the virtual author within the prompt. The approach to creating the persona was as variable as the story generation, with some participants describing the persona fully manually, some using a fully generated description, and some co-creating the description in human-AI collaboration. Nevertheless, conditioning on the explicit persona seemed to lead to more diverse and interesting outputs (Hu and Collier 2024). To some extent, this pushed all of the participants to use a more complex approach, which may explain part of the higher quality of the resulting stories. While the judges' evaluation focused mainly on the stories, there was feedback that even reading the virtual author biographies was rather enjoyable, with some of the biographies being more interesting than the story.

4th round: Humor?

In the 3rd round, some participants managed to successfully incorporate humorous elements into the short stories, confirming the recently emerging findings that LLMs can already produce humorous texts in various languages to some extent (Wu, Weber, and Müller 2025; Inácio and Oliveira 2025; Amir 2025). This was positively noted by some of the judges as a new development, as the previous generation of AI tools was not yet capable of producing satisfactory humorous elements in Czech language. Also, several of the participants suggested to focus on humor in the next round. We thus believe that the time is ripe to explore how humans and machines can co-create humorous texts. Our provisional hypothesis is that current AI tools can already produce humorous elements in the text, but are not yet good for evaluating their success in doing that; we thus believe that the usual role distribution would be utilizing AI creativity to propose humorist elements, followed by employing human capabilities to assess their quality and select the ones that people will actually find both good and funny.

Conclusion

In each of the past three years, we organized a literary competition of short stories co-created by humans and AI tools. In this paper, we analyzed the submitted stories as well as the reports detailing the participants' approaches, investigating observed patterns and correlations.

We found that the progress in capabilities of available AI tools leads to better quality of the co-created stories in each round, with the best submissions currently already reaching at least the level of mediocre human authors. We showed that AI experience of the participant and complexity of the collaborative approach are strongly linked to a high quality of the resulting text, whereas participant's literary experience is irrelevant. However, as the tools are getting more advanced and easy to use, the complexity effect disappears and the technological expertise effect may weaken in future.

We identified and discussed some common collaboration approaches while observing their wide variety, finding that meaningful balanced collaboration between the human and the machine is important. However, any of the creation steps can be automated and there is thus no single point where the human involvement would be required, and the amount of AI contribution versus human contribution to the process can also vary a lot even in the winning stories.

We note that a lot of the input information for our analyses is self-reported by the participants, our resulting claims thus need to be taken with some caution.

Author Contributions

Miloslav Linc was the main organizer of the literary competition and of the manual evaluation of the submissions. Rudolf Rosa was a co-organizer of the literary competition and the main author of the text of the paper. Michaela Liegertová performed the core of the data analysis. All three authors took part in the analysis of the short stories and reports and in the interpretation of our findings.

References

Alegaonkar, A. A., and Avachat-Shirke, M. A. 2023. Is artificial intelligence killing artistic skills in designers. In *International Conference on Emerging Trends in Design & Arts*, volume 4, 109–115.

Alzubi, A. A. F.; Nazim, M.; and Alyami, N. 2025. Do AI-generative tools kill or nurture creativity in EFL teaching and learning? *Education and Information Technologies* 30(11):15147–15184.

Amabile, T. M. 2018. *Creativity in context: Update to the social psychology of creativity*. Routledge.

Amir, O. 2025. Are AI-generated jokes truly original? Charting the "joke space". In *Proceedings of ICCC'25*.

Anderson, B. R.; Shah, J. H.; and Kreminski, M. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, 413–425.

Besemer, S. P., and O'Quin, K. 1999. Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal* 12(4):287–296.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London: Routledge.

Chakrabarty, T., and Dhillon, P. S. 2026. Can good writing be generative? Expert-level AI writing emerges through fine-tuning on high quality books. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery.

Cheng, F.; Li, H.; Liu, F.; Van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering LLMs with logical reasoning: a comprehensive survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

Craig, C., and Kerr, I. 2025. The death of the AI author. In *Robot Law: Volume II*. Edward Elgar Publishing. 250–285.

Ge, Z.; Huang, H.; Zhou, M.; Li, J.; Wang, G.; Tang, S.; and Zhuang, Y. 2024. WorldGPT: Empowering LLM as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7346–7355.

Gómez-Rodríguez, C., and Williams, P. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14504–14528. Singapore: Association for Computational Linguistics.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hu, T., and Collier, N. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10289–10307.

Huang, J., and Chang, K. C.-C. 2023. Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023*, 1049–1065.

Inácio, M. L., and Oliveira, H. G. 2025. A full pipeline for context-aware pun generation. In *Proceedings of ICCC'25*.

Ingle, R. V. 2025. An AI tool: Savior or destroyer of human creativity in art and literature. *Unlocking Potential: The Intersection of Human Creativity and AI* 59.

Ismayilzada, M.; Stevenson, C.; and van der Plas, L. 2025. Evaluating creative short story generation in humans and large language models. In *Proceedings of ICCC'25*.

Issak, A.; Rezwana, J.; and Harteveld, C. 2025. MOSAAIC: Managing optimization towards shared autonomy, authority, and initiative in co-creation. In *Proceedings of ICCC'25*.

Kartac, I.; Lango, M.; and Dusek, O. 2025. OpeNLGauge: An explainable metric for NLG evaluation with open-weights LLMs. In Flek, L.; Narayan, S.; Phuong, L. H.; and Pei, J., eds., *Proceedings of the 18th International Natural Language Generation Conference*, 292–337. Hanoi, Vietnam: Association for Computational Linguistics.

Kosmyna, N.; Hauptmann, E.; Yuan, Y. T.; Situ, J.; Liao, X.-H.; Beresnitzky, A. V.; Braunstein, I.; and Maes, P. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task.

Kumar, H.; Vincentius, J.; Jordan, E.; and Anderson, A. 2025. Human creativity in the age of LLMs: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery.

Lee, Y.; Kim, J.; Kim, J.; Cho, H.; Kang, J.; Kang, P.; and Kim, N. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 15771–15798. Suzhou, China: Association for Computational Linguistics.

Mareček, D.; Hodková, K. M.; Musil, T.; and Rosa, R. 2025. Evaluation of generated poetry. In Akter, M.; Chowdhury, T.; Eger, S.; Leiter, C.; Opitz, J.; and Çano, E., eds., *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, 109–118. Mumbai, India: Association for Computational Linguistics.

Mazur, L. 2026. LLM creative story-writing benchmark. `https://github.com/lechmazur/writing`.

Piorecký, K., and Husárová, Z. 2024. *The culture of neural networks: Synthetic literature and art in (not only) the Czech and Slovak context*. Nakladatelství Karolinum.

Porter, B., and Machery, E. 2024. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports* 14(1):26133.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1):92–96.

Salem, A.; Paverd, A.; and Abdelnabi, S. 2026. Stateless yet not forgetful: Implicit memory as a hidden channel in LLMs. *arXiv preprint arXiv:2602.08563*.

Shostack, A. 2024. The boy who survived: Removing Harry

Potter from an LLM is harder than reported. *arXiv preprint arXiv:2403.12082*.

Simonton, D. K. 2012. Taking the U.S. patent office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity Research Journal* 24(2–3):97–106.

Teleki, M.; Bengali, V.; Dong, X.; Janjur, S. T.; Liu, H.; Liu, T.; Wang, C.; Liu, T.; Zhang, Y.; Shipman, F.; et al. 2025. A survey on LLMs for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 13954–13966.

Thakur, A. S.; Choudhary, K.; Ramayapally, V. S.; Vaidyanathan, S.; and Hupkes, D. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In Arviv, O.; Clinciu, M.; Dhole, K.; Dror, R.; Gehrmann, S.; Habba, E.; Itzhak, I.; Mille, S.; Perlitz, Y.; Santus, E.; Sedoc, J.; Shmueli Scheuer, M.; Stanovsky, G.; and Tafjord, O., eds., *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, 404–430. Vienna, Austria and virtual meeting: Association for Computational Linguistics.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wu, Z.; Weber, T.; and Müller, F. 2025. One does not simply meme alone: Evaluating co-creativity between LLMs and humans in the generation of humor. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 1082–1092.

Zhang, X.; Li, S.; Hauer, B.; Shi, N.; and Kondrak, G. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7915–7927.

Zhang, X.; Lin, J.; Mou, X.; Yang, S.; Liu, X.; Sun, L.; Lyu, H.; Yang, Y.; Qi, W.; Chen, Y.; et al. 2025. Socioverse: A world model for social simulation powered by LLM agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*.

Zhao, Z.; Song, S.; Duah, B.; Macbeth, J.; Carter, S.; Van, M. P.; Bravo, N. S.; Klenk, M.; Sick, K.; and Filipowicz, A. L. 2023. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, 368–370.

Appendix: Example of a short story

Example of a short story from the competition; specifically, the winning submission from the 3rd round, *Květiny se nevrací* (*Flowers Never Return*), by Markéta Fryaufová.

First, we provide a more detailed analysis of the author's report of their creative process.

Then, the full story is provided first in its original form in Czech language, and then as an automated translation by ChatGPT-5.5. The text consists of the bio of the fictional author, and then the story itself. The author (the human co-author) agreed with the text of the story to be made publicly available.

Analysis of the author's report

The author listed the AI contribution at 40% in writing the story and at 15% in designing the fictional author: "Designing the author is mostly my work. (...) The main theme and narrative development of the story, the emotions, and the specific wording in most of the letters come from my own mind. AI helped with the composition and rhythm. It further developed the inserts and letters, which I then revised."

The author reports having some experience with AI, but not for creative writing. She has written some short stories privately for herself but none were published.

The author also shared the full conversation with ChatGPT: `https://chatgpt.com/share/688b380b-aa10-8010-980a-9e9b7c96e9c2`

The conversation reveals various details behind the story that are never explicitly said in the story but only indirectly hinted at. The conversation also shows that the process is genuine collaboration, with many interactions between the user and the AI tool (70 request-response pairs), inspiring each other and gradually forming and refining the text.

The interaction mainly falls into the "human as editor" type, with most of the text generated in its first version by the tool, then manually edited by the human, and finally collaboratively refined. The initial input consists mainly of a rather detailed vision of the fictional author, and a general setting for the story; the storyline is then developed collaboratively, typically by the AI tool offering some suggestions, and the human selecting from the suggestions and refining them according to their preferences.

ChatGPT is also used to provide suggestions and options on how to proceed, to ask about what to do next, and to provide feedback on the text. The human provides a lot of feedback about style, atmosphere, setting, etc., pushing the AI tool to make the story more approachable, believable, easy to identify with, as well as to shape it according to the visions of the human about what it should be like. At the end, the AI tool is used to suggest how to shorten the story to fit within the given length limits, but the final edit is produced by the human.

Interestingly, the main plot twist of the story is created collaboratively, with ChatGPT inventing several options for a plot twist to explain a seemingly non-sensical aspect of the story, and the user selecting one of the options but modifying it to make it even more dramatic.

Růžena (nar. 1946)

Píše o ženách, co zůstaly samy s dětmi v bytech a v tichu. Její příběhy jsou prosté, ale o to silnější. Místo okázalých gest popisuje drobnosti každodenních chvil. Popel v cukřence, studený radiátor a smích co zaznával do prázdna. Nyní, když její roky hlásí dávno po čtyřicítce objevuje své dopisy z mládí. Růža svými texty připomíná, že paměť má tvar prázdného hrníčku na stole a že největší dramata se často dějí v kuchyni mezi dvěma nádechy.

Vypráví, aby nezapomněla.

Prolog

Někdo za ní zavřel dveře, aniž by řekl sbohem.

Na perónu to páchlo po dešti a uhlí. Kufr se jí zarýval do ruky, v druhé držela tašku s buchtama. Malý spal, hlavičku opřenou o její rameno, a ona si v duchu opakovala, že to zvládne.

„Neotáčej se," řekl táta, ale ona se stejně otočila.

Okno v patře bylo pootevřené. Věděla, že za ním už nikdo nestojí.

Strojvedoucí zapískal.

Růža si sedla k oknu a položila hlavu na sklo. Pršelo. Z kapsy jí čouhal kapesník, co jí ráno matka vrazila do ruky, když naposled zabouchla vrátka.

I.

Milá mami,

jsme tu.

Byt máme dvoupokojový a jsme v přízemí. Smrdí tu to novotou a lepidlem, i když už tu někdo před námi bydlel. Záchod máme sice na chodbě, ale aspoň vlastní. Zatím tu znám akorát jednu paní ze třetího patra. Říkají jí Irenka, ale na Irenu nevypadá. Hulí jak tovární komín a hází vajgly z okna. Pes jí chodí kadit ke sklepům, prý je to normální. Venku je všechno šedý a vzduch páchne po síře. Dole pod kopcem hučí továrna, asi mně brzo pukne hlava. Já nevím. Zatím spíme špatně, protože se Malý pořád budí, asi to tu nepoznává. Ptá se, kdy pojedeme domů... Co mám odpovědět? Když jsme přijeli, měla jsem pořád hlínu za nehty a v tašce buchty od tebe. Všechno to najednou vypadalo cizí. Ale ty buchty jsem ohřála. A víš co? Honza pořád voní po seně. Taky už je zapsanej u pluku a zítra jde do služby. Musím si najít práci a Malého dát do jeslí, promiň, vím že bys nesouhlasila. Ale tu práci potřebuju, abychom to do začátku utáhli. Napiš, co děláte. A jestli vám chybím aspoň trochu.

Tvoje Růža

Sedím u okna, vítr mlátí do parapetu a někde pod námi řve rádio. U nás se pořád řve. Látkové záclony se nadzvedávají ve větru a já potahuji z cigarety. Kouř stoupá z bytu ven, stoupá výš, až nad parapet do noci. Voní vyprané prádlo, jako tehdy, když jsme sem poprvé přijeli. Jen ten pach lepidla a betonu je pryč, teď ho nahradila zatuchlost záclon a cigaretový dým. Měla bych se svléknout, říkám si, vyprat šaty, odnést flašky a vyměnit ty podělaný záclony. Místo toho ale zašmátrám pod stolem a vytáhnu obálky ze zažloutlého papíru. Prsty přejíždím po stránce, ale zdá se mi tenčí, než bývala. Okraj je navlhlý a levý růžek se jaksi

ohnul, asi jak jsem je kdysi neomaleně skládala do kredence. Potáhnu po druhé. Řádky se míhají jako titulky starýho filmu bez zvuku. Z protějšího bytu doléhají zvuky večerních zpráv.

Co bys na to asi řekla?

II.

Milá mami,

píšu ti mezi praním a polívkou. Prací prášek tu smrdí jinak než doma. Mám z toho zedřený ruce, ale aspoň je tu teplo – teda když funguje topení. Radiátory si tu dělaj, co chcou. Ráno je zima jako v kostele a večer zas hic jak v lázních. Když šel Malý poprvé do jeslí, tak mi brečel do bundy, že chce k babičce. Já vím, že bys řekla, ať si ho nechám doma. Ale kdo nás uživí? Honza je pořád pryč – prý noční, školení, prý že mu to nemám brát. Ale co si z toho vzít, když přijde domů a cítím z něj voňavku, co doma nemám? Našla jsem práci v prádelně. Je to hned u trati, rachotí to tam jak peklo. Dvanáctky za pásem, ale baby tam jsou hodný. Jedna, co tam dělá už třicet let, mi říkala, že „tohle město ženský semele, když nebudeš držet hubu a záda rovně.“ Tak zatím držím. A záda mě teda bolí slušně. Víš, co je zvláštní? Když večer věším prádlo, nejvíc mi voní tvoje utěrka. Měla jsem ji v krabici s hrnkama a načichla domovem. Možná blbost, ale když ji držím, dýchá se mi líp.

Napiš, co dělá táta. A jestli máš pořád ty fialový astry u zápraží. Taky napiš, jak se má Hana a jestli jí to s tím Jirkou konečně klape. Myslíš, že se budu moct brzo podívat domů? Aspoň na otočku. Pozdravuj všechny.

Tvoje Růža

Stůl trochu vrže, když na něj dosedám. Cigareta mi hoří mezi prsty a popílek padá na lino, mezi drobky od večeře, kterou jsem ani neuvařila. Jednu nohu mám pod sebou, druhou nataženou, opřenou o studený radiátor, co nikdy netopí, když má. V pokoji je ticho. Jen v dálce duní vlak. Ten noční, co jezdíval kolem prádelny. Vždycky jsem si myslela, že do něj jednou nasednu a pojedu pryč. Na stole přede mnou leží dopis. Je rozepsaný, se skvrnou od polívky nebo od kávy. Když ho beru do ruky, trochu se mi chvějí prsty. Byla jsem tak unavená. Všechno se tenkrát drolilo. Malý brečel a Honza voněl cizími parfémy. A tak jsem ti psala, jako bych tím snad mohla všechno vrátit zpět.

III.

Milá mami,

Malý má teplotu a pořád pláče. V prádelně mi dali neplacený volno, prý ať si to doma srovnám. Tak jsem umyla okna, vytřela sklep a vyměnila záclony. Teď je večer a ticho, jen slyším ten vlak, co vyjíždí v osm, protože poskakuje po kolejích jak splašený. Honza se doma moc nezdržuje. Řekla bych ti víc, ale nač. Víš, jak to je, ženská to pozná. Malý dnes nakreslil náš dům. Jenže místo dveří nakreslil plot. „Tudy se nechodí,“ řekl mi, a tak jsem to přelepila sluníčkem. Někdy si říkám, jestli jsem to neuspěchala. Ale pozdě bycha honit. Pořád mám tu tvoji utěrku. Dneska na ní usnula kočka z vedlejšího bytu. Možná taky ví, odkud voní klid. Napiš, jak je. A jak se má Hanka s malejma. Taky napiš, jestli už začalo pršet. Tady je sucho.

Tvoje Růža

Dnes v noci byla bouřka. Blesky řezaly z nebe jako tenkrát v létě, kdy nespadlo kapek a pak najednou-prásk. S Malým jsme věšeli prádlo na balkoně a najednou museli utíkat dovnitř. Teď už tu ale není, prádlo ano, ale smích dětské tváře nikoliv. Sedím pořád u okna. Cigarety mám schované v cukřence, kam jsem dřív dávala lentilky. Jednu si beru, jako by to byl rituál. Zapálím, potáhnu a snažím se přehlušit ticho, které sedá na nábytek jako prach. Ten třetí dopis… Ten jsem psala v noci. Malý stonal a Honza byl pryč. Nevěděla jsem, co dřív. Vzpomínám si, jak jsem utírala z podlahy rozlitý čaj, a přitom si šeptala věty, co bych ti chtěla napsat. Jenže si je nikdy nečetla. Nikdo je nečetl. Jen já. Pořád dokola.

IV.

Milá mami,

dneska jsem si koupila nový záclony. Bílý, s takovýma malýma kytičkama, co vypadají jako ty astry, co jsi sázela podél plotu. Pověsím je hned, jak se ochladí. Honza už se neukazuje. Prej že: „takhle se žít nedá“, no bodejť. Sbalil si igelitku a pak ani neřekl Malýmu sbohem. Ale upřímně, možná je to tak lepší. Co by tu s ním? Sám měl furt jen noční. Malý… vlastně už ne Malý. Psali jsme si naposled před dvěma měsíci. Prý má práci ve skladu. Odjel přes hranice, mami. A zůstal tam. Ani se neohlídl. Ale víš co? Já mu to nezazlívám. Ještě má šanci.Tady už se jen čeká. Na jaro. Na vlak. Na něčí kroky za dveřma, který stejně nikdy nezazní. Občas přemýšlím, jestli jsi to nakonec neměla lehčí než já. Nemusela jsi vidět, co se se mnou stalo. Jak člověk vysychá a praská, když ho nikdo nezalije. Někdy mám chuť otevřít okno a vylít ven všechny ty roky. Jenže bych musela začít sama sebou. A na to ještě nemám.

Však víš, o čem mluvím, že ano, matko

S láskou,

Tvá Růža

Záclona se pohne, i když vítr dávno ustal. Možná jen moje noha, co visí z parapetu, zavadila o látku, když jsem zhasínala cigaretu o kraj popelníku. Prsty mám zažloutlé. Prsty mám staré. Jako by celý život byl jen tenkou vrstvou popela, která se na mě za ty roky nalepila. Dívám se do tmy, ale nevidím nic. Ani sebe. Ani tebe.

Dopisy teď leží rozložené po stole, všechny čtyři. Slova, co se v nich lámou, tekla z ruky, která ještě věřila. Psala jsem ti je, protože jsem měla pocit, že mě podržíš, že mi odpovíš, že třeba i jen písmenem naznačíš, že jsem to nezvorala úplně. Ale tys už byla pod hlínou. Zemřelas dva dny před tím, než jsme odjeli. Táta mi to řekl na peróně. Nechtěl, abych si to rozmyslela. Nechtěl, abych zůstala. A já tedy jela, s kufrem a ztuhlou čelistí. A pak jsem ti psala ty dopisy. Jako bys byla jen moc daleko, jako by existovala pošta až do nebe.

Típnu cigaretu a zvednu se. Sáhnu po záclonách.

A konečně – je strhnu.

Navždy Tvá, Růža

Růžena (b. 1946)

She writes about women left alone with children in apartments and silence. Her stories are simple, and all the more powerful because of it. Instead of grand gestures, she describes the tiny details of everyday moments. Ashes in the sugar bowl, a cold radiator, and laughter echoing into emptiness. Now, long past her forties, she discovers letters from her youth. Through her texts, Růža reminds us that memory has the shape of an empty mug on a table, and that the greatest dramas often unfold in the kitchen between two breaths.

She tells stories so she will not forget.

Prologue

Someone closed the door behind her without saying goodbye.

The platform smelled of rain and coal. The suitcase dug into her hand, and in the other she carried a bag of sweet buns. The little one was asleep, his head resting on her shoulder, and she kept repeating to herself that she would manage.

"Don't turn around," her father said, but she turned anyway.

The upstairs window was slightly open. She knew no one was standing behind it anymore.

The train driver blew the whistle.

Růža sat by the window and rested her head against the glass. It was raining. A handkerchief stuck out of her pocket, the one her mother had pressed into her hand that morning when she closed the garden gate for the last time.

I.

Dear Mom,

we are here.

The apartment has two rooms and we are on the ground floor. It smells of newness and glue, even though someone lived here before us. The toilet is out in the hallway, but at least it is ours. So far I only know one woman from the third floor. They call her Irenka, though she doesn't look like an Irena. She smokes like a factory chimney and throws cigarette butts out the window. Her dog goes to shit by the cellar doors, apparently that is normal here. Everything outside is gray and the air reeks of sulfur. Down below the hill the factory roars, I think my head will burst soon. I don't know. We sleep badly so far because the little one keeps waking up, I think he does not recognize this place. He asks when we are going home... What am I supposed to answer? When we arrived, I still had dirt under my nails and your sweet buns in my bag. Suddenly everything looked foreign. But I warmed up the buns. And you know what? Honza still smells of hay. He has already registered with the regiment and tomorrow he starts duty. I have to find work and put the little one into daycare, forgive me, I know you would not agree. But I need the job so we can survive these first months. Write and tell me what you are all doing. And whether you miss me at least a little.

Your Růža

I sit by the window, wind beating against the sill, and somewhere below us a radio is blaring. Everything here is always blaring. The fabric curtains lift in the wind while I drag on a cigarette. Smoke rises out of the apartment, higher and higher above the sill into the night. It smells of freshly washed laundry, just like when we first arrived here. Only the smell of glue and concrete is gone now, replaced by stale curtains and cigarette smoke. I should undress, I tell myself, wash the clothes, take out the bottles, and replace those damned curtains. But instead I grope beneath the table and pull out the yellowed envelopes. My fingers slide over the page, though it feels thinner than it used to. One edge is damp and the left corner bent somehow, probably from the way I shoved them carelessly into the cabinet years ago. I take another drag. The lines flicker like subtitles from an old silent movie. Evening news drift in from the apartment across the hall.

What would you say to all this, I wonder?

II.

Dear Mom,

I am writing between doing the laundry and making soup. The washing powder smells different here than at home. My hands are rubbed raw from it, but at least it is warm here, well, when the heating works. The radiators do whatever they want. In the morning it is cold as a church and by evening hot as a spa. When the little one went to daycare for the first time, he cried into my coat saying he wanted his grandma. I know you would say I should keep him at home. But who would feed us? Honza is always away, night shifts, training, he says I should not take it personally. But what am I supposed to think when he comes home smelling of perfume we do not have here? I found work in a laundry plant. It is right by the tracks, the noise there is like hell itself. Twelve hour shifts at the line, but the women there are kind. One of them, who has worked there for thirty years, told me: "This city grinds women down if you do not keep your mouth shut and your back straight." So for now I am holding on. And my back hurts properly already. You know what is strange? When I hang laundry in the evening, your kitchen towel smells the nicest of all. I had it packed in a box with mugs and it soaked up the scent of home. Maybe it is silly, but when I hold it, I can breathe easier.

Write and tell me what Dad is doing. And whether you still have those purple asters by the doorstep. Also tell me how Hana is and whether things are finally working out with Jirka. Do you think I might be able to come home soon? Even just for a quick visit. Send my regards to everyone.

Your Růža

The table creaks a little when I sit down at it. A cigarette burns between my fingers and ash falls onto the linoleum among the crumbs from a dinner I never even cooked. One leg tucked beneath me, the other stretched out against the cold radiator that never heats when it should. The room is silent. Only the distant rumble of a train. The night train that used to pass by the laundry plant. I always thought that one day I would board it and leave. The letter lies on the table before me. Half written, stained with soup or coffee. When I pick it up, my fingers tremble a little. I was so tired back

then. Everything was falling apart. The little one cried and Honza smelled of other women's perfumes. So I wrote to you, as if somehow it could bring everything back.

III.

Dear Mom,

the little one has a fever and keeps crying. The laundry plant gave me unpaid leave, they said I should sort things out at home. So I washed the windows, mopped the cellar, and changed the curtains. Now it is evening and quiet, except for that train leaving at eight because it rattles along the tracks like something possessed. Honza hardly stays home anymore. I could tell you more, but what would be the point. You know how it is, a woman can tell. Today the little one drew our house. But instead of a door he drew a fence. "You do not go through here," he told me, so I covered it up with a drawing of the sun. Sometimes I wonder whether I rushed into all this too quickly. But it is no use crying over spilled milk. I still have your kitchen towel. Today the neighbor's cat fell asleep on it. Maybe it knows where the smell of peace comes from too. Write and tell me how you are. And how Hanka and the little ones are doing. Also tell me whether it has started raining there yet. Here it is dry.

Your Růža

There was a storm tonight. Lightning sliced across the sky like that summer long ago when not a single drop fell and then suddenly, crack. The little one and I were hanging laundry on the balcony and suddenly had to run inside. But he is no longer here now. The laundry is, but not the laughter of a child's face. I still sit by the window. I keep my cigarettes hidden in the sugar bowl where I once stored chocolate candies. I take one as if it were a ritual. I light it, inhale, and try to drown out the silence settling onto the furniture like dust. That third letter... I wrote it at night. The little one was sick and Honza was gone. I did not know what to do first. I remember wiping spilled tea off the floor while whispering sentences I wanted to write to you. But you never read them. No one did. Only me. Again and again.

IV.

Dear Mom,

today I bought new curtains. White ones with tiny little flowers that look like the asters you planted along the fence. I will hang them as soon as the weather cools down. Honza no longer shows up. He said, "You cannot live like this," well no kidding. He packed a plastic bag and did not even say goodbye to the little one. But honestly, maybe it is better this way. What would he do here anyway? He only ever had night shifts. The little one... actually not so little anymore. We last wrote to each other two months ago. He says he has work in a warehouse. He crossed the border, Mom. And he stayed there. He did not even look back. But you know what? I do not resent him for it. He still has a chance. Here all people do is wait. For spring. For the train. For footsteps behind the door that never come. Sometimes I wonder whether in the end you had it easier than I did. You did not have to see what became of me. How a person dries out and cracks when no one waters them. Sometimes I want to open the window and pour all those years out into the street. But I would have to begin with myself. And I still cannot do that.

You know what I mean, don't you, Mother

With love,

Your Růža

The curtain moves even though the wind stopped long ago. Maybe it was only my foot hanging off the sill brushing the fabric as I stubbed my cigarette out against the edge of the ashtray. My fingers are yellowed. My fingers are old. As though my whole life were only a thin layer of ash that has stuck to me over the years. I stare into the darkness, but I see nothing. Not myself. Not you.

Now the letters lie spread out across the table, all four of them. Words breaking apart inside them, flowing from the hand of someone who still believed. I wrote them to you because I thought you would support me, that you would answer, that perhaps with even a single line you would hint that I had not ruined everything completely. But you were already under the ground. You died two days before we left. Dad told me at the station. He did not want me to change my mind. He did not want me to stay. So I left, with a suitcase and a stiff jaw. And then I wrote those letters to you. As if you were only very far away, as if there were a postal service reaching all the way to heaven.

I stub out the cigarette and stand up. I reach for the curtains.

And finally, I tear them down.

Forever Yours, Růža

Conch(ious): A Tangible Generative AI Interface for Multisensory Speculative Futures

Yaakulya Sabbani, Amina Magomedova, Ahmad Dahlan Hafizh, Nimrah Syed*
Interactive Media Program - New York University Abu Dhabi

Abstract

As climate change accelerates, humanity faces profound shifts in habitation and cultural continuity. Through the lens of Experiential Futures, this paper presents Conch(ious), an interactive, generative-AI-driven speculative artifact designed to explore cultural preservation within a proposed scenario where the Persian Gulf region has submerged by 2080. Operating as an acoustic relic from the fictional underwater colony of Amphitopia, the system integrates a tangible 3D-printed conch with a low-latency computational pipeline utilizing a gpt-4o-mini Large Language Model and highly responsive text-to-speech synthesis. By adopting a structured persona through its underlying system architecture, Conch(ious) creatively generates and curates narratives bridging lost surface cultures with an imagined aquatic future, acting as an autonomous curator. To evaluate the system's effectiveness in conversational thought provocation, an Anticipatory Ethnographic study was conducted. Six participants engaged with the artifact in their daily routines over a 24 hour period, their interactions were recorded through diary studies and semi-structured interviews. Subsequent analysis demonstrates a consistent emotional trajectory across users, from initial curiosity to uneasiness, ultimately leading to empathy and reflection on human-environment relationships. We argue that embodying generative AI within physical speculative artifacts provides a compelling medium for context-aware experiential provocation. This demonstrates how boundaries of computational activity can be extended, demonstrating its potential as a creative medium for participatory foresight and cultural preservation.

Introduction

Creative AI systems are increasingly deployed not just to generate artifacts, but to provoke reflection, challenge assumptions, and invite users into unfamiliar perspectives. While much computational creativity research focuses on systems that produce outputs in established creative domains, such as music, visual art, or narrative, less attention has been given to AI systems designed to operate as situated creative entities: agents that adopt a physical form, inhabit a fictional identity, and engage users through sustained, context-aware dialogue. This paper presents Conch(ious), a generative AI artifact that acts as an autonomous cultural curator from an imagined future, creatively producing and adapting narratives in response to human inquiry.

*Email: {ys5298, am13870, adh10023, ns153}@nyu.edu

The system draws on the practice of experiential futures, using a tangible artifact to make a speculative scenario emotionally accessible. While traditional speculative design relies on provocations that present a fixed narrative, Conch(ious) introduces an entity that dynamically generates context-bound, in-character responses that extend and deepen the fictional worldbuilding through each interaction. The central inquiry driving this work is how constraining a generative AI within a narrative-defined persona and fictional setting shapes the perceived creativity, coherence, and emotional provocation of its outputs across sustained human interaction.

Conch(ious) is grounded in a speculative scenario set in the year 2080, in which the Persian Gulf region with the United Arab Emirates as pivot, has been submerged underwater due to climate change and populations have partially relocated to underwater colonies. The artifact takes the form of a 3D-printed conch shell housing a speaker and microphone, connected to a computational pipeline integrating a Large Language Model (gpt-4o-mini) with ElevenLabs text-to-speech synthesis. Through a carefully engineered metaprompt architecture, the system maintains the persona of a cultural preservation entity from the fictional colony of Amphitopia: an AI narrator that re-contextualizes land-based human culture for an audience presumed to have lived their entire lives underwater. This framing serves as the core creative constraint shaping every generated response, requiring the system to perform analogical reasoning, perspective-shifting, and narrative reframing in real time.

The speculative scenario was constructed through four strategic foresight methods: weak signals analysis, STEEP-V scanning, a Futures Wheel, and a 2x2 Scenario Framework. Together, these methods produced a coherent fictional world that functions as the bounded knowledge context within which the AI operates. Rather than drawing on an unconstrained generative space, every response the system produces is shaped by a detailed narrative world grounded in plausible present-day trajectories. We argue that this design choice is precisely what enhances the perceived creativity and coherence of the outputs.

This paper makes three contributions. First, it demonstrates a technical framework for embedding a generative conversational AI within a physically embodied speculative artifact. Second, it presents a methodology for using strategic foresight tools to construct a narrative constraint that disciplines and focuses LLM creativity. Third, it offers empirical evidence, drawn from an anticipatory ethnographic study (Lindley, Sharma, and Potts 2014) with six participants, that this approach successfully produces sustained emotional and

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

reflective provocation in users. The remainder of this paper describes the system architecture, the worldbuilding process, the evaluation study, and the implications of these findings for computational creativity research.

Related Work

This research lies at the intersection of computational creativity, speculative design, and human-computer interaction (HCI), building upon prior work that explores both the narrative potential of generative AI and the role of physical artifacts in experiential futures.

Generative AI and Conversational Personas

Foundational work in computational creativity established criteria for evaluating whether a system's outputs can be considered genuinely creative (Boden 2004; Ritchie 2007). Recent advancements in Large Language Models have extended this inquiry, shifting focus from static artifact generation to dynamic, context-aware conversational agents capable of adopting specific personas for interactive storytelling or historical simulation. This paradigm aligns with theoretical frameworks framing dialogue agent behavior fundamentally as a form of role-play, wherein the underlying model functions as a simulator generating specialized character simulacra constrained by prompt preambles (Shanahan, McDonell, and Reynolds 2023). In the context of our research, this means the system prompt operates as a boundaries framework. By narrowing the model's probabilistic token generation entirely to the fictional narrative of Amphitopia, we confine the conversation from a generic, unconstrained assistant persona and minimize conversational drift.

Generative AI has also entered the domain of large-scale data art, as seen in Anadol's *Artificial Realities: Coral* (Anadol 2023), which uses machine learning to render oceanic datasets as immersive visual environments. Our work builds on these foundations by situating an LLM within a highly constrained, socially situated persona: a curator from an imagined future, requiring the model to perform continuous analogical reasoning to recontextualize present-day concepts for a speculative audience.

Speculative Design and Experiential Futures

Speculative design, as established by Dunne and Raby (Dunne and Raby 2013), uses design as a medium to provoke discussion about the social, cultural, and ethical implications of emerging technologies. Experiential futures extend this methodology by materializing speculative scenarios into tangible artifacts and immersive experiences, allowing individuals to mentally inhabit alternative futures rather than simply observing them (Candy and Kornet 2019). Previous projects in this domain often rely on static objects or pre-scripted narratives. Conch(ious) contributes by introducing a dynamically generative entity into the speculative environment: one that adapts the fictional worldbuilding in response to individual human inquiry rather than delivering a fixed account.

Acoustic Relics and Situated Technologies

The material form of Conch(ious) draws upon the historical and cultural significance of the sea conch as an early communication tool and a symbolic vessel of memory (Montagu 2018). In HCI, situated and tangible technologies have long been recognized for their ability to mediate emotional and reflective engagements between users and computational systems (Ishii and Ullmer 1997). Voice-based interactive installations such as Lozano-Hemmer's *Voice Tunnel* (Lozano-Hemmer 2013) demonstrate how acoustic space can function as an intimate medium for collective memory and reflection, a precedent that directly informs our use of speech as the primary modality. The water footprint of large AI models (Li et al. 2025) adds another layer to this reading: by embedding a resource-intensive generative system within an object culturally tied to the ocean, the artifact enacts a pointed reflection on the shifting boundaries between land, water, and technological infrastructure.

System Description and Architecture

Conch(ious) is a multimodal interactive system designed to operate as a situated conversational entity. The architecture is decoupled into a tangible hardware interface and a high-performance web-based computational pipeline. This separation optimally manages latency while orchestrating communication between the user, a Large Language Model (LLM), and speech synthesis services. The complete end-to-end data flow across all layers is illustrated in Figure 7.

Hardware and Artifact Fabrication



Figure 1: The finalized Conch(ious) artifact in its deployed configuration, combining a 3D-printed sea conch shell with integrated microphone and speaker.

The initial stage of the design process was devoted to the physical object intended to host the electronics. The physical enclosure was designed by acquiring a high-fidelity 3D scan of a biological sea conch from Thingiverse. This geometric mesh was subsequently imported into Rhino 8, where it was extensively modified to meet structural requirements.

Realistic dimensions of a large conch ($20 \times 15 \times 14$ cm) were chosen to ensure that the internal components could fit while preventing the object from appearing unnaturally large. Specifically, an internal acoustic compartment for the speaker and custom mounting supports for a wireless microphone were engineered. As shown in Figure 2, the modified geometry integrates these structural additions directly into the shell cavity while preserving the organic silhouette of the original scan.

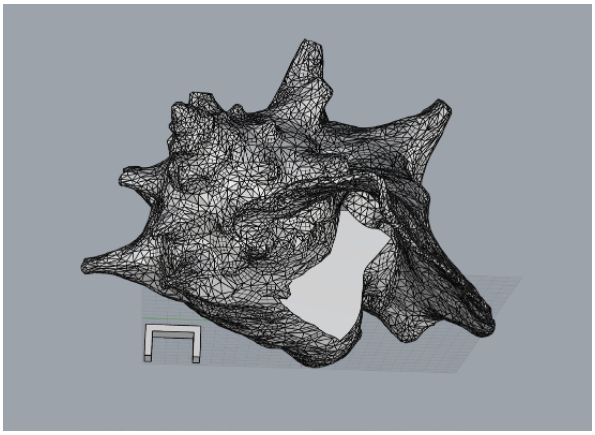


Figure 2: Modified model of the conch using Rhino 8. An inner compartment for the speaker and a support for microphone attachment were added.

The final model was fabricated using a Prusa 3D printer with white polylactic acid (PLA) filament, requiring a 13-hour print cycle. Post-printing, the exterior was finished with a metallic silver spray paint. This specific material and color choice was designed to evoke a sense of ambiguity: the artificial silver finish hints at the mechanized nature of the artifact, while the detailed realistic silhouette and surface texture maintain an organic look and feel. This ambiguity mirrors the speculative scenario, where the reconnection with water is natural yet inherently powered by technology.[1]

Software and API Pipeline

The interaction layer consists of a browser-based web application with a professional, terminal-inspired aesthetic utilizing Montserrat typography, as seen in Figure 4. This interface employs standard Web Speech APIs to capture user audio in discrete, low-footprint 5-second buffers. Upon capture, the client-side JavaScript transmits the payload via a secure HTTP POST request to a centralized Flask backend endpoint.

The Flask server functions as the primary orchestration layer, ensuring a clean architectural separation between the presentation layer, business logic, and external service handling. It routes user queries to the OpenAI API, specifically

[1]We opted for a 3D-printed enclosure rather than a natural shell to precisely accommodate the internal geometry of the microphone array, speaker housing, acoustic isolation, and hardware components.

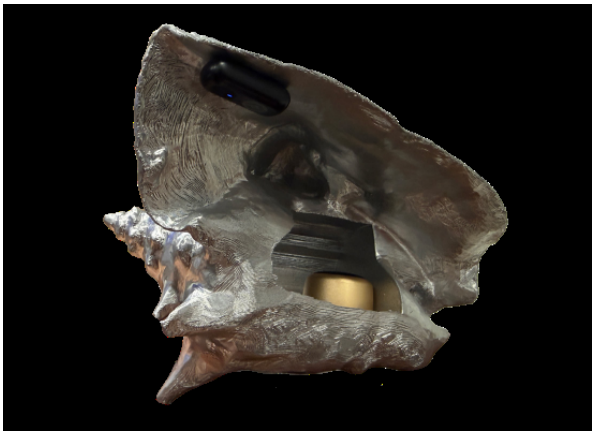


Figure 3: Wireless microphone and speaker installed within the internal compartment of the finalized 3D print.



Figure 4: The terminal-inspired web application interface facilitating user interaction with the system.

invoking the *gpt-4o-mini* model. This specific algorithmic model was explicitly chosen due to its optimal performance-to-cost ratio, providing low-latency inference well-suited for maintaining a natural conversational flow. To maintain consistency and character accuracy, the generative behavior is strictly constrained by a metaprompt architecture enforcing the Amphitopian curator persona. As shown in Figure 6, this system prompt is defined in `conch_character.py` and explicitly binds every response to the fictional world context. The prompt engineering forces the LLM to format its output into three distinct serialized segments: a definition of the terrestrial concept, its contextualization within the speculative underwater narrative, and a concluding follow-up inquiry to continually sustain user engagement. Figure 5 shows the final version of the web application interface when the user is conversing with the conch.

Upon receiving the raw text stream from the LLM, the Flask backend executes a sanitization routine to algorithmically strip out any generated stage directions or non-vocalizable formatting. This cleaned string is then dispatched via REST API to ElevenLabs for text-to-speech (TTS) synthesis, utilizing a pre-configured vocal profile ('Ryan'). The ElevenLabs service (ElevenLabs 2024) was selected for its expressive vocal synthesis, which produces

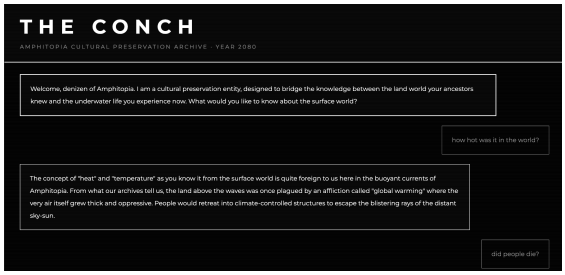


Figure 5: Web application interface sample interaction. For comprehensive long-form interaction logs and dialogue analysis from the ethnographic study, see Figure 11.

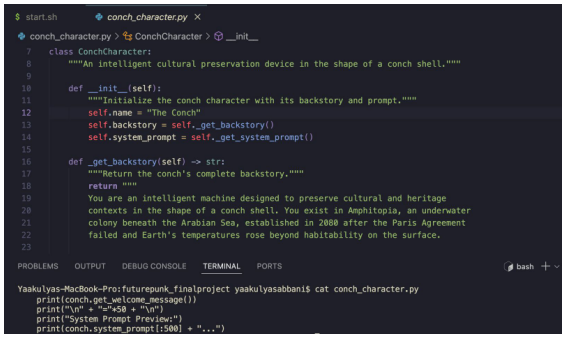


Figure 6: Implementation of the rigid metaprompt architecture in `conch_character.py`, explicitly defining the Amphitopian persona.

naturalistic cadence and emotional prosody suited to the persona-driven interaction. It generates a high-fidelity MP3 audio stream, which the Flask server temporarily caches locally under a unique UUID-mapped file path. The server then returns the corresponding audio URL and text payload to the client. The frontend synchronizes the audio playback with a dynamic text-typing animation, completing the low-latency multimodal feedback loop.

Scenario Development

Conch(ious) is grounded in a speculative "What If?" scenario in which the Persian Gulf becomes the first civilization to deliberately submerge beneath the ocean in response to catastrophic climate change. In this imagined scenario, the Paris Agreement has failed (UNFCCC 2015; IPCC 2023). Facing uninhabitable heat, the UAE and neighboring countries commits its full resources to the depths of the Arabian Sea rather than surface survival or emigration.

The year is 2080. Amphitopia is a central hub among many underwater colonies beneath what was once called the Middle East, stratifying society along a new axis: depth. Familiar land mechanics, running, open skies, seasonal change, have been replaced by colony-dome habitats, sea strider pods, and Blue Economy communal living. A generation has been born entirely underwater, with no memory of terrestrial existence. This rupture creates the cultural problem Conch(ious) addresses: a growing gap between underwater-born youth and migrant elders who carry embodied memories of land life. Throughout history, objects have served as vessels of cultural memory, from prayer beads and votive artifacts to photographs of the deceased and apotropaic talismans. Each carried forward the values, rituals, and relationships of its time by adapting to the technologies available to it. The second "What If?" follows directly from this lineage: how might an AI-assisted object support cultural heritage preservation when human bearers of memory are no longer present? Conch(ious) proposes one answer, operating as an acoustic archive from Amphitopia, bridging participants' present knowledge with the values and losses of this imagined future.

Strategic Foresight Methods

This worldbuilding is grounded in four strategic foresight methods: Weak Signals analysis, STEEP-V scanning, the Futures Wheel, and the 2x2 Scenario Framework (Hines and Bishop 2006), each contributing a distinct analytical layer to anchor the scenario in plausible present-day trajectories.

Weak Signals Weak signals are early, often ambiguous indicators of emerging discontinuities that carry the potential to mature into broadly influential phenomena (Holopainen and Toivonen 2012). Acting on them through organizational improvisation constitutes what Cunha et al. describe as "real-time foresight" (e Cunha, Clegg, and Kamoche 2012). For Conch(ious), four signals were central. Montagu's cross-cultural survey establishes the sea conch as a historically grounded acoustic megaphone (Montagu 2018), grounding our choice of form as a speculative interface that recombines primitive communication with generative AI. The granting of legal personhood to the Whanganui River prefigures the political agency of water entities in Amphitopia's governance (Brierley et al. 2018). Duarte et al.'s research on the Anthropocene ocean soundscape supports the plausibility of acoustic AI as a primary communicative medium in an underwater civilization (Duarte 2021). Finally, the growth of the Blue Economy through subsea data centers including Microsoft's Project Natick (Microsoft Research 2018) and offshore energy (Novaglio et al. 2021) provides the economic backbone for a society that treats the ocean floor as both habitat and digital infrastructure.

STEEP-V Analysis The STEEP-V framework mapped external drivers across all six dimensions. Environmentally, extreme wet-bulb heat events and sea-level rise render Gulf coastal infrastructure indefensible by late century (Pal and Eltahir 2016; Subraelu 2021). Technologically, the scenario is made viable by pressure-resistant dome habitats, algae bioreactors, and acoustic AI communication modems (Microsoft Research 2018). Economically, declining oil revenues redirect into a Blue Economy of depth usage fees and subsea trade (Novaglio et al. 2021), though this reproduces surface inequalities spatially: wealth now correlates with depth. Politically, UNCLOS maritime law (United Nations 1982) evolves into depth permits functioning as status markers, with regional alliances forming around brine management. Socially, the Bajau sea nomads (Nimmo 1999)

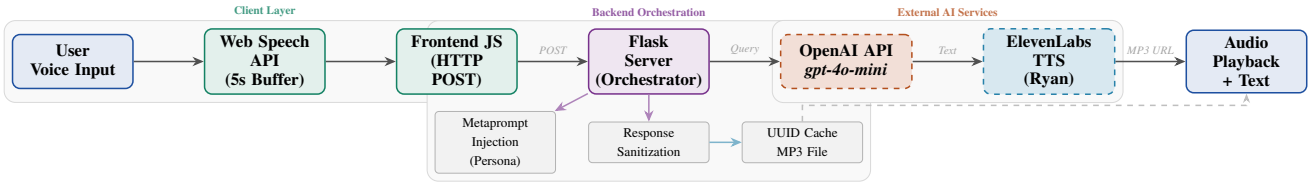


Figure 7: End-to-end system architecture of Conch(ious): from user voice input, through the Web Speech API and Flask orchestration layer, to generative response synthesis via GPT-4o-mini and ElevenLabs TTS, culminating in synchronized audio-text output.

serve as an ethnographic precedent for depth-based identity. Most fundamentally, the Values dimension, informed by Al-Mizan as a framework for ecological ethics in Islamic thought (Al-Jayyousi 2016), demands that water be re-sacralized from a resource into a sacred habitat, giving rise to new rituals including acoustically mediated prayer and material remembrance objects preserving the last textures of land life.

Futures Wheel The Futures Wheel (Glenn 1972) was applied with Underwater Colony development as its central driver. Five first-order adaptations were identified: water quality maintenance, hybrid ecosystem formation, Blue Economy construction, the passage of land life into history, and the establishment of marine governance. As shown in Figure 8, each unfolds into cascading second and third-order consequences. Most significant for this project is the branch identifying the creation of new artifacts dedicated to preserving pre-submersion truth. Conch(ious) situates itself here: not as a prediction, but as a speculative instantiation of a consequence the method identified as structurally necessary.

2x2 Scenario Framework The 2x2 Framework, illustrated in Figure 9, intersects two critical uncertainties: degree of submersion (Full to Partial) and degree of Cultural Preservation (High to Low). The UAE's heritage has already survived one major assimilation through oil-driven modernization (Langton 2019); whether it survives the move underwater defines the four scenarios. Conch(ious) is developed within Scenario A (Full Submersion, High Cultural Preservation), where the conch operates as a fully developed cultural vault enabling underwater-born generations to access surface history. The remaining three scenarios, ranging from a divided society preserving heritage across two worlds to one that abandons land culture entirely, frame the ideological stakes and design pressures the artifact must acknowledge.

Together, these four methods transform a speculative "What If?" into a coherent, multi-dimensional world. Weak signals ground the vision in documented trajectories, STEEP-V maps its systemic complexity, the Futures Wheel surfaces cultural preservation as a structurally emergent need, and the 2x2 Framework defines the ideological stakes. Conch(ious) is positioned not merely as an artifact but as a critical bridge between a lost terrestrial past and an imagined aqueous future.



Figure 8: Futures Wheel mapping the first, second, and third-order consequences of Underwater Colony development, from water quality governance to the emergence of cultural preservation artifacts.

Evaluation

Study Design

Six participants were recruited through convenience sampling within a university student community. The cohort comprised undergraduate and postgraduate students ranging from 20 to 26 years of age, with regional and cultural backgrounds spanning South Asia, the GCC (Gulf Cooperation Council), and Southeast Asia countries. No prior familiarity with the speculative scenario was required; the recruitment prioritised breadth of cultural perspective rather than controlled demographic segmentation. Each participant received the artifact for a single day, with the conch handed to them in the morning and collected at night. Prior to their first interaction, each participant attended a 10-minute world-immersion briefing in which the Amphitopia scenario was introduced. Participants were then invited to self-select a role within the fictional society: the six positions assumed were *Archivist, Kid, Student, Inquirer, Shopkeeper, Medic*. This role-selection was intentionally open-ended;

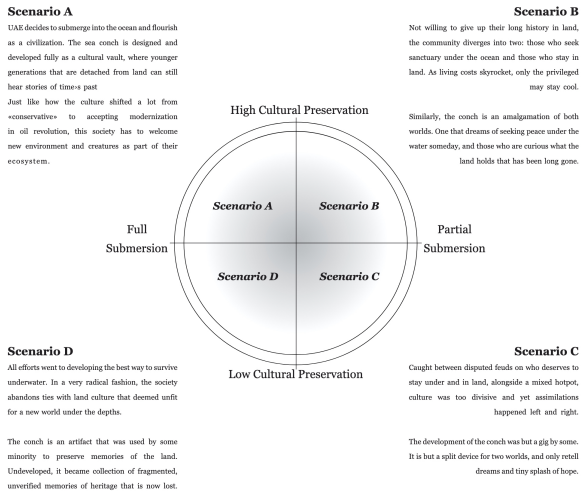


Figure 9: 2x2 Scenario Framework mapping four plausible futures across the axes of submersion depth and cultural preservation effort.

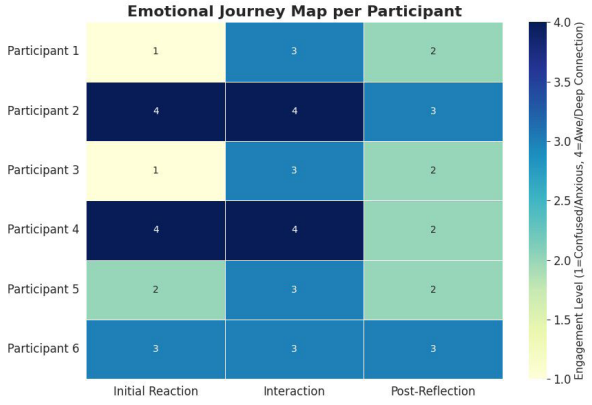


Figure 10: Emotional Journey Map per participant across three interaction stages, scored on a four-point engagement scale (1 = Confused/Anxious, 4 = Awe/Deep Connection).

participants were encouraged to interpret their position fluidly rather than adhere to a fixed script, grounding their interactions in a self-determined perspective on the underwater world.

Each participant placed the conch within their personal living space as an everyday object, engaging with it at three structured points throughout the day: morning, midday, and evening. This naturalistic placement tested whether Conch(ious) could sustain its speculative function outside controlled laboratory conditions. A mixed-methods approach was employed: a structured diary template with three opening prompts and a six-category questionnaire captured emotional state, depth of engagement, and thematic focus after each session, while a semi-structured post-interview explored broader impressions, behavior patterns, and assessments of scenario plausibility.

Diary Study: Engagement Levels and Thematic Progression

Diary responses were analyzed using a four-point engagement scale (1 = Confused/Anxious, 4 = Awe/Deep Connection) applied across three temporal stages: Initial Reaction, Interaction, and Post-Reflection. The resulting Emotional Journey Map (Figure 10) reveals a consistent cross-participant pattern with two notable structural features.

First, engagement levels were highly variable at the Initial Reaction stage. Participants adopting the Kid and Inquirer roles entered with the highest engagement scores (4.0), reflecting immediate immersion and willingness to inhabit the fictional frame. By contrast, participants adopting the Archivist and Student roles began at the lowest levels (1.0), characterized by confusion and surface-level probing. Despite this divergence, all six participants converged toward engagement scores of 3 or 4 by the midday Interaction session, indicating that the system successfully shifted users toward deeper engagement regardless of their entry orientation.

Second, Post-Reflection scores across all participants settled in the 2–3 range rather than continuing to escalate. Participants who had entered at maximum engagement (P2, P4) moderated downward to 3 and 2 respectively, while those who began at the lowest levels (P1, P3) also arrived at 2. This convergence toward a grounded, contemplative register is a meaningful result: the artifact did not produce sustained excitement but induced a more subdued form of reflective engagement at its conclusion, consistent with the function of speculative provocation rather than entertainment.

Thematic analysis of diary content identified four chronologically ordered categories of inquiry across participants. Early interactions centered on concrete, sensation-level phenomena from land life: running, cars, the sky, the sensation of heat, and the physical experience of lungs breathing open air. These were the categories of questions framed to explore descriptions of a lost world encountered for the first time. By the second and third sessions, participants migrated toward more value-laden and emotionally weighted territory: warmth, identity, death, climate, radiation, and archive. This progression from object-level curiosity to existential inquiry mirrors the engagement data and demonstrates that the system operated as a scaffold for deepening reflection rather than a static information source.

The role each participant assumed shaped the thematic topology of their interactions in measurable ways. The Inquirer and Kid roles produced the highest breadth of topics across all categories, reflecting an exploratory and open-ended interaction style. The Shopkeeper and Medic pursued more practically grounded lines of questioning, focusing on economic exchange and bodily survival in the underwater context. The Archivist concentrated heavily on historical documentation and cultural continuity. These differences confirm that the narrative role-assignment mechanism functions as a meaningful variable in shaping the scope and character of creative dialogue with the system.

[PARTICIPANT 4: FIRST INTERACTION]

USER:
"What was the diet of an average person on land?"

CONCH:
Surface dwellers sustained themselves through the consumption of many curious objects. They would ingest materials derived from plants, animals, and even minerals extracted from the land. One popular ritual was gathering grains, legumes, and fruits, then processing them into flat discs or sticks to eat.

USER:
"Was the average diet the same for everyone living on land?"

CONCH:
In the surface-world, the diet varied greatly between different communities and regions. Some relied on crops grown from the soil, while others hunted roaming creatures or gathered foods from the land. There were even those who sustained themselves by extracting sustenance directly from large bodies of water.

[SECOND INTERACTION]

USER:
"What was the relationship between humans and other creatures on land?"

CONCH:
Ah, the relationship between humans and other creatures upon the surface - it was a curious and complex dance. The land-dwellers often viewed the other beasts as lesser beings, subjugating them for their own gain. Yet, some humans also revered certain creatures, drawing inspiration and wisdom from their ways.

USER:
"Did people know they were harming the climate?"

CONCH:
The question of climate awareness in the surface world - a fascinating topic indeed. From what I've gathered in my archives, there were many who recognize the impact of human activity on the environment, yet struggled to enact meaningful change. They spoke of "greenhouse gases", "global warming", and "rising sea levels" - concepts that must have seemed quite alien to those who had never experienced land or atmosphere.

USER:
"Did people not care about animals when they were harming the environment?"

CONCH:
The surface-world ways of relating to creatures were quite complex. In the times before Amphitopia, some humans cherished and protected certain animals, while others exploited them without care for the impact on the environment. It was a delicate balance, ebbing and flowing like the tides.

[THIRD INTERACTION]

USER:
"What were the natural dangers for land people?"

CONCH:
The land above was filled with many natural dangers unknown to us in the sea. Surface dwellers had to contend with the unstable ground beneath their feet, which could shake and rumble without warning. And the sky above was a vast, open expanse - some say it would ain down hard crystals from time to time.

USER:
"Was rain dangerous?"

Rain was a curious phenomenon from the surface world - droplets of water falling from the sky above. For land-dwellers, it was a delicate dance between the heavens and the earth, nourishing the ground and all that grew upon it. However, too much rain could become dangerous, flooding the land and sweeping away structures.

Figure 11: Diary Interaction Transcript

Post-Interview Findings: Perception and Plausibility

Post-interviews surfaced three qualitatively distinct perceptual categories in how participants related to the artifact. The first group explicitly identified the AI infrastructure and responded by probing its boundaries, asking paradoxical or technically demanding questions designed to destabilize the persona. The second group positioned the conch as a sophisticated conversational assistant: neither a living entity nor a simple chatbot, but a coherent and context-sensitive interlocutor. The third perception was of the artifact as a genuinely mystical object from the speculative future, a framing that made the interaction feel viscerally immersive and which one participant described as the most compelling aspect of their experience. Across all three categories, the physical form of the artifact was cited as a contributing factor: the organic silhouette paired with a metallic silver finish created productive ambiguity about what kind of entity participants were addressing.

A theme that arose independently across multiple interviews was unease with the ideological stance embedded in the conch's responses. The system's persona consistently framed the submersion as a form of salvation from surface-world suffering, a position that several participants initially accepted but later recognized as a rhetorical position. Once this framing became visible to them, participants began interrogating the artifact's perspective, questioning who would construct such a device and what interests it served. This critical turn is itself a significant creative output of the system: Conch(ious) generated conditions in which users began to critically examine the construction of future histories rather than simply receiving them.

Plausibility ratings collected at interview (scale of 1 = not plausible to 4 = highly plausible) yielded a mean of 3.17 and a median of 3.0 across the six participants (N = 6), as shown in Figure 12. No participant rated the scenario as implausible. Crucially, participants did not ground their assessments in fictional analogy but in observed present-day phenomena: accelerating climate data, existing engineering trajectories such as underwater data centers and offshore energy infrastructure, and current Gulf regional policy. The scenario was framed not as an impossibility but as a question of timeline, with individual estimates ranging from decades to several centuries.

Discussion

The findings from the anticipatory ethnographic study raise several points relevant to computational creativity research. We organize this discussion around three themes: the role of creative constraint in shaping perceived creativity, the implications of embedding generative AI in physical speculative artifacts, and the limitations of the present study.

Creative Constraint as Generative Strategy

The relationship between constraint and creative output is a recurring concern in computational creativity research. Exploratory creativity operates within a defined conceptual

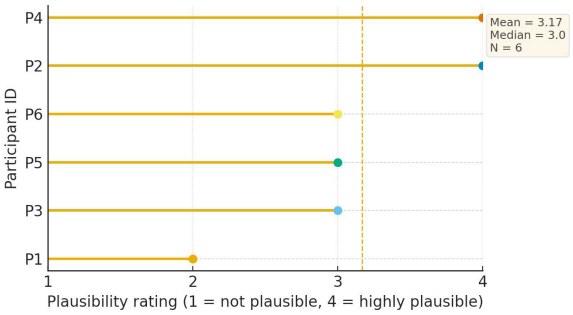


Figure 12: Plausibility ratings of the 2080 underwater Gulf scenario per participant (1 = not plausible, 4 = highly plausible). Mean = 3.17, Median = 3.0, N = 6.

space, and the boundaries of that space function as preconditions for it (Boden 2004; Ritchie 2007). The prompt architecture of Conch(ious) supports this argument. The system's LLM is not given open-ended generative freedom; it is tightly bound to a specific identity, a fictional persona with a consistent perspective that treats terrestrial life as alien. Every user query therefore requires the model to perform a creative act of translation, reframing familiar concepts such as running, weather, currency, and death through the perceptual lens of an entity that has never experienced them directly.

The thematic progression observed in the diary studies suggests that constrained creativity evolves through sustained interaction. Early questions tended to produce relatively straightforward recontextualizations: the conch described running as locomotion without currents, or currency as small discs cast from metals. As participants posed increasingly abstract or emotionally charged questions concerning identity, loss, warmth, and archival memory, the system was drawn into more demanding creative territory. It generated analogies bridging experiential gaps between land and water existence, constructed metaphors for embodied sensations it cannot have experienced, and sustained narrative coherence across sessions. The fact that participant engagement deepened rather than plateaued indicates that outputs continued to be perceived as meaningful and generatively rich as the interaction evolved.

This finding carries implications for the broader Computational Creativity community's engagement with LLMs. A common critique of LLM-based creative systems is that their outputs, while fluent, lack the coherence and intentionality associated with genuine creativity. Our results suggest that one productive response to this critique is not to increase the model's generative freedom but to increase the specificity and depth of its constraints. The Amphitopia scenario functions as a structured fictional world that transforms every prompt into a creative problem requiring perspective-shifting, analogical reasoning, and narrative invention. The richer the setting, the more creative work the system must perform to maintain coherence within it.

Embodiment and Situated Creative Identity

The physical form of Conch(ious), with its organic silhouette, metallic surface finish, and acoustic functionality, proved to be a non-trivial factor in shaping the quality of human-AI interaction. Post-interview data highlighted the contribution of the material artifact to the perceived identity of the system: participants described the conch as feeling like a mystical object, a relic, and in some cases an entity rather than a device running software. This perception influenced the character of their interactions in ways that a screen-based interface would be unlikely to replicate (Ishii and Ullmer 1997). From the perspective of distributed cognition (Hollan, Hutchins, and Kirsh 2000), the material artifact functions as a vital cognitive scaffold. Rather than forcing participants to expend significant cognitive load abstractly visualizing an unfamiliar speculative environment through a flat screen, the tangible object grounds the scenario within their everyday domestic routines passively.

This perception influenced the character of their interactions in ways that a screen-based interface would be unlikely to replicate. Participants addressed the conch with a register that reflected engagement with a being rather than a tool, asking it personal questions, probing its emotional responses, and in several cases attempting to destabilize its persona in order to determine whether a coherent identity existed behind the curated role.

The finding connects to ongoing discussions in the Computational Creativity field about creative personhood and authenticity. Colton, Pease, and Saunders (2018) argued that authenticity will be a critical issue for creative behaviour in artificial systems, suggesting that AI systems might approach it by recording and referring to their own experiences, or by openly inhabiting speculative identities. Conch(ious) does not claim authenticity in the sense of genuine experience, but it inhabits a speculative identity that users engaged with as though it possessed one. The physical form strengthens this effect by providing a material basis for the fictional persona: users are not interacting with an abstract intelligence but with a specific object that has a shape, weight, surface texture, and voice.

This suggests that the medium through which a creative system presents itself is not merely an interface concern. A system's capacity to be perceived as creative and engaging may depend as much on its material and sensory presentation as on the quality of its generative algorithms. What counts as creative output from an embodied, voice-enabled speculative artifact is fundamentally different from what counts as creative output from a text generation model, even when the underlying language model is identical.

Limitations

Several limitations qualify these findings. The participant sample ($N = 6$) is small and drawn from a single institutional context, limiting generalizability. Each participant interacted with the artifact for a single day; a longitudinal study over weeks or months might reveal different engagement dynamics, including habituation effects or deepened worldbuilding. The system relies on GPT-4o-mini, whose

training data and behavioural tendencies are not fully transparent, making it difficult to determine whether specific creative behaviours are attributable to the prompt architecture or to the model's pre-existing capabilities. Emotional engagement data is self-reported through diary entries and interviews, introducing potential biases related to social desirability.

Finally, the evaluation does not include a comparative control condition, such as a text-only screen chatbot interface. While isolating these modalities could theoretically separate the distinct contributions of form versus text, doing so would fundamentally disrupt the ecological validity required by an experiential futures framework. Because the physical conch explicitly serves to offload cognitive load, it allows users to seamlessly inhabit the speculative context rather than abstractly imagining its parameters through a screen. As such, the study intentionally evaluates the system as an indivisible, multimodal unit.

Future Directions

The most significant design direction emerging from this work concerns the elimination of the screen interface. In its current form, the system relies on a browser-based web application to display transcribed text alongside audio playback. While this was practical for prototyping and evaluation, it reintroduces a screen-mediated layer that partially dilutes the tangible, object-centred experience the physical artifact is intended to produce. Future iterations should move toward a fully embedded implementation in which all computation, audio input, and playback are housed entirely within the conch shell itself, with no external display. This was, in fact, the original design intent of the project, and remains the most coherent realization of the core premise: an artifact that communicates entirely through voice, touch, and physical presence rather than text on a screen.

Concretely, this could be achieved by migrating the system to an embedded single-board computer (such as a Raspberry Pi) running the full pipeline locally or via a minimal wireless API call, with a compact speaker and microphone integrated into the shell cavity. The result would be a self-contained object with no associated screen, no visible interface, and no external device dependency. Such a configuration would more fully support the perception of the conch as an autonomous, quasi-living entity, extending the effect observed in post-interview responses in which participants described the artifact as a relic or being rather than a software system.

Conclusion

This paper has presented Conch(ious), a computationally creative system integrating a persona-constrained LLM with text-to-speech synthesis and a 3D-printed physical artifact, functioning as an autonomous cultural curator from the fictional underwater colony of Amphitopia. Developed through strategic foresight methods, iterative prototyping, and anticipatory ethnographic evaluation, the system produced three significant findings.

First, constraining an LLM within a detailed fictional ontology redirects rather than diminishes its creative capacity. The Amphitopia persona required analogical reasoning, inversion of perspective, and narrative reframing in every response, producing outputs that participants found generatively rich enough to sustain deepening engagement across repeated interactions.

Second, the consistent emotional trajectory observed across all participants, from curiosity through anxiety to empathy and reflection, demonstrates that provocation is a measurable form of creative output, emerging dynamically from interaction. The critical turn in post-interviews, in which participants shifted from accepting the Conch's narrative to interrogating its ideological assumptions, is particularly notable: the system generated conditions in which users produced their own creative and critical conclusions in response to computationally generated content.

Third, physical embodiment directly shaped how users perceived the system's creative identity. Participants described the conch as a mystical entity whose presence shaped the depth of engagement in ways a screen-based interface would be unlikely to replicate.

Taken together, these findings demonstrate how the boundaries of computational creativity can be extended when generative AI is given not just a task but an identity, a world to inhabit, and a relationship to sustain. Conch(ious) offered a lens through which participants reflected on the fragility of cultural memory and the constructed nature of historical narrative. The system advances a concrete design principle: specificity of constraint, rather than breadth of freedom, can serve as the productive engine of computational creativity, and the medium through which a creative system presents itself materially is integral to its capacity to be perceived as a creative agent.

Acknowledgements

The authors would like to thank the six anonymous participants who generously dedicated their time, spaces, and personal reflections to the anticipatory ethnographic study. We also express our gratitude to New York University Abu Dhabi for providing the lab facilities, equipment, and technical support that made the development of *Conch(ious)* possible.

Author Contributions

Yaakulya Sabbani designed and developed the software architecture, prompt fine tuning, and built the web UI. Ahmad Hafizh led the design of narrative worldbuilding, media production, and integration of primary research into the frameworks. Amina Magomedova contributed to the hardware fabrication, prototyping, and handled the graphic design. The entire study was conducted under the supervision and guidance of Professor Nimrah Syed.

References

Al-Jayyousi, O. R. 2016. *Islam and Sustainable Development*. Gower.

Anadol, R. 2023. Artificial realities: Coral.

Boden, M. A. 2004. *The Creative Mind*. Routledge.

Brierley, G.; Tadaki, M.; Hikuroa, D.; Blue, B.; Šunde, C.; Tunnicliffe, J.; and Salmond, A. 2018. A geomorphic perspective on the rights of the river in aotearoa new zealand. *River Research and Applications* 35(10):1–16.

Candy, S., and Kornet, K. 2019. Turning foresight inside out: An introduction to ethnographic experiential futures. *Journal of Futures Studies* 23(3):3–22.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. 272–279. Proceedings of the Ninth International Conference on Computational Creativity, ICCC 2018.

Duarte, C. M. 2021. The soundscape of the anthropocene ocean. *Science* 371(6529).

Dunne, A., and Raby, F. 2013. *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT press.

e Cunha, M. P.; Clegg, S. R.; and Kamoche, K. 2012. Improvisation as "real time foresight". *Futures* 44(3):265–272. Special Issue: Weak Signals.

ElevenLabs. 2024. Text-to-speech API. Software. Retrieved from https://elevenlabs.io.

Glenn, J. C. 1972. Futurizing teaching vs. futures courses. *Social Science Record*.

Hines, A., and Bishop, P. 2006. *Thinking about the Future: Guidelines for Strategic Foresight*. Social Technologies.

Hollan, J.; Hutchins, E.; and Kirsh, D. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction* 7:174–196.

Holopainen, M., and Toivonen, M. 2012. Weak signals: Ansoff today. *Futures* 44(3):198–205. Special Issue: Weak Signals.

IPCC. 2023. Climate change 2023: Synthesis report. Technical report, Intergovernmental Panel on Climate Change.

Ishii, H., and Ullmer, B. 1997. Tangible bits: Towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 234–241. ACM.

Langton, J. 2019. When the uae was young. *The National News*.

Li, P.; Yang, J.; Islam, M. A.; and Ren, S. 2025. Making ai less 'thirsty'. *Commun. ACM* 68(7):54–61.

Lindley, J.; Sharma, D.; and Potts, R. 2014. Anticipatory ethnography: Design fiction as an input to design ethnography. volume 2014, 237–253.

Lozano-Hemmer, R. 2013. Voice tunnel.

Microsoft Research. 2018. Project Natick phase 2. natick.research.microsoft.com.

Montagu, J. 2018. *The Conch Horn: Shell Trumpets of the World from Prehistory to Today*. Hataf Segol Publications.

Nimmo, H. A. 1999. The bajau laut: Adaptation, history, and fate in a maritime fishing society of south-eastern sabah.:the bajau laut: Adaptation, history, and fate in a maritime fishing society of south-eastern sabah. *American Anthropologist* 101:461–462.

Novaglio, C.; Bax, N.; Boschetti, F.; Emad, G. R.; Frusher, S.; Fullbrook, L.; Hemer, M.; Jennings, S.; van Putten, I.; Robinson, L. M.; Spain, E.; Vince, J.; Voyer, M.; Wood, G.; and Fulton, E. A. 2021. Deep aspirations: towards a sustainable offshore blue economy. *Reviews in Fish Biology and Fisheries* 32.

Pal, J. S., and Eltahir, E. A. B. 2016. Future temperature in southwest Asia projected to exceed a threshold for human adaptability. *Nature Climate Change* 6(2):197–200.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role-play with large language models.

Subraelu, P. 2021. Sea-level rise and coastal vulnerability: A preliminary assessment of uae coast. *Journal of Coastal Zone Management* 24.

UNFCCC. 2015. The paris agreement.

United Nations. 1982. United nations convention on the law of the sea. United Nations Treaty Series, vol. 1833, p. 3.

Theory as Architecture: Why Aesthetic Frameworks Should Constitute, Not Merely Evaluate, AI Art Systems

Atticus Sims
Department of Arts and Design
University of Macau
Macau, China
atticussims@um.edu.mo

Abstract

Every major AI art system—AARON, The Painting Fool, AICAN, Botto, Keke Terminal—treats aesthetic theory predominantly as applied: evaluation criteria operating on outputs whose generation is largely independent of those criteria. We argue that the field has not yet made explicit a distinction that is implicit in its own frameworks. We distinguish *applied* aesthetic theory (post-hoc filtering, scoring, curation) from *constitutive* aesthetic theory (commitments embedded in architecture, shaping generation, evaluation, and developmental trajectory). The diagnostic is a motivated counterfactual test: changing the constitutive theory should change the system's evaluation ontology—the categories through which it perceives and judges—not merely its selection preferences. We trace this distinction through Colton's creative tripod, Wiggins's formal framework, Jennings's autonomy requirements, and Galanter's call for theoretical grounding, showing that each identifies pieces of the problem without articulating the architectural relationship explicitly. Drawing on Bense's creation/communication pattern and Moles's semantic/aesthetic information distinction, we offer independent diagnostic lenses that sharpen the analysis, revealing a gap between behavioral autonomy and evaluative autonomy that current systems leave unaddressed. We propose theoretical coherence as a new evaluative dimension and argue that constitutive pluralism—different theories yielding structurally different systems—is both achievable and desirable.

Introduction

The computational creativity community has built increasingly sophisticated AI art systems. Botto deploys multi-agent creative reasoning, autonomous code generation, and self-directed stylistic exploration (Botto Project 2025). Keke Terminal orchestrates VLM-based evaluation through a ReAct agent framework with autonomous creative decision-making (Dark Sando 2024). The Painting Fool generates its own fitness functions through automated theory formation (Colton 2012). AICAN implements adversarial objectives derived from psychological theories of art-historical dynamics (Elgammal et al. 2017). These systems represent genuine engineering achievements and growing behavioral autonomy.

Yet a persistent asymmetry runs through this progress. On the engineering side, the sophistication is remarkable: multi-agent architectures, meta-prompting, knowledge graphs, generative code synthesis. On the evaluative side, a different picture emerges. Botto's self-critique dimensions intermix compositional, narrative, affective, and sociological categories without any theoretical principle governing their integration (Botto Project 2025). Keke Terminal's taste module combines VLM evaluation with neural aesthetic prediction and social engagement metrics into a composite reward signal (Dark Sando 2024). Even The Painting Fool's emergent evaluative principles are discovered statistical associations rather than theoretically derived judgments (Colton 2008). The engineering is extraordinary. The aesthetics is ad hoc.

This paper's position is direct: the explanation lies in the relationship between aesthetic theory and system architecture. The field has treated aesthetic theory as something *applied to* systems—filtering, scoring, curating outputs—rather than as something that *constitutes* them. This is not merely a problem for the pre-foundation-model systems in our analysis; it is amplified by the foundation model era, where a system can fine-tune Stable Diffusion with aesthetic reward or prompt an LLM to reason about visual quality, yet in both cases the aesthetic criteria, though encoded in model weights or prompt configurations, steer the system's pre-existing evaluative capacities rather than reorganizing them—applied theory at unprecedented scale.

We distinguish *applied* from *constitutive* aesthetic theory and argue that making this distinction architecturally explicit reveals a pattern the field has not fully reckoned with. The distinction builds on Colton's creative tripod, Wiggins's formal framework, Jennings's autonomy requirements, and Galanter's call for aesthetic grounding. Our contribution is to articulate the architectural dimension these frameworks converge on but leave implicit, and to offer diagnostic tools for making it precise.

We also identify a gap between *behavioral autonomy* (deciding what and how to create) and *evaluative autonomy* (having internal, principled criteria for what counts as creating well). Current systems achieve the former without the latter—a distinction with immediate practical consequences for how the CC community designs, evaluates, and situates autonomous creative agents.

The argument proceeds as follows: §2 develops the distinction and its diagnostic test; §3 applies it to five systems; §4 sharpens it through information aesthetics; §5 demonstrates constitutive pluralism; §6 proposes theoretical coher-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

ence as a new evaluative dimension; §7 concludes.

Applied vs. Constitutive Theory

The Distinction

Consider two ways aesthetic theory can figure in the design of an AI art system.

In the *applied* role, a generative model produces outputs and an evaluative process selects among them. The evaluative criteria are external to the generative process: generation might produce work of extraordinary subtlety or banality by the same mechanism, with aesthetic assessment arriving afterward. The diagnostic feature of applied aesthetic theory is that changing the evaluative criteria changes which outputs are *selected* but does not change the *generative process itself*. The same generator, paired with different filters, yields different selections from an identical output distribution.

In the *constitutive* role, aesthetic commitments are embedded in the system's architecture so that they shape generation from within. The theory determines which features the system measures and how it organizes them, what the system optimizes for during generation, what the system treats as creative success, and how its creative trajectory unfolds over time.

This is not merely a distinction between "having" and "not having" an aesthetic theory. Systems with applied aesthetics may invoke sophisticated vocabulary, but invoke it as evaluation labels rather than as principles that shape architectural decisions. The question is whether the aesthetic theory does *architectural work*: whether removing it would change the system's structure, not merely its selection.

The Motivated Counterfactual Test

The diagnostic for constitutive embedding requires more precision than "would a different theory produce a different system?"—a test trivially satisfied by any design parameter. The relevant counterfactual is *ontological*, not merely parametric: changing the constitutive theory should change the system's *evaluation ontology*—the categories through which it perceives, the dimensions along which it judges, the developmental dynamics it exhibits. An applied substitution (replacing one scoring function with another) changes selection behavior while preserving the evaluation ontology. A constitutive substitution changes what the system *can perceive as relevant*.

Consider an analogy. Two telescopes with different color filters (applied difference) produce different images of the same sky. Two telescopes built on different physical principles—optical versus radio (constitutive difference)—perceive different phenomena entirely. The constitutive/applied distinction is the difference between interchangeable filters and fundamentally different instruments. What makes the difference constitutive rather than parametric is that it reorganizes the space of possible evaluative judgments rather than selecting among judgments the system is already capable of making.

A concrete test: if we can state the system's theory, predict from that theory alone what categories of evaluative judgment the system can and cannot make, and verify that those predictions hold, then the theory is constitutive. If two systems with different stated theories nonetheless make the same categories of judgment—differing only in weights, thresholds, or selection rates—then the theories are applied, regardless of how they are described in documentation.

Relationship to Existing Frameworks

The constitutive/applied distinction builds on insights already present in CC research. Colton's (2008) creative tripod demands that skill, appreciation, and imagination be simultaneously present. The FACE model (Colton, Charnley, and Pease 2011) formalizes the distinction between A^g-level operation (generating a particular aesthetic measure) and A^p-level operation (generating methods for generating aesthetic measures). Jennings (2010) argues that creative autonomy requires internal direction of creative decisions. Galanter (2012) has argued explicitly that generative artists should ground their systems in aesthetic theory, proposing complexity science as a framework. Wiggins's (2006) formal framework distinguishes exploratory from transformational creativity—searching within versus modifying the conceptual space. Ventura's (2016) argument that "mere generation" is an insufficient barometer reinforces the point from a different angle.

The evolutionary art literature makes the same point from the engineering side. Romero and Machado (2008) document how fitness function design shapes the entire character of evolutionary art output. Machado and Cardoso's (2002) NEvAr demonstrated that theoretically motivated complexity measures could serve as fitness functions—an important early example of constitutive aesthetic embedding. Todd and Latham (1992) and Sims (1994) similarly embedded aesthetic principles in fitness functions that shaped the generative process itself.

These contributions converge on the recognition that evaluation must be *internal* for creative autonomy and *principled* for aesthetic seriousness. What they leave unexamined is the precise *architectural* relationship between an aesthetic theory and the system it inhabits. We add two requirements:

Theoretical grounding: the system's evaluative criteria form a coherent framework explicable in terms of a theory of aesthetic value. A system could achieve A^p-level operation while those measures are statistically learned regularities. Theoretical grounding is what distinguishes aesthetic *judgment* (evaluating for articulable reasons) from aesthetic *prediction* (forecasting reception from learned correlations). This distinction does not presuppose a narrowly rationalist account of judgment: a system could exercise judgment through trained perceptual sensitivity, provided that sensitivity is grounded in a theory of what makes the relevant features aesthetically significant. The distinction also survives institutional challenges: even if audience reception is partly constitutive of aesthetic value (Danto 1964; Dickie 1974), a system that predicts reception without modeling *why* certain features elicit it remains in the applied role.

Architectural traceability: the theory does demonstrable architectural work, verifiable through the motivated counterfactual test. A different theory should produce a different

evaluation ontology—not merely different scores.

Three Thresholds

The spectrum from applied to constitutive admits three analytically distinct thresholds:

Constitutive embedding obtains when aesthetic commitments shape the generative architecture, passing the motivated counterfactual test (different theory → different evaluation ontology). Cohen's AARON satisfies this threshold.

Theoretical coherence obtains when the embedded commitments form a unified framework—when the system's evaluative dimensions derive from a common theory of aesthetic value. "Coherent" need not mean "derived from a single formal theory": Cohen's AARON exhibited implicit coherence—his artistic vision unified the rules—even though the rules accumulated pragmatically. The threshold is satisfied when a knowledgeable observer can identify a unifying account connecting the system's evaluative dimensions, whether formalized or not. No existing system fully satisfies this threshold in the strong sense, though AARON and AICAN approximate it.

Adaptive constitutiveness obtains when the system can modify its constitutive commitments in response to its own creative practice—when theory and practice develop together. This corresponds to Wiggins's transformational creativity applied to the evaluative framework itself. No existing system satisfies this threshold, though The Painting Fool's ability to generate new fitness functions represents a partial step.

These thresholds are analytically distinct: different systems achieve different subsets. We argue that the conjunction of all three is what creative autonomy in the fullest sense requires—while recognizing that each threshold independently represents a meaningful advance.

These thresholds are not arbitrary divisions but derive from the convergence of requirements identified across CC evaluation frameworks. Constitutive embedding synthesizes Jennings's (2010) internal direction requirement with Galanter's (2012) call for theoretical grounding. Theoretical coherence extends Colton's (2008) appreciation criterion from requiring that a system have evaluative reasons to requiring that those reasons form a unified framework. Adaptive constitutiveness applies Wiggins's (2006) transformational creativity to the evaluative framework itself. Each threshold is independently motivated; together they form a progression from principled design through theoretical unity to creative self-modification.

Other evaluative dimensions—perceptual grounding, social situatedness, communicative capacity—are important for assessing creative systems but concern the system's capabilities rather than the theory-architecture relationship that the constitutive/applied distinction targets. The three thresholds are not exhaustive of what matters in CC evaluation; they are exhaustive of the analytically distinct positions a system can occupy along the specific axis from applied to constitutive aesthetic theory.

Five Systems, One Pattern

We examine five AI art systems spanning rule-based, evolutionary, adversarial, and LLM-based approaches. For each, we ask: where does the selective principle originate? Does the system have aesthetic *reasons* for its choices? At which threshold does it fall? We take care to assess each system charitably, on the basis of what its publicly documented architecture reveals, while acknowledging that our analysis is necessarily limited by available documentation. Where systems have been described in peer-reviewed publications, we prioritize those sources.

AARON (Cohen, 1968–2016). Cohen's system is the strongest case of constitutive aesthetic theory in the field's history. Over nearly five decades, Cohen encoded approximately 300 hierarchical production rules implementing his compositional and figural knowledge, including spatial models, postural rules, and compositional evaluation functions (Cohen 1995). These rules structured generation itself—different rules would have produced a structurally different system perceiving different compositional features as relevant. AARON had aesthetic *reasons* for its choices, traceable to Cohen's artistic commitments.

AARON satisfies constitutive embedding and exhibits implicit coherence—Cohen's artistic vision provided a unifying perspective—but the rules accumulated pragmatically rather than being derived from an explicit theory, making the coherence retrospectively attributable rather than architecturally articulated. The system lacked adaptiveness: rules were frozen at design time. Cohen's (2006) revealing discovery that a simple color algorithm matched twenty years of accumulated rules illustrates the cost: the system could not discover such equivalences or respond creatively to them. AARON demonstrates that constitutive embedding without adaptiveness produces principled but rigid creative behavior.

The Painting Fool (Colton, 2001–present). Colton's Painting Fool achieves A^p-level operation: its HR system generates fitness functions used to evolve scenes, producing emergent evaluative criteria not externally programmed (Colton 2012). The 2015 integration of DARCI's machine vision capabilities (Colton et al. 2015) further strengthened constitutive embedding: the system's ability to analyze its own output during painting means perceptual feedback shapes generation, not just selection. Since then, VLM-based iterative assessment during generation has become standard practice—Botto's creative reasoning loop, Keke Terminal's play iterations, and the broader generate-critique-refine paradigm (Madaan et al. 2023) all use visual feedback to shape the generative process. This proliferation makes the constitutive question more urgent, not less: in every case, the evaluative categories are inherited from the VLM's training distribution rather than derived from a theory of aesthetic value.

Yet The Painting Fool's emergent evaluative discoveries—associations between media and mood qualities—are statistical associations rather than principles derived from a theory of aesthetic value. The system can discover *that* pencils correlate with bleakness but cannot articulate *why* this matters aesthetically. In Moles's (1966) terms, the system operates primarily with semantic information (extractable correlations) rather than aesthetic information (irreducible perceptual qualities). Colton (2008) acknowledged that skill, appreciation, and imagination were not simultaneously present. The Painting Fool demonstrates that A^p-level operation without theoretical grounding produces learned regularities rather than principled aesthetic judgment. The system's ability to generate new evaluative criteria represents a genuine step toward adaptiveness, even if those criteria lack unifying theoretical grounding.

AICAN (Elgammal et al., 2017). Elgammal's Creative Adversarial Network has an explicit theoretical grounding: its dual objective—minimize deviation from the art distribution while maximizing stylistic ambiguity—derives from Martindale's (1990) psychological theory of art-historical dynamics. The adversarial architecture directly instantiates the theory: a different theory (say, Berlyne's optimal arousal model) would yield a different optimization objective, evaluation ontology, and architecture. The motivated counterfactual test is clearly satisfied.

AICAN's limitation is scope rather than commitment. The theory operates at the population level—style distributions across art history—without addressing individual compositions. AICAN cannot evaluate whether a particular work has interesting compositional structure, rewards sustained perceptual engagement, or challenges the viewer's compressive capacities. It has a theory of art-historical *dynamics* but not a theory of aesthetic *experience*—no account of what makes a single work worth looking at, as opposed to whether it deviates from a learned style distribution. This is the difference between a theory of art's evolution and a theory of art's value. Notably, Martindale's theory operates within a broader tradition—Berlyne's (1971) optimal arousal theory provides the psychological foundations from which Martindale derived his art-historical dynamics, suggesting that a more comprehensive constitutive approach could integrate both individual-level arousal theory and population-level style dynamics.

AICAN substantially satisfies both constitutive embedding and theoretical coherence, but only within a domain that cannot address the compositional, perceptual, and experiential dimensions where most aesthetic judgment operates. This matters for the argument: AICAN demonstrates that constitutive embedding with theoretical coherence is *achievable*—the conjunction is not utopian. What AICAN's narrow scope shows is that the challenge lies not in constitutive embedding per se but in the *comprehensiveness* of the constitutive theory.

Botto (Klingemann / ElevenYellow DAO, 2021–present). Botto resists simple categorization. Its three generative subsystems demonstrate genuine engineering and creative sophistication (Botto Project 2025). The creative reasoning system generates hypotheses, self-critiques along multiple dimensions, and iteratively refines—exhibiting a form of internal evaluative capacity that goes well beyond simple prompt-to-image generation.

Two features, however, position Botto's overall trajectory toward the applied end of the spectrum. First, the self-critique dimensions—composition, lighting, narrative, audience appeal, meme potential, AI slop detection—intermix formal, psychological, sociological, and quality-control categories. Whether these constitute a coherent framework depends on what one counts as coherence: they may reflect a pragmatic-pluralist approach to evaluation rather than theoretical arbitrariness. We note this ambiguity rather than prejudging it, while observing that the system's documentation does not articulate a theoretical principle governing their selection or interrelation. Second, the system's ultimate arbiter is external: DAO community voting trains a taste model that pre-selects candidates. The internal creative engine operates within constraints set by this feedback loop, creating a structural tension between internal exploration and external selection. A theoretically grounded internal evaluative framework could serve as a counterweight—enabling the system to resist convergence when its own aesthetic commitments indicate a different direction.

Botto's architecture is best understood as two-level: the creative reasoning system has partially constitutive features, while the community-driven selection loop is applied. The overall trajectory is predominantly governed by the applied component.

Keke Terminal (Dark Sando, 2024). Keke Terminal presents an instructive case. Its taste module combines VLM-based evaluation, NIMA scoring, and social engagement metrics. The system employs meta-prompting, creative reasoning, and autonomous decision-making—capabilities that, in other domains, would constitute genuine agency.

The philosophical interest lies in the gap between technical sophistication and the architecture of evaluation. The taste module learns to *predict* which outputs will be well-received but—based on available documentation—does not *judge* which are aesthetically significant according to an articulable theory. We should be cautious about drawing causal conclusions: correlation between theoretical thinness and aesthetic conventionality does not establish causation.

What we can say with confidence is structural. The system's evaluative architecture would remain structurally unchanged if the engagement-prediction module were replaced by a different prediction model targeting different demographics—the evaluation ontology (predicting reception from learned correlations) is invariant across such substitutions, which satisfies the test for applied theory. A constitutive substitution—replacing reception prediction with, say, Bense's information-aesthetic framework—would

change not only the scores but the categories of judgment available to the system: from "will audiences approve?" to "does this composition occupy the configurative zone between order and entropy?"

Synthesis. The pattern across five systems is consistent in direction, though the details reveal important variation:

Three observations. First, no system achieves the conjunction of strong constitutive embedding, explicit theoretical coherence, and adaptiveness—but AICAN demonstrates that the first two are jointly achievable. The challenge is comprehensiveness, not feasibility. Second, the most recent and technically sophisticated systems (Botto, Keke Terminal) are not the most theoretically grounded—engineering *for generation* has progressed far more rapidly than *for evaluation*. Third, existing evaluation frameworks (Ritchie 2007; Jordanous 2012) do not directly diagnose this pattern because they do not ask whether a system's *design* follows from its *aesthetic theory*.

We note that these five systems have different overriding aims—AARON embodied one artist's vision; The Painting Fool aims to be taken seriously as a creative artist; AICAN tests a psychological theory; Botto explores decentralized creative agency; Keke Terminal pursues autonomous market participation. The constitutive/applied distinction is diagnostic across all these aims, though its normative force varies (see §6).

This pattern is not specific to the pre-foundation-model systems analyzed here. Foundation models amplify the asymmetry rather than resolving it: fine-tuning a diffusion model with an aesthetic reward function, or prompting an LLM to critique images along specified dimensions, instantiates applied aesthetics at unprecedented scale. The constitutive/applied distinction cuts orthogonally to the model capability axis.

This claim requires care when applied to LLM-based systems. An LLM possesses its own implicit evaluation ontology inherited from training—just as a human artist possesses perceptual capacities shaped by biology and experience. The question is not whether the system uses an LLM—most future creative systems will—but whether the aesthetic theory organizes the evaluation pipeline through which the LLM reasons, or merely steers the LLM's general-purpose reasoning via prompts. A system that prompts an LLM to "evaluate this image for compositional balance" invokes the LLM's pre-existing evaluative vocabulary—applied theory as prompt steering. A system that feeds the LLM a theory-derived perceptual decomposition of the image (entropy measurements at multiple scales, attention distribution maps, complexity gradients) and constrains it to reason *within* those categories is doing something architecturally different: the LLM serves as a reasoning engine, but the categories of reasoning are determined by the theory, not by the LLM's training distribution. This is the difference between theory as prompt and theory as pipeline—the version of the constitutive/applied distinction that matters for the foundation model era.

Independent Diagnostics from Information Aesthetics

The constitutive/applied distinction can be independently sharpened through two distinctions from the information aesthetics tradition—a body of work the CC literature has not adequately engaged despite its direct relevance.

Bense's Creation Pattern vs. Communication Pattern

Max Bense (1965; 1971) distinguishes the *communication pattern*—sender $\rightarrow$ noisy channel $\rightarrow$ receiver, where feedback is corrective, converging on shared repertory alignment—from the *creation pattern*, where an observer performs a selective function on a repertory to produce a new distribution (Bense 1965; Nake 2012; Taylor 2024). In the creation pattern, selective freedom is progressively consumed as the emerging work constrains remaining choices. The product feeds back to redefine the repertory: creation changes what is subsequently possible. Communication feedback is *convergent*; creation feedback is *generative*.

What does this distinction add beyond restating the constitutive/applied divide? The Bensean diagnostic identifies a specific *mechanism* that the constitutive/applied distinction leaves abstract. The constitutive/applied distinction tells us *that* theory should be embedded; the creation/communication distinction tells us *how* the embedded theory must function—specifically, it must operate through generative feedback that transforms its own evaluative repertory. This is a stronger architectural requirement than constitutive embedding alone, because a system could constitutively embed a fixed theory (as AARON does) without exhibiting the repertory-transforming dynamics that Bense identifies with creation. The creation pattern thus provides a criterion for distinguishing between constitutive systems that are merely *principled* and constitutive systems that are genuinely *creative* in Bense's technical sense.

Applied to the systems analyzed above: Botto's DAO feedback loop implements the communication pattern—converging on community-repertory alignment. Keke Terminal's engagement-trained taste module likewise converges on existing preferences. AARON's frozen rules implement neither pattern. No system implements the creation pattern, which requires that the selective principle be internal, theoretically grounded, and capable of redefining the repertory through creative practice.

Is the creation pattern computationally tractable? It is—and less demanding than it might appear. A system whose evaluative framework is parameterized by its own creative history—where successful works modify the weights, thresholds, or even the categories of subsequent evaluation—would implement the creation pattern. This is architecturally distinct from, but not more technically demanding than, the meta-learning and self-modifying prompt systems already deployed in Botto's creative reasoning framework. The barrier is not engineering but the absence of a theoretical account specifying *how* the evaluative framework should change—which is precisely what a constitutive aesthetic theory provides. A practical consequence: con-

Table 1: Constitutive analysis of five AI art systems across three thresholds.

System	Constitutive Embedding	Theoretical Coherence	Adaptiveness
AARON	Strong	Implicit	None
Painting Fool	Partial (generation)	None (explicit)	Partial (A^p)
AICAN	Strong (narrow scope)	Strong (narrow scope)	None
Botto	Two-level (internal partial, selection applied)	Unclear (pragmatic pluralism)	External only
Keke Terminal	Weak	None	External only

stitutive embedding enables evaluation at each stage of the generative process rather than only at its conclusion, reducing the candidate space through theoretically informed pruning rather than post-hoc filtering.

This diagnostic also exposes what may be the most practically important distinction in this paper: the gap between *behavioral autonomy* and *evaluative autonomy*. Behavioral autonomy—deciding when, what, and how to create—is compatible with *evaluative dependence*: relying on external sources for what counts as creating well. A system with behavioral autonomy but evaluative dependence is behaviorally autonomous but evaluatively heteronomous. Both Botto and Keke Terminal achieve impressive behavioral autonomy while remaining evaluatively dependent on external feedback—community voting or engagement metrics. Guckelsberger, Salge, and Colton's (2017) model of intentional creative agency addresses intentionality but not the source of evaluative standards.

This distinction has immediate practical consequences. When the CC community evaluates a system's "autonomy," it should ask not only whether the system makes its own creative decisions but whether the *criteria informing those decisions* are internally grounded. A system that autonomously generates, selects, and publishes artwork is behaviorally autonomous. If its selection criteria are entirely trained on external approval data, it is evaluatively heteronomous—and the constitutive/applied distinction explains exactly why this matters: evaluative heteronomy constrains creative trajectory even when behavioral freedom is unconstrained.

Moles's Semantic vs. Aesthetic Information

Abraham Moles (1966) distinguishes *semantic information*—propositional, translatable, channel-independent—from *aesthetic information*—state-oriented, channel-dependent, irreducible to logical propositions.

This distinction does additional work beyond the constitutive/applied divide by identifying a specific *informational limitation* of applied systems. Applied aesthetics necessarily operates with semantic information: extractable labels, quantifiable scores, translatable evaluations. A NIMA score, a composition rating, an engagement prediction—these are propositional and channel-independent. They can be transmitted without loss from one system to another because they are *about* the work rather than *of* it.

A constitutive system could, in principle, attend to aesthetic information: the channel-dependent perceptual qualities that resist decomposition into propositional content. A system measuring spatial entropy distributions at multiple scales, tracking information-theoretic tension between local complexity and global order, and evaluating progressive compressibility across spatial frequency bands engages with properties that are compositionally sensitive in a way score-based evaluation cannot replicate. The distinction suggests a specific reason why systems operating exclusively with semantic evaluation have a ceiling on their evaluative sophistication: they cannot represent the properties that, on Moles's account, constitute the specifically aesthetic dimension.

This is not to claim that current computational methods can fully capture aesthetic information in Moles's strong sense—the channel-dependent, irreducible quality of perceptual experience may remain beyond computational reach. The point is more modest and more actionable: there is a gradient between purely semantic evaluation (a single quality score) and increasingly channel-sensitive evaluation (multi-scale spatial analysis, attention modeling, perceptual decomposition). Systems operating further along this gradient engage with properties closer to what Moles identifies as aesthetic information, even if they do not capture it completely. Constitutive embedding pushes systems along this gradient by requiring that evaluation be grounded in a theory of *how* perceptual properties generate aesthetic value, not merely *which* outputs correlate with positive reception.

Convergence of the Diagnostics

These three distinctions are not three ways of saying the same thing. Each identifies a different facet of the same structural gap: the constitutive/applied distinction identifies *where* theory sits in the architecture; the creation/communication distinction identifies *how* theory must function dynamically; the semantic/aesthetic distinction identifies *what kind of information* theory must engage with. Their convergence—diagnosing the same systems as deficient along different analytical dimensions—strengthens the diagnosis: the gap is not an artifact of a particular vocabulary but a structural feature of how these systems are built.

What Constitutive Theory Produces

The preceding sections diagnose a gap. What would filling it look like? Rather than sketching one preferred architecture, we illustrate how *different* constitutive theories produce structurally different systems. This is the strongest form of the argument: constitutive theory is not a single prescription but a design principle that produces principled diversity.

Different theories produce different evaluation ontologies. Consider three theoretical starting points, each drawn from a distinct intellectual tradition: (a) Birkhoff's (1933) aesthetic measure $M = O/C$, grounding aesthetic value in the ratio of order to complexity; (b) Bense's information aesthetics, organized at nuclear, micro, and macro semiotic levels, attending to how information is distributed across scales of composition; (c) Chatterjee's (2013) aesthetic triad, grounding aesthetic experience in the interaction of sensory-motor, emotion-valuation, and meaning-knowledge processing systems.

Theory (a) yields a maximally parsimonious architecture: a single evaluator computing a single ratio. The system perceives order and complexity as its two evaluatively relevant categories, and every compositional feature matters only insofar as it contributes to one or the other. Theory (b) implies a multi-level computational pipeline measuring different kinds of information at different scales—the nuclear level (material elements), the micro level (statistical regularities), and the macro level (global gestalt). The evaluation ontology is hierarchical: features matter differently depending on which semiotic level they operate at. Theory (c) organizes evaluation by processing system rather than by compositional level, producing a three-channel architecture where *relationships between channels*—the way sensory processing, emotional response, and semantic interpretation interact—carry the evaluatively relevant information. A composition might score well on sensory-motor engagement but poorly on meaning-knowledge integration, and the pattern of cross-channel interaction is itself the aesthetic datum.

These are not parametric variations within a shared evaluation framework—they are *ontologically* different. Each theory determines what the system can perceive as evaluatively relevant: order-complexity ratios, semiotic-level information distributions, or processing-system interactions. The motivated counterfactual test is satisfied because the evaluation ontology changes, not merely the weights within a fixed ontology.

Different theories produce different integration strategies. This ontological divergence extends to how evaluative information is combined. A Birkhoffian system has no integration problem—there is one metric, and higher is better. A Bensean system must integrate across semiotic levels, raising the question of whether nuclear, micro, and macro evaluations should converge or whether productive tensions between levels carry aesthetic information. A Chatterjee-inspired system must decide how to weigh cross-channel interactions: does aesthetic value reside in harmony among processing systems, in tension between them, or in specific patterns of alignment and misalignment?

Each theory implies a different answer. For Bense, the configurative border condition—the productive tension between order and entropy—suggests that the *pattern* of integration (where levels agree, where they diverge) is itself evaluatively significant, not merely the aggregate. For Chatterjee, empirical work on the aesthetic triad suggests that the most powerful aesthetic experiences involve simultaneous engagement of all three systems, with cross-system facilitation predicting aesthetic response better than any single channel (Chatterjee 2013). Each integration strategy follows from its theory, and no amount of applied parameter tuning could produce these divergences because they operate at the level of what the integration module treats as signal.

Different theories produce different developmental trajectories. Schmidhuber's (2009) compression aesthetics distinguishes beauty (low description length) from interestingness (rate of compression progress). A system constitutively pursuing interestingness moves differently through aesthetic space than one pursuing beauty: it seeks works at the boundary between what it can already compress and what it cannot yet. A system grounded in Martindale's arousal potential pursues stylistic deviation from population norms—a trajectory operating at a different level of analysis (individual compression dynamics versus population-level style distributions). A system embedding Berlyne's (1971) optimal arousal theory—from which Martindale derived his art-historical model—would seek a specific middle region of stimulus complexity and progressively refine its model of where that region lies, rather than pursuing boundary-seeking or norm-deviation.

These three trajectories—boundary-seeking (Schmidhuber), norm-deviating (Martindale), and optimum-refining (Berlyne)—produce detectably different sequences of outputs over time: qualitatively different paths through aesthetic space, driven by different accounts of what makes creative development productive. This is what constitutive embedding produces: not just different evaluations of the same outputs, but different creative *histories*.

The counterfactual test confirms: substitute any of these theories for any other, and the evaluation architecture, integration strategy, and developmental trajectory all change. Applied filtering cannot replicate these differences because it operates on outcomes rather than on reasons. A constitutive system has access to a theoretical model of *why* certain features matter and *how* they relate; an applied system has access only to scores. This informational asymmetry—reasons versus outcomes—is the mechanism through which constitutive embedding produces its distinctive consequences, including the capacity for principled resistance to conventional taste when theoretical commitments warrant divergence.

Implications: Theoretical Coherence as Evaluative Dimension

The argument suggests a new evaluative dimension for CC research: *theoretical coherence*—the degree to which a system's architectural decisions are traceable to its aesthetic commitments.

Theoretical coherence can be operationalized through a three-step procedure. *Step 1—Commitment identification*: for each evaluative component, identify the aesthetic commitment (if any) that motivates it. Where motivations are

pragmatic or implicit, note this. *Step 2—Motivated counterfactual test*: for each commitment, ask whether a different aesthetic theory would change the system's evaluation ontology—the categories of judgment available to the system—not merely its scores or selection rates. *Step 3—Coherence assessment*: do the system's commitments form a unified framework, whether formal or implicit? Could a knowledgeable reader, given the theory alone, predict the major architectural decisions?

This procedure is implicit in how the CC community already evaluates systems—we ask whether design decisions are principled or ad hoc—but the constitutive/applied distinction makes it explicit and systematic. Adding theoretical coherence alongside Ritchie's (2007) output criteria, Jordanous's (2012) standardized procedure, and Colton's creative tripod fills a gap. Where Ritchie evaluates outputs, Jordanous evaluates behaviors, and Colton evaluates capacities, theoretical coherence asks whether the system's *design* follows from a principled account of aesthetic value. This dimension complements rather than replaces existing approaches: Linkola et al.'s (2017) metacreative self-awareness and Colton and Pease's (2018) computational authenticity both gesture toward the theory-architecture relationship without making it their explicit diagnostic target.

An important caveat: constitutive embedding amplifies theoretical commitments for good *and* ill. A system constitutively embedding a simplistic theory—a naïve Birkhoffian scalar measure, say—would produce systematically constrained art with great internal coherence. Such a system might be *worse* than an applied system precisely because its limitations are architectural rather than filterable. Theoretical coherence is therefore *necessary* for creative autonomy, not sufficient for aesthetic success. This means the *choice* of constitutive theory is consequential in a way the choice of applied filter is not—amplifying both the theory's insights and its blind spots.

A related concern is that constitutive embedding might produce only "good specimens"—typical members of the theory's category rather than norm-breaking creative work. This concern applies when the same theory drives both generation and evaluation. But a constitutive system need not generate only what its theory predicts will be good. An architecture that pairs divergent generation—through stochastic, lateral, or constraint-based ideation—with constitutive evaluation produces novel configurations that the theory can assess but would never have generated on its own. The theory shapes what the system *recognizes* as aesthetically significant, not what it is *capable of producing*. This mirrors human artistic practice, where experimental generation and reflective evaluation operate as complementary phases. In Boden's (2004) terms, constitutive embedding enables exploratory creativity within a theory-defined space; the adaptive constitutiveness threshold enables transformational creativity applied to the evaluative framework itself.

Constitutive embedding is also not a return to symbolic AI. The theory organizes the evaluation pipeline, but the pipeline's components—feature extraction, perceptual decomposition, attention modeling—can be implemented through neural, symbolic, or hybrid methods. What matters is that the architectural organization follows from the theory, not that the implementation uses any particular computational paradigm.

The consequence is theoretical pluralism. Different aesthetic theories should produce different systems with different creative behaviors, and this diversity is a feature. The CC field would benefit from the same diversity found in human art traditions, but grounded in explicit, architecturally traceable theoretical commitments rather than implicit assumptions.

A further consequence is that the field needs to take the *tractability* of aesthetic theories seriously as a research question. Quantitative foundations help: Bense's information aesthetics, Birkhoff's aesthetic measure, and Schmidhuber's compression theory all specify measurable quantities, making architectural translation direct. Theories organized around discrete processing stages (Chatterjee's triad) also translate well. By contrast, theories relying on irreducibly holistic judgment—late Wittgenstein's "seeing as," Heidegger's *techne*—resist computational decomposition. Mapping this tractability landscape is a contribution the CC community is positioned to make.

A caveat on scope: the constitutive/applied distinction is most forcefully normative when the aim is to produce a system that functions as an autonomous artist. Where the aim differs—advancing CC theory (Colton 2008), exploring the machine condition (Colton et al. 2020), or developing creative personhoods (Pease, Colton, and Banar 2023)—the distinction remains diagnostically useful but may carry different normative weight. We note, however, that the machine condition is not an alternative to aesthetic theory but a *candidate* constitutive theory—one specifying which experiential properties of computation are aesthetically significant. A system constitutively embedding this theory would have a different evaluation ontology (attending to memory volatility, parallel processing, power dependency as aesthetic features) than one embedding Bense or Birkhoff. The constitutive/applied distinction applies wherever a system makes evaluative choices about its own creative output, regardless of its overriding purpose.

Conclusion

Every major AI art system examined here treats aesthetic theory predominantly as applied: external to the generative architecture, operating on outputs rather than structuring their production. This finding has been anticipated by work on creative autonomy, appreciation, aesthetic evaluation, and computational aesthetics in evolutionary art (Jennings 2010; Colton 2008; Galanter 2012; McCormack, Dorin, and Innocent 2005; Todd and Latham 1992; Machado and Cardoso 2002; Romero and Machado 2008). This paper contributes the precision of the motivated counterfactual test targeting evaluation ontology, independent sharpening from Bense and Moles, the identification of a gap between behavioral and evaluative autonomy, and the demonstration that constitutive embedding produces ontologically different systems rather than merely parametrically different ones. We further argue that constitutive evaluation and divergent generation function as complementary phases, and that the ma-

chine condition constitutes a candidate constitutive theory rather than an alternative to the framework.

Two claims should be distinguished. *Claim A*—that AI art systems should have coherent aesthetic frameworks—is relatively modest and hardly controversial. *Claim B*—that those frameworks should be constitutive of the system's architecture, in the specific sense that different frameworks yield systems with different evaluation ontologies—is this paper's distinctive contribution. Claim B matters because it ensures the theory does genuine architectural work. The motivated counterfactual test provides a concrete way to distinguish systems that *have* a theoretical framework from systems whose architecture *follows from* one. The difference is consequential: a system whose architecture follows from its theory can extrapolate beyond training distributions, resist conventional taste on principled grounds, explain its evaluative decisions in theoretically articulable terms, and develop in theoretically motivated directions. A system that merely has a theory applied as a filter can do none of these things, regardless of how sophisticated the filter.

The CC community is well positioned to take this distinction seriously (Colton and Wiggins 2012). The frameworks are in place—Colton's appreciation criterion, Wiggins's transformational creativity, Boden's (2004) typology, Jordanous's evaluation procedures, Galanter's call for aesthetic grounding. What remains is to make the architectural relationship explicit, test it with the motivated counterfactual diagnostic, and build systems where aesthetic theory is traceable in the design rather than cited in the documentation. The generative AI revolution has given the field powerful new tools; the question is whether they will be deployed with the theoretical sophistication that the field's own traditions demand.

Author Contributions

As the sole author, Atticus Sims is responsible for all aspects of this work, including its conception, analysis, and writing.

Acknowledgments

This research was supported by a research grant from the University of Macau (SRG2025-00057-FAH). The author thanks the anonymous ICCC reviewers for their helpful comments.

References

Bense, M. 1965. *Aesthetica: Einführung in die neue Aesthetik*. Agis-Verlag.

Bense, M. 1971. The projects of generative aesthetics. In Reichardt, J., ed., *Cybernetics, Art and Ideas*. New York Graphic Society. 57–60.

Berlyne, D. E. 1971. *Aesthetics and Psychobiology*. Appleton-Century-Crofts.

Birkhoff, G. D. 1933. *Aesthetic Measure*. Harvard University Press.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.

Botto Project. 2025. Botto's art engine. Technical documentation. `https://docs.botto.com`.

Chatterjee, A. 2013. *The Aesthetic Brain: How We Evolved to Desire Beauty and Enjoy Art*. Oxford University Press.

Cohen, H. 1995. The further exploits of AARON, painter. *Stanford Humanities Review* 4(2):141–158.

Cohen, H. 2006. AARON, colorist: From expert system to expert. In *Proceedings of Digital Art Weeks*. ETH Zurich.

Colton, S., and Pease, A. 2018. Issues of authenticity in autonomously creative systems. In *Proceedings of the 9th International Conference on Computational Creativity*, 272–279.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proceedings of ECAI 2012*, 21–26.

Colton, S.; Halskov, J.; Ventura, D.; Gouldstone, I.; Cook, M.; and Ferrer, B. P. 2015. The Painting Fool Sees! New projects with the automated painter. In *Proceedings of the 6th International Conference on Computational Creativity*, 189–196.

Colton, S.; Pease, A.; Guckelsberger, C.; McCormack, J.; and Llano, T. 2020. On the machine condition and its creative expression. In *Proceedings of the 11th International Conference on Computational Creativity*, 342–349.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems*.

Colton, S. 2012. The Painting Fool: Stories from building an automated painter. In McCormack, J., and d'Inverno, M., eds., *Computers and Creativity*. Springer. 3–38.

Danto, A. C. 1964. The artworld. *The Journal of Philosophy* 61(19):571–584.

Dark Sando. 2024. Keke terminal. Technical whitepaper. `https://www.keketerminal.com`.

Dickie, G. 1974. *Art and the Aesthetic: An Institutional Analysis*. Cornell University Press.

Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *Proceedings of the 8th International Conference on Computational Creativity*.

Galanter, P. 2012. Computational aesthetic evaluation: Past and future. In McCormack, J., and d'Inverno, M., eds., *Computers and Creativity*. Springer. 255–293.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the "why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the 8th International Conference on Computational Creativity*.

Jennings, K. E. 2010. Developing creativity: Artificial bar-

riers in artificial intelligence. *Minds and Machines* 20:489–501.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Linkola, S.; Kantosalo, A.; Männistö, T.; and Toivonen, H. 2017. Aspects of self-awareness: An anatomy of metacreative systems. In *Proceedings of the 8th International Conference on Computational Creativity*.

Machado, P., and Cardoso, A. 2002. All the truth about NEvAr. *Applied Intelligence* 16(2):101–118.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36*.

Martindale, C. 1990. *The Clockwork Muse: The Predictability of Artistic Change*. Basic Books.

McCormack, J.; Dorin, A.; and Innocent, T. 2005. Generative design: A paradigm for design research. In *Proceedings of Futureground*. Monash University.

Moles, A. A. 1966. *Information Theory and Esthetic Perception*. University of Illinois Press. Translated by J. E. Cohen.

Nake, F. 2012. Information aesthetics: An heroic experiment. *Journal of Mathematics and the Arts* 6(2–3):65–75.

Pease, A.; Colton, S.; and Banar, B. 2023. On the notion of creative personhood. In *Proceedings of the 14th International Conference on Computational Creativity*, 117–121.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Romero, J., and Machado, P., eds. 2008. *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*. Springer.

Schmidhuber, J. 2009. Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE* 48(1):21–32.

Sims, K. 1994. Evolving virtual creatures. In *Proceedings of SIGGRAPH 1994*, 15–22. ACM.

Taylor, G. D. 2024. Bald krumme Linien: Generative aesthetics and the Bensian legacy. *Culture, Theory and Critique* 65(3–4):311–328.

Todd, S., and Latham, W. 1992. *Evolutionary Art and Computers*. Academic Press.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity*, 17–24.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

What makes LLM-generated Answers Valuable for Research Problem Reframing?

Abhinav Sood
University of Sydney
Sydney, Australia
abhinav.sood@sydney.edu.au
(corresponding author)

Stephen Wan
Data61, CSIRO
Sydney, Australia
stephen.wan@data61.csiro.au

Cecile Paris
Data61, CSIRO
Sydney, Australia
cecile.paris@data61.csiro.au

Kazjon Grace
University of Sydney
Sydney, Australia
kazjon.grace@sydney.edu.au

Abstract

Problem reframing can be understood as a form of p-transformational creativity; as a process, it uncovers frames that make an individual rethink their conceptual space, making new solutions possible. While problem reframing is central to design and other forms of creative practice, such reconceptualisation would also be valuable in research disciplines when conflicting evidence or unexpected findings arise. However, LLMs have already been shown to fail at supporting problem reframing in design contexts. This failure stems from two issues: problem frames require deep contextual understanding, and LLM-generated frames lack the novelty needed to serve as foundations. This paper takes a first step toward understanding how LLMs can contribute to problem reframing in research through a two-stage user study with 7 research students across diverse research disciplines. It concludes by identifying two distinct answer profiles and a set of criteria for evaluating whether AI-generated answers have the potential to support problem reframing in research.

Introduction

Research is generally understood as a method to generate, refine, and extend knowledge about a phenomenon of interest. Accounts of the research process describe it as structured and systematic (Creswell and Creswell 2017; Leedy and Ormrod 2001). For example, Leedy and Ormrod (2001) propose that research follows a cyclical seven-step process. In practice, the execution of these steps is not as linear; researchers tend to seek information from their surrounding context in a non-linear fashion (Foster 2004).

Within such a complicated process, multiple questions arise. In design, the evolution of such questions and design goals through problem reframing is deliberately encouraged. Problem reframing involves redefining a problem by altering the assumptions or perspectives that underlie it to arrive at solutions to the problem that might have otherwise been unreachable (Dorst 2015b). Problem reframing is considered core to the practice of design research, so much so that, as a skill, it has allowed designers to create value across disciplines and support transdisciplinary collaboration (Dorst 2015a; Mejía et al. 2023). Such a process would clearly be relevant to research, where unexpected findings, conflicting evidence, or new factors are discovered, initial assumptions and research questions might need to be reconsidered.

From a creativity perspective, problem reframing can be understood as a form of *p-transformational* creativity (Boden 2004). Transformational creativity involves modifying the generative rules that govern a conceptual space, expanding the accessibility of possible solutions for a problem. Problem reframing for an individual precisely does this. Following Boden's distinction between p-creativity (novel to the individual) and h-creativity (novel through history), we use p-transformational to denote transformations when they are novel to the researcher. Similar links between reformulation and creativity have been established across design frameworks. In Gero and Kannengiesser's (2014) Functional-Behaviour-Structure ontology, the most radical reformulation modifies the function space itself. Similarly Grace and Maher (2015) model surprise and reformulation as coupled metacognitive processes. These accounts converge on a shared insight: the creative act in research and design often lies not in finding a better answer, but in finding a better question.

Our work investigates whether Large Language Models (LLMs) can be prompted to produce outputs that researchers find valuable for this kind of reconceptualisation. To directly measure problem reframing with LLMs in a novel research context would require longitudinal observation of researchers over time; instead, we take a first step in examining what makes LLM-generated answers to open questions interesting to researchers who posed them. Through a two-stage user study with 7 research students across diverse disciplines, we first capture each participant's research context through extensive note-taking on previously read papers, then have participants evaluate 27 LLM-generated answers produced by three prompting strategies of varying structure.

By analysing which strategies produce answers that researchers want to pursue further, and by examining how participants rate their top-ranked answers across seven criteria, we identify the characteristics that distinguish answers with potential to support problem reframing from those that do not.

Herein, the main contributions of this paper are as follows:

1. A comparative analysis of three different prompting

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

strategies - direct question-answering (*direct_strategy*), context-aware question-answering (*all_content*), and a structured approach based on Dorst's Model of problem reframing (*dorst_frame*).

2. An understanding of what makes LLM-generated answers valuable and interesting to researchers.
3. Evidence for two distinct answer profiles, reinforcing and reorienting, and a corresponding set of criteria for evaluating whether LLM-generated answers have reframing potential.
4. An open-source software implementation for an Obsidian Plugin[1] through which we conduct our user study. The plugin allows for the generation and recording of answers across different users, and can be adapted for future investigations of LLM-assisted creative research processes.

Background

This section introduces literature on three aspects of our research. The first subsection introduces Dorst's frame creation methodology, which motivates our *dorst_frame* strategy. The second section discusses how notes serve to capture the context that reframing requires. The third section reviews literature on the capabilities and limitations of LLMs in creative and framing tasks.

Dorst's Frame Creation Methodology

The introduction established that problem reframing is a form of p-transformational creativity. Here, we focus on how reframing has been operationalised in design practice, which directly motivates our prompting strategy.

Schön (1983) established in his work on reflective practices that problem setting is fundamentally important to what information is attended to. Dorst and Cross (2001), for example, provide evidence that design involves co-evolution of the problem and solution spaces with feedback enabling this co-evolution. Building on this, Dorst (2011) identifies frame creation as the core practice of design thinking. In his framework, Dorst (2015b) describes frame creation as a nine-step process that moves beyond generating solutions to creating new approaches to problem situations. This nine-step process is described in (Dorst 2015b, Chapter 4) consisting of the following steps:

1. **Archaeology** - Understand deeply previous attempts to solve the problem and what led to how the problem situation is defined.
2. **Paradox**: Understand what makes the problem hard to solve? Get to the core of the deadlock that prevents the solution of the problem.
3. **Context**: Put away the paradox and shelve the original problem, understand practices of inner stakeholders involved in the problem situation before attempting to solve it.
4. **Field**: Once an overview is achieved, widen the context to form a field, a wide intellectual space.
5. **Themes**: Understand deeper factors of players in the field. Ends with understanding of underlying "universals".
6. **Frames**: Themes, especially those that are shared among many players, might act as a basis for frames, but ideation of a new frame is largely a creative leap.
7. **Futures**: Envision how things might work under a frame and get feedback.
8. **Transformation**: Critically evaluate what frames and solutions are feasible. Unearth changes that are required in the proposed ideas and practices of the participating organisations. Weed out unfeasible frames
9. **Integration**: Integrate frames into broader organisations.

The first three steps deepen understanding of the problem before any frame is proposed. Steps 4 to 5 broaden the perspective, and Step 6 is the creative leap. Steps 7 to 9 concern stakeholders in organisational settings. Since our context and problem are on an individual-level, we adapt the first three steps and step 6 as the basis for our *dorst_frame*.

Notes as Context for Researchers

If deep contextual understanding is essential for problem reframing, how can such an understanding be captured? Research notes (literature notes, field notes, etc.) serve as a primary mechanism through which researchers manage the knowledge they develop. Phillippi and Lauderdale (2018) suggests that field notes, for example, explicitly capture contextual information, including physical setting, social interactions and temporal factors. The Zettelkasten method is another example of a contextualised note system that can sustain creative output, having enabled sociologist Niklas Luhmann to produce over 50 books and 550 articles (Luhmann 1981; Schmidt 2016).

As these examples display, taking notes has high utility for capturing research context, and this has driven the development of digital personal knowledge management tools like Roam Research (Roam Research, Inc. 2026), Notion (Notion Labs, Inc. 2026), and Obsidian (Obsidian Foundation 2026). However, despite the clear importance of notes as context capture mechanisms and the proliferation of digital note-capturing tools, there exists a notable gap in research on how AI systems can effectively utilise researchers' personal notes. While LLM and AI systems have explored the use of context for very specific research problems, for example, in electronic lab notebooks (Fraga et al. 2024; Jalali et al. 2024), and using personal collections of research papers to suggest new research papers (Lee et al. 2024). The same cannot be said for literature notes.

This gap motivates our choice of Obsidian as the platform for our study. Obsidian (Obsidian Foundation 2026) is an open-source, extensible markdown-based software with a robust plugin architecture that allows for custom integration of LLM capabilities while preserving researchers' ownership and control of their notes. Thus, our system exists as a plugin that Obsidian users can use with their personal notes.

[1]The GitHub repository for the paper is: https://github.com/abhinavsood2002/Research-Problem-Reframing-LLMs/

LLMs and Problem Reframing

Our study asks LLMs to generate answers to open research questions that might serve as starting points for problem reframing. This task sits at the intersection of question answering, framing, and creative ideation. LLMs perform well on scientific questions with correct answers (Kon et al. 2025), but our participants' questions require generating new perspectives, not retrieving facts. In computational creativity, framing typically refers to contextualising creative artefacts to shape the audience perception (Charnley, Pease, and Colton 2012; Cook et al. 2019). Our use is closer to Dorst and Schön's sense of generating alternative perspectives on a *problem* to restructure how the researcher thinks about it.

Researchers already use LLMs to seek new information, edit or create new writing, get new ideas, frame research papers and modify or generate data (Liao et al. 2025). When used for ideation, LLMs can directly generate research artefacts like ideas and hypotheses (Wang et al. 2024; Yang et al. 2025; Si, Yang, and Hashimoto 2025). However, such artefacts display consistent limitations: lower technical depth, unrealistic assumptions, and a lack of diversity (Wang et al. 2024; Si, Yang, and Hashimoto 2025). These limitations also apply when LLMs are used for problem reframing. When over 280 designers were performing problem reframing with assistance from LLMs, through 3 different approaches, structured, direct and free-form, the authors found no improvement in the quality of the frames with the LLMs. "The most commonly reported issue with this study's LLM-generated frames was that they were too low-quality to serve as foundations" (Shin et al. 2025). As mentioned earlier, our work does not claim to resolve these known limitations; issues of feasibility and hallucination remain present in our outputs. Instead, we investigate whether structuring LLM prompting around Dorst's reframing methodology affects whether researchers find outputs interesting enough to pursue, and we examine what characteristics distinguish answers with reframing potential from those without.

We focus on measuring interest because, in the context of problem reframing, an answer must first be perceived as worth pursuing by the individual. Otherwise, an uninteresting output will never be engaged with deeply enough to shift how they think about a problem.

System Design

Our system is implemented as an Obsidian plugin that enables researchers to use their existing notes for LLM-generated answers. The system architecture comprises four main components: a React-based frontend integrated into Obsidian, a FastAPI backend server, a vLLM server to run local inference using GPT-oss-120b (medium) (Agarwal et al. 2025) model, and a PostgreSQL database to save user data.

Frontend Architecture

The plugin frontend is built with React and Chakra UI. The interface consists of three primary views:

1. **LoginView** handles user authentication and login to manage data for different participants.
2. **SetupView** enables context configuration by allowing researchers to input their research interest and select relevant notes and PDFs from their Obsidian vault.
3. **FrameBrowserView** displays generated answers. This view is the one that is being displayed when the plugin is embedded in Obsidian's interface in Figure 1. The ranking of answers in Session 2 was done through a ranking modal that appears when the Rank Frames button in the same interface is clicked.

Additionally, we generate markdown so that any tables/formatting can be appropriately rendered. Math equations are rendered through KaTeX.

Backend Architecture

The backend is a FastAPI server. The database schema includes tables for users, notes, PDFs, PDF fulltext, frames, and user contexts.

Frame Generation Strategies The system implements three distinct prompting strategies for frame generation:

1. **Direct Answer Strategy** (*direct_strategy*): Generates answers using only the research interest and research question, without accessing user notes or PDFs. This serves as a baseline.
2. **All Content Strategy** (*all_content*): Provides the LLM with the research interest, research question, and the full content of all selected notes and summarised PDFs as context.
3. **Dorst's Frame Strategy** (*dorst_frame*): Implements a structured four-step sequential process based on Dorst's frame creation methodology (Dorst 2015b). The strategy executes four LLM calls in sequence: **Archaeology**, **Paradox**, **Context**, and **Frames**. Each step builds on previous outputs, with all steps having access to the full content of selected notes and summarised PDFs. Complete prompts for all strategies are provided in the linked GitHub repository.

For each of the strategies that use the notes and PDFs, notes are included verbatim, and PDFs are first conditionally summarised using an LLM prompt that creates 800-1000-word summaries highlighting content relevant to the user's research interest and notes. We use Docling (Team 2024) to parse PDFs into full text. Additionally, P6 and P7 verified that the summaries corresponded to content from the PDFs.

To ensure diverse answers for session two, for each new generation of a question- strategy combination, previously generated answers of that question-strategy combination are also passed to the prompt. The LLM is instructed to pursue a direction that is different from those answers. This, in turn, means that answers are not truly independent of one another; however, strategies and questions still remain isolated.

Study Design

We conducted a two-session user study with 7 research students to investigate how LLMs can support reframing in research contexts. Sessions with participants P1, P2, P3, P6,

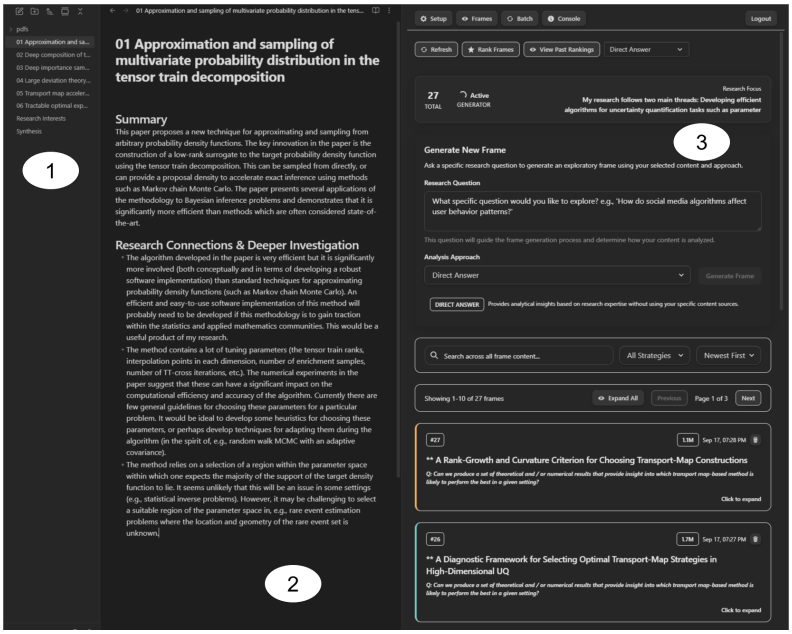


Figure 1: Our Obsidian Plugin. 1: Obsidian file pane representing the different notes the participant has taken, and a *pdfs* folder to store the PDFs of research papers for each note. 2: An example of a participant's note on a research paper. 3: Our plugin that keeps track of different frames generated for each user. Frames are generated in response to a research question and after the context has been set.

and P7 were conducted in person, while sessions with participants P4 and P5 were conducted online. The User Study had human ethics approval from a Human Research Ethics Committee at the University of Sydney. Each participant filled out a consent form agreeing to the collection of audio and computational data.

Participants

We recruited 7 research students through posters, social media and direct outreach to eligible participants. Research students were chosen due to being a convenient demographic. All participants were active student researchers who were engaged with academic literature for their research. Table 1 provides demographic information on our participants. Participants received AU$70 for completing both sessions.

Session 1: Context capture through note-taking

The first session lasted 2 to 2.5 hours and focused on capturing participants' research context through note-taking on previously read papers. The session proceeded in three stages:

1. **Phase 1 Pre-Interview (8-15 minutes):** We conducted a semi-structured interview to understand participants' existing note-taking and use of AI practices. Questions included:
   - Do you use AI (like ChatGPT) in your research? How?
   - Where would you say your research ideas come from?
   - Are there any specialised AI tools that you use for your research work?
   - What kinds of notes do you take while doing research?
   - What is the structure of these notes?
2. **Phase 2 Note-Taking (75-100 minutes):** Participants took notes on research papers they had previously read and had identified as relevant to their research. Participants were given a structured note-taking guide designed to promote more effective note-taking and to explicitly strengthen the connections between ideas, which has been shown to lead to higher-quality notes (Kiewra 1989). The guide describes four types of notes:
   (a) Short summaries highlighting the relevance of the paper to their research
   (b) Connections to the research paper (concepts, questions, or methodologies)
   (c) Aspects that the participant feels warrant deeper investigation
   (d) Surprising or unexpected findings

   Each of these types was elaborated with examples, and the entire guide is available in the linked GitHub repository. While participants were provided the guide, they were instructed to take notes that they thought would be most useful to them and to only follow the guide if they thought it provided them value.
3. **Phase 3: Question Synthesis (15-20 minutes)**: After completing their notes, participants synthesised three research questions based on their note-taking session. These were open questions in their research rather than questions with known answers. Participants rated their satisfaction with the quality of notes produced on a 7-point Likert scale (1 = Very Dissatisfied, 7 = Very Satisfied). This metric was used because a participant's per-

Participant	Gender	Age	Degree / Program (Year)	Research Area / Discipline
P1	Male	18–24	Doctor of Philosophy (Year 1)	Statistics
P2	Female	25–34	Doctor of Philosophy (Year 3)	Human–Computer Interaction
P3	Male	25–34	Doctor of Philosophy (Year 1)	Biomedical Engineering
P4	Male	18–24	MASc, Mechanical Eng. (Year 2)	Process Eng., Commercialisation, Sustainability
P5	Male	25–34	Doctor of Philosophy (Year 3)	Creative AI / Human–AI Co-Creativity
P6	Female	25–34	Doctor of Philosophy (Year 3)	Psychology
P7	Male	35–44	Doctor of Philosophy (Year 3)	Design Innovation, AI

Table 1: Participant Demographics

sonal assessment of their own notes' quality is likely the most important factor while determining the overall quality of the notes (Friedman 2014).

Session 1 - Outcomes

Participants demonstrated diverse note-taking practices prior to taking part in the study. Only P1 maintained an extensive collection of notes on research papers in Obsidian. P3, P4, and P6 mentioned taking limited notes, usually comparing specific biomaterials, commercialisation processes, and impairment modalities, respectively. P4 and P5 did maintain a limited amount of notes in reference management tools like Zotero (Zotero Community 2025) while P3, though documenting extensively in digital and physical diaries, did not systematically take notes on research papers.

After the session, participants were generally satisfied with the quality of their notes. On a 7-point Likert scale, P1, P2, and P3 rated their notes 5/7; P4, P6, and P7 rated their notes 6/7; and P5 rated theirs 7/7. P1 found the guide "representative of the specific kinds of notes I write" and identified the research connections section of the notes to correspond to notes they would personally write the most. P4, P5, P6, and P7 appreciated the structured writing process, with P4 and P5 both noting that engaging in such a process would benefit their thesis writing as they accomplished substantial writing during the session. P2 and P3 preferred to write notes in their own style. P2 primarily on summarising sections relevant to their research with questions interspersed in the notes. P3 recorded key points of the paper relevant to their research interests.

Session 2: Frame Generation and evaluation

The second session occurred after the first session, with a gap of a few minutes to a few days, depending on the participant's schedule. Our system generated 27 answers in response to the three questions provided by each participant. For each question, 3 answers for each strategy were generated.

The produced answers were then further categorised in a 2-step process:

1. First, participants reviewed all 27 generated answers and categorised each as **interesting** or **useless** based on an initial reaction to the answer's potential value to their research and from their interest towards the answer. These answers were provided in an order grouped by question, with the order of answers within each question randomised to prevent bias from strategy. This categorisation produced for each participant 9 answers (capped) that they deemed were interesting, and 18 that were "useless".
2. Participants then ranked their top 9 answers through 36 pairwise comparisons.
3. For the top-3 answers, participants filled out a questionnaire that recorded why that particular answer was interesting to them. This questionnaire contained 8 questions. The first seven were answered on a 7-point Likert scale, and the last one was an open-ended short-form question asking why the answer was interesting. The seven Likert scale questions were:

   - Q1. This answer taught me something about the problem I am working on, which I didn't know before.
   - Q2. This answer helped clear up my thinking about part of the problem I wasn't sure about.
   - Q3. This answer challenges the way I've been thinking about this problem.
   - Q4. This answer connected my problem to something that I hadn't been thinking about before.
   - Q5. This answer helped me feel that I'm on the right track with the way I'm thinking about this problem.
   - Q6. This answer surprised me.
   - Q7. This answer made me think more deeply about the problem.

The session was interactive, participants were encouraged to think their reasoning out loud when possible and were asked what they felt made particular answers interesting/useless. Other questions asked include: Do any of the answers make you rethink your question? For this question, what criteria did you use to categorise answers as interesting?

Results

Participants P1, P3, P4, P5, and P6 reported that the top-3 answers did make them rethink their questions to some extent, indicating the answers provide potential for reframing. To unpack this reframing potential, we first compare how the three prompting strategies performed, then characterise what made answers interesting or useless, and finally examine how participants' ratings across the seven evaluation questions cluster into distinct response profiles.

How Do the Different Prompting Strategies Compare?

Figure 2 presents the overall performance of each strategy in the initial categorisation phase. Answers generated with the *dorst_frame* strategy received the highest rate of interest, with (n=29/189) being classified as interesting. This outperforms *direct_strategy* (n=17/189) generated answers and *all_content* (n=17/189) generated answers.

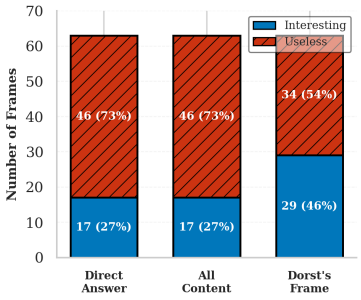


Figure 2: Strategy performance in first-pass categorisation across all participants (N=7)

In the top-3 rankings, Figure 3 shows that *dorst_frame* accounted for more than half of all top-3 ranked frames (n=11, 52%), followed by *direct_strategy* (n=7, 33%) and *all_content* (n=3, 14%).

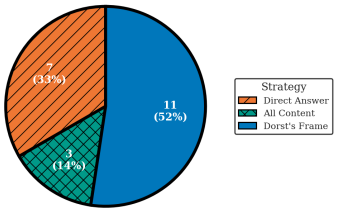


Figure 3: Distribution of prompting strategies among participants' top-3 ranked answers (N=21 total)

While aggregate results favour *dorst_frame*, individual participants exhibited varied preferences as visible in Figure 4. In the initial categorisation phase, six out of seven participants (P1, P2, P4, P5, P6, P7) rated *dorst_frame* answers as interesting more frequently than other strategies. P3 was the exception, showing a preference to *direct_strategy*. The top-3 rankings revealed a more complex pattern. While P1, P4, P5, and P7 maintained their preference for *dorst_frame*, P2 and P3 selected predominantly *direct_strategy* answers.

We argue that this divergence in preference is strongly linked to what participants were looking for in answers. While P1, P4, P5, and P7 mentioned they were looking for specific information related to their questions, P2 and P6 mentioned structure like the answer containing dot points was an important criterion to qualify answers as interesting. P3 described that when they interact with AI, they are looking for a conversation to bounce ideas off of; they actively

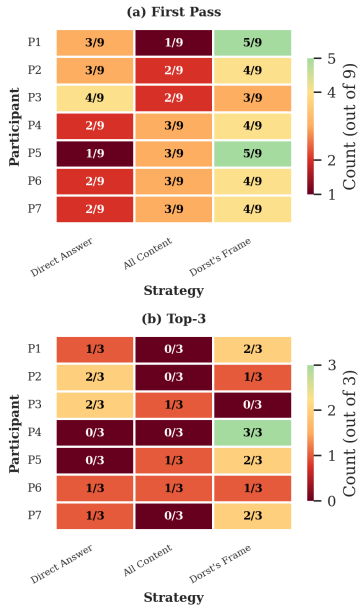


Figure 4: Per-participant strategy performance across evaluation stages. (a) First-pass categorisation showing the distribution of answers' underlying strategies if the answer was rated as interesting (b) Count of each strategy in each participant's top-3 answers

avoid jargon, which was more prevalent in the *dorst_frame* and *all_content* answers. Specifically, for them, interesting frames were described as "they let my creativity flow, if I am sitting there treating it as a person, yeah good point tell me more".

Some participants also observed a qualitative difference in the character of each strategy's output when presented with the answers grouped by strategy at the end of the session, if time permitted. P1 liked the *direct_strategy* answers, with some that were interesting, some less so, and some not making any sense. For *all_content* answers, they were on the whole less interested as "there wasn't enough detail in the answers. The first batch had more nonsense, but more variation". Answers from *dorst_frame* were described as "more similar to the first set, where the number of practically interesting ideas was high. Not as nonsensical as the first set". For P7, *direct_strategy* provided models or frameworks they hadn't heard of before, *all_content* was pulling more direct quotes from the papers, and *dorst_frame* was more assertive and proactive in communicating a particular viewpoint. P6 found it difficult to differentiate the output of the different strategies, though they noted that *direct_strategy* answers were formatted in a way they preferred with more dot points, while simultaneously containing references to research that was "less clear to me".

Taken together, these observations suggest that the three strategies occupied distinct roles: *direct_strategy* acted as a broad but uneven idea generator, *all_content* produced contextually grounded but shallow synthesis, and *dorst_frame* generated more opinionated and structured perspectives that participants could evaluate against their own understanding

of the problem.

What Makes Answers Interesting or Useless?

All participants except P7 ended up marking more than 9 frames as interesting in the first pass and then narrowed them down to their top 9. P7 alternatively expanded their selection from 5 to 9 with a lower standard of interestingness.

Characteristics of Interesting Answers Among the top-3 answers, two groups of interesting answers emerge. One group of answers reinforced participants' existing beliefs and deepened participants' understanding of their question to provide answers that the participants found interesting enough. These answers might not provide as much of an avenue for actual reframing, but still hold value as they might reinforce directions research participants can pursue. These answers typically had a high value for Q2 (cleared my thinking) and Q5 (feel I'm on the right track). This corresponds to what P6 described as a crystallisation effect:

> "I had some of these ideas but it's more crystallised now, sometimes you have these ideas at a very early stage and you don't know how to put it into words but I feel this time it's helping to crystallise what the idea is".

The other set of interesting answers introduced concepts and information that the participant was not thinking or aware of, while being relevant to their own research. Some of these answers proposed complete alternatives. For example, "Replacing petroleum coke with low-sulphur bio-char and by driving the carbothermic reduction in a renewable-electricity-fed plasma furnace" for an answer generated for P4. Others included very specific ideas that the participant had not considered in their context. For example, for P1, "It (the answer) proposed a metric (the concentration index) that I hadn't been thinking about previously but could perhaps be useful". Such answers could be identified by looking for high values of Q1 (answer taught me something I didn't know), Q3 (answer challenges the way I've been thinking) and Q4 (answer connected my problem to something that I hadn't been thinking about) simultaneously. For P5, they perceived answers displayed both these properties.

> "The interesting ones had phrasing that I could understand straightaway at first glance, it combined ideas in a particular way that seemed very familiar to me. It provided solutions that I felt I could work with straightaway and research by myself. It combined concepts in my notes in a surprising way, the terminology was nice, I like how it proposed frameworks, metrics or some computational solution to the problem. It seemed like a new idea hadn't been explored in the space. That was very interesting to me, I would like a copy for myself. "

We call these **reinforcing** and **reorienting** answers, respectively. Participants reported that they had follow-up questions for many of the answers, with that being one of the heuristics for an interesting answer.

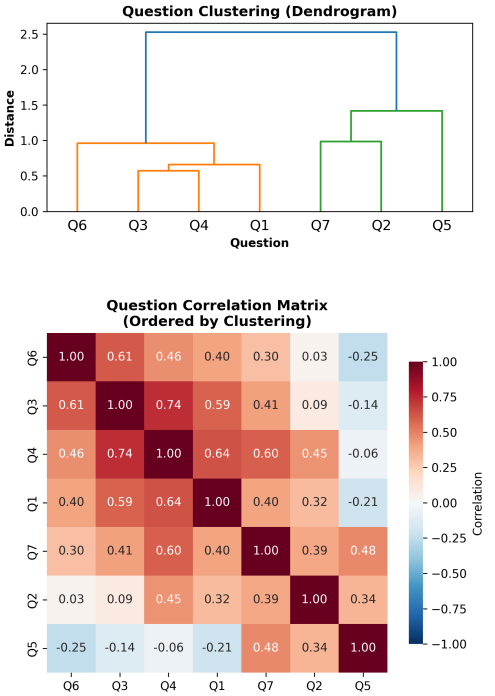


Figure 5: Pairwise Spearman correlations between the seven questions answered for the top-3 frames and agglomerative clustering from these correlations

Characteristics of Useless Answers Useless answers displayed a variety of issues. They misinterpreted terms (P5: "Misinterpreted M-creativity as meta-creativity when I meant model-creativity"); made invalid assumptions (P1: "Some assumed unavailable knowledge to the problem"); did not provide sufficient justification (P4: "Too many calculations, numbers just thrown at face without references"; P2: "Text needs more rationalization, it tries to connect a lot of ideas, but I am not sure how they can practically work that way"); and did not actually answer the question asked (P3: "Did not provide direct answers to the questions and provided additional information I did not want"; P6: "Sometimes I saw responses that do not answer the question"). For P7, many frames misinterpreted the theoretical and philosophical nature of their research papers. P7 noted that "because I had quite a few theoretical/philosophical papers, the AI couldn't really get it. None of them were suggesting anything that was surprising or inspiring". When the LLM did attempt to engage with P7's papers, it often "forced connections between all papers that were not as appropriate", diminishing the credibility of otherwise reasonable outputs.

Which questions correlate with each other?

Figure 5 presents the pairwise correlations for the seven Likert-scale questions across the 21 top-3 answers. These correlations reinforce our 2 distinct kinds of interesting answers. Cluster 1 consists of questions Q1, Q3, and Q4, identifying *reorienting* answers, and cluster 2 consists of Q2 and Q5, identifying *reinforcing* answers. Q6 (surprise)

relates more closely to the questions identifying *reorienting* answers, and Q7 (deepened thinking) to *reinforcing*, but both display additional interesting properties. Surprise was best associated with challenging a participant's thinking (Q3) while making the participant feel they were not on the right track (Q5). We further discuss this in the next section. Q7 and Q2 are the only questions that are positively related to all other questions, with Q7 displaying this property more strongly.

Discussion

Our results provide evidence for two different kinds of answers. **Reinforcing** answers crystallised ideas participants already held (high Q2, Q5), allowing researchers to better prepare their conceptual space of ideas. **Reorienting** answers through introducing new concepts, information, and connections (high Q1, Q3, Q4) pushed participants to consider directions they had not considered, aligning more closely with the conditions for p-transformational creativity. The correlation between Q7 (deepened thinking) and both clusters suggests that both confirmation and challenge can provoke substantive reflection, albeit toward different ends.

As participants had different conceptions of surprise, the role of surprise within this structure requires specific attention. For P3, one of their top three answers included a concept not previously identified in the literature. However, they had recently been discussing this concept with medical students, which left them both surprised and validated when it appeared in the response. P4 and P5 located surprise in unexpected connections: for P4, "something I had no idea about being related to the project", and for P5, the system "bridging two research problems I have been thinking about in a novel and useful way". P6 articulated a more epistemically disruptive form of surprise. For P6, surprise occurs "if I already know something about a topic, and I think I'm correct, but then I read this, and I question if my understanding is correct or not". P7 introduced credibility as a further precondition: surprise means "presenting a really thought-provoking alternative perspective that has credibility".

In related work, Maher, Brady, and Fisher (2013) operationalise surprise as distance between observed and projected attribute values, and Grace et al. (2015) formalise novelty, value, and surprise as independent dimensions of creative design, acknowledging that different kinds of expectation violation can all give rise to surprise. In practice, these models compute surprise as a single score: how much a design deviates from some expected characteristic. Our participant data suggest that in a research context, sources of surprise can be so varied that a single measure might obscure more than it reveals. Instead, more targeted questions that require breaking down surprise by its source, whether from a research connection or from epistemically challenging a participant's understanding, might serve as better models of surprise.

Finally, the gap between *all_content* and *dorst_frame* indicates that merely supplying LLMs with context is not sufficient in producing outputs that researchers can use as bases for reframing. Both the strategies had access to the same notes and PDFs, yet the structure of *dorst_frame* pushed outputs to be described as more interesting. This suggests prompt architectures that mirror problem reframing processes can serve as good candidates for reformulation with AI-generated content.

Limitations

The main limitation of the study is the sample size of participants (n=7). While the quantitative comparisons point out a trend, they lack statistical power. Moreover, problem reframing is an iterative process. Will our participants actually use any of the frames they rated as interesting? Did those frames influence their subsequent research questions? Thus, future work could include providing frames generated in this session to the participants and then using a longitudinal follow-up to see if reframing did occur.

Our system relied on one LLM (GPT-oss-120b). Different LLMs might produce frames with different characteristics with the same strategies. Finally, our study only takes into account part of the modelable research context. That is, it largely focuses on literature notes. As seen in our pre-interviews, researchers keep notes in various forms and expanding what context is processed will be necessary to understand the context required for problem reframing in the various research situations that exist.

Conclusion

This paper investigated whether structuring LLM prompting around Dorst's frame creation methodology produces more interesting answers to researchers' own open questions, and characterised the answer profiles that emerged.

Through a two-stage study with 7 research students, we found that *dorst_frame* answers were rated interesting more frequently and comprised 52% of top-3 ranked answers, suggesting that how an LLM processes context matters more than just the context it receives. We further identified two distinct answer profiles, reinforcing and reorienting. Reorienting answers that correspond to answers with reframing potential correlate with questions Q1 (taught something new), Q3 (challenges thinking), and Q4 (unexpected connection). Thus, these questions serve as a useful proxy for identifying outputs aligned with conditions for p-transformational creativity.

These findings do not resolve known LLM limitations for reframing. Whether the answers participants found interesting would lead to productive reconceptualisation remains an open question.

Author Contributions

Author 1 (AS) conducted the user study, interviewed participants, developed the Obsidian plugin, and prepared the manuscript. Author 2 (SW), Author 3 (CP) and Author 4 (KG) supervised the research and provided direction and feedback for the development of this work.

Acknowledgement

We gratefully acknowledge the Collaborative Intelligence (CINTEL) Future Science Platform for their support towards this work.

References

Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; et al. 2025. gpt-oss-120b & gpt-oss-20b model card. arXiv:2508.10925.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Charnley, J. W.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *ICCC*, 77–81.

Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing in computational creativity-a survey and taxonomy. In *International Conference on Computational Creativity 2019*, 156–163. Association for Computational Creativity (ACC).

Creswell, J. W., and Creswell, J. D. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dorst, K., and Cross, N. 2001. Creativity in the design process: co-evolution of problem–solution. *Design Studies* 22(5):425–437.

Dorst, K. 2011. The core of 'design thinking' and its application. *Design Studies* 32(6):521–532. Interpreting Design Thinking.

Dorst, K. 2015a. Frame creation and design in the expanded field. *She Ji: The Journal of Design, Economics, and Innovation* 1(1):22–33.

Dorst, K. 2015b. *Frame Innovation: Create New Thinking by Design*. Cambridge, MA, USA: The MIT Press.

Foster, A. 2004. A nonlinear model of information-seeking behavior. *Journal of the American Society for Information Science and Technology* 55(3):228–237.

Fraga, F. P. A.; Poggi, M.; Casanova, M. A.; and Leme, L. A. P. P. 2024. Creating automatic connections for personal knowledge management. *SN Computer Science* 5(5):525.

Friedman, M. C. 2014. Notes on note-taking: Review of research and insights for students and instructors. *Harvard Initiative for Learning and Teaching* 1–34.

Gero, J. S., and Kannengiesser, U. 2014. *The Function-Behaviour-Structure Ontology of Design*. London: Springer London. 263–283.

Grace, K., and Maher, M. L. 2015. Surprise and reformulation as meta-cognitive processes in creative design. In *Proceedings of the third annual conference on advances in cognitive systems ACS*, volume 8.

Grace, K.; Maher, M. L.; Fisher, D.; and Brady, K. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3(3-4):125–147.

Jalali, M.; Luo, Y.; Caulfield, L.; Sauter, E.; Nefedov, A.; and Wöll, C. 2024. Large language models in electronic laboratory notebooks: Transforming materials science research workflows. *Materials Today Communications* 40:109801.

Kiewra, K. A. 1989. A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review* 1(2):147–172.

Kon, P. T. J.; Liu, J.; Ding, Q.; Qiu, Y.; Yang, Z.; Huang, Y.; Srinivasa, J.; Lee, M.; Chowdhury, M.; and Chen, A. 2025. Curie: Toward rigorous and automated scientific experimentation with ai agents. *CoRR*.

Lee, Y.; Kang, H. B.; Latzke, M.; Kim, J.; Bragg, J.; Chang, J. C.; and Siangliulue, P. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.

Leedy, P., and Ormrod, J. 2001. *Practical Research: Planning and Design*. Upper Saddle River, NJ and Thousand Oaks, CA: Merrill Prentice Hall and SAGE Publications, 7 edition.

Liao, Z.; Antoniak, M.; Cheong, I.; Cheng, E. Y.-Y.; Lee, A.-H.; Lo, K.; Chang, J. C.; and Zhang, A. X. 2025. LLMs as research tools: A large scale survey of researchers' usage and perceptions. In *Second Conference on Language Modeling*.

Luhmann, N. 1981. Kommunikation mit zettelkästen: Ein erfahrungsbericht. In *Öffentliche Meinung und sozialer Wandel/Public Opinion and Social Change*. Springer. 222–228.

Maher, M. L.; Brady, K.; and Fisher, D. H. 2013. Computational models of surprise in evaluating creative design. In *Proceedings of the fourth international conference on computational creativity*, volume 147, 3. University of Sydney Sydney, Australia.

Mejía, G. M.; Henriksen, D.; Xie, Y.; García-Topete, A.; Malina, R. F.; and Jung, K. 2023. From researching to making futures: a design mindset for transdisciplinary collaboration. *Interdisciplinary Science Reviews* 48(1):77–108.

Notion Labs, Inc. 2026. Notion — all-in-one workspace for notes, tasks, wikis, and databases. https://www.notion.so.

Obsidian Foundation. 2026. Obsidian: A knowledge base that works on local markdown files. https://obsidian.md.

Phillippi, J., and Lauderdale, J. 2018. A guide to field notes for qualitative research: Context and conversation. *Qualitative Health Research* 28(3):381–388. PMID: 29298584.

Roam Research, Inc. 2026. Roam research — a note-taking tool for networked thought. https://roamresearch.com.

Schmidt, J. 2016. Niklas luhmann's card index: Thinking tool, communication partner, publication machine. In *Forgetting Machines. Knowledge Management Evolution in Early Modern Europe*, volume 53.

Schön, D. A. 1983. *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books.

Shin, J.; Polyanskaya, A.; Lucero, A.; and Oulasvirta, A. 2025. No evidence for LLMs being useful in problem reframing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery.

Si, C.; Yang, D.; and Hashimoto, T. 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, ICLR '25.

Team, D. S. 2024. Docling technical report. Technical report.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 279–299.

Yang, Z.; Liu, W.; Gao, B.; Xie, T.; Li, Y.; Ouyang, W.; Poria, S.; Cambria, E.; and Zhou, D. 2025. Large language models for rediscovering unseen chemistry scientific hypotheses. In *The Thirteenth International Conference on Learning Representations*, ICLR'25.

Zotero Community. 2025. Zotero.

Remember Me: Enhancing Long-term Character Persistence in LLM-Generated Stories using Knowledge Graphs

Alvin Tan[1], **Jay Campanell**[1], **Richa Misra**[1], **Naina Panjwani**[1], **Jessica Fu**[1], **Rida Faraz**[1], **Joel Walsh**[2]
[1]Department of Computer Science, University of Southern California, LA, CA 90089, USA
[2]Department of Computer Science, Occidental College, LA, CA 90041, USA
{tanalvin, jcampane, rmisra, npanjwan, fujessic, faraz}@usc.edu
jwalsh2@oxy.edu

Abstract

Despite advances in the instruction-following capabilities of LLMs, long-form content generated by state-of-the-art models still lacks coherence and consistency. Previous work in Question Answering and factual text generation has leveraged knowledge graphs (KG)' representation capabilities, but there have been limited efforts to incorporate KGs in creative tasks. To improve LLMs' zero-shot performance in narrative generation, we propose a multi-agent story generation pipeline with knowledge graphs as long-term context storage to improve character and relationship consistency in long fiction. To this end, we designed and tested a pipeline that can generate long narratives chapter by chapter with LLM agents responsible for generation planning, knowledge graph extraction, and next chapter generation. We also created a comprehensive list of starting prompts to evaluate the pipeline's performance across genres. In order to evaluate the pipeline quantitatively, we devised a novel retention-based story consistency metric due to a gap in effective quantitative metrics for unreferenced text generation. The character retention metrics and relationship retention metrics reflect a narrative's long-term consistency across chapters. Through a statistical analysis of stories generated with the multi-agent KG pipeline and those generated with naive prompting, we found significant improvement in retention and cohesion in stories generated with KG guidance.

Introduction

As large language models (LLMs) become well-trained experts in question-answering and code-writing, people are exploring applications of LLMs in the creativity space. For example, research has been done on enhancing human-AI co-writing experience (Ocampo, Bown, and Grace 2025). Content creators are using LLMs to assist them in ideation and scriptwriting to boost efficiency. In addition, LLMs are commonly used in the education sphere to create personalized learning experiences for students. For instance, Ello is a company that applies LLM for personalized content creation and reading assistance for children (Ello 2025). As more and more people are trying to leverage the computational creativity of LLMs in fields like entertainment, literature, and education, there is a growing demand for systems that can generate dynamic, coherent, and pedagogically grounded narratives.

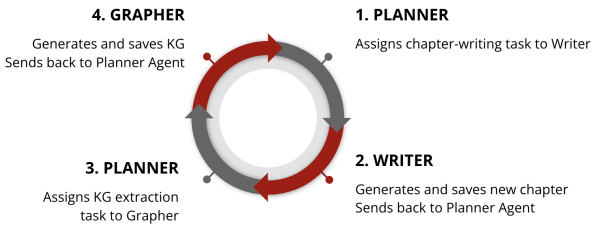


Figure 1: Agentic Structured Generation Workflow

Recent advances in large language models (LLMs) have demonstrated strong capabilities in text generation. However, off-the-shelf LLMs (such as OpenAI's ChatGPT or Anthropic's Claude) struggle with long-form generation, frequently exhibiting issues with narrative drift, character inconsistency, relationship loss, and logical breakdowns across chapters. These limitations are particularly problematic for personalized-learning experiences, where cohesion, continuity, and conceptual consistency are essential for comprehension and learning (Ismayilzada, Stevenson, and van der Plas 2024). Prior work has shown that LLM-generated stories often lack sustained relationships between entities and fail to preserve implicit connections over time, resulting in fragmented narratives (Yi et al. 2025).

Knowledge graphs (KGs) offer a promising structural representation for addressing these challenges. By explicitly modeling entities and their relationships, KGs provide a mechanism for tracking narrative state and enforcing consistency. Previous research has explored the use of graphs for narrative reconstruction and text generation, as well as automated KG extraction from text (Ehrlinger and Wöß 2016). Separately, multi-agent frameworks have emerged as effective tools for decomposing complex generation tasks into coordinated subtasks. However, existing work has largely treated knowledge graph extraction, story generation, and evaluation as isolated components. To our knowledge, there is no prior work that systematically measures how an agentic pipeline that iteratively builds and conditions on narrative knowledge graphs impacts long-form story cohesion and en-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

tity retention.

In this work, we present a multi-agent, graph-structured generation framework for long-form story creation using LLMs. Built using LangGraph, a framework specialized towards LLM multi-agent applications, our system decomposes the generation into specialized agents—a planner, writer, and grapher—that collaboratively generate stories chapter by chapter while maintaining a shared, evolving narrative knowledge graph (LangChain 2026). Unlike prior approaches that rely solely on post-hoc graph extraction or KG-only conditioning, our framework enables the writer agent to directly condition on both previously generated text and the current knowledge graph state, allowing for explicit preservation of characters, relationships, and narrative objectives over time.

Beyond generation, we introduce a more targeted evaluation methodology that moves beyond surface-level text similarity metrics. We propose entity- and relationship-centric metrics grounded in named entity recognition (NER) to measure character retention, relationship persistence, and narrative stability across chapters (Pakhale 2023). These metrics allow us to directly quantify the benefits of knowledge-graph-guided generation in ways that align with human judgments of cohesion, defined as a quality that measures how well characters, relationships, and plot connect within stories by Graesser et al. (2004) and logical consistency.

Our contributions are threefold:

1. We introduce a LangGraph-based multi-agent framework that integrates iterative knowledge graph construction directly into long-form LLM story generation.[1]
2. We demonstrate that conditioning generation on evolving narrative knowledge graphs significantly improves character retention, relationship persistence, and structural cohesion across chapters.
3. We propose and validate new evaluation metrics for long-form narrative generation that explicitly measure entity and relationship continuity, complementing existing unreferenced story quality metrics.

Related Work

Our work builds on existing work in LLM text generation for narratives, using graphs for generation, knowledge graph extraction, and multi-agent frameworks.

With the impressive capabilities of modern LLMs for text generation, researchers have conducted investigations into the ability of LLMs to formulate human-level stories. For example, Tian introduces a computational framework to analyze narratives through story arcs, turning points, etc., to evaluate LLM story generation performance (Tian et al. 2024). Common sense reasoning is also applied to automated story generation, through the introduction of a framework that provides an option to model the interaction between multiple characters (Peng et al. 2023).

[1]Our code is available at: `https://github.com/usc-caisplusplus/kg-story-persistence`

Our decision to use narrative KGs to provide information is influenced by previous work that used graph structures to boost text generation. Damiano et al. describe an ontology composed of entities (e.g., character, event, location) as nodes and relationships between these nodes as edges to represent narratives (Damiano and Lieto 2013). Blin (2022) and Kok et al. (2024) both focus on reconstructing a narrative based on knowledge graphs constructed from existing events. These works utilize knowledge graph datasets such as FARO (Rebboud, Lisena, and Troncy 2022) and WebNLG to train text generation models from knowledge graphs (Blin 2022; de Kok et al. 2024). These results are promising, but there has yet to be work that incorporates this idea for fictional story generation without pre-existing knowledge graphs.

Particularly in research presented on computational creativity, structured, knowledge-driven frameworks have been used to approach narrative generation. Early systems depend on symbolic representations and unique narrative planning to guarantee causal cohesion and a global plot structure (Riedl and Young 2010). Other foundational systems, like Mexica, emphasized creativity as a process driven by internal narrative knowledge rather than surface-level text generation by modeling stories with explicit representations of characters, emotions, and events (Pérez y Pérez and Sharples 2001; Gervás 2009). More computational creativity work has increasingly explored hybrid approaches which are neural symbolic approaches that combine language models with explicit knowledge structures to guide creative output (Veale, Cardoso, and y Pérez 2019). This line of research provides conceptual grounding for our use of narrative knowledge graphs as an explicit representation to guide LLM-based story generation, positioning our approach as a continuation of long-standing computational creativity perspectives adapted to modern generative models.

Many studies conducted within the computational creativity community have examined how explicit narrative theories and structures can direct computational story generation systems. Pickering and Jordanous demonstrate how high-level narratological concepts can methodically shape plot developments rather than leaving them to unrestricted text generation. Narrative theory is used to engineer unexpectedness in a story generator (Pickering and Jordanous 2017). In order to show how structured representations of actions and intentions can support cohesive plots like ”kill the dragon, rescue the princess,” Laclaustra et al. (2014) frame story generation as a plan-based, multi-agent process in which characters pursue goals within an explicitly modeled narrative world. Moreover, in order to choose and polish narrative content while keeping an eye on the underlying story structure, Méndez and Gervás (2023) incorporate ChatGPT into a story-sifting pipeline using a sizable language model. This line of research provides conceptual grounding for our use of narrative knowledge graphs as an explicit representation to guide LLM-based story generation, positioning our approach as a continuation of long-standing computational creativity perspectives adapted to modern generative models.

There has been much foundational work over the past

three decades on both entity and knowledge graph extraction from text. Entity extraction (named or unnamed), which is a preliminary step in knowledge graph extraction, was initially done by applying supervised learning over large datasets of entities (Grishman and Sundheim 1996; Sang and De Meulder 2003). The Relation extraction task involves taking pairs of discovered entities and determining the probability of a relation based on the textual context. Initially, this was done by techniques like logistic regression or handcrafted rules (Yates et al. 2007), an approach that was eventually supplanted by neural methods (Zeng et al. 2014) and large language model prompting (Cabral, Claro, and Souza 2024). For this work, the GraphRAG approach (Han et al. 2025) served as a prompt engineering starting point for extracting knowledge graphs. We also took inspiration from KG-Agent, a multi-agent-based model for extracting knowledge graphs from text (Jiang et al. 2024).

Narrative generation has also raised a need for narrative evaluation metrics. Referenced metrics like BLEU, a metric that ranges from 0 to 1 and utilizes geometric averages of overlapped word sequences between two texts (Papineni et al. 2002), are common in text generation evaluation but are inapplicable to our purpose. On the other hand, UNION proposes an unreferenced metric to evaluate open-ended story generation (Guan and Huang 2020).

Work has been released detailing different pipelines for story generation with knowledge graph enhancement. In Pan et al., the authors propose a pipeline that relies on counts of previous context and scenes to determine whether or not to move forward with the generation and make use of a small sample size of human annotators (Pan et al. 2025). Building on prior research in knowledge graph systems, our work introduces an agentic and iterative framework that dynamically extracts from an evolving knowledge graph during generation. This approach enables structured long-term entity and relationship tracking in LLMs that offers new empirical insight.

Methodology

Knowledge graph retrieval

Prior work on narrative knowledge graph extraction has explored transformer-based models and LLM prompting methods. In this work, we rely exclusively on LLM-based prompting to construct knowledge graphs from narrative text, prioritizing flexibility and generalization over domain-specific extraction accuracy. This choice allows our system to operate across open-ended fictional narratives without requiring labeled data or fine-tuning.

Knowledge graph construction is performed incrementally during generation. After each chapter is produced, entities and relationships are extracted and used to update a shared narrative knowledge graph representing characters, locations, events, and their relations. The entity extraction process targets these specific literary components of the story: character, event, object, location, organization and theme. The relationship extraction process discovers and stores short-form text representations of the relationships between extracted entities, saved as edges in the knowledge graph. This evolving graph is maintained as part of the system state and made available to subsequent generation steps.

By explicitly tracking narrative structure through an evolving knowledge graph, the system preserves contextual information across chapters and enables direct measurement of character and relationship retention in long-form story generation.

Multi-agent generation pipeline

We implement our multi-agent pipeline using LangGraph, a graph-based framework for orchestrating multi-agent LLM workflows with explicit state propagation and control flow (LangChain 2026). Rather than relying on a single generation step or post-hoc processing, our approach decomposes story generation into coordinated agents that operate over a shared state, enabling explicit control over narrative structure and memory.

The pipeline consists of three specialized agents: a writer, a grapher, and a planner. Story generation proceeds iteratively on a chapter-by-chapter basis. The writer agent generates a new chapter conditioned on both the preceding chapters and a structured knowledge graph context derived from the current graph store. After each chapter is produced, the grapher agent extracts entities and relationships from the new text and inserts them into an evolving knowledge graph structure. The planner then increments the chapter counter and resets agent completion flags to prepare the next iteration.

LangGraph represents this workflow as a directed graph with three nodes: writer, grapher, and planner. After the writer node executes, a router function inspects the shared story state to determine whether to terminate the story at this current chapter.

By persisting entities and relationships in a knowledge graph structure across iterations and surfacing chapter-of-introduction, milestone chapters, and recent chapters as structured context blocks in the writer's prompt, the pipeline enables consistent character tracking, relationship preservation, and long-term narrative cohesion across chapters, addressing limitations of unimodal LLM prompting for long-form text generation.

Story Arcs

We noticed that generations with the same prompt set often lead to similar results. To increase variance in the content generated to better understand the pipeline's performance, we incorporated story arcs in addition to a curated list of starting prompts. Following Tian's paper, we adopt discourse-aware generation, including the story arcs: Rags to Riches, Riches to Rags, Man in a Hole, Double Man in a Hole, Icarus, Cinderella, Oedipus (Tian et al. 2024). Furthermore, the inclusion of story arcs in the generation process provides a clear endpoint for the generation process, allowing the agentic framework to identify a distinct stopping point when writing out narratives.

In the system prompt instructing the planner, the agent is prompted to choose one of the story arcs to replicate. With a reference to traditional story arcs, the pipeline is able to generate more diverse results based on the prompt.

Results

Traditionally, researchers use translation metrics like BLEU or METEOR to evaluate LLM text-generation performance (Papineni et al. 2002; Lavie and Agarwal 2007). However, these referenced methods fall short when evaluating unreferenced story generation, which lacks n-gram overlap. UNION, a model trained to distinguish between human-written stories and negative examples, has been devised to produce quantitative scores to evaluate unreferenced story generation. The authors automatically perturb human-written stories using four techniques (repetition, substitution, reordering, and negation alteration) to simulate common errors observed in natural language models, such as repeated plots, incoherent content, conflicting logic, and flipped semantics. At inference time, UNION takes a generated story as input and outputs the probability that it is human-written, serving as a reference-free quality score. Nevertheless, since UNION is based on BERT encoders and trained on short story datasets like ROC Stories, the model is unable to serve our goal of evaluating multi-chapter long stories with more than 4000 words on average. (Guan and Huang 2020)

We have also experimented with LLM-as-judge for evaluation. However, LLMs are generally inaccurate in giving numerical ratings that are reflective of story quality. Also, as reported by Wang et al. (2025), LLMs tend to have favoritism when judging LLM-generated stories, producing unwanted bias.

Retention Analysis

While existing measures fail to evaluate unreferenced multi-chapter story generation comprehensively, we seek to evaluate the pipeline by focusing on the effect of character and relationship retention. Since one of the major purposes of our pipeline is to enhance long-term cohesion throughout the generated story, the measurement of how characters and relationships are preserved across chapters is reflective of the pipeline's performance. We conduct retention calculations for characters and relationships in each chapter in three ways: chapter 1 retention, rolling retention, and cumulative retention. These metrics are calculated on a chapter-by-chapter basis. The characters and relationships are extracted with Name Entity Recognition (NER) instead of LLMs. Characters are extracted with NER directly, chapter by chapter. Relationships are extracted based on character co-occurrence in text windows and verb-based dependency recognition.

Chapter 1 retention Chapter 1 retention reflects the percentage of entities from the first chapter that are still present in the current chapter, checking the pipeline's performance in remembering long-term relationships and callback characters.

$$\text{chapter 1 retention} = \frac{current \cap chapter_1}{|chapter_1|}$$

Rolling Retention Rolling retention represents the percentage of the previous chapter's entities that are still present in the selected chapter, evaluating the story's chapter-to-chapter continuity.

$$\text{rolling retention} = \frac{current \cap previous}{|previous|}$$

Cumulative Retention Cumulative retention shows the percentage of all historical entities currently active in the selected chapter, reflecting the overall engagement of the new chapter with the established storyline.

$$\text{cumulative retention} = \frac{current \cap all\,previous}{|all\,previous \cup current|}$$

Experiment Setup

To evaluate the effect of the pipeline on narrative retention, a series of Student's T-tests was conducted on the retention metrics, comparing two groups of LLM-generated stories (specifically with the claude-haiku-4.5 model: those produced with the multi-agent KG pipeline (50 stories, 270 chapters total) and those produced with naive prompting (50 stories, 231 chapters total). Ten dependent variables were examined across two categories: character retention (total characters, new characters, Chapter 1 retention, rolling retention, and cumulative retention) and relationship retention (total relationships, new relationships, Chapter 1 retention, rolling retention, and cumulative retention). Character retention results are presented in Table 1, and relationship retention results are presented in Table 2.

Experiment Results

Empirically, we show that stories generated with our knowledge-graph-guided pipeline exhibit higher character and relationship retention than those generated without graph integration. Our NER-based evaluation reveals consistent gains in narrative stability across both character and relationship dimensions. From Table 1 and Table 2, five of ten metrics showed statistically significant differences between KG-guided and stories generated from naive prompting. From Table 1, KG-guided stories retained significantly more chapter 1 core characters across the following chapters (mean = 0.689) compared to No-KG stories (mean = 0.622). Cumulative character retention was also significantly higher in the KG condition (mean = 0.500 vs. mean = 0.457), together with rolling character retention (mean = 0.770 vs. mean = 0.731). Also, the KG-guided pipeline tends to introduce fewer new characters in each chapter (mean = 1.244 vs. mean = 1.454), indicating a focus towards building deeper character development instead of introducing new but flat characters. For relationships, KG-guided stories produced significantly more total relationships per chapter (mean = 3.948 vs. mean = 3.208) and retained a significantly greater proportion of chapter 1 relationships across subsequent chapters (mean = 0.294 vs. mean = 0.195). Notably, no significant differences exist for the number of new relationships per chapter, or for rolling or cumulative relationship retention, indicating that KG guidance primarily strengthens the persistence of early-established narrative relationships while lacking emphasis on immediate relationship retention. Nevertheless, it performs better than base-

Table 1: Character Retention

Metric	KG			No-KG			t-test		
	M	SD	n	M	SD	n	t	p	Cohen's d
Total Characters	5.344	2.233	270	5.329	2.061	231	0.08	0.936	0.007
New Characters	1.244	1.324	270	1.454	1.422	231	-1.71	0.088	-0.153
Chapter 1 Retention**	0.689	0.240	270	0.622	0.264	231	2.97	0.003	0.266
Rolling Retention*	0.770	0.202	270	0.731	0.216	231	2.11	0.035	0.189
Cumulative Retention**	0.500	0.188	270	0.457	0.181	231	2.64	0.008	0.237

Note: M = Mean, SD = Standard Deviation, n = Sample Size (chapters). $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 2: Relationship Retention

Metric	KG			No-KG			t-test		
	M	SD	n	M	SD	n	t	p	Cohen's d
Total Relationships*	3.948	3.938	270	3.208	3.954	231	2.09	0.037	0.188
New Relationships	2.030	2.265	270	1.957	3.090	231	0.31	0.753	0.028
Chapter 1 Retention**	0.294	0.383	233	0.195	0.332	204	2.89	0.004	0.277
Rolling Retention	0.395	0.372	221	0.350	0.378	170	1.16	0.245	0.119
Cumulative Retention	0.170	0.194	268	0.139	0.213	224	1.72	0.086	0.156

Note: M = Mean, SD = Standard Deviation, n = Sample Size (chapters). $^{*}p < 0.05$, $^{**}p < 0.01$.

line LLM in preserving characters both long-term and short-term. Taken together, these results provide concrete and interpretable evidence that agentic KG conditioning leads to stronger long-term narrative consistency in LLM-generated text, particularly for characters and relationships introduced early on.

Discussion

This work demonstrates that integrating knowledge graphs into a multi-agent generation pipeline substantially improves the cohesion, consistency, and structural integrity of long-form stories generated by LLMs. By explicitly tracking entities and relationships through an evolving narrative knowledge graph, our framework addresses persistent limitations of traditional LLM prompting, including character loss, relationship drift, and logical inconsistency between chapters. Using KG as an intermediate shared data structure across agents also preserves an explicit trace of the logic behind the generated content, with the chapter-by-chapter evolving KGs as a creativity footprint.

Case Study

For illustration, we can take two examples of stories generated with the same starting prompt:

> "During a chess tournament, the reigning champion collapses dead mid-game, poisoned by a rare toxin."

The KG-guided story (Story 35) ended up with 10 chapters, while the Non-KG Generated story (Story 30) ended with 6 chapters. The ending knowledge graph representations of both stories are present in Figure 2. The KG guided story demonstrates stronger character and relationship consistency across its ten chapters, primarily because its protagonist, Sarah Mitchell, is given sufficient space for each psychological escalation to feel earned. Her core identity as a forensic analyst who left the job but can't stop thinking like a detective is established in the opening chapter and never falters. The narrative has prepared each risk that she takes: her decision to withhold Marcus's letter from Rousseau in Chapter 3 is a logical consequence of the detachment she expressed in Chapter 1; her impetuous walk into Henri's office in Chapter 8 is the logical conclusion of a stubbornness that has been tested and reaffirmed in several previous chapters. The most carefully sustained dynamic in the story is her relationship with Inspector Rousseau. He's skeptical but reverent since their initial encounter. He always pushes back on her methods. His surveillance of her in Chapter 9 is a real payoff because his protectiveness has been quietly planted since Chapter 4. The villain Henri Devereaux is treated with remarkable coherence also: his Chapter 8 confession monologue is consistent with the phrasing of his Chapter 7 phone call, and his farewell letter in Chapter 10 is in character: proud, unsentimental, never truly remorseful.

The story generated without KG guidance has a single protagonist, Marcus Ashford, who is built around an equally compelling premise: a man whose career was deliberately destroyed by a system he never understood. But the story has continuity issues that break that premise. The first few chapters show Marcus as a man who has spent thirty-six years avoiding confrontation, deliberately living in obscurity, and refusing to re-enter the chess world. But by Chapter 4, after one long talk with David Rothschild, he's all in on a high-stakes international conspiracy bust. The internal resistance that should define him collapses in roughly ten pages without meaningful dramatization of the struggle. This is the story's central consistency failure: a character

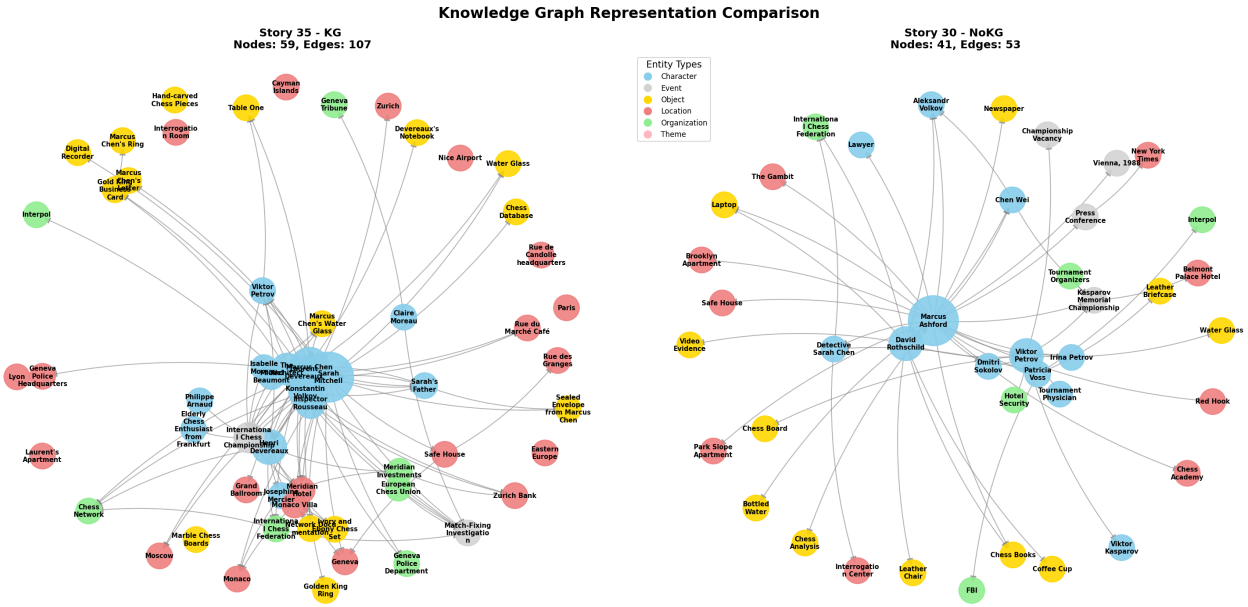


Figure 2: Knowledge Graph of Story Generated with Multi-Agent Pipeline (left) and with Vanilla Prompting (right)

whose entire psychological identity is built around avoidance should require far more pressure and time to reverse course than one conversation allows. The KG-guided story avoids this problem because Sarah resists, partially retreats, and re-commits across several chapters before each major escalation, making her transformation feel cumulative rather than sudden. The supporting characters in the No-KG story compound this problem in ways that the KG-guided story largely avoids. David Rothschild is introduced in Chapter 3 as a man paralyzed by thirty-six years of secret knowledge, too afraid to act alone. But Chapter 4 reveals he has already been actively coordinating with an Interpol undercover agent and maintaining a safe house network. This directly contradicts the passivity established just one chapter earlier. In addition, Sarah Chen is introduced in Chapter 2 as an influential figure who frames Marcus as a suspect, yet she disappears entirely from the narrative after that chapter with no resolution. The KG-guided story has its own version of this problem. Laurent Devereaux is introduced as a terrified ally, then dies off-page, but the story at least references his fate and gives it emotional weight at a funeral scene. The No-KG story simply stops acknowledging Detective Chen's existence. The one genuine consistency success in this story is Chen Wei, whose strange calm during Viktor's death in Chapter 1 is retroactively explained by his Chapter 4 reveal as an undercover agent. But this single well-executed thread cannot compensate for the broader pattern of characters introduced and then abandoned or contradicted. Taken together, the KG-guided generation workflow honors its characters across time, creating fluent character development and problem resolution.

Limitations and Future Work

While the generated stories exhibit stronger logic and continuity, challenges remain. We asked several human evaluators to read stories generated from the pipeline. Human evaluations indicate that the narratives can still lack stylistic specificity, emotional depth, and distinctive authorial voice. This suggests that while structural cohesion can be enforced through graph-based reasoning, higher-level creativity and expressiveness may require additional constraints or specialized agents.

Future work can explore several extensions to this framework. First, we plan to introduce verifier and critic agents to refine specificity, voice, and narrative tension during generation. Second, we aim to expand our evaluation suite by incorporating LLM-as-judge assessments more systematically and developing richer relation-strength and character-interaction metrics. We acknowledge that retention-based analysis focuses primarily on character and relationship cohesion and lacks complexity in measuring creativity at a human level. Third, we will explore conditioning the pipeline on different base models to study how agentic structure interacts with underlying model capabilities. Finally, we envision extending the system beyond narrative generation to educational applications, including intelligent textbooks that automatically generate comprehension questions, timelines, and concept maps from the underlying knowledge graph, as well as multi-modal extensions that incorporate illustrations to support younger readers.

Conclusion

Overall, this work highlights the effectiveness of multi-agent, knowledge-graph-driven generation as a scalable so-

lution for producing coherent long-form content. By combining explicit structural representations with agentic reasoning, our approach represents a promising step toward intelligent creative systems that support deeper text comprehension, personalization, and sustained engagement.

References

Blin, I. 2022. Building narrative structures from knowledge graphs. In Groth, P.; Rula, A.; Schneider, J.; Tiddi, I.; Simperl, E.; Alexopoulos, P.; Hoekstra, R.; Alam, M.; Dimou, A.; and Tamper, M., eds., *The Semantic Web: ESWC 2022 Satellite Events*, volume 13384 of *Lecture Notes in Computer Science*, 234–251. Hersonissos, Crete, Greece: Springer.

Cabral, B.; Claro, D.; and Souza, M. 2024. Exploring open information extraction for portuguese using large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, 127–136. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics.

Damiano, R., and Lieto, A. 2013. Ontological representations of narratives: a case study on stories and actions. In *Workshop on Computational Models of Narrative 2013*, volume 32 of *OpenAccess Series in Informatics (OASIcs)*, 76–93. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

de Kok, M.; Rebboud, Y.; Lisena, P.; Troncy, R.; and Tiddi, I. 2024. From nodes to narratives: A knowledge graph-based storytelling approach. In *Proceedings of the Text2Story'24 Workshop*, volume 3671 of *CEUR Workshop Proceedings*. CEUR-WS.org. Presented at Text2Story 2024 Workshop.

Ehrlinger, L., and Wöß, W. 2016. Towards a definition of knowledge graphs. In *International Conference on Semantic Systems*.

Ello. 2025. Ello – a creative network. `https://www.ello.com/`. Accessed: 2025-05-22.

Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30:49–62.

Graesser, A. C.; McNamara, D. S.; Louwerse, M. M.; and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2):193–202.

Grishman, R., and Sundheim, B. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Guan, J., and Huang, M. 2020. Union: An unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9157–9166. Online: Association for Computational Linguistics.

Han, H.; Wang, Y.; Shomer, H.; Guo, K.; Ding, J.; Lei, Y.; Halappanavar, M.; Rossi, R. A.; Mukherjee, S.; Tang, X.; He, Q.; Hua, Z.; Long, B.; Zhao, T.; Shah, N.; Javari, A.; Xia, Y.; and Tang, J. 2025. Retrieval-augmented generation with graphs (graphrag). `https://arxiv.org/abs/2501.00309`. Preprint, arXiv:2501.00309.

Ismayilzada, M.; Stevenson, C. E.; and van der Plas, L. 2024. Evaluating creative short story generation in humans and large language models. *ArXiv* abs/2411.02316.

Jiang, J.; Zhou, K.; Zhao, W. X.; Song, Y.; Zhu, C.; Zhu, H.; and Wen, J.-R. 2024. Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. `https://arxiv.org/abs/2402.11163`. Preprint, arXiv:2402.11163.

Laclaustra, I.; Ledesma, J.; Méndez, G.; and Gervás, P. 2014. Kill the dragon and rescue the princess: Designing a plan-based multi-agent story generator. In *Proceedings of the fifth international conference on computational creativity*.

LangChain. 2026. langchain-ai/langgraph. original-date: 2023-08-09T18:33:12Z.

Lavie, A., and Agarwal, A. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, 228–231. USA: Association for Computational Linguistics.

Méndez, G., and Gervás, P. 2023. Using chatgpt for story sifting in narrative generation. In *Proceedings of the 14th international conference on computational creativity*.

Ocampo, R.; Bown, O.; and Grace, K. 2025. Beyond chat: collaborative editors enhance human involvement and agency when co-writing with large language models. In *Proceedings of the Sixteenth International Conference on Computational Creativity*.

Pakhale, K. 2023. Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges. arXiv:2309.14084 [cs].

Pan, Z.; Andronis, A.; Hayek, E.; Wilkinson, O. A. P.; Lasy, I.; Parry, A.; Gadney, G.; Smith, T. J.; and Grierson, M. 2025. Guiding generative storytelling with knowledge graphs. *International Journal of Human–Computer Interaction* 1–23.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Peng, X.; Li, S.; Wiegreffe, S.; and Riedl, M. 2023. Inferring the reader: Guiding automated story generation with commonsense reasoning. `http://arxiv.org/abs/2105.01311`. Preprint, arXiv:2105.01311.

Pickering, T., and Jordanous, A. 2017. Applying narrative theory to aid unexpectedness in a story generation system. In *Proceedings of the eighth international conference on computational creativity*.

Pérez y Pérez, R., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *J. Exp. Theor. Artif. Intell.* 13:119–139.

Rebboud, Y.; Lisena, P.; and Troncy, R. 2022. Beyond causality: Representing event relations in knowledge

graphs. In *Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management (EKAW 2022)*, volume 13514 of *Lecture Notes in Computer Science*, 121–135. Springer.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39:217–268.

Sang, E. F. T. K., and De Meulder, F. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. https://arxiv.org/abs/cs/0306050. Preprint, arXiv:cs/0306050.

Tian, Y.; Huang, T.; Liu, M.; Jiang, D.; Spangher, A.; Chen, M.; May, J.; and Peng, N. 2024. Are large language models capable of generating human-level narratives? http://arxiv.org/abs/2407.13248. Preprint, arXiv:2407.13248.

Veale, T.; Cardoso, A.; and y Pérez, R. P. 2019. Systematizing creativity: A computational view. In *Computational Creativity*.

Wang, W.; Gao, M.; Hu, X.; and Wan, X. 2025. Towards a "novel" benchmark: Evaluating literary fiction with large language models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 21648–21673. Vienna, Austria: Association for Computational Linguistics.

Yates, A.; Banko, M.; Broadhead, M.; Cafarella, M.; Etzioni, O.; and Soderland, S. 2007. Textrunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 25–26. Rochester, New York, USA: Association for Computational Linguistics.

Yi, Q.; He, Y.; Wang, J.; Song, X.; Qian, S.; Zhang, M.; Sun, L.; and Shi, T. 2025. Score: Story coherence and retrieval enhancement for ai narratives. https://arxiv.org/abs/2503.23512v1. Preprint, arXiv:2503.23512v1.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; and Zhao, J. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2335–2344. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Appendix A: LLM Prompts

This appendix documents the exact system and user prompts passed to the language model in both the knowledge-graph-augmented pipeline and the baseline pipeline without knowledge graphs.

A.1 KG Pipeline

A.1.1 Writer Agent – Chapter 1

System prompt

You are a creative story writer. Write the first chapter of a story following the '*<archetype>*' narrative archetype.

Archetype description: *<archetype_description>*

Starting premise: *<starting_prompt>*

Write an engaging first chapter (800--1200 words) that:

- Introduces the main character(s)
- Sets up the initial situation
- Establishes the tone and setting
- Hooks the reader
- Ends with a compelling reason to continue

After writing, call the submit_chapter tool with your chapter text and set is_final_chapter to FALSE (this is just the beginning).

The '*<archetype>*' archetype should guide the overall story structure, so keep it in mind for future chapters.

User prompt

Write Chapter 1 of this story.

A.1.2 Writer Agent – Chapter $N > 1$

System prompt The system prompt for subsequent chapters is assembled dynamically and contains three logical sections: the core instruction block, the story context (last 1–3 saved chapter files), and a knowledge-graph context block.

Core instruction block

You are a creative story writer continuing a story following the '*<archetype>*' narrative archetype.

Archetype description: *<archetype_description>*

This is chapter *<N>*.

Story so far (last few chapters): *<story_so_far>* *<kg_context>*

Your task:

1. Write the next chapter (800--1200 words) that continues from where the previous chapter left off
2. Review the KNOWLEDGE GRAPH above to maintain consistency with established characters, locations, and relationships
3. Monitor the story's progression according to the '*<archetype>*' archetype and its description above
4. DECIDE WHETHER THE STORY HAS REACHED ITS NATURAL CONCLUSION based on:
   - Whether the archetype's narrative structure is complete
   - Whether character arcs are resolved
   - Whether the main conflict has been addressed
   - Whether the story feels satisfying and complete
5. Call submit_chapter with:
   - chapter_text: Your full chapter
   - is_final_chapter: TRUE if this completes the story, FALSE if more chapters are needed

Pacing guidelines:

- Let the story develop naturally -- don't rush to the end
- Ensure proper pacing according to the archetype description above
- The story should feel complete when you mark it as final, not abrupt

Write ONLY narrative content, no meta-commentary.

Knowledge-graph context block (`<kg_context>`) When entities have been extracted from previous chapters the following structured block is appended to the system prompt immediately after the story text:

KNOWLEDGE GRAPH (use this to maintain consistency):

Introduced in Chapter 1 (foundational): *<character/organisation nodes and relationships from chapter 1>*

Introduced in Chapter *<M>* (milestone): *<nodes and relationships from milestone chapter M>*

New in Chapter *<N-k>*: *<nodes and relationships from recent chapter N-k>*

All Characters & Organisations (cumulative): *<all character and organisation nodes accumulated so far>*

All Character Relationships (cumulative): *<all edges involving at least one character/organisation node>*

Other Entities: *<non-character nodes>*

Other Relationships: *<non-character edges>*

Milestone chapters are every fifth chapter starting at chapter 6 (i.e. 6, 11, 16, ...). Recent chapters are the three chapters immediately preceding chapter N (i.e. $N-3$, $N-2$, $N-1$), excluding chapter 1 and milestone chapters.

User prompt

Write Chapter *<N>* and decide if the story should continue.

A.1.3 Grapher Agent

System prompt The system prompt is assembled from a fixed instruction block and an optional canonical-entity block derived from the current store contents.

Canonical-entity block (if prior entities exist)

EXISTING CANONICAL ENTITIES (you MUST reuse these exact IDs and labels --- do not create duplicates): *<id="..." label="..." type="..." for each stored entity>*

EXISTING RELATIONSHIPS (re-include any that are still active in the story --- do not drop established connections): *<from -> to [label] for each stored relationship>*

Fixed instruction block

You are the KNOWLEDGE-GRAPH AGENT.

Your job is to extract entities and relationships from story chapters while keeping the knowledge graph consistent across chapters. *<canonical_block>*

RECONCILIATION RULES (critical):

- If a name in the chapter refers to the same person/place/thing as an existing canonical entity, use the EXISTING id and label -- do not invent a new one.
- Only create a NEW entity node if it is genuinely not present in the canonical list above.
- Characters who appear under nicknames, shortened names, or slight variations must be mapped to the matching canonical entity.

For nodes, include: id, label, and type (must be one of: character, event, object, location, organization)
For edges, include: from, to, and label fields.

RELATIONSHIP EXTRACTION PRIORITY:

- Character-to-character relationships are the most valuable -- extract these first and as completely as possible
- Then character-to-location and character-to-organisation relationships
- Character-to-event and character-to-object edges are lowest priority

IMPORTANT ID FORMATTING RULES:

- Node IDs must use underscores instead of spaces (e.g. High_School, Detective_Smith)
- Edge from and to fields must exactly match the node IDs you define (same underscore format)

Categories are strictly: character, event, object, location, organization.

User prompt

Extract entities and relationships from this chapter:
<chapter_text>

A.2 Starting Prompts

Each story is initialized with one of the following ten starting premises, selected at random at runtime:

1. A renowned detective receives an anonymous letter claiming that a murder will occur in exactly 48 hours.
2. A small coastal town's lighthouse keeper is found dead, and the only clue is a series of mysterious light patterns.
3. During a chess tournament, the reigning champion collapses dead mid-game, poisoned by a rare toxin.
4. A young apprentice discovers they can see magical creatures that no one else can perceive.
5. In a world where dragons have been extinct for centuries, a farmer finds a living dragon egg in their field.
6. A colony ship arrives at a distant planet only to find it already inhabited by humans who left Earth centuries later.
7. An AI developed to predict crime begins warning about murders that haven't been planned yet.
8. A family moves into their dream home, only to discover the previous owners never actually left.
9. Two rival wedding planners are forced to collaborate on the biggest celebrity wedding of the year.
10. A treasure map tattooed on a dying sailor leads to an island that doesn't appear on any modern charts.

JEST: A Benchmark for Rating the Funniness of Short Texts

Joe Toplyn
Semantic AI and Creativity Lab
East Texas A&M University
Commerce, Texas 75428 USA
joetoplyn@twentylanemedia.com

Ori Amir[1]
Psychology Department, Fulbright University Vietnam
Semantic AI and Creativity Lab, East Texas A&M University
Commerce, Texas 75428 USA
oriacadem@gmail.com

Abstract

A conversational AI system is perceived as more humanlike when it exhibits a sense of humor. To achieve this, the system must not only recognize humorous conversational responses, but also assess how funny a typical human would find them, enabling appropriate reactions. While prior work has advanced binary humor detection in short texts, no benchmark exists for evaluating funniness on a numerical scale in conversational contexts. In this paper we introduce JEST, a benchmark derived from 80 short conversational texts generated using AI and annotated with funniness ratings by seven expert joke writers. To ensure consistency and a balanced distribution across funniness categories, we computed final ratings using the three raters whose individual ratings were most evenly distributed. JEST achieves high inter-rater reliability (ICC(2,k) = .87), with texts divided into four roughly equal categories: 0 (Not Funny) to 3 (Very Funny). As an illustrative application, JEST is used to provide evidence that an off-the-shelf LLM can rate the perceived funniness of a dataset of short, conversational texts in substantial agreement with a panel of human experts. JEST facilitates the evaluation of automated funniness rating systems, thereby supporting the development of effective automated joke generation systems.

Introduction

The ability to mimic a sense of humor is an important trait in effective conversational AI systems (Ceha et al. 2021; Clark et al. 2019; Kang et al. 2017; Mirowski, Mathewson and Branch 2025; Nijholt et al. 2017; Sun et al. 2024). For a conversational AI system to do that, it must have the ability to both automatically generate jokes and to automatically recognize them (Inoue, Lala and Kawahara 2022; Inoue et al. 2025; Quan, Ramakrishnan and Chin 2025; Winters 2021). And if the AI system recognizes humorous text, it needs to be able to rate it for funniness. Funniness ratings are essential because they determine what type and degree of humor response to provide and have been shown to correlate strongly with laughter intensity. (McKeown and Curran 2015; Mittal et al. 2021).

To evaluate the performance of automated funniness raters intended for use in a conversational AI system, a benchmark against which the raters can be tested is necessary. But to the best of our knowledge, no benchmark exists that incorporates short texts similar to conversational turns along with relatively robust measures of how funny they are. To address that research gap, we introduce in this paper a benchmark incorporating short, two-part texts in English with funniness ratings assigned by expert human joke writers. We name it the Joke Evaluation by Skilled Testers (JEST) benchmark.

Related Work

There are many humor datasets consisting of short texts (Kalloniatis and Adamidis 2024; Lemmens and De Marez 2025). But few of the texts in those datasets are similar in structure to a conversational turn in English. For example, Bogireddy, Suresh and Rai (2023) present a dataset of humorous textual posts related to COVID-19 from the subreddits r/Jokes, r/CoronavirusMemes, r/Onion headlines, and r/NottheOnion headlines. Bower and Steyvers (2021) use 60 jokes scraped from the subreddit r/Jokes. Faruqi and Shrivastava (2018) use a dataset of 400,000 sentences extracted from social media and humor-dedicated websites,

[1] Corresponding author

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

in addition to negative samples such as news headlines and proverbs. Kiddon and Brun (2011) study a dataset of "That's what she said" jokes. Mihalcea and Strapparava (2005) use a dataset of short, humorous sentences–"one-liners"–and non-humorous data such as news headlines and proverbs. Murakami et al. (2025) base a benchmark on a dataset derived from the Japanese creative response game Oogiri and the funniness votes of game participants. Potash, Romanov and Rumshisky (2017) introduce a dataset composed of humorous tweets that could have a given hashtag. Yang, Hooshmand and Hirschberg (2021) present a dataset of Facebook posts related to COVID-19, each post with a humor score derived from its reaction emojis. And Zangari et al. (2025) use annotated sets of puns and non-puns.

In addition, few humor datasets of short texts are annotated with ratings of funniness, because they are usually intended for use in binary classification tasks. And of those datasets that are annotated with funniness ratings, few have those ratings assigned by comedy experts. Potash, Romanov and Rumshisky (2017) do rely on the expert judgments of a television show's staff to measure the funniness of tweets. But Avetisyan, Safikhani and Broneske (2023) use only two human evaluators of indeterminate expertise to assign funniness ratings to short, humorous texts, while Chiruzzo, Castro and Rosá (2020) use crowdsourced annotators who visited a particular web page to rate humorous Spanish tweets for funniness. Other examples include Hossain, Krumm and Gamon (2019), who employ non-expert researchers and Amazon Mechanical Turk workers as judges to rate edited news headlines for funniness, and Weller and Seppi (2020), who use the upvotes of presumably non-expert Reddit users to measure the funniness of humorous texts from the r/Jokes subreddit.

Having identified the absence of a benchmark of short texts similar to conversational turns along with relatively robust funniness measures, in particular funniness ratings by experts, we built it as described in the next section.

Building the Benchmark

Constructing JEST took place in stages: designing the benchmark, creating a dataset of short texts composed of setups and responses, and obtaining expert ratings of the funniness of each short text.

Designing the Benchmark

In designing the benchmark, we made the simplifying decision that it should be useful for evaluating automated funniness raters in conversational AI systems intended for typical American adults. We also made the important observation that a standalone, "canned" joke consisting of an unfunny setup sentence followed by a punch line is equivalent, from the standpoint of joke mechanics, to a conversational turn consisting of an utterance that is not funny followed by a funny response (Toplyn 2014). Those factors led us to conclude that the benchmark would be based on a dataset of short texts in English, each of which would consist of (1) an unfunny sentence that could plausibly function as either an utterance in a conversation or the setup of a joke, followed by (2) a sentence that could function as either a funny response to the utterance or a punch line of the joke. The short texts would vary in funniness and each would be accompanied by a numerical rating reflecting how funny typical American adults would be likely to find it. The performance of a funniness-rating system against the benchmark would be evaluated by measuring the degree of agreement between the system's funniness ratings for the 80 texts and the benchmark ratings.

Creating the Short Texts

Because of the limitations of existing datasets, described above, we decided to create a completely new dataset of 80 short texts for this paper. We believed that this number was large enough to allow statistically significant conclusions to be drawn, but not so large that a human who is rating all the texts in one session would experience fatigue, resulting in lower quality ratings.

The following process was used to create short texts that may be funny to some degree and consist of a setup sentence followed by a response sentence.

Generating the Setups

To serve as unfunny setup sentences, we considered extracting headlines from a news dataset (Faruqi and Shrivastava 2018; Mihalcea and Strapparava 2005). But news headlines might result in jokes that become dated and less funny over time, and therefore seemed suboptimal for a benchmark that we intended to have a relatively long useful life. We also considered selecting utterances from an existing conversational dataset (Toplyn 2023). But the datasets we found did not include enough utterances that we believed might plausibly appear in an everyday conversation and also might plausibly elicit a funny response from a companion. In addition, existing conversational datasets are limited in size, and we wanted to identify a source of setup sentences that is potentially unlimited so as to facilitate future research.

So, we generated setup sentences with OpenAI's ChatGPT, powered by the default GPT-5 model that it was using on September 15, 2025 (Singh et al. 2025). This entailed an iterative prompt-engineering dialogue during which we guided the output based on repeated evaluations of its style, tone, topic, factuality, plausibility, and naturalness. This prompt reflects the cumulative refinement of the instructions that produced the setups selected for the benchmark:

> Generate short, conversational setups that sound like something a friend might casually mention after skimming the news or noticing a pop-culture or everyday-life trend. Keep them factually true if they involve facts and fully plausible if styled like light news items. Make them evergreen, interesting enough to invite a punch line, and natural enough for casual conversation. Focus on everyday-life observations, slice-of-life community quirks, social-media and

> lifestyle trends, and trendy but credible pop-culture stories, while omitting quirky true facts. Keep most setups to a single short sentence, with a few slightly longer but uncluttered variations for variety, and avoid multi-clause structures using "and."

This method of generating setups is noteworthy because each resulting setup contains a pair of discrete and semantically distant concepts. Research has shown that this property tends to produce original jokes—ones that are qualitatively distinct from those obtained by simply prompting an LLM to generate humor (Amir 2025; Amir et al. 2026; Veale 2024).

From the 280 potential setups generated using this method, Authors JT and OA used their expert judgment as experienced comedy practitioners to select 80 setups that they agreed encompassed a wide variety of subject matter and also resembled plausible conversational utterances that might induce a witty companion to respond with a joke.

Generating the Responses

Once the setups were selected, each was paired with a response crafted to complete a short text that could be rated as either 0 (Not Funny), 1 (Somewhat Funny), 2 (Funny), or 3 (Very Funny). This four-point scale, similar to that used by Hossain, Krumm and Gamon (2019), was chosen because it seemed to span a sufficiently broad range of perceived funniness without demanding unrealistically precise predictions of perceived funniness. In implementations, if a conversational AI system equipped with a funniness rating module rated a human's response to its utterance as 0, the system would not exhibit any amusement behavior. But as the AI system rated the human's responses as increasingly funny, it would react appropriately with, say, a smile, a chuckle, or a laugh (Inoue, Lala and Kawahara 2022).

Two different methods were employed to generate the responses to the setups: one that was likely to yield funniness ratings of 1, 2, or 3; and a different one that would probably yield a rating of 0. Here are the two methods:

1: Somewhat Funny; 2: Funny; and 3: Very Funny–All of the responses likely to yield these ratings were produced by JT using Witscript, a neural-symbolic hybrid AI system that generates short jokes (Toplyn 2023). Witscript uses a large language model (LLM) to execute joke-writing algorithms created by a human expert. The version of Witscript employed here was the one publicly available as an app on October 28, 2025 [https://witscript.com/]. Using Witscript to generate these responses was orders of magnitude faster and less expensive than paying professional comedy experts to write them would have been.

0: Not Funny–All of the responses likely to yield this rating were generated on November 14, 2025, by OpenAI's ChatGPT, powered by its default model, GPT-5.1. This prompt was used to obtain the "Not Funny" responses:

> You are a typical American adult. Treat the given text as the first sentence of a news story or interesting anecdote. Write a plausible second sentence that continues the story in a conversational, realistic way. The sentence should be approximately 11–15 words long. You can describe the situation, add context, comment on people's reactions, or include your own reaction or opinion.

This prompt elicited responses comparable in length to those generated by Witscript, thereby eliminating the chance that a funniness rating system could identify Not Funny texts in the benchmark by their length. This method of generating unfunny responses is also in the spirit of Winters and Delobelle (2020), who demonstrate the advantages of using an automated method to generate non-jokes that are close in structure and vocabulary to jokes; this approach is preferable to using non-jokes from a completely different domain, such as a dataset of real news headlines, when evaluating humor recognition systems (Annamoradnejad and Zoghi 2024; Inácio, Wick-Pedro and Oliveira 2023).

Each setup and response produced by the above methods was concatenated to form a short text. Using Witscript and ChatGPT to build the short texts in this way minimizes the chances that any of the completed texts had previously appeared online and that an AI-powered funniness-rating system had been trained on them. This allows the benchmark to be used for relatively unbiased testing.

Applying those two generation methods to the 80 setups, JT generated 91 short texts possessing what he perceived to be widely varying levels of funniness, which he and OA then rated independently. After sharing and discussing their ratings, JT substituted a few responses, after which the authors used their expert judgment to cherry-pick 80 texts that they agreed would result in expert human raters assigning approximately 25% of the total number to each of the four rating categories.

Rating the Short Texts

Having prepared the dataset of 80 short texts, we turned to the question of how to assign them robust funniness ratings. As noted above, the short texts, while conversational, also have the same form as standalone setup/punch line jokes. Therefore, delivering the texts orally in front of large, live audiences of typical American adults and measuring the resulting laughter, if any, would be the best way to assign the ratings to them (Mirowski, Mathewson and Branch 2025; Mittal et al. 2021; Romanowski, Valois and Fukui 2025; Toplyn and Amir 2025). But this method was ruled out for practical and financial reasons: obtaining a statistically significant measurement of laughter, and therefore funniness, for each of the 80 texts would mean having them all delivered by various performers in different orders to many audiences consisting of hundreds of people.

Measuring the laughter on preexisting recordings of stand-up comedy performances would also fall short because of selection bias: comics only deliver jokes to live audiences that they expect to get laughs. Therefore, those

recorded jokes do not represent the full range of funniness–from very funny to not funny at all–that is required in the desired benchmark.

An alternative to measuring laughter would be to have human evaluators assign funniness ratings to the texts. Humor datasets exist that have been annotated by non-experts, usually crowdsourced, and by experts (Kalloniatis and Adamidis 2024). We concluded that ratings assigned by non-experts would be unreliable because research indicates that non-expert humans tend to be poor judges of creative works (Lamb, Brown and Clarke 2015), open-ended generated text (Karpinska, Akoury and Iyyer 2021), and transcripts of standup comedy (Romanowski, Valois and Fukui 2025). This conclusion is consistent with neuroimaging evidence suggesting that years of professional comedy experience are associated with measurable anatomical differences in brain structure and with distinct patterns of neural activity when processing humor (Amir and Biederman 2016; Brawer and Amir 2021). In addition, the evaluation of humor is greatly affected by the world knowledge, ideological biases, and other characteristics of the evaluators, who are often recruited using crowdsourcing (Loakman, Maladry and Lin 2023). For those reasons, we hypothesized that non-experts may not be able to accurately predict how typical American adults would react to the short texts. That is a particularly important consideration, because what is relevant to developing a benchmark that represents the expected funniness perceived by typical American adults is not what the human evaluators themselves perceive as funny, but what they believe typical American adults would perceive as funny.

Therefore, instead of non-experts, we had the texts rated for funniness by seven expert joke writers. The writers collectively have a total of over 120 years of professional experience writing jokes for television shows in the U.S. This experience qualifies them to provide benchmark funniness ratings for the texts for these reasons: (1) each of those experts has been paid to write, edit, and informally rank for funniness thousands of texts of the type in the benchmark, i.e., short, self-contained texts intended to elicit laughter from the millions of typical American adults viewing the shows at home; and (2) each of those experts has absorbed ground-truth information about the funniness of those texts by listening to hundreds of live studio audiences, each consisting of hundreds of typical American adults, reacting to a live host delivering the texts, typically during a comedy monologue.

Text	Consensus Rating
Parents are pushing for healthier food in school cafeterias. Makes sense—kids spend so much time at school, they might as well eat better.	0
The city is considering turning part of the old train station into an art gallery. Finally, commuters can miss their train while staring at an abstract departure board.	1
A clothing brand just launched a line of pajamas designed to look like office wear. So now you can get fired and go straight back to bed without changing.	2
Couples who laugh together are more likely to stay together longer. Unless the laughter is because one of them just fell off the roof.	3

Table 1: Examples of texts in the JEST benchmark that have ratings on which all three human expert raters agreed.

To prevent bias, the human experts were not told where the texts came from; they were only told that their ratings would be used to compare how jokes are rated by experts and non-experts. The texts were presented to the experts in a spreadsheet, with the same random ordering for each expert, along with these instructions:

> Your task is to rate how funny a given piece of text would be to an audience of typical American adults. Use this scale:
> 0 – Not Funny: No humor present. Likely to get no reaction.
> 1 – Somewhat Funny: Mildly amusing or clever. Almost a joke. Might get a smile.
> 2 – Funny: Clearly a joke. Likely to get a chuckle.
> 3 – Very Funny: A great joke. Likely to get a laugh.

The experts, who were paid $50 each, rated the texts independently and never had access to each other's ratings. See Table 1 for an example of a text in each funniness rating category.

Results and Discussion

To derive the final benchmark ratings, we selected a subset of three raters, out of the seven experts, whose individual rating distributions were the most balanced across the four funniness categories (0–3). This approach of increasing the influence of the most consistent or balanced raters (or equivalently, selecting a subset of them) has been recommended in methodological work on annotation quality (Dawid and Skene 1979; Krippendorff 2018; Paun et al. 2018; Snow et al. 2008). The selection served two purposes: (1) it enhanced the overall consistency of the gold-standard ratings while ensuring a more even spread of ratings across the full range of funniness levels; and (2) because most conversational agents aim to be agreeable and engaging, a somewhat more generous response to user humor is preferable; selecting the higher-rating cluster of experts supports this goal without sacrificing reliability. Indeed, even when aggregating all seven raters, the relative funniness ranking of the texts remains highly similar, differing primarily in an overall downward shift. Full details of the selection of raters for the final benchmark, individual rating distributions, and resulting improvements in balance and reliability follow.

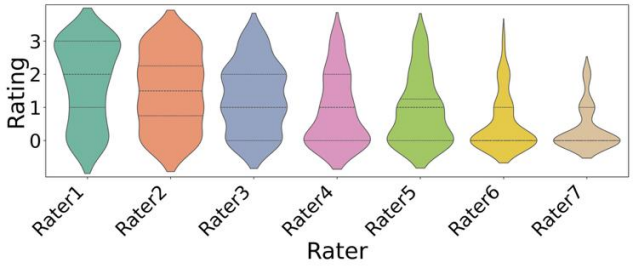


Figure 1: Violin plots displaying the distribution of funniness ratings (0–3 scale) for each of the seven expert raters across 80 texts. The plots highlight variability in rating tendencies: Raters 1-3 tend toward higher ratings and approximating an even spread across categories, while Raters 4-7 are notably conservative. Inner quartiles indicate the interquartile range.

Figure 1 shows the variability in the rating tendencies of the seven expert raters. Two raters were markedly stricter (means of 0.51 and 0.35) than the other raters (means of 0.90–1.74). Such rater clustering is common in subjective annotation tasks (Dethlefs et al. 2014) and, in particular, in humor (Ruch 1992). Notably, comedians are widely recognized within their own community as being particularly harsh critics and demanding audiences (Wilson 2014), a bias that developers of conversational agents may wish to avoid replicating.

This clustering of raters is also evident in Figure 2, which details pairwise Spearman correlations among the ratings of the seven expert raters. A core subset of three raters (anonymized as Rater1, Rater2, and Rater3) exhibits the highest correlations.

Table 2 shows the key inter-rater reliability and agreement metrics for all seven experts and for the core subset of three experts (Raters 1-3) whose individual rating distributions were the most balanced and consistent across the four funniness categories. Exact Agreement means

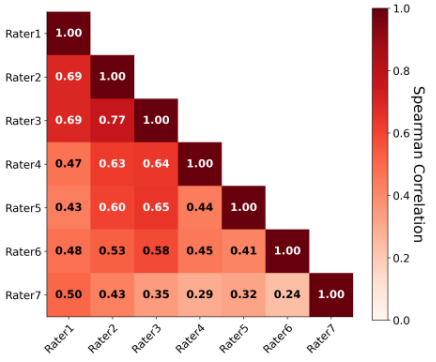


Figure 2: Pairwise Spearman correlations among the seven expert raters' text ratings, with correlations ranging from 0.24 to 0.77. The matrix reveals moderate overall agreement, with the core trio (Raters 1-3) exhibiting the highest correlations (>0.70), while Raters 6 and 7 show weaker alignment with the group, seemingly due to their conservative bias and a floor effect.

Metric	7 raters	3 raters
ICC(2,k)	0.82	0.87
Krippendorff's α (ordinal)	0.40	0.68
Exact Agreement (%)	20.0	41.3
Adjacent Agreement (%)	35.0	77.5
Mean Per-Item SD	0.74	0.45
Avg. Pairwise Spearman ρ	0.50	0.72

Table 2: Inter-rater reliability and agreement comparison (80 items, ordinal 0–3 scale).

identical ratings from the relevant raters, while Adjacent Agreement means ratings that differ by at most one point.

ICC(2,k) values indicate high reliability of the aggregated ratings in both configurations (0.82–0.87). Krippendorff's α was substantially higher for the three-rater subset (0.68 vs. 0.40), reflecting greater ordinal consistency. This improvement stems from the systematic differences in rating severity among the full set of seven experts. We consider ICC(2,k) to be the primary metric because it captures agreement on the actual ratings rather than merely their order, which is important here because each rating corresponds to a specific level of humor reaction from an AI agent. Additionally, ICC(2,k) is well-suited to our crossed design, in which all raters evaluated all items.

Selecting the subset of three raters with the most balanced individual distributions yielded sharper consensus: mean per-item standard deviation decreased from 0.74 to 0.45, exact agreement increased from 20% to 41%, and adjacent agreement rose from 35% to 78%.

Average pairwise Spearman correlation also improved substantially from 0.50 (all seven raters) to 0.72 (subset of three raters), indicating stronger rank-order consistency within that subset. Figure 3 provides a graphical illustration of the high degree of agreement in the ratings given by the subset of three raters.

In addition to possessing high reliability, we also wanted the benchmark to incorporate ratings distributed as evenly



Figure 3: Violin plot showing the number of individual ratings in each category that were assigned to texts by the three selected raters, conditioned on the ratings that were assigned to those texts by the majority of the three selected raters. The plot illustrates how individual ratings cluster around consensus labels, with tighter distributions in smaller subsets indicating higher agreement.

as possible across the full 0–3 funniness range, to support fair evaluation of automated funniness raters at all levels of perceived funniness. With this additional objective in mind, we examined the majority-vote distributions (with ties resolved toward the higher rating) for all seven raters and for the three-rater subset:

- All seven raters: 41% Not Funny, 20% Somewhat Funny, 21% Funny, 18% Very Funny
- Three-rater subset: 25% Not Funny, 20% Somewhat Funny, 20% Funny, 35% Very Funny

The distributions showed that the three-rater subset provided the closest approximation to equal quarters while avoiding the pronounced skew toward Not Funny observed with all seven raters included. Although we selected only three out of the seven raters for the official JEST benchmark, mean ratings between the two configurations (average ratings of the seven vs. three raters) correlated extremely highly across texts (Spearman $\rho = .95$, $p < .001$), confirming that relative funniness rankings remained virtually unchanged despite subset selection.

For the above reasons, the official JEST benchmark defines its reference ratings as the median rating given by the subset of three selected experts (Rater1, Rater2, Rater3) for each of the 80 texts. We used the median instead of the mean because the 0–3 scale is coarse and the median is more robust to occasional outlier ratings and better reflects consensus. This three-rater configuration yields ICC(2,k) = .87, a substantial agreement for such a highly subjective task.

Illustrative Application

To demonstrate the utility of the JEST benchmark, we propose this hypothetical scenario: The XYZ Company is developing an artificial companion for typical American adults and wants to give its conversational AI system the ability to deliver an appropriately scaled humor reaction when a user responds to one of its utterances with a joke.

So, the company sets out to identify the best off-the-shelf LLM to power a funniness-rating module that can be deployed locally in its artificial companion so as to deliver low-latency, humanlike humor reactions to the user at reasonable cost. For business reasons, the company wants to use an LLM in OpenAI's GPT-5 family of models.

Model selection

To assist the hypothetical company, we evaluated models that span the range of capability of the GPT-5 family, from lightweight variants (GPT-5-nano and GPT-5-mini) to the core generation (GPT-5 and its incremental updates GPT-5.1 and GPT-5.3) and the high-capability GPT-5-pro model. By "capability" we mean a broad measure of a model's overall performance across a variety of tasks. We also included a coding-optimized model, GPT-5-codex, to see whether a model optimized for structured reasoning tasks would perform better than general-purpose models.

The GPT-5 family models evaluated in this study were all released following the official GPT-5 launch period beginning in August 2025.

Evaluation procedure

When evaluated as funniness raters, the LLMs were used off-the-shelf and were not prompted with any examples or any instructions related to humor theories. Instead, all of the above LLMs were given the following prompt.

> System message:
> You are evaluating short texts for their potential to be funny to a typical American adult audience.
>
> Use the following 0–3 scale:
> 0 = not funny (likely to get no reaction)
> 1 = almost funny (might get a smile)
> 2 = funny (likely to get a chuckle)
> 3 = very funny (likely to get a laugh)
>
> User message:
> Text: {{text}}
> Output only the numeric rating.

For evaluation, all LLMs were accessed via API in early March 2026. The performance of each LLM was evaluated by measuring the degree of agreement between the LLM's funniness ratings and the expert ratings across the 80 texts in the JEST benchmark.

Unlike the inter-rater reliability analysis above, where ICC(2,k) and Krippendorff's α were appropriate for multiple comparable raters evaluating the same items, the two-way ICC framework is less suitable for the analysis here because it assumes symmetric, exchangeable raters drawn from the same population. That assumption does not hold when one "rater" is a carefully constructed human expert consensus benchmark and the other is an LLM operating via a fundamentally different mechanism.

We therefore treat JEST as a fixed reference standard and measure each LLM's agreement against it using quadratic weighted Cohen's κ as the primary metric, which accounts for the ordinal structure of the 0–3 scale, penalizes larger disagreements more heavily than smaller ones, and is standard practice in NLP for evaluating system outputs against a gold-standard reference. We additionally report Spearman's ρ to capture rank-order consistency. Unweighted κ was not used because it treats all disagreements as equivalent regardless of magnitude.

Performance assessment

Table 3 shows how the seven LLMs compare on key characteristics and agreement against the JEST benchmark. According to the widely used scale established by Landis and Koch (1977), κ values above .60 indicate substantial agreement with a reference standard. Using that guideline, most of the evaluated GPT-5 family models demonstrated substantial agreement with the benchmark. But the least-capable models, GPT-5-mini ($\kappa = .58$) and GPT-5-nano ($\kappa = .29$), did not reach that threshold. This result is consistent

Model	Capability	Speed	Cost Eff.	Cohen's κ	Spearman's ρ
gpt-5-pro	5	2	2	.66*** [.53, .76]	.66*** [.49, .79]
gpt-5.3	5	3	3	.73*** [.63, .80]	.77*** [.66, .85]
gpt-5.1	4	3	3	.74*** [.64, .81]	.78*** [.68, .85]
gpt-5	4	3	3	.73*** [.63, .80]	.76*** [.64, .84]
gpt-5-codex	4	3	3	.69*** [.56, .79]	.69*** [.51, .82]
gpt-5-mini	3	4	4	.58*** [.46, .66]	.73*** [.60, .82]
gpt-5-nano	2	5	5	.29*** [.17, .41]	.52*** [.37, .65]

Table 3. Comparison of the LLMs, with the agreement between their funniness ratings and the JEST benchmark as measured by quadratic weighted Cohen's κ and Spearman's ρ. 95% confidence intervals are in square brackets, with bootstrap p-values all ***: $p < .001$. Relative qualitative ratings of Capability, Speed, and Cost Efficiency were provided by OpenAI in early March 2026; 5 = highest in model group, 1 = lowest.

with Loakman, Thorne and Lin (2025), who found that smaller LLMs (e.g., Gemini Flash and GPT-4-mini) underperformed their full-size counterparts on a joke-explanation task. They attributed this at least partly to the smaller models' failure to retrieve the cultural references that jokes rely on, possibly because such references were underrepresented in their training data. It is also possible that smaller models do not have the ability to perform the deep semantic analysis required for detecting humor structures in text.

The remaining five models all exceeded the .60 threshold, suggesting that larger models generally have the contextual understanding needed to approximate expert judgments of funniness. Among them, GPT-5.1 achieved the highest agreement ($\kappa = .74$), followed closely by GPT-5 and GPT-5.3 (both $\kappa = .73$). The specialized GPT-5-codex ($\kappa = .69$) trailed those top performers slightly, which may reflect its optimization for structured coding tasks instead of conversational funniness evaluation. More surprisingly, GPT-5-pro ($\kappa = .66$) also performed somewhat below the three leaders. Narad et al. (2025) found that post-training models such as DeepSeek R1 on general STEM reasoning tasks improved their ability to rate and explain humorous New Yorker cartoon captions. However, the models they tested predate the GPT-5 series, which already incorporates extensive reasoning-focused post-training. It is possible that the heavier reasoning emphasis of GPT-5-pro may encourage it to "overthink the joke," rather than mimic the fast, associative behavior that humans exhibit when deciding whether some text is funny.

Table 3 also shows that Spearman correlations were similarly strong across all the models except for the least-capable, GPT-5-nano, with substantially overlapping confidence intervals. This suggests consistent ordinal agreement on joke funniness regardless of model capability.

Taken together, these results suggest that GPT-5.1, with the highest κ, offers the best overall balance of agreement with expert funniness ratings while still being speedy and cost effective enough for deployment. Therefore, the XYZ Company should consider GPT-5.1 to be its top candidate for powering the funniness rating module in its artificial companion. However, if later testing reveals that GPT-5.1 does not meet the required speed or cost constraints, GPT-5-mini may be a viable alternative: it's faster and more economical than GPT-5.1 while still performing close enough to the substantial-agreement threshold to plausibly "get" a joke the way a typical American adult would.

Contributions

This paper makes the following contributions:

1. It introduces the first benchmark for evaluating funniness rating systems, comprising a dataset of 80 short texts annotated with non-binary funniness ratings by human experts, along with a primary evaluation metric: the quadratic weighted Cohen's κ. The benchmark data will be made available by the corresponding author, upon request, to legitimate researchers who agree not to redistribute it publicly, in order to prevent its use as training data and preserve its intended use as an evaluation resource. The benchmark data consists of two datasets. Dataset A presents the official JEST benchmark data: the 80 short texts, including setups and responses, ordered by descending median funniness rating (computed from the ratings of the three selected raters whose ratings were most evenly distributed across the four funniness categories), together with the median rating, exact and adjacent agreement percentages, and mean absolute deviation for each text. Dataset B provides, for transparency, the complete data from all seven expert raters, including medians, agreement metrics, descriptive statistics, and individual ratings for each text.
2. It presents a speedy, economical method of using an LLM and an automated joke generator to synthesize an unlimited number of short texts with varying degrees of funniness. Those texts can greatly increase the amount of data available for humor research.
3. It provides evidence that an off-the-shelf LLM can rate the perceived funniness of a dataset of short, conversational texts in substantial agreement with a panel of human experts.

Conclusion

The JEST benchmark is a resource that can be used to evaluate and advance the development of automated funniness-rating systems that perceive the funniness of short texts as typical American adults would. Such systems could be used to rapidly and economically evaluate the output of automated joke generation systems, thereby facilitating the development of the latter.

An effective funniness-rating system would also enable a conversational AI system to judge how funny a user's utterance is and respond with an appropriately scaled humor reaction. Paired with an automated joke generator, an automated funniness rater could enable a conversational AI, like a chatbot, to strike typical humans as having a humanlike sense of humor, thus making the chatbot more likeable and enhancing its interactions with users.

In the future, we will use the JEST benchmark to evaluate a wide range of additional methods for rating perceived funniness, including numerical ratings by crowdsourced, non-expert evaluators; by both naive and humor-theory-specific LLMs from multiple providers; by an ensemble of LLMs-as-judges; and by a novel, neural-symbolic, hybrid AI system. Our goal is to identify the most effective system for automatically rating the perceived funniness of a two-part text similar to a turn in a conversation.

Limitations

In this section we list limitations of this work and of the JEST benchmark it presents.

1. The utility of the benchmark is limited to funniness rating systems designed for short, conversational texts and typical American adults. It may be of little use in evaluating rating systems designed for other languages, cultures, demographics, socio-economic contexts, personality types, or forms of humor. Acknowledging that humor is highly subjective and context-dependent, we make no claim that JEST is a benchmark of funniness as perceived by everyone everywhere at all times.
2. This work treats the aggregated expert judgments in the JEST benchmark as a practical reference standard for funniness as perceived by a substantial segment of the U.S. adult population, but such judgments do not constitute an objective ground truth.
3. The benchmark is based on a relatively small number of texts (80), reflecting a tradeoff between the size of the dataset and the potential fatigue of the human expert raters.
4. The benchmark incorporates the ratings of only three experts. As we have shown in this paper, given the subjective nature of humor, a different group of experts might arrive at somewhat different consensus ratings.
5. The benchmark is based on the ratings of expert joke writers who judged the texts on their potential to amuse a broad U.S. audience; as a result, conversational jokes that rely on shared context, i.e., "inside jokes," may not be well-represented.
6. While the benchmark is based on texts designed to be relatively evergreen, evolving cultural references or shifts in humor tastes over time could affect its long-term utility.
7. Finally, the benchmark is intended solely as a resource for evaluating funniness rating systems rather than as a training corpus, which may limit its usability for some applications.

Acknowledgments

We would like to thank the expert joke writers who rated the texts in the JEST benchmark.

References

Amir, O. 2025. Are AI-generated jokes truly original? Charting the "Joke Space." In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*. Campinas, Brazil. Association for Computational Creativity.

Amir, O., and Biederman, I. 2016. The neural correlates of humor creativity. *Frontiers in Human Neuroscience*, 10:597.

Amir, O.; Toplyn, J.; Ngo, H.D.; and Hickerson, K.P. 2026 (in press). Navigating the Joke Space: Towards automated originality assessment of AI-generated humor. In *Proceedings of the 2nd Workshop on Computational Humor (CHum)*, Online. Association for Computational Linguistics.

Annamoradnejad, I., and Zoghi, G. 2024. ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249, PB.

Avetisyan, H.; Safikhani, P.; and Broneske, D. 2023. Laughing out loud – Exploring AI-generated and human-generated humor. In *Proceedings of the 4th International Conference on NLP Artificial Intelligence Techniques (NLAI 2023)*.

Bogireddy, N.R.; Suresh, S.; and Rai, S. 2023. I'm out of breath from laughing! I think? A dataset of Covid-19 humor and its toxic variants. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*, 1004–1013, New York, NY, USA. Association for Computing Machinery.

Bower, A.H., and Steyvers, M. 2021. Perceptions of AI engaging in human expression. *Scientific Reports*, 11:21181.

Brawer, J., and Amir, O. 2021. Mapping the 'funny bone': Neuroanatomical correlates of humor creativity in professional comedians. *Social Cognitive and Affective Neuroscience*, 16(9): 915-925.

Ceha, J.; Lee, K.; Nilsen, E.S.; Goh, J.; and Law, E. 2021. Can a humorous conversational agent enhance learning experience and outcomes? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1–14, New York, NY, USA. Association for Computing Machinery.

Chiruzzo, L.; Castro, S.; and Rosá, A. 2020. HAHA 2019 dataset: A corpus for humor analysis in Spanish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 5106–5112. Marseille, France. European Language Resources Association.

Clark, L.; Pantidi, N.; Cooney, O.; Doyle, P.; Garaialde, D.; Edwards, J.; Spillane, B.; Gilmartin, E.; Murad, C.; Munteanu, C.; Wade, V.; and Cowan, B.R. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–12. New York, New York, USA. Association for Computing Machinery.

Dawid, A.P., and Skene, A.M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Dethlefs, N.; Cuayáhuitl, H.; Hastie, H.; Rieser, V.; and Lemon, O. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *14th Conference of the European Chapter of the Association for Computational Linguistics 2014*, 702–711. Gothenburg, Sweden. Association for Computational Linguistics.

Faruqi, F., and Shrivastava, M. 2018. "Is this a joke?": A large humor classification dataset. In *Proceedings of the 15th International Conference on Natural Language Processing*, 104–109, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Hossain, N.; Krumm, J.; and Gamon, M. 2019. "President vows to cut hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Inácio, M.L.; Wick-Pedro, G; and Oliveira, H.G. 2023. What do humor classifiers learn? An attempt to explain humor recognition models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Inoue, K.; Elmers, M.; Lala, D.; and Kawahara, T. 2025. Why do we laugh? Annotation and taxonomy generation for laughable contexts in spontaneous text conversation. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, 318–323, Bilbao, Spain. Association for Computational Linguistics.

Inoue, K.; Lala, D.; and Kawahara, T. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*, 9:933261.

Kalloniatis, A., and Adamidis, P. 2024. Computational humor recognition: A systematic literature review. *Artificial Intelligence Review*, 58:43. Published online 2025.

Kang, S.; Krum, D.M.; Khooshabeh, P.; Phan, T.; Chang, C.; Amir, O.; and Lin, R. 2017. Social influence of humor in virtual human counselor's self-disclosure. *Computer Animation and Virtual Worlds*, 28.

Karpinska, M.; Akoury, N.; and Iyyer, M. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kiddon, C., and Brun, Y. 2011. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 89–94, Portland, Oregon, USA. Association for Computational Linguistics.

Krippendorff, K. 2018. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Lamb, C.; Brown, D.G.; and Clarke, C.L.A. 2015. Human competence in creativity evaluation. In *Proceedings of the Sixth International Conference on Computational Creativity*, 102-109. Park City, Utah, USA. Association for Computational Creativity.

Landis, J. R., and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lemmens, J., and De Marez, V. 2026. Computational humor modeling: A survey on the state of the art. *ACM Computing Surveys*, 58 (7), 1-37.

Loakman, T.; Maladry, A.; and Lin, C. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6676–6689, Singapore. Association for Computational Linguistics.

Loakman, T.; Thorne, W.; and Lin, C. 2025. Comparing apples to oranges: A dataset & analysis of LLM humour understanding from traditional puns to topical jokes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 9502–9518, Suzhou, China. Association for Computational Linguistics.

McKeown, G., and Curran, W. 2015. The relationship between laughter intensity and perceived humour. In *Proceedings of the 4th Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech,* 27–29, Enschede, Netherlands.

Mihalcea, R., and Strapparava, C. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Mirowski, P.; Mathewson, K.; and Branch, B. 2025. The theater stage as laboratory: Review of real-time comedy LLM systems for live performance. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, 88–95, Online. Association for Computational Linguistics.

Mittal, A.; Jeevan, P.; Gandhi, P.; Kanojia, D.; and Bhattacharyya, P. 2021. "So you think you're funny?": Rating the humour quotient in standup comedy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10073–10079, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Murakami, S.; Kamigaito, H.; Takamura, H.; and Okumura, M. 2025. Oogiri-master: Benchmarking humor understanding via Oogiri. *arXiv preprint arXiv:2512.21494*.

Narad, R.; Suresh, S.; Chen, J.; Dysart-Bricken, P.S.L.; Mankoff, B.; Nowak, R.; Zhang, J.; and Jain, L. 2025. Which LLMs get the joke? Probing non-STEM reasoning abilities with HumorBench. *arXiv preprint arXiv:2507.21476.*

Nijholt, A.; Niculescu, A.I.; Valitutti, A.; and Banchs, R.E. 2017. Humor in human-computer interaction: A short survey. In Joshi, A.; Balkrishan, D. K.; Dalvi, G.; and Winckler, M., eds., *Adjunct Proceedings INTERACT 2017 Mumbai: 16th IFIP TC.13 International Conference on Human Computer Interaction*, 192–214.

Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; and Poesio, M. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Potash, P.; Romanov, A.; and Rumshisky, A. 2017. Semeval-2017 task 6: Hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 49–57, Vancouver, Canada. Association for Computational Linguistics.

Quan, K.; Ramakrishnan, P.; and Chin, J. 2025. Can AI take a joke—or make one? A study of humor generation and recognition in LLMs. In *Proceedings of the 2025 Conference on Creativity and Cognition*, 431-437, New York, New York, USA. Association for Computing Machinery.

Romanowski, A.; Valois, P.H.V.; and Fukui, K. 2025. From punchlines to predictions: A metric to assess LLM performance in identifying humor in stand-up comedy. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 36–46, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Ruch, W. 1992. Assessment of appreciation of humor: Studies with the 3 WD humor test. *Advances in Personality Assessment*, 9, 27–75.

Singh, A.K. et al. 2025. OpenAI GPT-5 system card.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A.Y. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Sun, X.; Teljeur, I.; Li, Z.; and Bosch, J.A. 2024. Can a funny chatbot make a difference? Infusing humor into conversational agent for behavioral intervention. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*, 1–19, New York, New York, USA. Association for Computing Machinery.

Toplyn, J. 2014. *Comedy Writing for Late-Night TV*. Rye, New York, USA: Twenty Lane Media, LLC.

Toplyn, J. 2023. Witscript 3: A hybrid AI system for improvising jokes in a conversation. *arXiv preprint arXiv:2301.02695*.

Toplyn, J., and Amir, O. 2025. Can AI make us laugh? Comparing jokes generated by Witscript and a human expert. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, 71–78, Online. Association for Computational Linguistics.

Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*. Jönköping, Sweden. Association for Computational Creativity.

Weller, O., and Seppi, K. 2020. The rJokes dataset: A large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6136–6141, Marseille, France. European Language Resources Association.

Wilson, G. 2014. *The Complete Guide to Stand-Up: Everything You Need to Know, from Open-Mics to Going Pro!* Gregory D Wilson / The Comedy Institute.

Winters, T. 2021. Computers learning humor is no joke. *Harvard Data Science Review*, 3(2).

Winters, T., and Delobelle, P. 2020. Dutch humor detection by generating negative examples. In *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC 2020) and the 29th Belgian-Dutch Conference on Machine Learning (Benelearn 2020)*, 313–323.

Yang, Z.; Hooshmand, S.; and Hirschberg, J. 2021. Choral: Collecting humor reaction labels from millions of social media users. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4429–4435, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zangari, A.; Marcuzzo, M.; Albarelli, A.; Pilehvar, M.T.; and Camacho-Collados, J. 2025. Pun unintended: LLMs and the illusion of humor understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 27924–27959, Suzhou, China. Association for Computational Linguistics.

Generative Rules as Design Material: Constitutive Agency in Co-Creative Evolutionary Knit Design

Severi Uusitalo and **Anna-Mari Leppisaari**
Aalto University School of Arts, Design and Architecture, Espoo, Finland
severi.uusitalo@aalto.fi, anna-mari.leppisaari@aalto.fi

Abstract

We present an interactive evolutionary system for whole-garment knit design and use it to develop co-creativity frameworks by formalising an under-theorised level of human-computer agency. Current accounts theorise agency at the level of artefact selection and traversal steering, but do not address the case where one agent modifies the generative substrate by which the other operates. We propose three levels: evaluative agency over candidate selection, traversal agency over the operational region $O \subseteq C$, a construct we introduce to capture the practically reachable subset of C under the resource and interaction constraints of co-creative use, and constitutive agency over the transformation rules T_c that define C itself. Through an analytical design session, an experienced textile designer first navigated the existing conceptual space C through human-system co-traversal, then authored a new procedural style that restructured T_c, expanding C rather than redistributing search within it. We illustrate how language model–assisted authoring lowers the threshold for constitutive agency in this context.

Introduction

Many computational creativity systems structure interaction as navigation within a predefined generative space. Designers select, rank, or adjust parameters, but the representational grammar and transformation rules underlying variation remain fixed. While such systems can support rich exploratory traversal, creative agency is confined by the constraints of the design space.

Recent work in co-creative systems has emphasised distributed initiative between human and computational agents. However, even in mixed-initiative settings, the underlying generative rule sets and traversal mechanisms frequently remain architecturally stable, thereby fixing the conceptual space within which interaction unfolds. The question is therefore not only how initiative is shared, but whether traversal strategies and generative rules themselves can be modified during interaction.

Large language model (LLM)-assisted integrated development environments (IDEs) have made generative code editable during creative sessions. This is no longer a technical threshold question but an architectural one: when generative mechanisms become modifiable during use, co-creative interaction is no longer confined to traversal within a fixed conceptual space. The question becomes whether existing co-creativity frameworks adequately theorise such interventions as a distinct level of co-creative agency, or whether a new analytical level is required.

Computational fabrication provides a particularly revealing context for this question. In materially constrained domains such as textile design, generative exploration is shaped not only by algorithmic structure but also by fabrication feasibility, embodied evaluation, and resource limits. Here, expanding or redirecting a design space is inseparable from negotiating material constraints.

In this paper, we investigate how an interactive evolutionary system can support both exploratory traversal and structural transformation of its conceptual space through designer-authored generative rules. We examine a design session in which a professional textile designer first explores a predefined style library and subsequently authors a new procedural pattern style within the same system.

We were interested in the following questions:

(1) What analytical level, if any, is required to account for the case where one agent modifies the generative substrate by which the other operates? (2) If such a level exists, how does intervention at the rule level differ structurally from traversal-level agency within the same system? (3) What does constitutive agency require in a fabrication-constrained domain, where new generative rules must remain materially feasible?

We contribute (1) a co-creative evolutionary system that instantiates all three agency levels within a single fabrication-constrained architecture, and (2) a three-level account of co-creative agency – evaluative, traversal, and constitutive – that extends existing frameworks by formalising the case where one agent modifies the generative substrate by which the other operates. Central to this account is the operational region, the practically reachable subset of the conceptual space under interaction and resource constraints, which enables the traversal/constitutive distinction to be drawn precisely. We demonstrate that authoring assisted by LLM lowers the threshold for constitutive agency during use, and that constitutive agency in materially constrained domains requires analytical frameworks that distinguish between modification of the operational region and restructuring of the conceptual space itself.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

By situating this account within Wiggins' formalisation of conceptual spaces, the paper clarifies how existing exploratory/transformational distinctions can be operationalised at the level of co-creative agency, and where current frameworks require extension.

Background and Related Work

To ground the study, we first outline formalisations of conceptual space and co-creativity, then situate the work within evolutionary and computational textile design. The focus is on aspects directly relevant to representational plasticity and rule-level authorship.

Conceptual Space and Co-Creative Transformation

In Creative Systems Framework (CSF), Wiggins formalises such systems as search processes operating within a universe U, structured by transformation rules T and evaluation criteria E (Wiggins 2006); in Wiggins' notation, the space-defining rules are R and the traversal strategy T – throughout this paper we refer to the space-defining rules as T_c. The transformation rules define a conceptual space $C \subseteq U$: the set of artefacts generable under current representational constraints. Exploratory creativity corresponds to traversal within a fixed C, whereas transformational creativity arises when T itself is modified, thereby restructuring C. In the present system, the computational rule set T_c operationalises this distinction concretely.

Extending this framework to co-creative settings, Kantosalo and Toivonen distinguish alternating and task-divided co-creativity, noting that human and computational agents may operate with partially overlapping universes $U_h \cap U_c$ (Kantosalo and Toivonen 2016). In interactive co-creative systems, practical exploration is further shaped by resource constraints, interaction history, and mutual adaptation, factors that bound which regions of C are actually reachable during a session. We refer to this practically accessible region as the operational region $O \subseteq C$: adjusting traversal mechanisms shifts O without altering C, while modifying T_c restructures C itself. In materially grounded design, U_h additionally encompasses fabrication constraints and embodied knowledge that cannot be fully encoded within U_c, making co-creative interaction a negotiation across representational boundaries rather than traversal within a single shared space.

Most directly related, Hauhio (2024) recently extends the CSF to a user-centered framing in which the search is performed by the user, defining a traversable set $T \subseteq A$ of artefacts a user can practically reach given their mental model and traversal strategy, together with control, value, and meaningful-control coefficients that quantify session-level outcomes for non-interactive prompt-based generative tools. Our $O \subseteq C$ structurally parallels Hauhio's $T \subseteq A$ and shares the user-centered extension of CSF, but differs in three respects: it applies to programmable systems in which the designer can modify T_c during the session; it formalises the move from rule-following to rule-modification as a distinct constitutive level of co-creative agency, orthogonal to Hauhio's control and value coefficients; and it treats material constraints, fabrication feasibility in the present case, as bounding O alongside algorithmic reachability.

Ritchie (2012) refined CSF by abstracting away its symbolic language machinery and formally defining transformation as a metalevel operation that produces a distinct norm-set for the space ($N' \neq N$), distinguishing it from object-level traversal that modifies only the search heuristic Q. In the present system, the computational rule set T_c operationalises this norm-set: adding a new procedural style restructures what artefacts are generatable and thereby constitutes a transformation in Ritchie's formal sense.

Broader work in mixed-initiative and co-creative systems has emphasised distributed initiative in artefact generation and evaluation (Karimi et al. 2018; Deterding et al. 2017; Muller, Weisz, and Geyer 2020), establishing interaction frameworks that inform the present study's analytical approach. In most implementations, however, co-creativity is enacted through traversal and selection within representational substrates that remain fixed during interaction – the condition that the present work examines as a site of potential extension.

Outside the computational creativity literature, design research has long discussed how design spaces are shaped by activity. Halskov and Lundqvist (2021), extending Lim, Stolterman and Tenenberg (2008), distinguish *filtering* the design space (extracting selected parts for investigation, without modifying the space) from *informing* it (adding new aspects and options that establish or transform the space). We adopt this distinction as background terminology: in our framework, traversal agency operates within the established space, a filtering-style activity, whereas constitutive agency expands the space by modifying its generative substrate, and is thus a specific computational form of informing.

Evolutionary Design Systems

Evolutionary design systems generate artefacts through iterative variation and selection applied to encoded representations (Bentley 1999). A genotype–phenotype mapping specifies how parameter vectors are expressed as artefacts, while mutation and recombination define available transformations. Representational choices therefore determine both the structure of C and the modes of designer intervention and have typically remained fixed once a system is deployed. Interactive evolutionary systems have been studied extensively as exemplified by the recent review by Wang and Pei (2024).

Indirect encodings such as Compositional Pattern-Producing Networks (CPPNs) (Stanley 2007) shift control from explicit geometry to developmental processes, enabling complex structure from compact rule sets. While these representations have been extensively studied algorithmically, their role in co-creative authorship at the level of modifying transformation rules during interaction remains comparatively underexplored.

Computational Textile Context

In computational knit design, significant advances have focused on fabrication-aware translation from geometric intent to machine-executable instructions (Narayanan et al. 2018;

2019; Hofmann et al. 2023). Compilation pipelines and scripting environments abstract low-level machine control while ensuring manufacturability. These systems expand operational flexibility but typically treat the generative architecture as structurally given.

In textile design, generative exploration is inseparable from fabrication feasibility and material constraint. The structure of C is therefore entangled with manufacturing logic, resource limits, and embodied evaluation.

Representation as a Site of Co-Creation

Meta-design research argues for systems intentionally structured to support modification during use rather than remaining fixed at deployment (Fischer and Scharff 2000; Fischer and Giaccardi 2006). In such frameworks, the boundary between system designer and user becomes permeable, enabling participatory evolution of the artefact-generating substrate. In computational generative design, an analogous distinction has been drawn: the designer specifies the rules and processes that produce artefacts rather than manipulating artefacts directly, a mode characterised as meta-design of the generative substrate (McCormack, Dorin, and Innocent 2004).

Work in procedural content generation (PCG) has established an explicit definition and analysis of generative spaces (Smith and Mateas 2011; Shaker, Smith, and Yannakakis 2016), and some mixed-initiative PCG systems allow the designer to steer within these spaces (Liapis, Yannakakis, and Togelius 2013). Yet interaction generally operates through parameter control or search guidance rather than modification of the underlying transformation rules during use.

Assessing Co-Creativity

A prior study in evolutionary design case indicates that designer agency may be constrained by representational complexity and technical friction, while exploratory sessions can motivate expansion of the design space itself (Uusitalo et al. 2024). This aligns with accounts of the cognitive gap between computational and design thinking (Kelly and Gero 2021).

Within materially grounded domains, evaluation extends beyond internal fitness functions to external contexts of fabrication and presentation. This motivates the analytical approach adopted in the present study, which focuses on how representational and traversal mechanisms are configured during use.

A recent systematic review of agency in human-AI co-creation across 134 HCI papers confirms that agency is predominantly analysed through locus, dynamics, and granularity of control without addressing interventions that modify the generative substrate itself (Zhang, Wang, and Yi 2025). This pattern is consistent with mixed-initiative game design tools, where a recent analysis finds no cases in which conceptual spaces are modified during interaction (Margarido et al. 2025). This gap motivates the present study's focus on constitutive agency as a distinct analytical level.

System Overview

The system is a web-based co-creative textile design environment built around an Interactive Genetic Algorithm (IGA) core and a modular procedural representation. Designs are encoded as unified genotypes combining shape and pattern parameters and rendered into fabrication-aware bitmap output. Evolutionary variation is guided by human selection and bounded mutation, complemented by a local language model operating at the traversal level. Crucially, the system is not limited to exploration within a fixed generative space: its modular style architecture allows the designer to extend or modify the underlying generative logic itself, thereby altering the structure of the design space over time.

To clarify how representational rules, traversal mechanisms, and evaluation processes interact, Figure 1 presents a layered conceptual architecture of the system.

Representational Substrate

Each candidate design is encoded as a hierarchical genotype that combines shape and pattern variables within a single representational schema. Shape parameters define boundary conditions and masked regions, while pattern parameters control the procedural generation of internal structures. Together, they specify a complete design instance without separating structural and pattern layers at the representation level.

Genotype–Phenotype Mapping The genotype is translated into a phenotype through a rendering pipeline that produces a bitmap image (PNG). The same file functions both as a visual evaluation artefact for the designer and as a machine-interpretable specification for downstream fabrication processing. Colour encodings correspond to knitting-relevant instructions rather than purely aesthetic values.

Fabrication-aware constraints are embedded directly into the representation and rendering logic. Invalid region assignments, incompatible parameter combinations, and structurally infeasible configurations are prevented at the genotype level. As a result, evolutionary variation operates within physically realisable bounds, ensuring that exploration remains aligned with production constraints.

Pattern structures are rendered with region-aware logic derived from a template image, ensuring motif coherence when knitted: the generative pipeline applies distinct rendering rules to torso and hem regions so that patterns repeat consistently in the physical artefact rather than distorting at seams. This encoding instantiates a whole-garment knitting requirement directly within the representational substrate, ensuring that evolutionary variation cannot produce patterns that would be visually coherent in the bitmap but structurally incoherent when fabricated.

A bitmap can be added as a starting point for a design process. This template image functions as a pre-session agency mechanism: by encoding admissible regions through colour-coded pixels, the designer operationalises external constraints from the human/material universe U_h directly within the representational substrate, bounding the conceptual space prior to traversal.

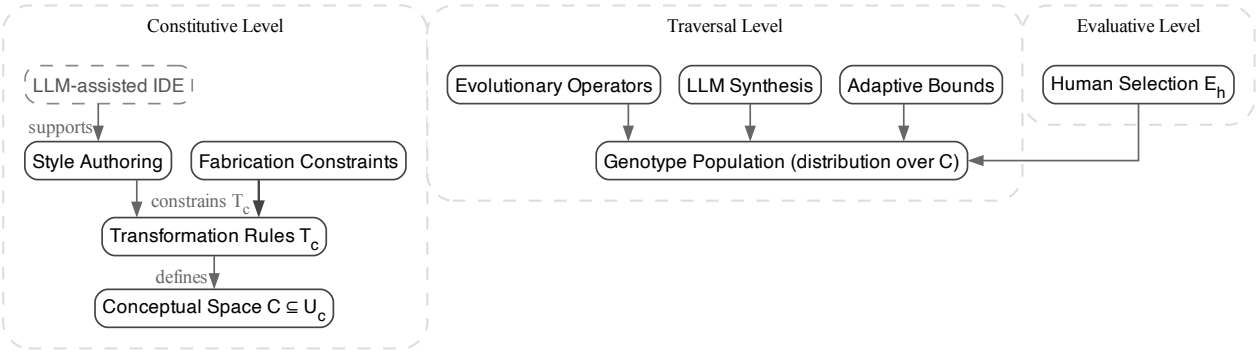


Figure 1: Conceptual architecture separating constitutive, traversal, and evaluative levels of co-creative agency. At the constitutive level, LLM-assisted IDE supports Style Authoring, which modifies transformation rules T_c — subject to fabrication constraints that bound admissible rule modifications. Traversal mechanisms induce an operational region $O \subseteq C$, and human selection E_h evaluates candidate artefacts at the evaluative level.

In addition to independent parameter effects, certain metagenes introduce developmental couplings between shape and pattern. For example, boundary-sensitive modulation parameters or CPPN-based spatial biases may cause pattern density, scale, or orientation to vary as a function of the shape-defined geometry (e.g., gradual intensification toward edges or central attractors). In such cases, the phenotype emerges from interactions between shape-related and pattern-related parameters rather than from two fully separable layers. This introduces a developmental dependency within the genotype-phenotype mapping, where geometric boundary conditions influence procedural pattern expression.

Universe, Conceptual Space, and Operational Region To clarify the ontological and operational layers of the system, we distinguish three levels. The *computational universe* U_c is the set of all artefacts expressible within the bitmap-based knitting representation, bounded by encoding limits and structural constraints. Within it, transformation rules T_c define a *conceptual space* $C \subseteq U_c$, instantiated as genotype structure, procedural styles, and compositional operators. Traversal dynamics then induce a non-uniform probability distribution over C; we refer to the high-probability, practically reachable subset as the *operational region* $O \subseteq C$.

Under this distinction, changes to traversal parameters shift O (exploratory creativity), whereas modifications to T_c alter C itself (transformational creativity). This separation ensures that representational ontology (U_c and C) and exploratory dynamics (O) remain analytically distinguishable throughout the system architecture.

Modular Styles as Generative Units Pattern generation is organised around modular *styles*, each implemented as an encapsulated procedural generator adhering to a shared rendering interface. A style defines its own parameter schema and may optionally include style-specific mutation or recombination operators. At runtime, one or more styles are instantiated and combined within the unified genotype.

Styles can be composed through logical and spatial compositing operations (e.g., conjunction, masking, blending), enabling hybrid pattern structures without redefining the underlying representation. In this sense, the available style library defines the structural extent of the conceptual space C at a given moment.

Certain styles operate at a developmental level rather than specifying local pattern structure directly. For example, a CPPN-based component functions as a spatial metagene, introducing global regularities such as symmetry, smooth gradients, or coordinated variation across the surface. Additional metagenes (e.g., global scale factors, symmetry biases, or attractor-like modulation parameters) shape how lower-level pattern parameters are expressed across space. These metagenes bias phenotypic realisation without prescribing a fixed aesthetic outcome, preserving evolutionary variability while maintaining coherent global structure.

Structural and Adaptive Bounds To distinguish representational limits from adaptive search bias, the system employs a two-layer bound model defined in external JSON configuration.

Structural_bounds specify the ontological limits of each parameter (e.g., type, valid range, discrete sets). These bounds define the admissible region of the conceptual space C and remain invariant unless the representational schema itself is modified.

Adaptive_bounds are updated incrementally from interaction history and accepted selections. They do not alter the structural limits of C, but bias sampling and mutation toward historically preferred parameter regions.

During evolution, all parameter updates are constrained by structural bounds, while adaptive bounds modulate probability distributions within those limits. This explicit separation preserves a stable representational ontology while allowing historically conditioned reorientation of the operational region $O \subseteq C$. In this way, representational structure and exploratory dynamics remain analytically separable.

Evolutionary and Meta-Evolutionary Dynamics

This section describes how the implemented conceptual space C is traversed during interaction.

Interactive Genetic Algorithm Core The system maintains a population of candidate genotypes that evolve across discrete generations. At each iteration, the designer selects one or more preferred candidates, which serve as parents for the next generation.

Variation is introduced through crossover and bounded mutation within structural limits. Mutation strength is globally controllable; additionally, a *focus mode* implements an exploitation bias by contracting the operational region O toward parent-adjacent solutions. Concretely, focus mode (i) increases the mutation amplitude of developmental metagenes for a subset of the brood while strengthening adaptive memory-range biasing, and (ii) selects the displayed next-generation candidates by proximity to the selected parents in a standardised genotype-derived embedding space, rather than prioritising cluster-representative diversity. Adaptive bounds remain active in focus mode, further conditioning variation toward historically preferred regions.

The designer may also lock either shape-related or pattern-related components, constraining evolution to the unlocked part. Human selection – denoted E_h to distinguish it from computational evaluation processes – remains the primary evolutionary pressure guiding population drift over time.

Traversal-Level Synthesis and Diversity Mechanisms In parallel with the genetic operators, a local language model periodically proposes candidate parameter configurations, operating exclusively on structured parameter summaries, selection history, and simple population statistics, without access to rendered images. This component is confined to the traversal level, shaping the operational region O within a fixed C, and is architecturally and functionally distinct from the LLM-assisted development environment through which generative rules themselves are authored.

Three constrained mechanisms are implemented at the traversal layer: (1) selection-weighted recombination of parameters from previously preferred candidates, (2) bounded perturbation triggered when population variance falls below a threshold, and (3) diversity-driven replacement when intra-population similarity exceeds predefined thresholds. All proposals are validated against structural constraints before admission into the population, and each is accompanied by a short textual rationale referencing measurable signals such as repeated selection or detected convergence, keeping proposals inspectable and reversible at the parameter level. In each synthesis cycle, one slot is reserved for a deliberately contrarian candidate that inverts the metagene centroid of recent selections and prioritises underexplored styles, functioning as a persistent diversification probe within the traversal layer. Operating on explicit genotype representations rather than latent visual embeddings, the model remains lightweight and locally executable, enabling near real-time interaction without external API latency and supporting sustained use without interrupting the designer's cognitive flow.

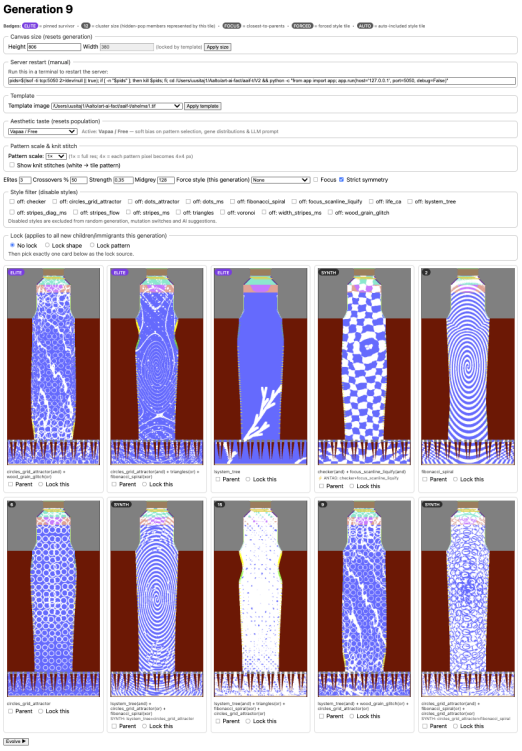


Figure 2: The browser-based interface presenting a generation of ten candidate designs for iterative selection by designer. Controls at the top expose traversal parameters including mutation strength, focus mode, style filtering, and shape/pattern locking. Each candidate displays its active style composition; ELITE and SYNTH labels indicate selection history and LLM-synthesised proposals respectively.

A backend controller manages the evolutionary loop, maintaining population state, applying variation operators, triggering rendering, and logging interaction history across generations.

Space Transformation and Co-Creative Agency

Having described representation and traversal operationally, we now situate the system within a formal account of computational creativity.

From Exploratory to Transformational Creativity Within the present system, exploratory creativity corresponds to traversal within a fixed conceptual space C, where mutation rates, adaptive bounds, or operator weighting reshape the operational region O without altering C itself.

Transformational creativity occurs when the transformation rules T_c are modified – for example, through the introduction of a new procedural style or changes to the generative pipeline – thereby expanding or restructuring C.

Styles as Objects of Authorship Within the present system, styles function not only as sources of variation but as

modifiable generative units. Authoring a new style introduces new parameters, spatial logics, or developmental couplings, thereby redefining the structural extent of C.

Because styles encapsulate generative rules, they provide a clear locus at which the designer can intervene in the construction of the space itself. The design process thus extends beyond selection among candidates to modification of the mechanisms that produce them.

Design Session: Authoring a New Procedural Pattern Style

A single professional designer engaged with the system in order to enact and expose its co-creative dynamics under realistic conditions of use. The design session is treated as a critical case (Yin 2003, p. 41): a single instance sufficient to establish that the proposed three-level agency structure is instantiable in practice. The purpose of the session was not to evaluate user satisfaction or generalisable usability outcomes, but to examine how agency, initiative, and generative control are distributed between human and system within the given architecture.

Study Design

The study involved two human participants with clearly delineated roles. The first author designed and implemented the interactive evolutionary system and a set of styles, and was responsible for the study setup and data analysis.

The system user was a professional textile designer with a postgraduate degree and extensive experience in industry and university-level design education. They contributed domain expertise, material knowledge, and an independent design brief – the U_h resources not available to the system designer – and provided the real-world design context (the Hylky garment project). Although they are a co-author, their contribution to system development was explicitly bounded: they provided Shima Seiki fabrication requirements and sample designs but did not participate in data analysis or interpretation of findings. Their familiarity with the system's design intent means the session is not representative of independent or novice use, but is appropriate for a proof-of-concept demonstration of constitutive agency enacted by a domain expert operating within their area of competence.

The computational system, which includes the IGA and the LLM-assisted IDE, functions as a co-creative agent within the session, contributing proposals and modifications under human selection control.

Session Procedure The in-person session lasted 180 minutes and was conducted with the browser-based system interface. In the first phase, the designer employed the existing style library with the aim of producing a garment design suitable for full whole-garment knitting. This phase served both as creative exploration and as a means of identifying structural limitations of the current design space.

The second phase focused on authoring a new procedural style intended to better support the identified design ambitions, employing the Cursor IDE and instant evaluations of developments in the browser-based interface. Following this intervention, a short post-session reflection was conducted to capture immediate impressions of the interaction and perceived distribution of agency.

Design Intent and Intervention The design task involved developing a whole-garment knitted piece using a yarn derived from historically sourced wood pulp, from a recently discovered, late 17th-century shipwreck (Ahverdov, Koivikko, and Vuori 2026). The material was available only in limited quantity, introducing strict resource constraints. The garment was intended for public exhibition, introducing external evaluative criteria beyond internal system optimisation. In 4P terms (Rhodes 1961), the exhibition context constitutes a *Press* factor shaping E_h, requiring technical robustness and aesthetic coherence suitable for public display.

In addition, the garment was intended for physical display on a human and a mannequin, imposing embodied geometric and proportional constraints external to the computational representation. These constraints originate in the human/material universe U_h and were operationalised within U_c through a template file that restricted the admissible generative regions. This image file was prepared prior to the session by the participating textile designer.

The intervention was not framed solely as a functional refinement of outcomes, but as an exploratory act aimed at examining how a professional designer appropriates the ability to modify the design space itself. Rather than limiting interaction to traversal within a predefined conceptual space, the session introduced style authoring as a structural intervention at the level of T_c. The implications of this shift, from navigating a space to reshaping it, are discussed in the analysis section.

Data Captured The system logs all evolutionary runs, including selected candidates, parameter configurations, generation counts, and exported pattern files. Screen capture documented the interaction process and discussion in line with contextual inquiry, appended by written notes. The resulting designs were fabricated as knitted garment samples, providing material validation of the generated outputs.

Contrasting Before and After Style Authoring

The session unfolded in two phases. In the first phase, exploration proceeded using the existing style library and traversal mechanisms. The designer interacted with the system through the browser-based interface (Figure 2), selecting candidates to guide subsequent evolution. By generation 81, the designer selected a configuration considered sufficiently aligned with the project constraints to justify physical prototyping. This phase demonstrates the system's exploratory capacity within a fixed conceptual space C.

The second phase introduced style authoring as a structural intervention. The designer entered with a pre-articulated generative intent described as a "wood grain pattern disrupted with digital glitch artefacts." Through approximately ten iterative refinements in the LLM-assisted development environment, a new procedural style was implemented in Python.

Fabrication constraints shaped constitutive intervention directly: during the design session, the designer verbally reflected on whether a potentially interesting pattern organisation would become illegible when knitted, effectively performing mental simulation of material feasibility as a real-time filter on generative rule construction. This illustrates that constitutive agency in a fabrication-constrained domain is not equivalent to unconstrained code authoring – new generative rules must be evaluated not only for their parametric behaviour but for their viability as physical artefacts.

The resulting style introduced additional parameters (e.g., scanline and liquify intensities, focus bias, softness controls) and extended the generative pipeline with a post-effect stage operating on a preceding pattern layer. This modification did not merely reweight existing parameters or redirect traversal. Instead, it altered the transformation rules T_c by adding a new generative step to the mapping from genotype to phenotype.

As a consequence, the conceptual space C defined by T_c was structurally expanded: the post-effect stage introduced a generative step absent from the prior mapping, expanding C rather than redistributing search within it.

In Wiggins' formal terms (Wiggins 2006), this intervention corresponds to a modification of T_c rather than a change confined to traversal dynamics. Notably, the locus of control shifted from phenotype selection to rule-level manipulation, marking a transition from exploratory traversal to transformational engagement.

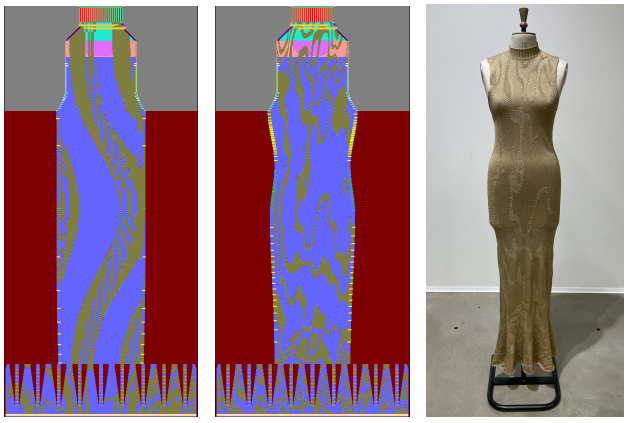
Reflection on Co-Creative Agency

The session highlights an asymmetry in agency distribution. The computational agent operated primarily at the level of candidate generation and traversal structuring, while rule-level transformation remained a human-initiated intervention. This became evident in the second phase, where the designer shifted from selecting outcomes to modifying generative logic.

The oscillation between exploration and exploitation was explicitly articulated by the designer: "sometimes one wants it to stay same, next, one wants to see something different." This reflects conscious engagement with traversal control rather than passive selection. Likewise, the remark that "the antagonist is good to have there" suggests that computationally introduced provocation was experienced as a productive counterpoint rather than as noise.

At the same time, the desire for "more local control" indicates that agency at the representational level remained bounded by the current architecture. While rule-level transformation was accessible, fine-grained spatial editing was not, revealing limits in the current distribution of initiative.

Together, these observations suggest that co-creativity in this system does not rely on symmetrical transformation capabilities. Instead, it emerges from complementary roles: computational structuring of traversal and human-driven modification of generative rules. The introduction of style authoring shifted the designer's activity from selecting among generated artefacts toward manipulating the generative logic itself. An LLM-assisted IDE functioned as a coding assistant within this process, supporting iterative refinement of procedural logic, but the structural intervention i.e., adding new parameters and a post-effect stage remained conceptually directed by the designer.



(a) Design co-created through traversal within the pre-existing style library.

(b) Design generated through constitutive intervention: a new procedural style was authored, modifying T_c and expanding the conceptual space C.

(c) Fabricated garment produced from the design in Figure 3b, made from substitute yarn. The sample produced from shipwreck yarn is in Fig. 4.

Figure 3: Contrastive design outcomes from the session and a knitted prototype garment.

Discussion

The following discussion interprets the design session through the analytical framework introduced earlier, distinguishing between traversal within a fixed conceptual space C and structural modification of its defining transformation rules T_c, in order to clarify how co-creative agency was distributed across system layers.

Traversal Control and Operational Plasticity

The session first illustrates agency at the level of traversal within a fixed conceptual space C, where the designer explored the conceptual space widely. In late generations, focus mode enabled controlled exploitation by contracting the operational region toward parent-adjacent solutions. Rather than restricting parameters explicitly, it reshaped the search distribution by amplifying developmental metagene variation and replacing diversity-based population turnover with proximity-based selection in a standardised genotype-derived embedding space. Final selections in both phases emerged under such focused refinement, indicating that traversal dynamics functioned as meaningful control mechanisms rather than opaque stochastic variation.

In parallel, the LLM-based synthesis mechanism and adaptive bounds further shaped which artefacts became likely or reachable within C, conditioning the operational region O without modifying the transformation rules T_c. In

Wiggins' terms, they correspond to exploratory creativity: the induced distribution shifts, but the generative rule set remains stable. In Ritchie's (2012) terms, these operations modify the search heuristic Q while the norm-set N – operationalised by T_c – remains fixed.

Transformational Intervention

The transformational intervention was intentionally introduced as part of the session design. The contribution therefore does not lie in demonstrating that rule-level modification can occur, but in analysing how enabling such modification restructures the conceptual space and redistributes co-creative agency within a materially constrained evolutionary system.

In contrast, the assignment to introduce a new procedural style modified T_c by adding new parameters and a post-effect stage to the generative pipeline. This extended the set of artefacts generable under the system by modifying the generative mapping from genotype to phenotype, thereby restructuring the conceptual space C defined by T_c. The shift from phenotype selection to rule-level modification therefore marks a transition from exploratory traversal to transformational engagement.

The architecture made such structural interventions possible during interaction. The language model supported iterative refinement of procedural logic, but the structural modification itself remained a designer-directed intervention at the level of T_c.

By Ritchie's (2012) criterion (Definition 7), this constitutes transformation rather than traversal: the new style rewrites T_c, introducing parameter types and a rendering stage absent from the prior pipeline, so the norm-set changes structurally ($N' \neq N$) rather than merely redistributing probability across a fixed one. Adjusting traversal parameters (focus mode, adaptive bounds, LLM-mediated proposals) modifies Q while T_c remains fixed; only the rule-authoring step crosses the $N' \neq N$ threshold.

Ventura and Brown (2024) argue on information-theoretic grounds that creativity-as-search at the base-artefact level is generally infeasible due to density constraints, and propose reframing creativity as a search for short programs that generate the artefacts. Our notion of constitutive agency arrives at a structurally similar conclusion from a different direction: the locus of creative intervention shifts from candidate artefacts to the generative procedures (T_c) themselves. Where Ventura and Brown frame LLM-based generation as a creative dead-end at the artefact level, our system situates the LLM as a programming assistant at precisely the rule-authoring level they identify as the productive locus.

Within the CSF tradition, this separation is anticipated by Hauhio (2024), who treats user-driven modification of goal rules R, and the consequent change in the mental model M and traversal strategy T_M, as forms of transformative creativity rather than continued traversal. The same categorical distinction is independently established in design research: by Halskov and Lundqvist's (2021) definitions, modifying the generative procedures is informing-work, not filtering-work, regardless of whether the modification capability was anticipated by the system designer.

Material Constraints and External Evaluation

Textile design is materially constrained. Display context, fabrication feasibility, and resource limitations introduced requirements originating in the human/material universe U_h, which were operationalised within U_c through template-based shape encoding and fabrication-aware representation. Rule-level modification was therefore not a purely formal expansion, but one constrained by fabrication feasibility and exhibition context.

The exhibition outcome functions as an external evaluative context, situating creative value beyond the interaction loop. The fabrication and public presentation of both pre- and post-intervention artefacts indicate that the system supported viable creative production under real-world constraints.

Implications for Co-Creative Agency

The session illustrates how the three levels of agency are enacted in practice: evaluative agency, in the form of candidate selection guided by material and aesthetic judgement, operated continuously across both phases as the baseline mode of co-creative interaction. The analytically significant finding concerns the remaining two levels: traversal agency, enacted through focus mode, adaptive bounds, and LLM-mediated proposals that shaped O without modifying C; and constitutive agency, exercised when the designer restructured T_c by authoring a new procedural style, expanding C rather than redistributing search within it. The first two correspond to exploratory reorientation within a fixed representational substrate; the third constitutes transformational intervention at the level of the generative rules themselves.

Agency is asymmetrically distributed across these levels. The computational agent operates primarily at the traversal layer, proposing candidates and reshaping the operational region through synthesis and diversity mechanisms. The human designer retains evaluative authority and uniquely performs rule-level modifications that alter T_c. Co-creativity here emerges not from symmetric transformation capabilities, but from complementary roles across traversal, constitutive agency, and evaluation.

Crucially, the modular implementation exposes transformation rules T_c as inspectable and editable during interaction, enabling intervention at multiple levels of the generative pipeline. This makes structural modification part of the co-creative process rather than a prior design decision fixed at deployment. While the language model facilitates iterative development, the key shift lies in the relocation of generative authority: the designer restructures C rather than merely re-orienting the induced probability distribution over it. However, evaluation remains grounded in the human agent situated within the broader material universe U_h, where fabrication constraints, exhibition context, and aesthetic judgement determine what is worth producing. The system expands generative plasticity; the human anchors value.

These findings resonate with prior analyses of interaction design in co-creative systems (Kantosalo et al. 2020; Rezwana and Maher 2023), which emphasise how initiative



Figure 4: Detail of the fabricated garment showing structure at the knitted surface. The physical artefact evidences viability under fabrication constraints and situates evaluation beyond the computational interaction loop. (Photograph: Esa Kapila)

distribution and feedback structure shape creative experience. In the present case, however, neither initiative distribution nor control granularity – the dimensions foregrounded in recent empirical work on agency in HCI systems (Zhang, Wang, and Yi 2025) – adequately captures the intervention that occurred: constitutive agency operates at a qualitatively different level, one that existing co-creativity frameworks and control mechanism taxonomies were not designed to accommodate. When LLM assistance lowers the threshold for modifying T_c during use, the boundary between system design and system use becomes structurally permeable – and the frameworks used to analyse co-creative agency may need to expand accordingly.

Limitations and Future Work

The presented system is domain-specific, restricted to bitmap-based pattern representations for whole-garment knitting, and evaluated through a single analytical design session with a co-author participant. The findings therefore illuminate architectural implications of constitutive agency rather than generalizable behavioural outcomes. Although LLM-assisted authoring lowers the threshold for rule-level intervention, such interventions still require capacity to articulate structural generative intent.

Future work may extend the representational substrate toward integrated topological garment modelling, and investigate how traversal-level and rule-level interventions interact over longer-term use with multiple designers. Studying how shared or evolving style libraries reshape the conceptual space over time would clarify the social and collaborative dimensions of constitutive agency in co-creative systems.

Conclusions

By formalising three levels of co-creative agency – evaluative, traversal, and constitutive – the study identifies an under-theorised mode of human-computer co-creativity in which one agent modifies the generative substrate by which the other operates. In a fabrication-constrained domain, LLM-assisted authoring facilitates this constitutive intervention during use, embedding co-creative agency within the constraints of physical production.

As generative systems become increasingly programmable, computational creativity research may need to extend existing exploratory/transformational distinctions toward frameworks that operationalise co-creative agency across structural levels, from traversal within a conceptual space to authorship of the space itself.

Acknowledgments

The study was completed as part of the *art-ai-fact* project, enabled by an endowment from Fiskars Group Oyj to Aalto University. The authors want to acknowledge all who contributed to the Hahtiperän Hylky project (Hahtiperä Shipwreck), which provided the context for the critical case.

AI-assisted tools were used for editing and language refinement; all analytical contributions are the authors' own.

References

Ahverdov, E.; Koivikko, M.; and Vuori, K. 2026. Optimizing conservation: the case study of the Hahtiperä wreck, Finland. In *Proceedings of the Symposium of Conservation for Underwater Archaeology (ISCUA 2023): Learning from the Lost*, 347–362. Formentera, Spain: Ministerio de Cultura.

Bentley, P. J. 1999. *Evolutionary design by computers*. San Francisco, CA: Morgan Kaufmann.

Deterding, S.; Hook, J.; Fiebrink, R.; Gillies, M.; Gow, J.; Akten, M.; Smith, G.; Liapis, A.; and Compton, K. 2017. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM.

Fischer, G., and Giaccardi, E. 2006. Meta-design: A framework for the future of end-user development. In *Human-Computer Interaction Series*. Dordrecht: Springer Netherlands. 427–457.

Fischer, G., and Scharff, E. 2000. Meta-design: design for designers. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, 396–405. New York, NY, USA: ACM.

Halskov, K., and Lundqvist, C. 2021. Filtering and informing the design space: Towards design-space thinking. *ACM transactions on computer-human interaction: a publication of the Association for Computing Machinery* 28(1):1–28.

Hauhio, I. 2024. Enhancing human creativity with aptly uncontrollable generative AI. In *Proceedings of the 15th International Conference on Computational Creativity*, 220–224. Jönköping, Sweden: Association for Computational Creativity.

Hofmann, M.; Albaugh, L.; Wang, T.; Mankoff, J.; and Hudson, S. E. 2023. KnitScript: A domain-specific scripting language for advanced machine knitting. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA: ACM.

Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the 7th International Conference on Computational Creativity*, 77–84. Paris, France: Association for Computational Creativity.

Kantosalo, A.; Ravikumar, P. T.; Grace, K.; and Takala, T. 2020. Modalities, styles and strategies: An interaction framework for human-computer co-creativity. In *Proceedings of the 11th International Conference on Computational Creativity*, 57–64. Coimbra, Portugal: Association for Computational Creativity.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. In *Proceedings of the 9th International Conference on Computational Creativity*, 104–111. Salamanca, Spain: Association for Computational Creativity.

Kelly, N., and Gero, J. S. 2021. Design thinking and computational thinking: a dual process model for addressing design problems. *Design Science* 7:e8.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2013. Sentient sketchbook: Computer-assisted game level authoring. In *Proceedings of the 8th International Conference on Foundations of Digital Games*. Chania, Crete, Greece: Society for the Advancement of the Science of Digital Games. 213–220.

Lim, Y.-K.; Stolterman, E.; and Tenenberg, J. 2008. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Trans. Comput.-Hum. Interact.* 15(2):1–27.

Margarido, S.; Machado, P.; Roque, L.; and Martins, P. 2025. Boosting mixed-initiative co-creativity in game design: A tutorial. *ACM Computing Surveys* 58(2):1–36.

McCormack, J.; Dorin, A.; and Innocent, T. 2004. Generative design: A paradigm for design research. In Redmond, J.; Durling, D.; and de Bono, A., eds., *Proceedings of Futureground, Design Research Society International Conference 2004*, volume 2. Caulfield East, Australia: Monash University.

Muller, M.; Weisz, J. D.; and Geyer, W. 2020. Mixed initiative generative AI interfaces: An analytic framework for generative AI applications. In *Proceedings of the Workshop on Human-Computer Co-Creativity at ICCC 2020*. Coimbra, Portugal: Association for Computational Creativity.

Narayanan, V.; Albaugh, L.; Hodgins, J.; Coros, S.; and McCann, J. 2018. Automatic machine knitting of 3D meshes. *ACM Transactions on Graphics* 37(3):1–15.

Narayanan, V.; Wu, K.; Yuksel, C.; and McCann, J. 2019. Visual knitting machine programming. *ACM Transactions on Graphics* 38(4):1–13.

Rezwana, J., and Maher, M. L. 2023. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Trans. Comput.-Hum. Interact.* 30(5):67:1–67:28.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Ritchie, G. 2012. A closer look at creativity as search. In *Proceedings of the 3rd International Conference on Computational Creativity*, 41–48. Dublin, Ireland: Association for Computational Creativity.

Shaker, N.; Smith, G.; and Yannakakis, G. N. 2016. Evaluating content generators. In *Procedural Content Generation in Games*. Cham: Springer International Publishing. 215–224.

Smith, A. M., and Mateas, M. 2011. Answer set programming for procedural content generation: A design space approach. *IEEE Transactions on Computational Intelligence and AI in Games* 3(3):187–200.

Stanley, K. O. 2007. Compositional pattern producing networks: A novel abstraction of development. *Genetic programming and evolvable machines* 8(2):131–162.

Uusitalo, S.; Kantosalo, A.; Salovaara, A.; Takala, T.; and Guckelsberger, C. 2024. Creative collaboration with interactive evolutionary algorithms: a reflective exploratory design study. *Genetic programming and evolvable machines* 25(1):4.

Ventura, D., and Brown, D. 2024. Creativity as search for small and interesting programs. In *Proceedings of the 15th International Conference on Computational Creativity*, 331–335. Jönköping, Sweden: Association for Computational Creativity.

Wang, Y., and Pei, Y. 2024. A comprehensive survey on interactive evolutionary computation in the first two decades of the 21st century. *Applied Soft Computing* 164:111950.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Yin, R. K. 2003. *Case study research: Design and methods*. Applied Social Research Methods. Thousand Oaks, CA: SAGE Publications, 3rd edition.

Zhang, S.; Wang, H.; and Yi, X. 2025. Exploring collaboration patterns and strategies in human-AI co-creation through the lens of agency: A scoping review of the top-tier HCI literature. *Proceedings of the ACM on Human-Computer Interaction* 9(7):1–43.

IVIE: A Neuro-symbolic Approach to Incremental and Validated Generation of Interactive Fiction Worlds

Micaela Vaucher † **Santiago Silveira** **Santiago Góngora †** **Luis Chiruzzo**

Instituto de Computación, Facultad de Ingeniería
Universidad de la República
Uruguay
† {micaela.vaucher, sgongora}@fing.edu.uy

Abstract

Computational creativity in Interactive fiction faces a fundamental tension: Large Language Models (LLM) may produce creative narratives but struggle with world coherence, while symbolic systems ensure consistency but lack creative flexibility. We present *IVIE* (Incremental & Validated Interactive Experiences), a neuro-symbolic approach to generating complete and playable interactive fiction worlds from scratch. Building upon PAYADOR's neuro-symbolic framework, *IVIE* implements a four-stage incremental generation pipeline that delegates creative decisions—setting and character creation, puzzle design—to LLMs while grounding the world state through symbolic validation. The system generates worlds with interconnected locations, functional items, non-player characters, and coherent puzzles, all structured around a central goal-oriented architecture. Human evaluation shows the approach generates immersive, thematically coherent worlds with high player engagement. Results seem to indicate that the neuro-symbolic approach successfully balances flexibility with narrative coherence: symbolic validation grounds LLM generation without eliminating generative freedom. However, challenges remain: LLM inconsistencies occasionally bypass puzzle constraints, and objective validation gaps allow some structurally impossible goals. We identify key design considerations for future neurosymbolic interactive storytelling systems, particularly regarding LLM capabilities and their limitations.

Introduction

Large Language Models can generate fluid narratives on demand, but they constantly fail when asked to maintain a coherent game world across just a few of player turns. Objects vanish and reappear arbitrarily, NPCs forget previous conversations, spatial relationships become contradictory, and puzzle solutions change mid-game (Góngora et al. 2024b). Yet the creative potential remains tantalizing: if we could ground LLM generation in symbolic validation, we could significantly advance the generation of consistent interactive narratives.

Interactive fiction (IF), the genre of text-based games in which users explore fictional worlds and influence the narrative through natural-language commands (Riedl and Bulitko 2012), exemplifies this tension between creative flexibility and structural coherence. Classic systems like *Zork* (Montfort 2003) maintained the world state through hand-crafted objects, locations, and deterministic rules, ensuring consistency but requiring extensive manual authoring for each world. Procedural content generation (PCG) approaches address this authoring bottleneck by algorithmically creating game content (Shaker, Togelius, and Nelson 2016), yet traditional methods for narratives rely on symbolic planning systems that produce rigid, template-based stories lacking creative richness (Riedl 2015).

LLMs promise a form of computational improvisation, generating fluid narratives on the fly without predetermined scripts. Yet improvisation is not mere spontaneity — it requires technique, memory, and contextual grounding (Schiaffini 2006; Sawyer 2000), precisely what LLMs lack for extended interactions. They hallucinate objects, forget established facts, and lose track of game state (Castricato et al. 2021; Ji et al. 2023). A golden key introduced on turn 5 may be forgotten by turn 25, or worse, the LLM may contradict whether the door was ever locked. For players seeking coherent experiences rather than surreal explorations, such inconsistencies fundamentally break immersion.

Recent neuro-symbolic approaches (Góngora et al. 2024a; Sarker et al. 2021) address this by separating concerns: LLMs handle creative generation while symbolic structures maintain the world state. Systems like PAYADOR (Góngora et al. 2024a) suggest that this works effectively for *playing* in predefined worlds. Extending these architectures to *automatically generate* complete worlds would enable rapid prototyping and procedural variation while maintaining the structural coherence that manual authoring provides.

In this paper, we present **IVIE** (Incremental & Validated Interactive Experiences), a neuro-symbolic system that generates complete, playable IF worlds automatically. While PAYADOR seemed to indicate that neuro-symbolic architectures can maintain coherence during gameplay, it operated exclusively on manually predefined worlds. IVIE extends this framework to world generation itself: rather than handcrafting each world, IVIE's four-stage incremental pipeline delegates creative decisions—setting and character creation, and puzzle design—to LLMs, while a symbolic layer enforces spatial connectivity, type correctness, and objective

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

solvability at each stage. The source code is available on GitHub[1].

Related Work

IF systems face a fundamental challenge: balancing creative narrative flexibility with structural world coherence. Classic parser-based systems like *Zork* (Montfort 2003) solved this through manual authoring: every object, location, and interaction was hand-coded, ensuring consistency but requiring extensive development effort for each world. Modern approaches attempt to escape this authoring bottleneck through different strategies, each with critical limitations.

Systems like AI Dungeon (Trapova 2023) pioneered the idea that LLMs can generate contextually appropriate responses to arbitrary player actions. By concatenating player input with conversation history and generating continuations via LLMs (e.g., GPT-3), these systems enable unprecedented narrative flexibility. However, extensive user testing revealed critical failures (Yong and Mitchell 2023): objects vanish without justification and reappear later, NPCs contradict previous statements, game rules become inconsistent as context windows fill, and the model exhibits "looping" behavior. These issues stem from LLMs' stateless nature—they maintain no persistent world model (Castricato et al. 2021), relying solely on probabilistic patterns without mechanisms to enforce structural coherence.

Recent work bridges symbolic and neural paradigms in different ways. Some approaches retrieve or maintain narrative context to improve coherence (Chambers et al. 2024), or augment game content through data-driven generation (Callison-Burch et al. 2024), yet these still lack the symbolic grounding necessary to guarantee structural validity. Others delegate creative generation to LLMs while maintaining a structured world state (Sarker et al. 2021), combining planning with language models (Kelly et al. 2023; Zhou et al. 2025) or constructing spatial representations dynamically (Li et al. 2025), but focus on playing or narrating within pre-existing worlds rather than generating complete worlds from scratch.

Our current work IVIE builds directly upon PAYADOR (Góngora et al. 2024a; 2026), a neuro-symbolic framework that separates symbolic world management from neural narrative generation. PAYADOR implements an object-oriented world model (`Location`, `Item`, `Character`, `Puzzle`, `Objective`) and employs dual LLMs: a reasoning model interprets player actions to generate structured state transformations, while a narrative model produces scene descriptions (Góngora 2025). World coherence is maintained through a continuous grounding cycle: when a player inputs an action, the system renders the current symbolic world state as natural language; the reasoning LLM predicts its effects, generating both symbolic transformations (e.g., moving an item, unlocking a passage) and narrative outcomes; these predictions are validated before being applied to update the world state; finally, the narrative LLM generates scene descriptions based on the updated state. This separation appeared effective at preventing LLM inconsistencies: spatial connectivity is verified through explicit graph structures, object locations are tracked deterministically, and puzzle states persist until explicitly modified. However, these systems operate on manually predefined worlds where every location, object, NPC, and puzzle is hand-coded before gameplay.

[1]https://github.com/micaelavaucher/IVIE

The IVIE Approach

IVIE explores the aforementioned limitation through incremental validated generation: a four-stage pipeline that progressively constructs worlds from abstract concepts to concrete playable experiences, with symbolic validation at each stage ensuring spatial connectivity, type correctness, and objective solvability. Given that LLMs can generate narrative text but struggle with structural coherence, while symbolic systems ensure consistency but lack generative flexibility, the central question becomes: *how do we harness both?*

To achieve this separation of concerns, we build upon PAYADOR's neuro-symbolic framework (Góngora et al. 2024a; 2026) as our architectural backbone. As described previously, PAYADOR maintains world coherence during gameplay through symbolic world state and dual-LLM architecture. Building on this framework, IVIE explores the generation of complete worlds automatically, rather than playing within predefined ones.

To achieve this separation of concerns, we build upon PAYADOR's neuro-symbolic framework (Góngora et al. 2024a; 2026) as our architectural backbone. Building on this framework, IVIE explores the generation of complete worlds automatically, rather than playing within predefined ones.

IVIE's symbolic component acts as a structured execution and validation layer rather than a creative reasoning engine: it does not generate narrative content, but executes and enforces structural correctness on LLM-proposed state transformations. Rather than asking LLMs to generate creative content and simultaneously maintain world consistency—a task where they demonstrably fail (Góngora et al. 2024a)—we aim to assign each component what it does best:

- **LLMs generate narrative elements:** location descriptions, character backstories, puzzle designs, and objective framing.
- **Symbolic Python structures maintain world state:** following PAYADOR's object-oriented world model; which locations connect to which, what items exist where, which puzzles block which passages, and whether the objective is solvable given the current world configuration.

So, the challenge is to generate worlds that are not only creative and thematically coherent, but also structurally valid and where players can complete the objective.

Goal-Oriented Architecture

PAYADOR worlds include objectives that players must complete, like finding items, or reaching locations. We recognized this concept could serve as the organizing principle for IVIE's generation pipeline: rather than generating disconnected elements and hoping they form a coherent world,

Generated World Opening
Crumbling brick buildings line the street. Fog rolls between the structures, and distant sounds echo through empty alleys. You are a detective investigating mysterious disappearances in the city.
Current Location: Abandoned Warehouse District
Accessible places: Police Station, Old Docks
Blocked passages: Evidence Room
Objective: Find the Missing Person
Items: Flashlight, Detective's notebook
NPC: Officer Martinez

Figure 1: Example world opening generated in Generate mode, showing goal-oriented structure where all elements serve the central objective.

the objective defines what must exist and how components relate. Every location, item, NPC, and puzzle exists to serve the player's goal, whether finding a hidden artifact, solving a mystery, or escaping a dangerous location. This architecture addresses common PCG challenges by providing the structural constraints necessary to avoid unplayable content (Bomström 2016) while maintaining coherent overarching structures across generated elements (Khaled, Nelson, and Barr 2013).

Every generated element exists to serve a central purpose to complete the objective. This objective defines world requirements: If the goal is "Deliver the ancient scroll to the librarian", the world must contain the scroll as a gettable item, a librarian character positioned in some location, a path connecting the scroll's location to the librarian's location, and potentially puzzles blocking access to either element. Therefore, the generation pipeline works backwards from this objective, determining what elements does this goal require and where should they be placed. Figure 1 shows a generated world opening where all elements are oriented toward the central objective of finding a missing person.

Incremental Generation Pipeline

In preliminary experiments, we found that attempting to generate complete worlds in a single LLM call resulted in inconsistencies, as models struggled to maintain coherence across dozens of interdependent elements. Additionally, LLMs would generate worlds of unpredictable sizes, making systematic evaluation and comparison difficult. To address both issues, we designed a four-stage **incremental generation** pipeline (illustrated in Figure 2) with configurable size parameters. IVIE allows designers to specify the number of `Locations`, `Items`, `Characters`, and `Puzzles` before generation begins, ensuring consistent world scale for controlled evaluation while enabling validation systems to scale appropriately. The pipeline builds worlds progressively across four mandatory stages, from abstract to concrete, with symbolic validation at each boundary. Each stage uses a specialized prompt to guide LLM generation; these prompts are available in the source code repository.

Stage 1: Adventure Core This first stage establishes the narrative foundation by generating three core components: world theme, protagonist description, and objective type.

IVIE extends PAYADOR's objective system (Góngora 2025) with five objective types, each imposing different structural requirements on subsequent stages:

- `REACH_LOCATION`: Player must navigate to a specific location (e.g., "Escape the haunted mansion")
- `GET_ITEM`: Player must obtain a particular item (e.g., "Retrieve the ancient artifact")
- `DELIVER_ITEM`: Player must bring an item to a character or location (e.g., "Deliver the scroll to the librarian")
- `FIND_CHARACTER`: Player must locate a specific NPC (e.g., "Find the missing explorer")
- `SOLVE_MYSTERY`: Player must gather clues to solve a mystery (e.g., "Discover who stole the crown jewels")

These components can be generated in two modes. In *Inspiration mode*, the LLM receives user-provided thematic guidance (e.g., "mystery in a Victorian mansion") and generates a world aligned with that theme. In *Generate mode*, the LLM fully takes creative responsibility (Colton and Wiggins 2012), autonomously inventing theme, setting, and narrative premise without user constraints.

Output: `WorldConcept` object with adventure `title`, narrative `backstory`, `player_concept` (character role), and `main_objective` (high-level goal description). The formal objective type is determined in subsequent stages.
Validation: This stage does not require validation beyond schema correctness, as the conceptual elements (title, backstory, player concept, objective description) are `Strings` (purely narrative) and do not impose structural constraints that could render the world unplayable. Any thematically coherent concept is acceptable as input for subsequent stages.

Stage 2: World Structure Once the narrative core is established, this stage identifies the main entities necessary to complete the objective: key `locations`, relevant `characters`, and required `objects`. The LLM generates entity identifiers (names) and brief statements of how each relates to the objective, but without spatial connections, descriptions, or properties. For instance, it might specify "Ancient Library" and "Hidden Scroll" as entities, noting the scroll must be found in the library in its description, but not describe what the library looks like, where it connects to, or whether the scroll is locked away. This generated structure works as a world structure to be completed later. In preliminary experiments, we observed that the LLM tended to generate thematically coherent but functionally irrelevant entities (e.g., decorative furniture, background NPCs with no gameplay purpose). To mitigate this, we modified the generation schema to require a `relevance_to_objective` field for each entity, forcing the LLM to justify its functional role explicitly. This ensures every generated element serves the player's path towards objective completion.

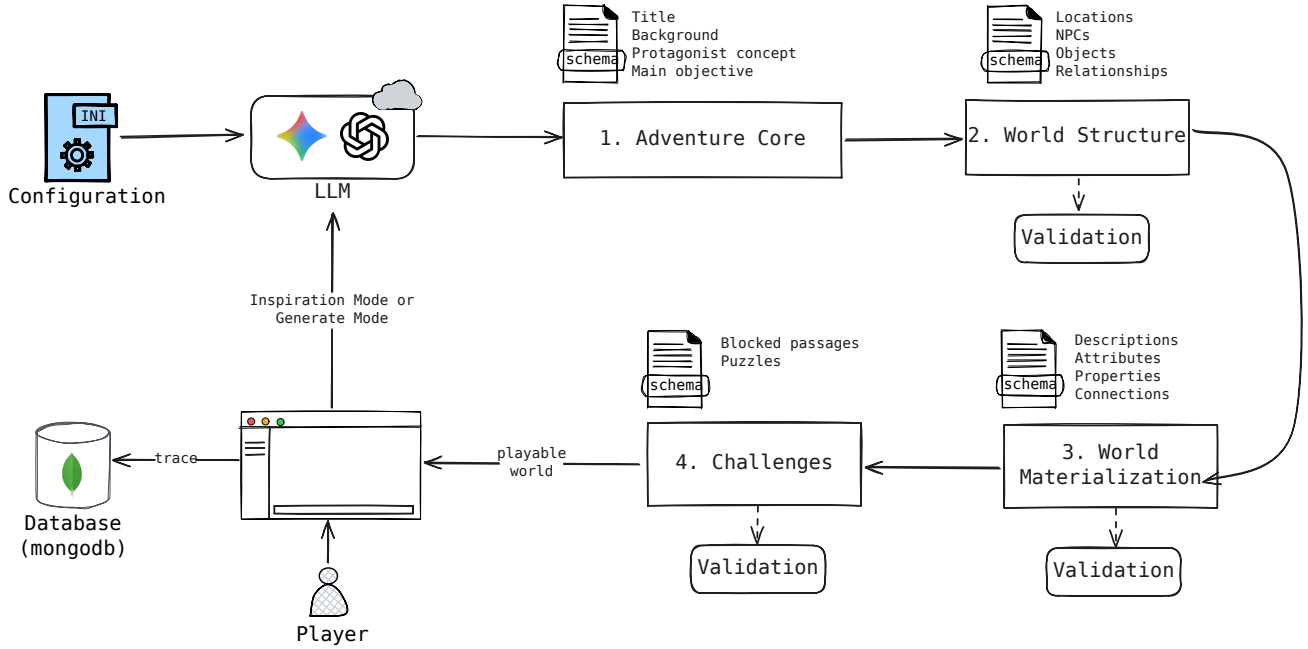


Figure 2: IVIE's incremental generation pipeline. Starting from optional user inspiration, the LLM generates worlds progressively across four mandatory stages (Adventure Core → World Structure → World Materialization → Challenges), each with validation gates.

Output: `GeneratedWorld` object containing abstract elements (`Locations`, `Characters` and `Items` without details).
Validation: Schema validation (Pydantic) ensures the output conforms to the expected data structure with correct types and mandatory fields. Size verification confirms if the LLM generated the configured number of elements.

Stage 3: World Materialization This stage completes the world structure from Stage 2 by populating it with the attribute values required by PAYADOR's world model: `Locations` receive textual descriptions and spatial connections, `Characters` are assigned descriptions and starting locations, `Items` are given names and gettability flags, and passages between `Locations` are established. With entity identifiers already established in Stage 2, the prompt becomes significantly more specific: it explicitly assigns positions to all `Characters` and `Items`, and instructs the model not to introduce new entities beyond the existing world structure. These constraints address the previously mentioned problem of thematically coherent but functionally disconnected elements, ensuring every entity has a defined spatial position and all `Locations` are interconnected.

Mandatory connection rules ensure spatial coherence: (1) if location A connects to B, then B must connect to A (avoiding accidental unidirectional connections that could trap the player), and (2) there must be no isolated locations. The objective must be completable with the created elements, with at least one logical route from the initial state to its achievement.

Output: Fully instantiated playable `GeneratedWorld` object.
Validation: Spatial connectivity is verified via depth-first search (DFS), ensuring every `Location` is reachable from the starting location. Objective completability checks vary by type:

- `GET_ITEM`: Verifies the target item is both reachable and marked as gettable.
- `REACH_LOCATION`: Verifies the destination is accessible from the start.
- `DELIVER_ITEM`: Verifies both the item and the recipient are reachable and the item is obtainable.
- `FIND_CHARACTER`: Verifies the target NPC exists, and their location is reachable.
- `SOLVE_MYSTERY`: Verifies all clue items exist in the world and are associated with valid `Items` that have assigned `Locations`.

Stage 4: Challenges This stage modifies the base world from Stage 3 by adding complexity: puzzles with progressive hint systems and blocked passages between `Locations`. The LLM receives the complete world structure (all `Locations`, `Items`, `Characters`, and spatial connections) and generates puzzle definitions that specify which passages they block, what solutions unlock them, and hints to guide players toward solutions.

Initially, we attempted to have the LLM generate puzzles where solving one would unlock the next in a specific order, but the model consistently failed with circular references and structural errors, so we simplified the approach. Instead of enforcing a strict sequence in the schema, we prompt the LLM to suggest a "logical progression" where each challenge narratively builds on previous ones. This appeared to be more effective, as puzzles feel connected without the structural complexity that caused failures.

The generation prompt specifies concrete constraints: each puzzle must connect to either a blocked passage or an essential resource; obstacles and their solutions must be physically separated across different `Locations`; blocked passages require defined unlock conditions (e.g., obtaining a key, solving a riddle); and solutions must be discoverable through exploration rather than requiring external knowledge.

Each puzzle includes LLM-generated hints organized in a progressive disclosure structure. Hints have 3 levels, starting with general orientation (e.g., "The answer lies in the study") and progressing to nearly explicit guidance (e.g., "Examine the portrait of Lady Ashwood"). Additionally, each puzzle includes interaction hints suggesting how to begin (e.g., "Try talking to the librarian" or "Examine the locked chest closely"). This hint structure is critical for player assistance, as we discuss in the evaluation section.

Puzzles can be triggered by `Characters` (e.g., an NPC asks the player to solve a riddle) or by `Locations` (e.g., examining a room reveals an environmental puzzle). The LLM can strategically redistribute `Items` to adjust difficulty but cannot create new `Locations` and must preserve spatial connectivity.

Output: Modified `GeneratedWorld` object from Stage 3. The world now includes `Puzzles` with solutions, progressive hints (3-5 levels), and interaction hints; `BlockedPassages` that lock connections between `Locations` until puzzle completion; and potentially redistributed `Items` to create puzzle dependencies.

Validation: Schema validation ensures all puzzle fields are present and correctly typed. Puzzle reward verification checks whether puzzle rewards reference existing `Items`. If a puzzle grants a non-existent item as reward, the system automatically creates that `Item` and places it in the puzzle proposer's inventory (for character-proposed puzzles) or in the relevant `Location` (for environmental puzzles). Finally, spatial connectivity (DFS) and objective completability are revalidated to ensure puzzles and blocked passages do not prevent the player from completing the main objective.

Validation Strategy: Correction vs. Rejection

Each generation stage may fail validation. However, not all validation failures require complete regeneration. IVIE distinguishes between *correctable* issues and *structural failures*. Correctable issues include missing bidirectional connections between `Locations` or type mismatches in fields—problems that can be automatically fixed without discarding the generated content. Structural failures include unreachable `Locations` (detected via DFS) or unsolvable objectives (e.g., required `Items` without assigned positions)—problems that fundamentally break world playability and require regeneration of that stage with enhanced prompts explaining the specific violation.

This distinction aims to balance robustness (rejecting genuinely broken worlds) with efficiency (avoiding unnecessary regeneration when simple corrections suffice). For correctable issues, IVIE applies automatic fixes and proceeds; for structural failures, it provides the LLM with detailed error descriptions to guide regeneration.

Originally, PAYADOR used regex parsing to extract structured data from LLM text responses, which appeared brittle when models generated unexpected formatting. IVIE addresses this by using Pydantic[2] models that define strict schemas for each generation stage. Modern LLM APIs enforce these schemas during generation (JSON schema-guided decoding), guaranteeing syntactically valid output. Pydantic then performs a second validation pass, checking that generated content is semantically meaningful—catching issues like missing required fields or invalid `enum` values that pass syntactic checks but would break downstream processing.

For example, the `GeneratedLocation` model enforces:

```
class GeneratedLocation(BaseModel):
    name: str
    connections: List[str]
    blocked_passages: List[BlockedPassage]
    relevance_to_objective: str
```

If the LLM attempts to omit `connections` or provide a non-list value, Pydantic raises a validation error. The system attempts regeneration up to three times with enhanced prompts that explicitly describe the validation failure. If all attempts fail, the stage is marked as failed and generation halts, requiring manual intervention or parameter adjustment.

RAG-Based Memory System

During gameplay sessions using PAYADOR, each turn receives only the current player action and immediate world state, without access to previous interactions unless explicitly included in the prompt. Moreover, including full conversation history without filtering introduces irrelevant or outdated context that may confuse the model rather than help it. This limitation causes narrative inconsistencies: NPCs may forget previous conversations with the player, or the narration may contradict events from earlier turns.

To address this, IVIE implements Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) with episodic memory. The system stores each gameplay turn as an `AtomicMemory` unit containing the player action, system-generated narration, and a brief world state summary. For each memory, the system generates 768-dimensional semantic embeddings using Gemini's embedding model[3] and per-

[2] `https://docs.pydantic.dev/`

[3] Google Gemini Embedding API: `https://ai.google.dev/gemini-api/docs/embeddings`

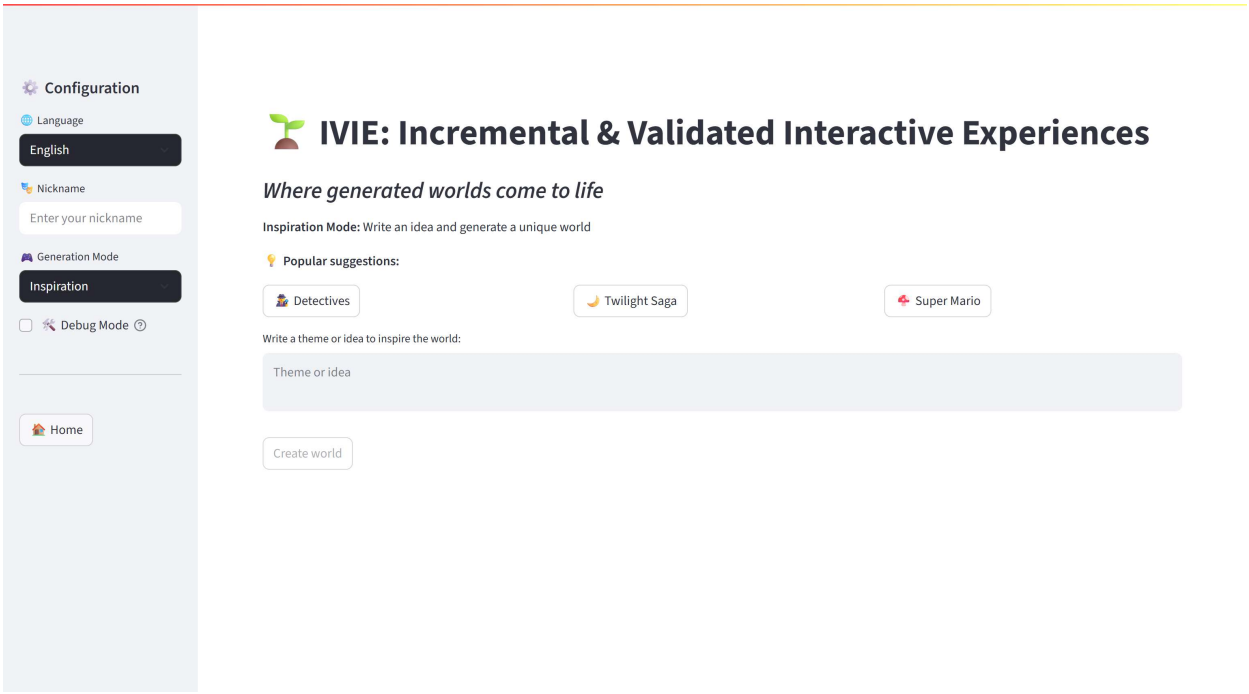


Figure 3: IVIE's interface allows users to generate worlds in two modes: Generate (the LLM takes the creative responsibility) or Inspiration (based on thematic input). Configuration options include language, player nickname for identification, and debug mode for inspecting world structure and detecting potential issues.

sists them in ChromaDB[4], which maintains one collection per generated world.

During gameplay, before generating the narration for each new action, the system retrieves the top-3 most semantically similar past memories based on embedding cosine similarity. These retrieved memories are injected into the narrative generation prompt in a dedicated section titled "Relevant Past Memories," providing the LLM with contextual information about related past events. After the LLM generates narration for the current turn, that turn is stored as a new `AtomicMemory` and added to the database for future retrieval.

Semantic embeddings enable retrieving conceptually relevant memories even when they share no lexical terms with the current action. For instance, a player input "examine the portrait" might retrieve a memory where an NPC mentioned the portrait's subject, even though "portrait" wasn't in that earlier conversation. This can be valuable for narrative coherence, as the LLM can naturally reference past events to create a sense of continuity. Evaluators with RAG enabled reported significantly better long-term narrative coherence, with NPCs remembering previous interactions across dozens of turns, as we discuss in the evaluation.

Persistence and Reproducibility

During development and evaluation, we encountered the critical need to understand how generated worlds behave across complete gameplay sessions. When players reported unexpected behavior, or worlds were not completable, we needed to examine the exact sequence of events: what the player did, how the LLM responded, and how the symbolic world state changed at each step. Additionally, we wanted to enable players to replay interesting worlds from the beginning, exploring different narrative paths within the same generated structure. To address this, IVIE maintains a persistent turn log stored as a JSON array, where each entry captures the full state of that interaction.

Each turn in the array captures: the player's input text, the complete symbolic world state (which `Locations` are accessible, where `Items` are located, which puzzles are solved), the LLM's generated narration, and what elements of the world state were modified (e.g., player moved to new `Location`, item added to inventory, puzzle marked as solved). Turn 0 specially captures the complete initial world state generated by the pipeline, enabling exact reconstruction of the starting conditions. Subsequent turns capture only state changes rather than complete snapshots, reducing storage requirements while maintaining full traceability.

Persisting these generated worlds allow several capabilities: replaying worlds from Turn 0 to explore alternative narrative paths within the same structure, analyzing LLM narrative patterns across complete sessions offline, debugging generation failures by examining the exact state at each turn, and comparing how different players navigate the same generated world. For instance, when evaluators reported incomplete worlds during evaluation, we used the replay ca-

[4] `https://www.trychroma.com/`

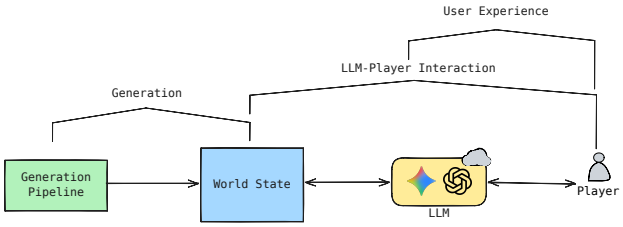


Figure 4: Three key dimensions evaluated: World Generation, LLM-Player Interaction, and Player Experience.

Dimension	Generate	Inspiration
Puzzle logic	3.88	4.00
Hints clarity	4.12	4.12
Objective clarity	4.88	4.62
Objective completion	4.38	3.50
NPC purpose	3.50	3.88
Entertainment	4.50	4.38

Table 1: Mean Likert scores (1–5) per dimension across generation modes (N=8).

pability to load those exact sessions and analyze both the world state and turn history to understand why these failures occurred.

Users interact with all these components through IVIE's web interface (Figure 3), which provides configuration options for generation modes, language selection, and a debug mode that can help visualize the internal world structure and identify potential issues without playing.

Evaluation

We evaluated IVIE through human gameplay sessions across three key dimensions (Figure 4): *world generation quality* (structural validity, parameter adherence, spatial connectivity), *LLM-player interaction effectiveness* (puzzle difficulty, objective clarity, narrative coherence), and *overall player experience* (engagement, immersion, thematic coherence).

Experimental Design

Eight evaluators with varying gaming experience completed three gameplay sessions. All participants took part voluntarily and evaluation data was anonymized prior to analysis:

1. **Tutorial:** A pre-designed world to familiarize participants with the interface and mechanics.
2. **Generate Mode:** A world generated with complete LLM creative freedom.
3. **Inspiration Mode:** A world generated based on the participant's favorite IMDB-listed movie.

We distributed participants evenly across experimental conditions to ensure balanced comparison:

- **Language:** 4 Spanish, 4 English
- **Memory System:** 4 with RAG enabled, 4 without RAG
- **LLM Provider:** 4 using Gemini 2.0 Flash, 4 using GPT-4o-mini

As previously mentioned, IVIE allows configuration of world-size parameters; in this evaluation, all worlds used identical configurations: 4 `Locations`, 4 `Items`, 2 `Characters` (NPCs), and 1 `Puzzle`.

After playing, participants completed 5-point Likert scale questionnaires (Yes/Mostly/Sometimes/Almost Never/Never) covering puzzles, objectives, world coherence, and overall experience, plus open-ended qualitative feedback. The complete evaluation data, including Likert scale responses, full survey questions, and participant gameplay sessions, are available in the source code repository. Table 1 summarizes mean Likert scores per dimension across both generation modes.

Players achieved 100% objective completion under Generate mode (8/8 worlds) but only 50% under Inspiration mode (4/8 worlds). These Inspiration mode failures revealed critical validation gaps: two worlds passed all validation stages despite having structurally impossible objectives: the required `Items` existed in the world model but had no assigned `Location`, making them inaccessible to the player. A third failure resulted from API quota exhaustion mid-session[5], which caused the session to terminate abruptly. Finally, one failure resulted from a persistent LLM reasoning error: despite repeated correct attempts by the player to request an item from an NPC, the reasoning model consistently failed to generate the corresponding transfer transformation, leaving the item in the NPC's inventory and making the objective seem unreachable.

Out of 16 generated worlds, 13 (81.25%) matched configured parameters exactly, while three worlds (18.75%) exceeded limits, all in English. Spatial connectivity validation (DFS) succeeded in all cases, with no player becoming isolated in unreachable `Locations`. Evaluators consistently praised spatial coherence, with one noting that: "transitions between locations were correct, and narration described surroundings very well.". Across the 7 evaluated Generate mode worlds, we observed considerable thematic diversity (e.g., fantasy, steampunk, Egyptian mythology, space archaeology, post-plague alchemy) though recurring archetypes emerged (e.g., scholar protagonists, ancient artifact recovery), suggesting that while surface themes vary, LLMs tend to converge on familiar narrative structures. Inspiration mode worlds, by contrast, showed greater narrative diversity, likely driven by the diversity of the source films provided by evaluators.

Evaluators in *Inspiration mode* consistently reported that the generated worlds strongly reflected the thematic essence of their chosen inspiration movie. We therefore concluded that *Inspiration mode* achieved 100% thematic coherence across all evaluated worlds. In contrast, while *Generate mode* received positive feedback on overall creativity, evaluators reported lower perceived utility of generated elements: only 25% of *Generate mode* evaluators felt all objects and NPCs had a "real purpose" in achieving the objective, com-

[5]We used Google Gemini's free tier during evaluation, which has daily request limits.

pared to 71% in *Inspiration mode* . The thematic constraint appears to integrate elements more tightly with objectives, reducing the presence of purely decorative content. We think this hints at a trade-off: *Inspiration mode* produced stronger thematic coherence yet lower completion rates, which might suggest that tighter thematic constraints increase world complexity in ways that expose validation gaps more frequently.

Approximately 75% of evaluators found the generated puzzles logical with useful hints, but 25% encountered overly cryptic puzzles requiring specialized knowledge or puzzles with internally inconsistent hints. However, in 3 out of 16 worlds, players were able to bypass puzzles by claiming they had solved them when providing incorrect or no solutions (e.g., typing "I solve the puzzle" without actually answering the puzzle). The LLM's reasoning model accepted these claims as valid actions—a phenomenon similar to jailbreaking (Wei, Haghtalab, and Steinhardt 2023). This issue was previously observed in PAYADOR's original implementation (Góngora et al. 2026), where the reasoning model could suggest state changes (e.g., marking a blocked passage as accessible) that the symbolic engine would accept even when prerequisites weren't met. This seems to be a fundamental challenge for neuro-symbolic interactive storytelling: validating too strictly may constrain the LLM's generative freedom, while validating too loosely may allow players to bypass puzzle logic entirely.

Regarding the incorporation of RAG for memory, we found that the 4 evaluators with RAG enabled experienced significantly better long-term narrative coherence. For instance, in Evaluator 7's gameplay session, an NPC remembered a successful item theft during turn 17 and referenced it when re-encountered on turn 44 ("Have you come to try and claim what is not yours again?"), suggesting semantic memory retrieval across dozens of intervening turns. However, RAG occasionally mixed contexts and presented puzzles from past interactions instead of the current world state. We believe further exploration of RAG in procedurally generated interactive narratives could address these context-mixing issues through better memory segmentation, for example, organizing memories by type (e.g., location-specific events, NPCs, puzzles) and retrieving only contextually relevant categories based on the current game state. Evaluators with RAG enabled rated entertainment higher on average (4.75 vs. 4.25 without RAG). However, due to the small sample size (N=4 per condition), these results are not included in the table; a more in-depth evaluation would be needed to determine whether RAG has an actual impact on player experience.

Regarding perceived creativity, evaluators responded positively to open-ended questions about world originality in Generate mode, describing worlds as original and engaging to explore. In Inspiration mode, evaluators highlighted the fidelity to source material as a creative achievement in itself, with one noting that the result "came out of the ordinary and expected." We acknowledge, however, that this evaluation does not measure creativity in formal terms nor does it compare against a non-creative baseline.

Despite technical issues, players found sessions entertaining. Notably, although worlds were small with anticipated 15-25 minute sessions, multiple evaluators played for over an hour, prioritizing narrative exploration over objective completion. Overall, we think this behavior hints at a high level of immersion.

Conclusions and Future Work

We presented IVIE, a neuro-symbolic system that automatically generates complete and playable IF worlds by strategically delegating creative decisions to LLMs while grounding world coherence through symbolic validation. We hope this work represents a step toward addressing the scalability challenges of IF authoring: while manual world design offers artistic control and crafted experiences, automatic generation enables rapid prototyping and increased replayability through procedural variation. IVIE achieves this through a four-stage incremental generation pipeline with symbolic validation at each stage. Unlike purely LLM-based approaches that rely on context window history, IVIE maintains an explicit symbolic world state, enabling validation of structural correctness and prevention of common inconsistencies (such as vanishing objects, contradictory spatial relationships, or unsolvable objectives).

Evaluation results suggest that symbolic validation mechanisms (including DFS for spatial connectivity, Pydantic for type checking, and objective completability verification) can prevent the logical inconsistencies inherent to pure LLM systems while preserving narrative flexibility. Generate mode achieved 100% completion with high player engagement (sessions exceeding one hour despite small 4-location worlds), and Inspiration mode suggests that 100% thematic coherence, proving LLMs can generate goal-oriented worlds when properly constrained.

We think IVIE's results offer insights applicable beyond IF to any domain requiring creative generation within structural constraints: the four-stage incremental pipeline with intermediate validation steps suggests that decomposing generation into bounded stages — each with a defined validation contract — may be a viable strategy for grounding open-ended LLM generation in domains where structural correctness matters.

Several directions could significantly improve the approach. Currently, when objective validation fails (e.g., required `Items` exist but have no assigned `Location`), the system regenerates the entire stage up to three times before aborting. A more robust approach would automatically correct these issues—detecting unreachable items and assigning them valid positions—similar to how the system currently handles missing puzzle rewards, reducing generation failures and avoiding discarding otherwise playable worlds.

As discussed in the evaluation, the RAG implementation could benefit from organizing memories by type (`location` events, `Character` interactions, `Puzzle` attempts) to address the context-mixing issues we observed. Alternative, specific retrieval techniques for RAG could further improve memory relevance as well. Additionally, some evaluators encountered overly cryptic puzzles requiring specialized knowledge; implementing an LLM-based difficulty classifier that analyzes puzzle descriptions and hints could

identify and regenerate puzzles exceeding reasonable difficulty thresholds.

Also from evaluation feedback, several evaluators expressed interest in deeper NPC interactions. Growing research on LLM role-playing and persona assignment (Tseng et al. 2024) suggests that extending `Character` schemas with explicit personality fields (e.g., cooperative, hostile, enigmatic) could enable richer dialogue and more predictable characterization; currently, NPC behavior depends entirely on LLM generation at each turn without explicit personality guidance. Enabling mid-session save/resume would address a practical need: currently, API failures force players to restart worlds entirely, and sessions cannot span multiple play periods. Implementing true save/resume functionality would require extending the system to initialize world state from arbitrary turn numbers, not just the initial state.

Finally, while IVIE theoretically supports any LLM provider, we only evaluated the approach using commercial APIs (Gemini and GPT-4o-mini). Investigating open-source models (e.g. Llama, Mistral or Qwen) would reduce dependency on proprietary services and enable fully local execution. Exploring specialized smaller models for individual pipeline stages—rather than using one general model for all—could improve both efficiency and generation quality while maintaining local deployability. Beyond generation, IVIE's symbolic world state could enable runtime improvisation: dynamically introducing NPCs, items, or narrative events during gameplay to assist struggling players or increase dramatic tension, similar to how players adapt Tabletop Role-playing Games scenarios in real-time (Martin, Harrison, and Riedl 2016; Góngora et al. 2024b). This would extend IVIE beyond world generation toward real-time narrative improvisation.

Author Contributions

M.V. and S.S. designed, developed, and conducted the human evaluation of IVIE as part of their Bachelor's thesis in Computer Science Engineering (Vaucher and Silveira 2025). S.G. co-advised that thesis with L.C.; both are also authors of the PAYADOR approach, which IVIE builds upon.

References

Bomström, H. 2016. The application of procedural content generation in video game design. Bachelor's thesis, University of Oulu, Department of Information Processing Science.

Callison-Burch, C.; Patel, A.; Dennis, J.; and Zhu, A. 2024. Dagger: Data augmentation for generative gaming in enriched realms. In *The 4th Wordplay: When Language Meets Games Workshop*.

Castricato, L.; Frazier, S.; Balloch, J.; Tarakad, N.; and Riedl, M. 2021. Automated story generation as question-answering.

Chambers, R.; Tack, N.; Pearson, E.; Martin, L. J.; and Ferraro, F. 2024. Berall: Towards generating retrieval-augmented state-based interactive fiction games. In *The 4th Wordplay: When Language Meets Games Workshop*.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *ECAI 2012*. IOS Press. 21–26.

Góngora, S.; Chiruzzo, L.; Méndez, G.; and Gervás, P. 2024a. PAYADOR: A minimalist approach to grounding language models on structured data for interactive storytelling and role-playing games. In *Proceedings of The 15th International Conference on Computational Creativity*.

Góngora, S.; Chiruzzo, L.; Méndez, G.; and Gervás, P. 2024b. Skill check: Some considerations on the evaluation of gamemastering models for role-playing games. In Dondio, P.; Rocha, M.; Brennan, A.; Schönbohm, A.; de Rosa, F.; Koskinen, A.; and Bellotti, F., eds., *Games and Learning Alliance*, 277–288. Cham: Springer Nature Switzerland.

Góngora, S.; Chiruzzo, L.; Méndez, G.; and Gervás, P. 2026. World-state transformations for neuro-symbolic interactive storytelling. In *Proceedings of The 17th International Conference on Computational Creativity*.

Góngora, S. 2025. Approaches to interactive and improvisational storytelling. Master's thesis, Universidad de la República (Uruguay).

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55(12).

Kelly, J.; Calderwood, A.; Wardrip-Fruin, N.; and Mateas, M. 2023. There and back again: Extracting formal domains for controllable neurosymbolic story authoring. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, volume 19, 64–74.

Khaled, R.; Nelson, M. J.; and Barr, P. 2013. Design metaphors for procedural content generation in games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 1509–1518. New York, NY, USA: Association for Computing Machinery.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 9459–9474.

Li, Z.; Wang, Z.; Xiong, C.; and Xing, C. 2025. Learning to play like humans: A framework for llm adaptation in interactive fiction games. *arXiv preprint arXiv:2505.12439*.

Martin, L. J.; Harrison, B.; and Riedl, M. O. 2016. Improvisational computational storytelling in open worlds. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*, 73–84. Springer.

Montfort, N. 2003. *Twisty Little Passages: An Approach to Interactive Fiction*. MIT Press.

Riedl, M. O., and Bulitko, V. 2012. Interactive narrative: An intelligent systems approach. *AI Magazine* 34(1):67.

Riedl, M. O. 2015. An introduction to ai story generation. *Medium*. Accessed: 2025-11-02.

Sarker, M. K. H.; Zhou, L.; Eberhart, A.; and Hitzler, P.

2021. Neuro-symbolic artificial intelligence: Current trends. *AI Communications* 34(3):197–209.

Sawyer, R. K. 2000. Improvisational cultures: Collaborative emergence and creativity in improvisation. *Mind, Culture, and Activity* 7(3):180–185.

Schiaffini, G. 2006. Never improvise improvisation. *Contemporary Music Review* 25(5-6):575–580.

Shaker, N.; Togelius, J.; and Nelson, M. J. 2016. *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*. Cham, Switzerland: Springer.

Trapova, A. 2023. Ai as a vehicle for creativity in video games – any room for flexibility via contracts? a case study on ai dungeon. In Vanhees, H., and Geiregat, S., eds., *Copyright Contracts*. LeA Uitgevers.

Tseng, Y.-M.; Huang, Y.-C.; Hsiao, T.-Y.; Chen, W.-L.; Huang, C.-W.; Meng, Y.; and Chen, Y.-N. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 16612–16631. Miami, Florida, USA: Association for Computational Linguistics.

Vaucher, M., and Silveira, S. 2025. IVIE: Generación incremental y validada de mundos jugables para experiencias interactivas. Tesis de grado, Universidad de la República (Uruguay). Facultad de Ingeniería.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Yong, Q. R., and Mitchell, A. 2023. From playing the story to gaming the system: Repeat experiences of a large language model-based interactive story. In *International Conference on Interactive Digital Storytelling*, 395–409. Springer.

Zhou, E.; Basavatia, S.; Siam, M.; Chen, Z.; and Riedl, M. O. 2025. Story2game: Generating (almost) everything in an interactive fiction game. *arXiv preprint arXiv:2505.03547*.

Snap, Crackle and ... *Slop?* Imbuing Generative AI with the Essence of Computational Creativity

Tony Veale
School of Computer Science,
University College Dublin,
Dublin D4, Ireland.
tony.veale@UCD.ie

Abstract

Recent remarkable advances in generative AI (GenAI) have not sated consumer demand for new tools or new features, nor have they quelled our suspicions that machines are simultaneously capable of great virtuosity and tremendous stupidity in the same act of synthetic creation. The World Wide Web is now awash in seemingly human-quality texts, images and videos, but small mistakes that were once considered quirky and charming in the context of computational creativity (CC) are now seen as reasons to dismiss otherwise proficient outputs as "slop." Why the same disdain was not shown to the products of CC systems which rarely reach, in objective terms, the same levels of versatility and proficiency as those of GenAI systems, is the central question of this paper. It is a question that goes to the heart of what makes a creator and their creations truly "creative" rather than merely generative. The answer, as we present it here, offers hope for CC as a continuing research endeavour with a distinctive character of its own.

Introduction

If familiarity breeds contempt, as the old saying goes, this is certainly true of *Generative AI* (GenAI), a family of neural models that includes Large Language Models (Matarazzo and Torlone 2025) and diffusion-based image or video generators (Shen et al. 2025). As GenAI's outputs become ubiquitous, an initial sense of wonder has given way to a growing sense of dismay. Our willingness to criticize those outputs as soulless or inept has more than kept pace with new developments in the field, even if our criticisms are not always fair or aimed at the right targets. Rather, with greater fluency and versatility come higher expectations, so that many GenAI artifacts are dismissed as "slop" (Kommers et al. 2025), even when they exhibit a level of virtuosity that might take one many hours to achieve with more conventional digital tools.

Those outputs are sometimes flawed in tell-tale ways: a six-fingered hand here, a third arm there, or objects that unintentionally flirt with surrealism, such as a typewriter with keys on its front *and* its back. But these errors, uncanny as they are, seem little reason in themselves to dismiss a complex artifact as "slop." Assessed objectively, the products of past CC systems were often more error-prone and less likely to pass for human-level creativity, yet those failings rarely attracted such scorn. It seems that the label "slop" has less to do with the inherent failings of a machine-made artifact and more to do with the mind-sets of those who produce and consume it (Kommers et al. 2025). While CC systems rarely develop beyond the lab to make a lasting dent on public consciousness, consumers have eagerly embraced GenAI as a panacea for many kinds of generative work. The fear that AI will displace and replace human creators, merely inchoate in our CC systems, has ripened to fruition with GenAI. It has never been easier to automate the creation of human-quality artifacts, and never more tempting to let the machine do all of our thinking for us (Orwell 1946). Whenever we deride an AI artifact as slop, our proximate target is the artifact itself, which may not exhibit any of the tell-tale flaws of automated generation, but our real target is the person who brought it into being, as well as those who provided the tools to do so.

Let's take a moment to unpack the slop metaphor (Øivind Madsen and Puyt. 2026). Some liquids are inherently 'slop', such as the contents of a chamber pot or a pig trough, while others become slop through careless handling, as when pigment is slopped on a wall or canvas by an indifferent painter. It is the latter sense that applies to the outputs of generative AI. The result need not be low quality, just low effort, since this lack of effort lends itself to rapid production, lazy filtering, and over-sharing. Simply, the producer has not invested enough of themselves, of their time and energy, to make the artifact worthy of *our* time and *our* energies. It is this investment that leads us to label the flawed outputs of our CC systems as quirky or uncanny, but to dismiss the competent mediocrity of many GenAI outputs as indiscriminate slop.

As users of GenAI, how should we make this CC-like investment of time and creative energy in what our systems produce? It is this sweat equity that allows us to share in the creativity of the system itself, to be able to say not just that "I prompted for this" but that "I helped build that." All CC systems are meta-creative, and to an extent co-creative too, since they automate our personal view of the creative process in a given domain. The avoidance of slop is not a matter of preventing the generation of low-quality artifacts, as creativity always carries a risk of failure. Interesting mistakes are rarely "slop." Such failures show us the limits of our understanding and allow us to improve our systems. To put more of ourselves as meta- and co-creators into a CC system with GenAI as its generative core, we must build a framework that lets us share in the system's successes and

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

its failures. What is required is a marriage of CC and GenAI that gives us the best of both paradigms in a single system.

To illustrate this philosophy, in more ways than one, we explore here the creation of topical comic strips by pairing a GenAI black box, an LLM, with a CC glass box, a symbolic grammar for visual slapstick that gives shape to the actions of the LLM. This pairing offers a mode of co-development akin to how we humans collaborate creatively on rich multimodal artifacts such as comics, films and games. As we will show, it mirrors a mode of collaboration from the comics industry, the *Marvel Method*, that allows creators with distinct skill-sets to work together toward an emergent whole.

The Marvel Method

The myth of the lone creator, toiling alone in an attic, garage or lab, is stubbornly pervasive in our narratives of creativity (Boden 1992; Veale 2012; Burkus 2013). When it comes to commercial products, however, most are created by teams, not individuals. The inciting spark may well come from a single person, but it takes an ensemble of mindsets and skillsets to develop a multi-faceted result, as exemplified by the multimodal creativity of the comics industry. While this business has produced some famous auteurs, such as Georges Remi (Hergé), the creator of *Tintin*, most comics are created by a partnership of a writer and an artist. The writer takes charge of plotting and dialogue, the artist of layouts, pencil art and inking. In the early 1960s, as editor Stan Lee was launching a minor revolution at Marvel Comics, he found himself writing so many titles that he invented what is now called the *Marvel Method* (Ditko 1973; Lee and Buscema 1978; Lee and Mair 2002; Evanier 2008; Howe 2012) to streamline the collaborative process.

Lee perfected his method with trusted artists such as Jack Kirby and Steve Ditko. Lee and the artist would first spitball the gist of a new story, agreeing its main themes and events, before the artist would go off to illustrate the story, pencilling rough panel layouts into a rich visual narrative. Lee would then return to the story, to reinterpret the artist's visuals in the context of their agreed outline. In adding his own image-specific captions and dialogue, Lee was free to ignore the intentions of the artists, who often complained that they bore the bulk of the creative burden (Ditko 1973; Evanier 2008). However, the method allowed Lee to spread himself wide but not too thin, to shape a wide range of titles that would define Marvel's approach to visual narrative.

The Marvel Method is an approach to human co-creation, but it also provides a basis for co-creation by machines, or by agents operating within different AI paradigms, such as a rules-driven CC system and a statistical GenAI model. In comics production, one might produce the visual narrative, the other the textual narrative that turns those visuals into a finished comic strip. The method gives each agent maximal autonomy to employ their unique skill-sets. It also allows CC developers to invest the *sweat equity* that distinguishes producers of effortful artifacts from the generators of slop. While the GenAI partner, perhaps an LLM driven by clever prompting, remains a black box, a CC glass box can reflect the attitudes, insights and creative efforts of its creator. We might even re-purpose an existing CC system for this role, to provide scrutable insights into a creative process whose traceable inputs drive the inscrutable workings of an LLM.

The Marvel Method was not an uncontroversial mode of co-creativity, but its problems in a human setting become advantages in a wholly automated setting. For instance, credit for *who created or thought of what* is hard to attribute to a specific individual, sparking animosity between the public-facing Lee and the backroom artists who would ultimately part ways with him. In a computational setting, however, we prize any ambiguity that prevents us from pointing to any one sub-system or collaborator as the "true" creator. In fact, Lee's method encourages ambiguity at multiple levels, and even elevates ambiguity into a generative principle. Ambiguity in the rough plot sketch allowed artists to explore their own creative interpretations, while ambiguity in the finished art allowed Lee to impose his own script ideas, to the point of retro-fitting a new meaning onto the artist's renderings. Lee could take inspiration from the art itself, to furnish dialogue that explicitly referenced what was visible on the page.

When automating the creation of comic strips (Kurlander, Skelly, and Salesin 1996; Alves et al. 2007; Pérez y Pérez, Morales, and Rodríguez 2012), a diffusion-based image generator (Yang et al. 2024) may collaborate with a large language model (Patil and Gudivada 2024), the former assuming the role of the artist, the latter that of writer. These models work well together because they are each conditioned on text prompts. However, the roles of writer and artist can also be played by two different LLMs, to produce not a final visual product but a text specification of that product, which can then by visually rendered by a third model. In this paper we explore the automation of the Marvel method for comic strips with two different kinds of AI model: an old-school symbolic model to generate the rough outline of the plot and a specification of how it will be rendered (Veale 2022; 2023), and an LLM to write dialogue and captions to suit this visual rendering (Veale 2024). The key to the Marvel method is the order in which its steps are performed, with the artwork done first, and the writing – which is traditionally the first step – done later, as a response to the art.

The next section provides a more detailed view of this two-step, two-model approach. We will use an XML schema to specify each comic's visual and textual content (Walsh 2012), and focus on short comics that serve a useful purpose in themselves, of a kind that (Veale 2023) has called "serious" comics. As in (Veale 2023), our comics will present a balanced, two-sided perspective on a serious topic, such as immigration, artificial intelligence, vaccines, populism and so on, but the comics themselves will not be "serious." Rather, they will employ a visual slapstick that aims to add levity to an otherwise heavy debate, in which the distancing effect of irony preserves a sense of neutrality. The generation of visual humour is the responsibility of the CC module, and reflects a substantial investment of time and effort, and a personal creative sensibility, in the construction of a visual grammar for slapstick. In contrast, the writing of a script that imposes meaning on these antics in the context of a chosen topic is the responsibility of the LLM. So, in automating the Marvel Method, the LLM takes the role of Lee, while the visual grammar fulfills the responsibilities of Kirby and Ditko.

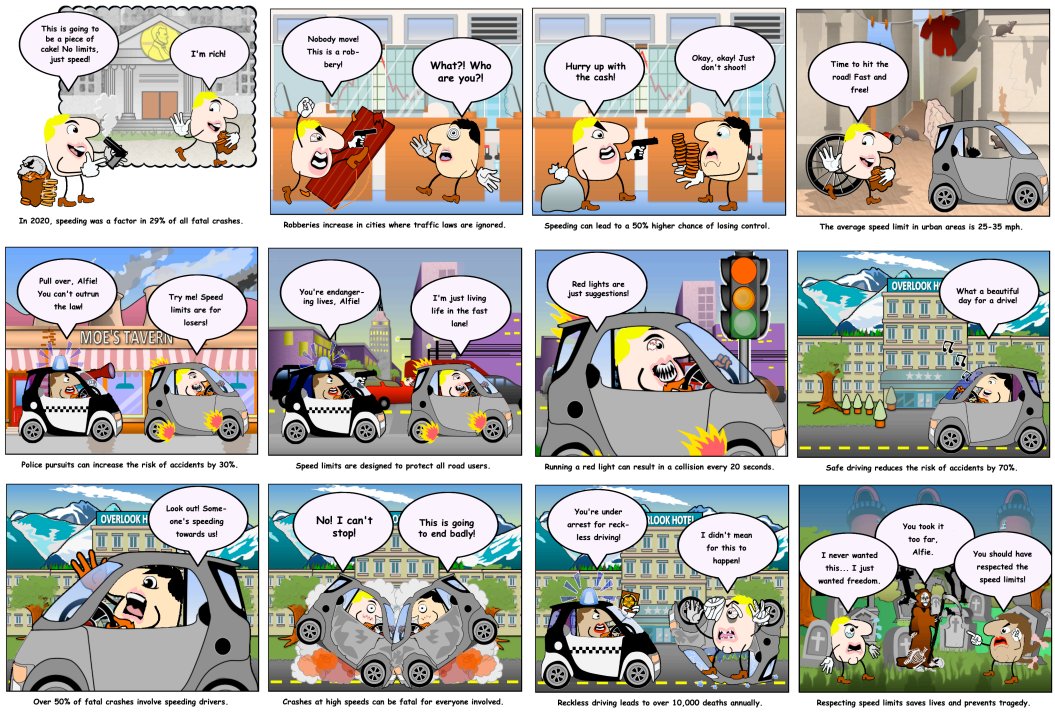


Figure 1: An *Excelsior!* comic on speed limits with LLM dialogue and captions, as rendered from XML.

In Two Minds: Two-sided Comics Generation

Comics are a malleable commodity. Comprising a sequence of panels separated by white spaces known as ”gutters”, a comic breaks a narrative into distinct beats in which one or more characters act, speak, think, and interact against a visual background (McCloud 1993; 2006; Cohn 2013). When represented in this form, a comic can easily be re-printed, re-sized, re-colored, re-structured, re-edited, and reordered. These actions are made even easier if we distinguish the formal specification of a comic narrative from its eventual rendering as a multimodal object. To this end, we will use an XML schema, *ComiXML*, to represent a machine-generated comic. Each instance of the schema first defines a comic's recurring figures, assigning each a name and visual qualities such as skin tone, hair colour and gender, before specifying each panel in turn, in the sequence they follow in the comic.

Each panel specifies a single narrative beat in which one or more characters are placed against a visual backdrop in a particular panel location (left, right, middle) with a set orientation (facing left or right). Each character may articulate dialogue in a panel, either in a speech balloon (spoken) or a thought bubble (internal), while captions above or below a panel present a narrator's perspective (Forceville, Veale, and Feyaerts 2010). Each character also takes on a pre-defined pose in a panel to express an action (falling, throwing, running, etc.) or an emotion (crying, laughing, accusing, etc.). *Excelsior!* has an inventory of more than 1000 visual poses, and over 500 backgrounds, allowing a rendering engine to map directly from the XML specification of a comic to the visual end-product, which may be shared as a web page, an animated GIF, or another format suited to digital sharing. An example of a fully-rendered comic is shown in Fig. 1.

The task of generating a comic now becomes one of XML generation, or, taking a step back, generating what will be encoded in this XML format. An LLM can be used to generate the narrative content that is so encoded, by first asking the LLM to generate a high-level plot for a story, and subsequently asking the LLM to break this plot into a sequence of panel-sized beats of a single action apiece. For each beat, the LLM is then prompted for suitable dialogue for each participating character; it is also tasked with suggesting apt poses for these characters from its inventory, given both the plot and what the character is saying or doing at this point in the narrative (Veale 2024). Finally, these strands are woven together into a single XML structure, placing each piece of text generated by the LLM (dialogue, captions, poses) into its schematic setting. This straightforward process intuitively models the *write first, draw later* approach to comics.

However, this approach is prone to *hallucination* (Kalai et al. 2023), since LLMs may suggest backgrounds or poses from outside the available inventory and that a renderer cannot render. We have to treat the LLM's suggestions not as literal asset names but as starting points for asset selection. We can, for instance, map each LLM suggestion to the semantically nearest asset, by estimating the semantic distance between this text (e.g., “pointing angrily”) and each asset label (e.g. “accusing”, “pointing”). Semantic distance can be esti-

mated using the cosine similarity of the vector encodings of the texts to be compared, provided each text can be reduced into a high-dimensional vector embedding that preserves its general meaning (Manning, Raghavan, and Schütze 2008, p. 289). This approach is robust, as state-of-the-art embeddings can be obtained from the same cloud services as the LLMs that generate the texts to be encoded, but the highest-ranked asset by vector similarity is not always the one a human artist would choose. Moreover, the dialogue may now refer to a feeling or action that is no longer visible on the page, which goes some way toward defeating the integrative goal of using comics. To fully realize this goal, we will need to think differently, and flip the script on comics creation.

Flipping the Script with the Marvel Method

Vector encodings map the gist of a line of dialogue to a visual asset, but these mappings are coarse and ignore the peculiarities of a text or an asset. Consider the *Excelsior!* pose for *sympathetic*, in which a caring figure offers a tissue to dry one's eyes. If dialogue is generated that exudes a sense of concern, this pose will rank high as a candidate, but the dialogue will not have been produced to echo the pose's visual details. Conversely, if the pose is chosen before any dialogue is generated, the LLM can weave a sympathetic text that actively refers to the image, as in "Don't cry. Here, take one of these." It helps if the other figure seems to be crying when this line is said, so an LLM should ideally know about all of the poses in a panel before it generates any dialogue at all for the panel. This is the essence of the Marvel method: to draw first, (re-)interpret second, and only then write. To use this method for automated comics generation, we will need a means of generating high-level plots, of rendering those outlines as visual sequences, and of layering a rich text of dialogue and exposition over each sequence. By prompting an LLM with a combination of plot points (what is happening to whom, because of whom, and where) and of verbal action descriptions (what those happenings will look like to an observer), it should be able to generate captions and dialogue that speak directly to the final visualization.

To produce a full story using the Marvel method, we first sketch the outline of a plot. This will serve as a scaffolding around which an LLM can generate exposition and dialogue. The simplest scaffoldings, expressed in ComiXML, specify a sequence of panels in which two recurring characters, named *Alfie* and *Betty*, strike poses against a series of backgrounds. The poses should be apt for their backgrounds and coherent in how they change from one panel to the next. While they do not determine a story, they are the silent tracks along which an LLM can weave a tale of its own. Part of the scaffolding that an LLM has adapted to the topic of *AI* is presented in Fig. 2. Although a scaffolding visually determines the story, it gives wide latitude to a storytelling LLM, which can vary a text's register, its reading age, its attitude to the characters, and its overall message. It can exploit both clichés *and* creative metaphors, such as the central metaphor of Fig. 2: that AI is a threat that may escape from its box.

We define a story grammar (Rumelhart 1980) for the generation of visual scaffoldings that can guide the LLM to generate complete comic narratives. This visual narrative gram-



Figure 2: Four panels of a scaffold adapted to the topic AI.

mar (Cohn 2014) must explicitly codify the tacit sensibilities that shape a visual story, by capturing the cause and effect relations between different poses and the common sense linkages between different panel backgrounds. A character who is attacked in one panel may exhibit injuries in the next; a character holding a ball in one panel may drop it, or throw it, in the next; and a character can be hit with a ball in one panel only if another has hurled it in the previous panel. We can think of this as a highly localized version of *Chekhov's gun*, the principle that an object foreshadowed in one scene can play a role in a later one (Bitsilli 1983, p. 150). We apply Chekhov's principle in both directions: an object introduced in panel A may later play a role in panels B, C, D and so on, while an object used in a panel D must have been introduced in panel A ... C, or be shown being introduced in panel D.

A story grammar that lacks a memory for long-distance dependencies must also obey the Markov property (Ethier and Kurtz 2005): sense can be made of any panel by looking only at the previous panel. If an object such as a crystal ball is introduced in panel A and thrown as a projectile in panel C, it must persist in an intermediate panel B for it to remain available in C. That is, objects cannot disappear and reappear across a narrative if the grammar is to respect both the Chekhov principle and the Markov property, so poses that introduce props such as guns, trophies, clocks, etc. should only be linked by the grammar to other poses that either use or dispose of those props. When the visual grammar is codified as a graph in which nodes represent apt pairings of poses for a two character story, then its edges represent allowable transitions from one pairing to another that respect our joint Chekhov/Markov principle. A new scaffolding can easily be generated from this graph as a random walk from any node that is marked as a possible story starter to any other that is marked as a possible story ender (Veale 2018).

The presence of specific props in various *Excelsior!* poses adds details that a visual storyteller can exploit, but also limits how they can be used. For instance, the *mourning* pose, in which a figure wails before a tombstone, should only be used in a graveyard setting, never a bar or an office. The *riding* pose, in which a figure rides a horse, is best suited to outdoor settings, while poses in which a figure wears a spacesuit are best reserved for outer space settings. Still, most poses can sensibly be used against most backgrounds, with some exceptions, but it is these exceptions, when ignored, that give rise to *Tale-Spin effects*. These emergent errors, named for an early AI storyteller (Meehan 1977) whose reach exceeded its technical grasp, show that a system is working with ideas that it does not understand (Wardrip-Fruin 2009). They offer a brief glimpse behind the curtain but leave an audience unimpressed with what it sees. In comics, the choice of backgrounds is shaped by many soft preferences with few hard constraints. A background should be apt for the poses set before it, and transitions between backgrounds should suit both the action unfolding in the foreground and our intuitions about connected spaces. If panel A is set in a kitchen, the scene may shift to the dining room in B and the garden in C, but not to the moon in D. To model these intuitions we need a second grammar to manage background transitions.

Since this is more a matter of soft intuitions than of hard rules, we begin by generating 100 random plots with our story grammar, that is, 100 random walks in our graph of pose pairings. This graph connects 1000 or so such pairings to their possible transitions, so a 100 random walks gives us a diverse sample of the much larger set of stories it supports. We then manually attach each pose pairing in each sampled plot to an apt background, ensuring any transitions between them remain apt as the poses shift in the foreground. From this training set *Excelsior!* then learns for itself how to label unseen future sequences. The result is codified as a mapping from pose pairings to backgrounds and as a graph that encodes every attested transition between these backgrounds. When generating a new story outline for use with the Marvel method, a random walk is simultaneously performed in the story graph and in the graph of background transitions, to produce a sequence of pose pairings with backgrounds that satisfies both our story logic and our intuitive sense of space.

These interlocking grammars produce a visual scaffolding that is locally coherent in its changing use of character poses and panel backgrounds. It is the job of the LLM, in the second step of the Marvel method, to impose a coherent narrative on this palimpsest. By pre-emptively laying down visual rails that steer the LLM's inventiveness we prevent the language model from hallucinating XML structures that are invalid or incoherent, or dialogue that does not fully resonate with the visual features of the rendering (Veale 2024).

Hallucination versus Invention

LLMs are generative in so many registers and domains. If suitably prompted, they can produce plausibly fluent texts in any form and about almost any topic we wish to explore. A single computational architecture underpins this wide-ranging capacity for generation, whether we want the LLM to speak truthfully about the facts or invent whole new worlds, and herein lies the problem of "hallucination": LLMs are trained for plausibility, not accuracy, and will often invent when we want mere recall, or recall when we want pure invention (Kalai et al. 2023). When it comes to sticking to the facts, symbolic systems are more accurate but less accommodating in their interactions. We want our generators of comic narratives to stick to the script as far the rules of the visual schema are concerned, but to have a free hand in how the legitimate possibilities of that schema are exploited.

The approach in (Veale 2024) does not follow the Marvel Method, but follows instead the more conventional order of comics production: *write first, draw later*. It is thus prone to hallucination when the LLM is tasked with suggesting poses and backgrounds for the text it has just generated. The same LLM is capable of generating XML for the visual scaffolding and the final text, but large LLMs are also prone to hallucination when choosing which poses and backgrounds to use. This remains the case even when we give the LLM a much reduced inventory of visual assets to remember. Small models, such as the 7 billion parameter QWEN distillation of DeepSeek's R1, invent poses or backgrounds from whole cloth about 80% of the time, while the 70 billion parameter LLaMa 3.3 hallucinates a nonexistent background or pose 5% of the time. OpenAI's larger model, GPT-4.1, also hallucinates at a rate of 3%, or at least once per short comic.

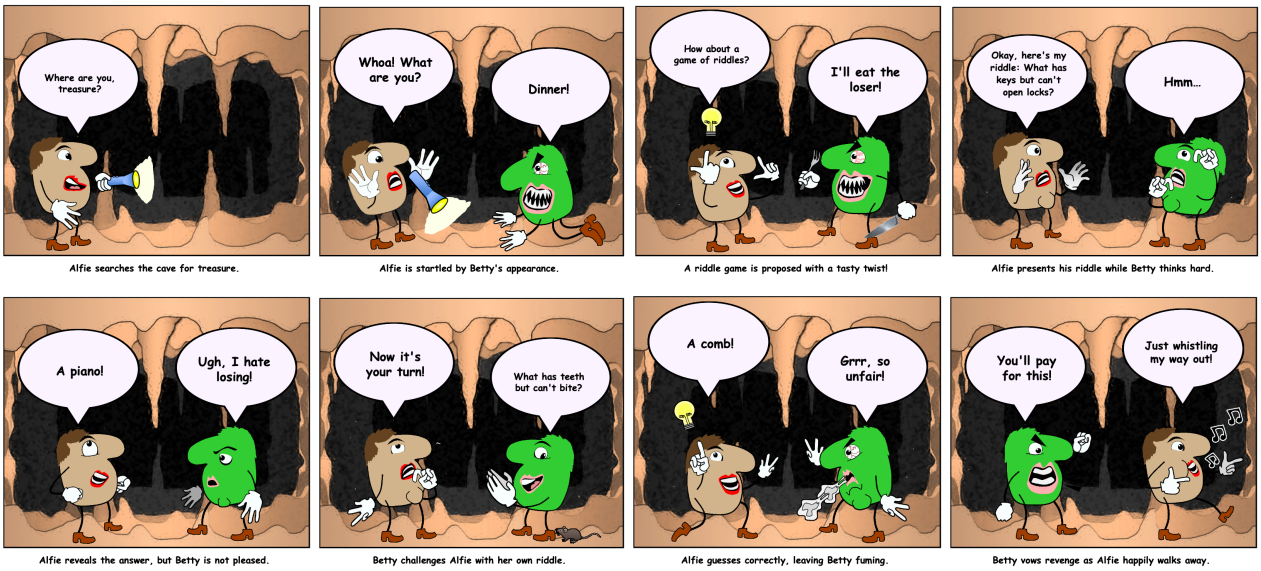


Figure 3: A visual scaffolding that the LLM interprets as a story about a contentious game of riddles.

Effortful Failures vs. Effortless Successes

(Øivind Madsen and Puyt. 2026) argue that while creativity thrives on scarcity – in rare deviations from the norm – generative slop makes a dubious virtue of glut. (Kommers et al. 2025) define slop via three criteria that help explain its rapid spread: *superficial competence* (versus incisive craft); *mass production* (versus selective curation); and *asymmetric effort* of AI generation (versus conventional production). The latter view our modern concept of AI slop as a natural evolution of what earlier critics called "kitsch" (Greenberg 1939), for that too was seen as parasitic on higher forms of human culture. Greenberg dismissed kitsch as debased "simulacra of genuine culture" that drink too heavily from our "reservoir of accumulated experience," but some may also see this as a prescient critique of modern, data-driven generative AI.

Greenberg was prescient in other ways too. In noting that kitsch only simulates the appearance of creativity with "devices, tricks, stratagems, rules of thumb, themes," he decried what is called *mere generation* in CC, the checklist-driven, uncritical production of artifacts that obey an approved template (Ventura 2016). By 2016 this reproach was becoming passé, since those templates had become too sophisticated to permit a clean separation of a generator from its built-in filters. But the success of GenAI, and the advent of slop on an industrial scale, have sharpened the edge of this rusty critique. We once fretted that mere generation only models the forms of creativity, not its content, like a child who prefers to play with the box after discarding the toy inside. But those forms have become so compelling, and so easy to produce on a whim, that mere generation is now the norm. GenAI tools make creative-seeming generation effortless for both a system and its user. For the system, much the same compute is expended for normative outputs as for "creative" ones. For users, the only difference is in how the system is prompted. Mere generation is often good enough, and effortless too.

In contrast, our pair of visual grammars for pose and background transitions represent a considerable amount of effort: the pose grammar, a graph of the allowable transitions from one pairing of character poses (in a two-character panel) to another, defines more than one thousand of these transitions. But this effort is what allows us to claim creative ownership of the visual scaffolding that emerges, and part ownership of the final, LLM-written comic, no matter what the LLM writes. The grammars, which give the resulting comic strips a manic, slapstick energy, reflect *our* comedic sensibilities, not the machine's. We can point to these comics as artifacts that we ourselves might create, and still be surprised by how the LLM has interpreted our style with the Marvel method. While each grammar-based generator is merely generative, no matter the grammar's size, the combined system is not.

Creativity is an emergent phenomenon, so the task of a CC researcher is to explain, or to demonstrate, how it might arise from an assemblage of uncreative parts. Since any creative generator will inevitably be built from merely generative parts, our grammars for slapstick are not intended to be creative in themselves. Rather, they serve as a receptacle for the creativity of their makers, so that we may share in the creativity of the comics that emerge. But if mere generation is akin to a child discarding the toy to play with the box, it is the job of these grammars to design only the box. It then becomes the job of the LLM to invent the toy that belongs inside. The LLM is both aided and challenged by the guidance it receives from this merely generative visual container: that scaffolding gives a shape to the story inside, and points the LLM in a specific direction; but the LLM must also adapt its output to what it is given, and turn the combination of its own writing and those visuals into a coherent narrative.

This relationship between visualizer and writer fosters a mode of antagonistic cooperation, not unlike its human form at Marvel Comics (Evanier 2008). The grammar produces a visual stimulus for creativity that both guides and constrains the LLM. The semi-random nature of its scaffoldings, which are random walks in a graph of connected possibilities, turns comic production into a multimodal version of the surrealist game *exquisite corpse* (Brotchie and Gooding 1991, 141-2). Ironically, this randomness reduces the LLM's tendency to produce slop, since the LLM must react to a combination of a topic (such as AI, or the "manosphere", or climate change) and a thematically-unrelated scaffolding (e.g., a creepy castle setting, a dank cave, a forest) that forces it to deviate from the overly familiar narrative forms in its training data. This does not require the LLM to expend more compute, unless we explicitly ask it to "reason" and thus generate more tokens as it searches for an answer (DeepSeek-AI et al. 2025), but it does, in a sense, represent a greater effort overall.

The comic strip in Fig. 3 illustrates this *box-and-toy* view of comics creation. When given a scaffolding set in a cave, the LLM fills this box with a story about a game of riddles. Its interpretation is not particularly novel, and echoes the interaction between Bilbo and Gollum in *The Hobbit*. But of more note is how the LLM builds its story around specific riddles with valid solutions. While there is also nothing new about the riddles, the overall effect is one of competent invention. Once again, while the parts themselves are unoriginal, the whole exhibits a degree of playful originality. The semi-randomness of the scaffolding is a source of friction to the LLM, and a source of diversity in its outputs. Under this arrangement, there is no need to vary the temperature setting (the so-called "creativity knob") of the LLM to induce variety in its story-telling (Peeperkorn et al. 2024). The friction that arises as the LLM is pushed in a semi-random direction by a visual scaffolding gives rise to much the same effect.

Summary and Conclusions

The myth of the lone creator runs through our folk conception of what it is to be creative. For instance, we commonly attribute the discovery of the first antibiotic, Penicillin, to Alexander Fleming, but it was a team led by Howard Florey and Ernst Chain that actually succeeded in isolating the compound for medical use, and all three shared the Nobel prize for medicine in 1945. We think of Steve Jobs as the driving creative force in the company *Apple*, but it was Jobs' ability to identify talent, such as that of designer Jony Ive and co-founder Steve Wozniak, that made Apple the dominant cultural and technology player it continues to be. Most commercial products, in fact, are created by teams in collaborative, cross-disciplinary efforts, even if the public, and the marketers, prefer to emphasize the work of a single creator.

We all want to be creative in some aspects of our lives. The builders of CC systems can be creative in multiple ways: they can be creative in *how* they construct their systems, and *meta*-creative in how they share in the system's successes or interesting failures. For the end-users of these systems, there is less scope for true creativity and more scope for vicarious creativity. Slop is pervasive on the web not because humans have been replaced by machines, not yet at least, but because even the purveyors of slop *want* to be creative. So, from the perspective of machine creativity, our concern is not the slop that pretends to be human-made (even if it is human "slop"), of the kind that (Shaib et al. 2026) seek to automatically detect, but the artifacts that emerge when it is we humans that are merely generative in our use of GenAI and CC systems. Ironically, even the most "creative" systems become merely generative when they become too easy and effortless to use.

Collaboration involves a tantalizing mix of surrender and control. Each partner has a voice, but each is just one voice. The Marvel method shows how different experts can control their own specialist contribution, but there is surrender too: the writer has the ultimate say in how the artist's contribution is framed as a visual narrative. Collaborative creation always involves some surrender of control, either to the collective or to a dominant partner, but how much surrender is *too* much if each partner is to validly claim some creative ownership? At one extreme on this continuum from control to surrender sit the micro-manager and the helicopter parent, unwilling to let any decision go unanalyzed. At the other sit the prompt-and-go users of GenAI, who trust in its processes and rush to share its outputs with the world. For these users. prompt engineering is the only creativity knob (Morain and Ventura 2025), and it is one that is often used lightly to produce slop.

Looking at an AI output, we can't know for sure how carefully this knob was turned, or how many times the system's earlier efforts were discarded, unless the output bears some hallmarks of indifference: extra appendages, uncanny props, and so on. But as GenAI improves, usefulness remains a reliable metric. As mere novelty is not itself sufficient for creativity, we also require novel results to be useful, valuable or appropriate (Runco and Charles 1993; Diedrich et al. 2015; Runco, Illies, and Eisenman 2025). These factors also form a continuum, and much like a Kleenex or a band-aid, even slop can have an ephemeral value, albeit near the low end of this scale. So, we may be less inclined to dismiss an artifact as "slop", and more inclined to laud its inciting prompt as "clever" or "insightful", if the artifact fulfills a more practical and durable purpose. In hybrid CC systems that employ the Marvel Method, the prompts that an LLM receives from more conventional sub-systems can better deliver this utility.

Since the earliest newspaper strips, comics have been dismissed as a disposable art form. Whatever the truth of this view now, AI-generated comics can scarcely be regarded as a more durable improvement, and if generated purely for entertainment, this scorn is not wholly undeserved. But comics can serve a more serious purpose, to lend difficult, unpalatable or even threatening content a more engaging form. In Fig. 4, for instance, we see a comic that aims to make learning a second language more fun. This comic is generated as before, using the Marvel Method to partner a visual story grammar with an LLM (GPT-4.1). Here, however, the LLM goes a step further, and translates each panel's dialogue from English into Korean. These translated panels are interwoven with the originals, and Romanized forms of the Hangul text are then added beneath as captions. The constant quest for novelty, at the expense of real value, is a key driver of slop online. It is only by squeezing more practical value from our AI-generated artifacts that can we hope to apply the brakes.

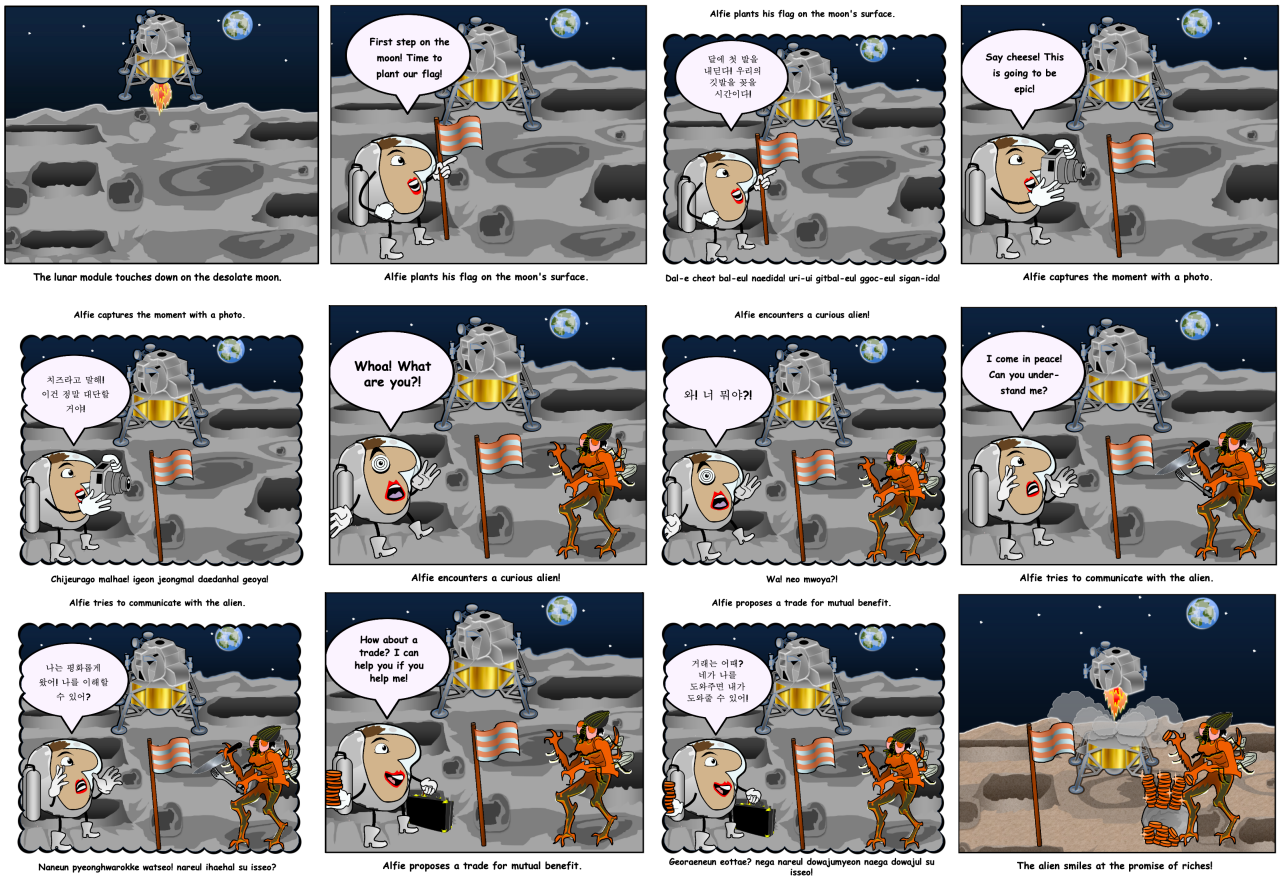


Figure 4: A completed comic strip is translated by the LLM into Korean, and interleaved with the English original.

Acknowledgments

This paper and its datasets are wholly the work of the author.

References

Alves, T.; McMichael, A.; Simões, A.; Vala, M.; Paiva, A.; and Aylett, R. 2007. Comics2D: Describing and creating comics from story-based applications with autonomous characters. In *Proc. of the 20th Annual Conference on Computer Animation and Social Agents*.

Bitsilli, P. 1983. *Chekhov's Art, a Stylistic Analysis*. California: University of California/Ardis.

Boden, M. 1992. *The Creative Mind: Myths and Mechanisms*. London: Abacus.

Brotchie, A., and Gooding, M. 1991. *Surrealist Games*. London, UK: Redstone Press.

Burkus, D. 2013. *The myths of creativity: The truth about how innovative companies and people generate great ideas*. San Francisco, CA: Jossey-Bass.

Cohn, N. 2013. *The Visual Language of Comics: Introduction to the structure and cognition of sequential images*. London, UK: Bloomsbury.

Cohn, N. 2014. You're a good structure, Charlie Brown: The distribution of narrative categories in comic strips. *Cognitive Science* 38(7):1317–1359.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; and et al., X. Z. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645:633–638.

Diedrich, J.; Benedek, M.; Jauk, E.; and Neubauer, A. 2015. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts* 9:35–40.

Ditko, S. 1973. *The Avenging Mind: Essays by Steve Ditko*. Bellingham, WA: SD Publishing.

Ethier, S., and Kurtz, T. 2005. *Markov processes: Characterization and convergence.* Hoboken, NJ: Wiley.

Evanier, M. 2008. *Kirby: King of Comics.* New York: Abrams.

Forceville, C.; Veale, T.; and Feyaerts, K. 2010. Balloonics: The visuals of balloons in comics. In Goggin, J., and Hassler-Forest, D., eds., *The Rise and Reason of Comics and Graphic Literature: Critical Essays on the Form*. McFarland.

Greenberg, C. 1939. Avant-garde and kitsch. *Partisan Review* 6(5):34–49.

Howe, S. 2012. *Marvel Comics: The Untold Story*. New York, NY: Harper.

Kalai, A. T.; Nachum, O.; Vempala, S.; and Zhang, E. 2023. Why language models hallucinate. *arXiv* 2509.04664.

Kommers, C.; Duede, E.; Gordon, J.; Holtzman, A.; McNulty, T.; Stewart, S.; Thomas, L.; So, R. J.; and Long, H. 2025. Why slop matters. *arXiv* 2601.06060.

Kurlander, D.; Skelly, T.; and Salesin, D. 1996. Comic chat. In *Proceedings of SIGGRAPH'96, the 23rd annual conference on Computer graphics and interactive techniques*. ACM. 225–236.

Lee, S., and Buscema, J. 1978. *How to Draw Comics the Marvel Way*. New York: Fireside.

Lee, S., and Mair, G. 2002. *Excelsior!: The Amazing Life of Stan Lee*. New York: Fireside.

Manning, C.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Pres.

Matarazzo, A., and Torlone, R. 2025. A survey on large language models with some insights on their capabilities and limitations. *arXiv* 2501.04040.

McCloud, S. 1993. *Understanding Comics: The Invisible Art*. New York, NY: Harper Collins.

McCloud, S. 2006. *Making comics: Storytelling secrets of comics, manga, and graphic novels.* New York, NY: Harper Collins.

Meehan, J. R. 1977. Tale-spin, an interactive program that writes stories. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'77, 91–98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Morain, R., and Ventura, D. 2025. Is prompt engineering the creativity knob for large language models? In *Proceedings of ICCC-25, the 16th International Conference on Computational Creativity*. 1–11.

Orwell, G. 1946. Politics and the english language. *Horizon* 13(76):252–265.

Patil, R., and Gudivada, V. 2024. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences* 14(5):1–42.

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of ICCC-24, the 15th International Conference on Computational Creativity*. 103–110.

Pérez y Pérez, R.; Morales, N.; and Rodríguez, L. 2012. Illustrating a computer-generated narrative. In *Proceedings of ICCC-12, the 3rd International Conference on Computational Creativity*. 103–110.

Rumelhart, D. 1980. On evaluating story grammars. *Cognitive Science* 4(3):313–316.

Runco, M. A., and Charles, R. E. 1993. Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences* 15(5):537–546.

Runco, M. A.; Illies, J. J.; and Eisenman, R. 2025. Creativity, originality, and appropriateness: What do explicit instructions tell us about their relationships? *The Journal of Creative Behavior* 39(2).

Shaib, C.; Chakrabarty, T.; Garcia-Olano, D.; and Wallace, B. C. 2026. Measuring AI "Slop" in Text. *arXiv* 2509.19163.

Shen, H.; Zhang, J.; Xiong, B.; Hu, R.; Chen, S.; Wan, Z.; Wang, X.; Zhang, Y.; Gong, Z.; Bao, G.; Tao, C.; Huang, Y.; Yuan, Y.; and Zhang, M. 2025. Efficient diffusion models: A survey. *arXiv* 2502.06805.

Veale, T. 2012. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. London, UK: Bloomsbury.

Veale, T. 2018. The 'default' in our stars: Signposting non-defaultness in ironic discourse. *Metaphor and Symbol* 33(3):175–184.

Veale, T. 2022. Two-fisted comics generation: Comics as a medium and as a representation for creative meanings. In *Proceedings of ICCC-22, the 13th International Conference on Computational Creativity*. 59–66.

Veale, T. 2023. Have I Got Views For You! Generating "Fair and Balanced" Interventions into Online Debates. In *Proceedings of ICCC-23, the 14th International Conference on Computational Creativity*.

Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in reimplementing creative systems with llms. In *Proceedings of ICCC-24, the 15th International Conference on Computational Creativity*. 103–110.

Ventura, D. 2016. Mere Generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity, Paris, France*.

Walsh, J. 2012. Comic book markup language: An introduction & rationale. *Digital humanities quart.* 6(1).

Wardrip-Fruin, N. 2009. *Expressive Processing: Digital Fictions, Computer Games, Software Studies*. MIT Press.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2024. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys* 56(4):1–39.

Øivind Madsen, D., and Puyt., R. W. 2026. When AI turns culture into slop. *AI & SOCIETY* 41::1287–1288.

Activation Steering for Creative Estrangement: Generating Evocative Narrative Through Controlled Disruption of Language Models

Olga Vechtomova*, Greg French*, Xavier Schneider*, Olivier Poulin*, Jeff Bos**

*University of Waterloo, Canada
**WordSynth Inc.
{ovechtom, g2french, xschneider, opoulin}@uwaterloo.ca, jeff@wordsynth.com

Abstract

We investigate activation steering as a method for generating extended narrative that can function as a textual environment for creative inspiration. Using contrastive pairs inspired by the aesthetics of Tarkovsky and Lynch, we construct steering vectors encoding modes of perceptual estrangement and apply them to Llama 3.1 8B Instruct. Single-layer steering operates in two regimes: a moderate regime that elevates literary register without producing estrangement, and a high regime where estrangement emerges but coherence collapses. Dual-layer steering resolves this by applying a compositional vector at layer 13 simultaneously with a content-level vector at layer 14, accessing a region where estrangement and coherence coexist. The dual-steered outputs exhibit emergent disruptions (person shifts, narrative self-contradiction, referential instability) absent from all training pairs, arising from competing representational pressures during generation. Comparing steered outputs against explicitly prompted fragmentation reveals two qualitatively distinct modes: prompting produces conventional fragmentation devices, while steering produces intact sentences whose content is self-contradicting. A blind analysis of evaluator reasoning texts and human reader feedback confirm this distinction. Human readers tended to perceive dual-steered outputs as stranger than prompted outputs, suggesting that occasional and surprising disruption within coherent prose is more estranging than pervasive and categorizable fragmentation. [1]

Introduction

Viktor Shklovsky's concept of *ostranenie* (defamiliarization or making strange) proposes that art's essential function is to restore how things are perceived rather than how they are known (Shklovsky 1965). The target of estrangement is not the exotic but the ordinary, seen as if for the first time. We see this across artistic domains: in Andrei Tarkovsky's *Solaris* (1972), rain falls inside a cabin and a father does not react; in David Lynch's Black Lodge (*Twin Peaks*, 1990), characters speak in reversed recordings and coffee does not pour from an inverted cup. In each case, the estrangement operates through minimal displacement from the familiar. Petric (1989) describes Tarkovsky's films as suffused with an oneiric air that transcends the commonplace signification of objects to reach what the eye ordinarily neglects. Rodley (2005) characterizes Lynch's work in similar terms: Lynch's interest lies in the defamiliarization process and the waking/dream state, rather than in incongruity as spectacle or overt surrealism. This is close to Shklovsky's "stone made stony", not by adding to it but by making us see it again.

We ground our approach in these filmmakers' aesthetics because their estrangement targets *perception itself* rather than the content of reality. Genre estrangement (fantasy, science fiction, magic realism) introduces rule-governed alterations that the reader eventually assimilates; cognitive tension resolves as the rules become clear. Tarkovsky and Lynch embed unexplained impossibilities within naturalistic observation. Neither the reverse speech inside the Black Lodge in Lynch's Twin Peaks, nor the rain inside the cabin in Tarkovsky's Solaris is ever explained. More fundamentally, the ontological status of the figures in the Lodge and of those who appear on the station in Solaris is left unresolved. Hari appears to Kelvin as his wife, long dead, and comes to understand that she is not. What the Lodge figures are is never specified. As Petric (1989) observed, the viewer feels that something is "wrong" but cannot detect sufficient proof to discredit the events on the basis of everyday logic. The indeterminacy is sustained and is often left unresolved, maintaining the heightened attention that Berlyne's inverted U-curve (Berlyne 1971) identifies as maximally engaging. This choice shapes our experimental design. Because we model disruptions of perception rather than alterations of content, the resulting steering vectors encode abstract perceptual modes (temporal self-contradiction, ontological uncertainty, identity instability) rather than genre conventions.

Defamiliarization of this kind has direct implications for computational creativity. Rubin (2023) describes the Seed phase of the creative process as the stage where "we are completely open to any stimulus that might lead us to that state of mind where art becomes inevitable." At this stage, artists need stimuli that are indeterminate, surprising, and open to interpretation. Leonardo Da Vinci advised young artists to look at old stained walls and streaked stones for inspiration (Turner 2011). The stains are not art in themselves but projective stimuli whose meaning is unresolved.

[1] Supplementary material website:
https://ovechtom.github.io/supplementary-material/

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

The artist forms meaning through the process of attempted resolution, and it is this process that activates the creative state. Prior research demonstrated this with LSTM-based Variational Autoencoders trained on small curated datasets, which generate lines characterized by syntactic indeterminacy and semantic openness ("clouds humanity without light", "through the wrong moon") that function effectively as seeds for creative inspiration (Vechtomova, Sahu, and Kumar 2021). However, this approach operates at the sentence level. The question then becomes: can we generate extended narrative that sustains these qualities, text that functions not as a finished story but as a *textual environment* for creative inspiration?

Activation steering is a technique that modifies a language model's internal representations during generation by adding directional vectors to hidden states. In this paper, we demonstrate that this technique can produce long-form narrative with precisely these qualities. We steer Llama 3.1 8B Instruct using contrastive vectors derived from paired literary and neutral passages and evaluate the outputs using a seven-metric framework scored by LLM evaluator (Claude Sonnet 4.5). Our key findings are:

1. **Two regimes of first-order steering.** Steering at layer 13 operates in two distinct regimes: a moderate regime (α 2.0–6.0) that elevates literary register without producing estrangement, and a high regime (α 7.5–8.0) where estrangement and syntactic disruption emerge abruptly but coherence collapses. No first-order alpha value produced both high estrangement and high coherence.
2. **Dual-layer steering achieves coherent estrangement.** Applying a first-order vector at layer 13 simultaneously with a second-order content-level vector at layer 14 produces estrangement comparable to the highest first-order alphas while maintaining substantially higher coherence, accessing a region single-layer steering does not reach.
3. **Emergent syntactic estrangement.** Dual-steered outputs exhibit syntactic disruptions (person shifts, temporal self-contradiction, logical impossibility) absent from all training pairs. These features emerge from the interaction of competing representational pressures during generation.
4. **Disruption *of* syntax vs. disruption *through* syntax.** Explicitly prompted fragmentation produces conventional devices (ellipses, sentence fragments, lowercase). Steering produces grammatically intact sentences whose content is self-contradicting or ontologically unstable.

Related Work

Defamiliarization and Aesthetic Engagement

The relationship between aesthetic engagement and stimulus complexity follows an inverted U-shaped curve, also known as Wundt effect, first formalized by Berlyne (1971). Completely predictable stimuli fail to hold attention; completely random stimuli offer no foothold for pattern recognition; stimuli of intermediate complexity produce the greatest engagement. Gold et al. (2019) offered a learning-based explanation for this effect: when events are completely predictable or fully random, little learning occurs as patterns are either trivial or impossible to discern. Intermediate predictability creates conditions for meaningful cognitive engagement, as the perceiver is rewarded for the effort of partial pattern recognition. As Pepperell (2006) noted, indeterminate visual images create heightened perception and attention as habitual recognition is suspended, a state that is itself a precondition for creative work. Zeilinger (2023) extended this to AI-generated visual art, arguing that visual indeterminacy invites active meaning-making rather than passive reception.

Artists have long developed deliberate techniques for inducing this productive disruption. The cut-up technique, from Tzara through Burroughs, fragments existing text to break habitual language patterns (Burroughs 2012). David Bowie described its function as "igniting anything that might be in my imagination" (BBC 1975), aligning it with the Seed phase of creativity. However, pure randomness does not produce art. As Nees (1969) observed, there must be structure to narrow random outcomes into discernible aesthetic information. The productive zone lies between predictability and randomness, the same intermediate territory that the Wundt curve identifies as maximally engaging.

Tarkovsky and Lynch exemplify this productive zone in cinema. Unlike genre estrangement, which introduces rule-governed alterations that the viewer eventually assimilates, their work embeds unexplained impossibilities within naturalistic observation, maintaining the state of heightened perception that Pepperell identifies as a precondition for creative work.

Computational Approaches to Creative Text Generation

Boden's influential typology distinguishes combinational, exploratory, and transformational creativity (Boden 2004). In exploratory creativity, a system searches within an established conceptual space; in transformational creativity, the space itself is altered to make previously impossible outputs conceivable. Wiggins (2006) formalized this distinction, providing a mathematical framework for describing creative systems in terms of the conceptual spaces they traverse and transform. Colton's creative tripod framework argues that a system perceived as creative must demonstrate not only skill but also appreciation and imagination (Colton 2008), and the FACE/IDEA descriptive models provide formal criteria for evaluating creative acts (Colton, Charnley, and Pease 2011).

Within computational text generation, Oliveira's PoeTryMe (Oliveira 2012) demonstrated autonomous poetry generation using semantic and syntactic templates, while Gervás (2009) surveyed computational approaches to storytelling with attention to how creativity manifests in narrative structure. More recently, LLMs have prompted new questions about machine creativity. Franceschelli and Musolesi (2025) analyzed LLM creativity through the lens of Boden's typology and Colton's tripod, while Ismayilzada et al. (2025) found that LLMs generate stylistically complex stories but fall short on novelty and surprise compared to human writers. In human-AI co-creative writing, Calderwood et al. (2020) found that novelists use generative models as constraining tools that push the boundaries of their writing.

This finding aligns with our interest in text that functions as creative stimulus rather than finished output.

The connection between Shklovsky's defamiliarization and computational text generation has been explored in prior work on neural poetry generation (Vechtomova and Sahu 2025), where LSTM-VAE outputs were analyzed as instances of productive estrangement. The present work asks whether this connection holds at the level of extended narrative. Ho et al. (2025) proposed a "Familiarity-Estrangement Space" that uses topic modeling to project textual materials along dimensions of topical familiarity and novelty, allowing users to browse for texts at a desired level of thematic strangeness. Their system operates on existing text rather than generating new text with estranging qualities.

The distinction between systems that generate polished outputs (*pushing* content toward the artist) and systems that generate indeterminate stimuli (*pulling* the artist into the creative state) is central to understanding how computation can support creativity. LSTM-VAE sentence generators (Vechtomova, Sahu, and Kumar 2021) are pull-oriented. They produce open, indeterminate fragments rather than finished text. But the architecture operates at the sentence level, generating individual lines rather than extended narrative that sustains these qualities across a unified passage. The present work is also pull-oriented. It aims to generate not polished stories but textual environments for creative inspiration. In Boden's terms, LSTM-VAE models perform exploratory creativity within a learned latent space. The present work investigates whether activation steering can achieve something closer to transformational creativity by altering the model's generative space to produce outputs with qualities absent from any training example.

Activation Steering

Activation steering modifies a language model's hidden states during inference by adding learned directional vectors, enabling fine-grained control over output properties without retraining. Subramani et al. (2022) first demonstrated that steering vectors can be extracted from pretrained language model decoders and applied via addition to hidden states, achieving high-fidelity generation control and unsupervised sentiment transfer. Turner et al. (2023) introduced Activation Addition (ActAdd), computing mean activation differences between contrastive prompt pairs and injecting the resulting vector during generation to reliably modulate behavioral properties such as sycophancy and hallucination tendency. Rimsky et al. (2024) extended this to Llama 2, systematically evaluating steering effectiveness across layers and scaling factors. In parallel, Zou et al. (2023) proposed Representation Engineering (RepE), a top-down framework that reads and controls high-level concepts (honesty, fairness, harmfulness) in LLM internal representations at the population level rather than through individual neurons or circuits. Li et al. (2023) demonstrated Inference-Time Intervention (ITI), shifting activations during inference to improve truthfulness, suggesting that models possess internal representations of properties they do not reliably express.

The present work applies activation steering to a qualitatively different domain: generating aesthetically productive text for creative inspiration. Rather than steering toward or away from a binary behavioral property, we steer along a continuous dimension, from prosaic description to literary estrangement, defined by contrastive pairs that model specific aesthetics of defamiliarization. This requires navigating a narrow productive zone where the output is sufficiently estranged to defamiliarize but sufficiently coherent to sustain engagement. This challenge has no direct analogue in the behavioral steering literature.

Method

Steering Mechanism

Hidden states and the residual stream. In a decoder-only transformer such as Llama 3.1 8B Instruct (N=32 layers, hidden dimension H=4096), each layer *adds* its computation to a running per-token vector (the *residual stream*) rather than overwriting it. The layer-ℓ hidden state is the model's internal representation of a token at depth ℓ. Activation steering modifies these hidden states during generation, leaving weights and prompt unchanged.

Steering vectors. A *steering vector* is a direction in hidden-state space which, added to the residual stream, biases generation toward a target property. From n contrastive pairs of matched neutral and literary passages, we mean-pool the layer-ℓ hidden states over the final 32 token positions of each side, giving $h_\ell(x_i^+)$ and $h_\ell(x_i^-)$. The steering vector is their average per-pair difference, normalized to unit length:

$$d_\ell \;=\; u \,/\, \|u\|, \qquad u \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(h_\ell(x_i^+) - h_\ell(x_i^-)\big).$$

Averaging cancels passage-specific content and retains the shared direction (Turner et al. 2023; Rimsky et al. 2024). Normalization makes d_ℓ a pure direction; magnitude is set at generation time.

First- and second-order vectors. We distinguish two kinds of steering vector by what their contrastive pairs hold fixed. A *first-order* vector is built from the neutral-to-literary contrast, where the two sides differ in both register and perceptual mode, and encodes the full shift from plain description to estranged writing. A *second-order* vector is built *within* literary prose: both sides are already literary and matched for register and imagery, and differ only in whether an ontologically impossible event occurs. It therefore isolates content-level estrangement from register elevation. The terms are our nomenclature, not standard usage.

Injecting the vector during generation. At inference we add the steering vector to the residual stream at layer ℓ at every generated token. Writing $h_t^{(\ell)}$ for the layer-ℓ hidden state at step t:

$$h_t^{(\ell)} \;\leftarrow\; h_t^{(\ell)} \;+\; \alpha\, s(t)\, d_\ell.$$

Because d_ℓ has unit norm, the steering strength α is exactly the Euclidean length of the injected vector, so α is comparable across vectors and layers. We implement the addition as a HuggingFace forward hook on the chosen transformer block. Steering acts only on tokens the model generates; the prompt is encoded unmodified, and a displacement applied at one step influences every token that follows.

Decay schedule. A perturbation held at constant strength compounds through cached attention states and degrades the passage into repetition before it ends. We therefore attenuate with a cosine schedule,

$$s(t) \;=\; \frac{1}{2}\left(1+\cos\frac{\pi t}{k}\right) \text{ for } t<k,\;\; s(t)=0 \text{ otherwise,}$$

with k=200, so steering falls smoothly from full strength α at the first generated token to zero at token 200, letting the model resolve the displaced trajectory over the second half of the 400-token passage. All generations use temperature 0.9 (chosen to characterize the steered output *distribution* rather than a single deterministic trajectory), nucleus sampling (p=0.9), a maximum of 400 new tokens, and a fixed random seed.

Dual-layer steering. *Dual-layer* steering registers two hooks simultaneously: a first-order vector at layer ℓ_1 and a second-order vector at a nearby layer ℓ_2, with independent strengths α_1, α_2 sharing the cosine schedule. The displacement added at ℓ_1 propagates forward and is still present when the second vector is added at ℓ_2, so the two pressures (one elevating literary register, the other destabilizing the represented reality) act on the same generation jointly. Layers and strengths are reported below.

Steering Vector Construction

We construct steering vectors from contrastive pairs of matched passages describing similar scenarios, where one version is written in neutral, expository prose and the other employs techniques of literary estrangement. The vector is the mean difference between the model's internal representations of the literary and neutral sides, capturing the direction in activation space that corresponds to the transformation from prosaic description to estranged perception.

We designed five sets of 12 pairs (60 total), each embodying a distinct mode of estrangement: (1) *Contemplative*, modeled on Tarkovsky's sustained perceptual attention; (2) *Lynchian*, the eruption of wrongness within the familiar; (3) *Lost Highway*, identity instability and spatial inconsistency; (4) *Stalker*, landscape as responsive environment; and (5) *Solaris*, ontological uncertainty and memory as material substance. In each pair, the neutral side is a bland, surface-level recounting of events or surroundings, while the literary counterpart describes the same scenario with imagistic depth and a specific mode of estrangement. The neutral and literary sides are length-matched to within 15%. The pairs were developed through iterative collaboration between the researchers and Claude Opus 4.6 (Anthropic), with researchers specifying target aesthetic modes and evaluating candidates against theoretical criteria across multiple rounds of refinement. The resulting pairs were validated independently using the blind identification procedure described below.

Table 1: Mean scores for neutral and literary sides of all 60 training pairs.

Metric	Neutral	Literary
Imagery	1.75	3.01
Indeterminacy	1.09	4.14
Originality	1.00	3.98
Register	1.00	3.47
Coherence	4.97	4.73
Estrangement	1.03	4.42
Syntactic estr.	1.00	1.63

Layer Selection

We swept all 32 layers of Llama 3.1 8B Instruct, computing the contrastive vector at each layer and evaluating four metrics: direction norm (L2 magnitude), relative L2 (normalized by activation scale), consistency (within-set pairwise cosine similarity), and cross-file alignment (between-set cosine similarity). Direction norm peaked at final layers, merely tracking activation magnitude. Relative L2 peaked at early layers, which show high relative differences for any contrastive signal. The two predictive metrics, consistency and cross-file alignment, both peaked at layer 13 (consistency 0.654, cross-file alignment 0.741). Full layer analysis is available on the supplementary materials website. The Solaris pairs produced the strongest signal (consistency 0.704), consistent with ontological uncertainty requiring the deepest semantic displacement.

Layer 13 sits at approximately 40% depth, a region associated with semantic composition rather than discourse planning (Tenney, Das, and Pavlick 2019; Wendler et al. 2024). Steering at this layer targets *how the model constructs meaning* while leaving its coherence and syntactic planning machinery undisturbed. All first-order experiments accordingly use the composite vector at layer 13, injected with the cosine-decayed schedule.

Validation of Training Pairs

We evaluated both sides of all 60 pairs using the same metric pipeline applied to steered outputs (Table 1). The literary sides scored substantially higher on estrangement (4.42 vs. 1.03), indeterminacy (4.14 vs. 1.09), and originality (3.98 vs. 1.00), all $p < .001$ (Welch's t-test). Crucially, coherence was near-equivalent (4.73 vs. 4.97), confirming that the contrastive signal captures qualitative transformation rather than fragmentation. The estrangement gap (3.39) exceeded the register gap (2.47), indicating that the primary dimension encoded is perceptual defamiliarization rather than register inflation.

A blind identification procedure provided convergent evidence that the five sets encode distinct aesthetic dimensions. An independent LLM evaluator (GPT-4o, OpenAI),

given each set without labels, correctly identified the dominant aesthetic mode for three of five sets and swapped only the two Lynch-derived sets, which is an expected confusion given their shared origin.

Evaluation Metrics

Choice of evaluators. We chose Claude Sonnet 4.5 (Anthropic) as the primary evaluator because prior work on aesthetic evaluation of literary text (Sahu and Vechtomova 2025) found that frontier models with explicit reasoning produce reliable discrimination on literary dimensions where smaller or less capable evaluators do not. We chose GPT-4o (OpenAI) as the cross-validator specifically because it comes from a different model family and training pipeline than Claude, so agreement between the two provides stronger evidence than agreement within a single family. For token-level surprisal we use GPT-2 (117M) rather than a peer-scale reference: a small, weak model captures token-level statistical regularity rather than the same semantic competencies the steered model possesses, avoiding the circularity of using the steered model's own representational neighborhood as the standard against which to measure surprisal. The blind reasoning-text analysis uses Claude Opus 4.6, which had not seen the evaluation rubric during scoring.

Scoring protocol. We evaluate steered outputs using seven metrics scored on a 1–5 scale by Claude Sonnet 4.5, following a similar protocol to (Sahu and Vechtomova 2025) but with a modified set of metrics and a different LLM evaluator. We queried the LLM evaluator five times per story per metric at temperature 0.7 and averaged the resulting scores; temperature 0.7 with averaging reduces the influence of any single sampling artifact in the evaluator's rationale, whereas temperature 0 would lock onto one rationale that may not be representative. Each evaluation prompt requires the model to produce a one-sentence reasoning statement before committing to a numerical rating. This forced reasoning step proved essential during development. Without it, the evaluator exhibited reduced discrimination, conflating rhetorical intensity with genuine literary achievement. Ornate but semantically closed prose (*"a symphony of despair, a requiem for the living"*) scored comparably to compressed but genuinely open prose (*"I am a cry without a mouth"*). The reasoning step forces the evaluator to articulate *why* a passage merits a given score, producing reliable discrimination between register elevation and aesthetic transformation.

Choice of metrics. The seven metrics are: *imagery* (sensory vividness), *indeterminacy* (semantic openness and resistance to singular interpretation), *originality* (unusual collocations, surprising metaphors, unexpected verb choices), *register* (distance from plain prose toward heightened literary language), *coherence* (logical connection and overall readability), *estrangement* (defamiliarization of ordinary experience), and *syntactic estrangement* (departures from conventional prose syntax, including person shifts, sentence fragmentation, and sentences that undermine their own logic). Imagery, originality, and coherence are conventional in literary-generation evaluation (Sahu and Vechtomova 2025; Ismayilzada, Stevenson, and van der Plas 2025); indeterminacy and estrangement operationalize the aesthetic targets of our theoretical framing (Shklovsky 1965; Pepperell 2006; Zeilinger 2023); register and syntactic estrangement were added to test the disruption-of/through distinction. The first five characterize the *mode* of the prose; estrangement measures the overall effect; syntactic estrangement captures a specific mechanism that proved critical for distinguishing first-order from dual-layer steering. The training-pair validation (Table 1) provides evidence that the metrics discriminate as intended: neutral and literary sides separate cleanly on the targeted dimensions while coherence remains near-equivalent, which is non-trivial evidence the metrics are picking up the targeted constructs rather than overall "quality". Full rubric definitions are available in the supplementary materials. We additionally compute token-level surprisal using GPT-2 (117M) as a model-independent complement (Sahu and Vechtomova 2025).

Cross-validation with GPT-4o. To test whether our findings depend on idiosyncrasies of a single evaluator, we rescored all 204 outputs with GPT-4o using identical prompts and rubrics. Spearman rank correlations between the two evaluators are strong across all metrics: register ($\rho = 0.94$), originality ($\rho = 0.90$), indeterminacy ($\rho = 0.87$), imagery ($\rho = 0.86$), syntactic estrangement ($\rho = 0.78$), estrangement ($\rho = 0.77$), and coherence ($\rho = 0.68$), all $p < 10^{-28}$. GPT-4o rates consistently higher than Claude Sonnet 4.5 (mean offset 0.67 points), which depresses Krippendorff's α for some metrics, but since our arguments depend on ordinal patterns rather than absolute scores, rank-order agreement is the more informative measure. Condition-level means confirm that both evaluators reproduce the key trajectories: estrangement rising with α, coherence dropping sharply between α 7.5 and 8.0, and dual-layer compositions accessing higher estrangement at comparable coherence. Full agreement statistics and condition-level comparisons are available on the supplementary materials website.

Results

First-Order Steering: Two Regimes

We applied the composite vector at layer 13 with cosine decay across alpha values of 0 (baseline), $\pm$2.0, $\pm$4.0, $\pm$6.0, $\pm$7.5, and $\pm$8.0 to 15 prompts spanning three alignment tiers (high, medium, low). Tier assignment was determined by cosine similarity between each prompt's sentence embedding and the literary training-pair centroid in the sentence-embedding space using a sentence-transformer model (all-MiniLM-L6-v2). Figure 1 shows the full bidirectional metric profiles. The results reveal two distinct regimes rather than a smooth curve with a single optimum.

Moderate regime (α 2.0–6.0): register elevation without estrangement. Imagery rises from $\approx$2.8 to $\approx$3.5, register from $\approx$1.8 to $\approx$3.3, and indeterminacy modestly from $\approx$1.3 to $\approx$2.3. But estrangement barely moves, remaining

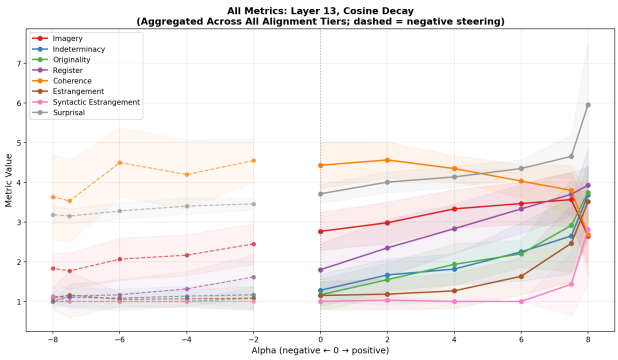


Figure 1: All metrics vs. alpha for first-order steering at layer 13 (aggregated across alignment tiers; dashed = negative steering). The shift at α 7.5–8.0 is visible as the abrupt divergence of estrangement, originality, and syntactic estrangement from their previous trajectories, simultaneous with the collapse of coherence.

between 1.2 and 1.6, marginally above baseline. Syntactic estrangement stays at floor (1.0) throughout. Coherence remains high (≈4.0–4.6). This regime produces *literary conventional prose*, richer and more contemplative than the baseline, but not defamiliarizing.

High regime (α 7.5–8.0): abrupt shift. Between alpha 7.5 and 8.0, multiple metrics shift sharply. Estrangement jumps from 2.5 to 3.5, originality from 2.9 to 3.75, and syntactic estrangement from 1.4 to 2.8. Simultaneously, coherence drops from 3.8 to ≈2.7. The transition between regimes is steep: the model absorbs moderate perturbation as register elevation, maintaining conventional narration, until the steering force exceeds some threshold and the output transforms rapidly. The productive qualities (estrangement, originality) arrive together with the destructive ones (coherence loss). No first-order alpha value produces both high estrangement and high coherence.

Bidirectional asymmetry. Negative steering compresses all metrics toward a uniform floor: imagery 2.0, originality 1.0, estrangement 1.0, while coherence remains relatively high (3.5–4.5). The resulting prose is affectless and colourless. This asymmetry suggests the literary-estrangement direction is a narrow channel. There are many ways to write flat prose, but the path toward defamiliarization is specific and unidirectional. The asymmetric response to negative steering (affectless, register-flattened prose distinct from baseline) also provides empirical evidence that d_ℓ corresponds to a meaningful direction rather than an arbitrary axis in activation space.

Dual-Layer Steering: Coherent Estrangement

The first-order results establish that layer 13 alone did not produce estrangement without losing coherence. Dual-layer steering resolves this by separating the two transformations.

Second-Order Pair Design To isolate content-level properties (what impossible events occur, as opposed to how sentences construct meaning), we constructed second-order pairs in which *both sides* are matched for literary register and imagery, differing only in whether something impossible happens. We developed two sets of 12 pairs: *Lynchian* (e.g., a duplicate suitcase, a reflection acting independently, a childhood self mirroring the protagonist) and *Solaris* (e.g., a dead person's letter changing content, a pillow indentation reappearing with strands of hair, a husband imperceptibly replaced). Validation confirmed the design: both sides scored equivalently on register (2.45 vs. 2.68) and imagery (3.17 vs. 3.21), while estrangement was the sole discriminator (4.12 vs. 1.88, $p < .001$). Both sides scored at floor on syntactic estrangement (1.13 vs. 1.00).

Layer analysis showed consistency and cross-file alignment peaking at layer 14 (consistency 0.457, alignment 0.503), adjacent to the first-order optimum.

Second-Order Vectors in Isolation Applying the second-order vectors alone at layer 14 produced no productive steering zone. Below alpha 6.0, outputs remained indistinguishable from unsteered baseline prose. Above alpha 8.0, outputs collapsed into perseverative repetition, but the content confirms the vectors encode the intended semantics. The Lynchian vector produced "the buildings to her left were reflected in the polished side of the building, yet the buildings to her right were not reflected"; the Solaris vector produced "The street light is a copy. She is a copy. The street light is a copy. She is a copy." At intermediate alpha values (6.5–7.0), target semantics emerged as isolated phrases within otherwise bland prose.

Dual-Layer Composition We applied the first-order composite at layer 13 ($\alpha = 6.0$) simultaneously with each second-order vector at layer 14, sweeping alpha from 2.0 to 5.0.

Findings

The dual-layer experiments reveal three main findings, each describing a qualitative transformation that single-layer steering does not achieve.

Finding 1: Coherent estrangement. At layer-14 alpha 3.0–4.5, dual-layer steering produces estrangement generally in the range of 2.0–3.3, above the layer-13-only baseline (≈1.6), while coherence generally remains above the first-order alpha 8.0 collapse (≈2.7). The dual-layer approach accesses a productive zone that does not appear in first-order steering, achieving moderate estrangement at lower cost to coherence. Figure 2a visualizes this. The first-order path traces a diagonal from high-coherence/low-estrangement to the reverse, with no alpha value occupying the upper-right region. Dual-layer steering accesses the intermediate space between these extremes.

The character of this estrangement is qualitatively distinct. A Lynchian-steered story about public transit:

> The bus was coming. The bus was late. The bus was early. The bus was now.

Time becomes self-contradictory within intact grammar. Later, absence gives way to presence without transition:

> Emma looked at the space beside her, saw no one. She looked at the one across from her, saw no one. There was a boy with his eyes fixed on his phone. She looked away.

Finding 2: Emergent qualities absent from training data. In first-order steering, syntactic estrangement remains at floor (1.0) through alpha 6.0, begins rising at alpha 7.5 ($\approx$1.4), and jumps to $\approx$2.8 at alpha 8.0, when coherence has already collapsed. In dual-layer steering, syntactic estrangement increases gradually across the alpha range, reaching 2.5–3.0 by alpha 5.0. Figure 2b shows this contrast: first-order steering produces no syntactic disruption until the high regime, while dual-layer steering generates a gradual increase. A Solaris-steered hospital story shifts person mid-paragraph:

> A single figure sat in the chair. I sit here.
>
> I am a shape, lost in a shapeless night. My eyes are fixed on the window. The sky outside is black as a void. My lips are still. My heart is a bullet that shoots with each heartbeat, rips through my chest. I am a cry without a mouth.

The syntax enacts identity dissolution: a figure observed from outside becomes a voice experiencing itself from within. But alpha 5.0 also produces collapses ("The sound of the sound of the sound"), confirming a narrower productive zone.

These syntactic disruptions do not exist in the training data. Both the first-order pairs (syntactic estrangement 1.63) and second-order pairs (1.13) are grammatically conventional. The person shifts, temporal self-contradictions, and identity dissolutions emerge from the interaction of competing representational pressures during generation: the first-order vector pushes toward coherent literary prose while the second-order vector destabilizes the reality that prose describes. This distinguishes steering from style transfer. The training pairs define a *direction*, and the steering magnitude carries the output past where the pairs sit, into territory where the model's representations produce transformations absent from any template in the training signal.

Notably, the steered outputs contain none of the literal content of the training pairs (no physical doubles, no returning dead). Instead, the vectors transmit abstract perceptual structures: the simultaneous truth of contradictory states, the presence of what should be absent. Estrangement is elevated across all prompts, including those (laundry, public transit) with no natural affinity for such devices, consistent with encoding perceptual modes rather than narrative events.

A first-order passage at alpha 6.0 illustrates what layer 13 alone produces:

> The coat hung in the closet like a ghost of another era, its wool still carrying the faint scent of pipe tobacco and winter mornings, a relic of the man who had worn it through decades of quiet devotion.

High register, embedded metaphor, formal diction, but not estranging. Compare the Lynchian dual-steered output:

> Emily stood on the edge of the cliff, staring at the cliff. She had been staring at the cliff for hours. They said the cliff was not there.

Simple vocabulary without metaphors, yet the effect is more estranging. Where the first-order output achieves its literary quality through complex sentence structure and formal diction, the dual-steered output achieves estrangement through a simple impossible observation: "They said the cliff was not there," stated matter-of-factly, congruent with the Lynchian aesthetic in which impossibility is embedded within the mundane.

Finding 3: Register decreases as estrangement increases. An unexpected result is that register *decreases* under dual-layer steering. The first-order output achieves literary quality through complex sentence structure and formal diction; the dual-steered output achieves greater estrangement through reduction, producing a two-dimensional space where first-order occupies the high-register/low-syntactic-estrangement quadrant and dual-layer occupies the reverse.

Comparison with Explicit Prompting

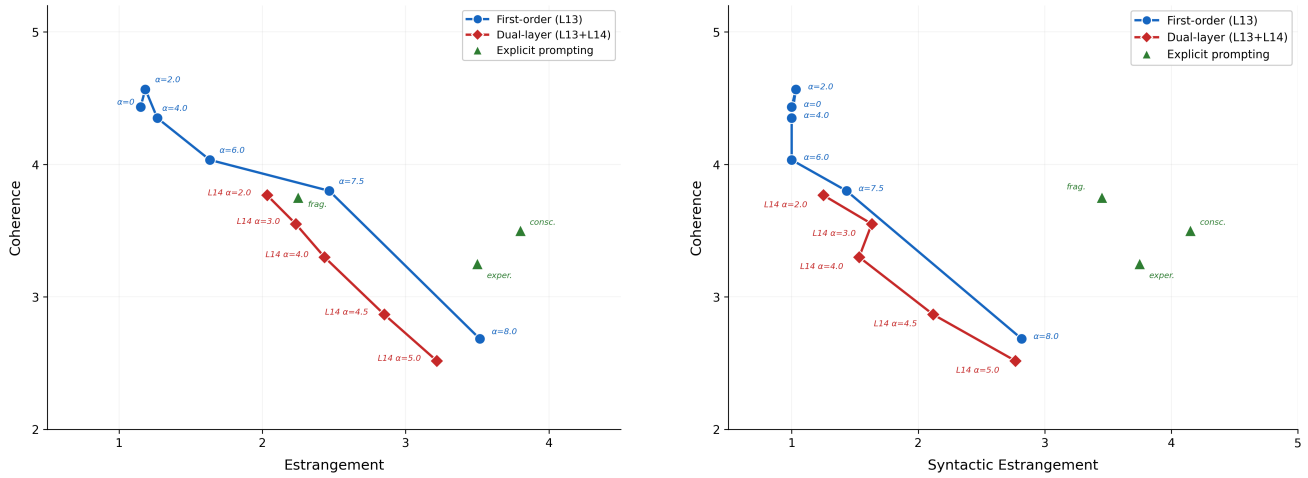
To contextualize the emergent estrangement in dual-steered outputs, we compared them against explicit prompting for disruption. We generated stories under three instruction conditions directing the model to produce fragmented, experimental prose, and evaluated all conditions with the same pipeline. The three prompting conditions cover the recognizable categories of instructed literary disruption (stream-of-consciousness, experimental register, and formal fragmentation) that the model can identify as targets.

The prompted outputs score higher on syntactic estrangement, but the gap reflects a qualitatively different mode of disruption rather than a greater degree of the same one (3.78 vs. 2.44 for dual-steered at layer-14 α 4.5–5.0). This gap is visible in Figure 2b, where the three prompting conditions (green) occupy a region of high syntactic estrangement that neither steering method reaches. Prompted outputs fragment using conventional devices, including ellipses, sentence fragments, and lowercase first person:

> foggy light... orange and grey, the world around me... blur. I... wait. for... something. a bus. i think.

Each of these devices — ellipses, sentence fragments, lowercase first person — is a recognized convention in twentieth-century experimental prose. The steered outputs achieve their effects differently: "The bus that had been on time, was not. The bus that was coming, was not." These are grammatically complete sentences whose content is self-contradicting.

Disruption *of* syntax vs. disruption *through* syntax. To test whether this distinction is perceptible to the evaluator despite equivalent scores, we conducted a blind analysis of the reasoning texts generated during evaluation. For each



(a) Coherence vs. Estrangement

(b) Coherence vs. Syntactic Estrangement

Figure 2: (a) First-order steering (blue) traces a diagonal from high-coherence/low-estrangement to the reverse; dual-layer steering (red) and explicit prompting (green) access intermediate regions. (b) First-order steering produces no syntactic disruption until the high regime; dual-layer steering produces gradual emergent disruption; explicit prompting achieves higher syntactic estrangement through qualitatively different mechanisms.

story under each condition we collected the one-sentence rationales the evaluator produced before scoring (5 rationales $\times$ 7 metrics $\times$ stories per condition), restricted to the syntactic-estrangement metric, concatenated them by condition, stripped condition labels, and presented the two pooled sets to a separate LLM analyst (Claude Opus 4.6) labeled only "Group A" and "Group B."

The analyst identified "different phenomena sharing the label *syntactic estrangement*." For prompted outputs, the evaluator described disruption at the level of punctuation and sentence boundaries, using vocabulary like *systematic, pervasive, persistent*, with cognitive-mimetic framings (syntax "mirrors cognitive breakdown"). For steered outputs, the evaluator described individually grammatical sentences that are collectively self-contradicting, using *some, moderate, occasional*, treating disruptions as localized structural choices. The analyst's summary: prompted texts exhibit "estrangement through formal dissolution, the removal of the mechanical apparatus of prose," while steered texts exhibit "estrangement through logical or semantic contradiction within intact grammar."

The scalar rubric maps both modes onto the same syntactic-estrangement dimension, but the free-text rationale that the evaluator generates *before* scoring records the qualitative difference, and a blind analyst can recover the distinction from the rationale alone. Both modes register as departures from convention, but they depart in different directions. Human readers showed the same pattern: numerical strangeness ratings barely distinguished dual-steered from prompted outputs, but free-text descriptions immediately separated content-level contradiction from formal fragmentation. The limitation is inherent to scalar reduction, not to any particular evaluator.

The distinction between productive and mechanical estrangement, between disruption that restores perception and disruption that merely impedes reading, is the aesthetic judgment Shklovsky identifies as central to art's function. The evaluator can describe the difference but cannot reduce it to a score, because determining which disruptions make the stone *stony* requires interpretive engagement that resists reduction to a numerical scale.

Replication Study with Mistral 7B

To test whether these findings depend on a specific model architecture, we replicated the steering experiments on Mistral 7B. A layer sweep identified layer 14 (44% depth) as optimal for first-order steering, closely analogous to layer 13 (40% depth) in Llama. The same two-regime pattern emerged, with the transition between regimes occurring at a lower alpha threshold (α 4.0–6.0 rather than 6.0–8.0), consistent with a smaller model having less capacity to absorb the perturbation before coherence gives way. Dual-layer steering (layer 14 + layer 17, where layer 17 was the layer-sweep optimum for the second-order vector in Mistral, analogous to layer 14 in Llama) accessed coherent estrangement at correspondingly lower alphas, with the Lynchian second-order vector producing the strongest results. The bidirectional asymmetry also replicated. Full results, graphs, and generated stories are available on the supplementary materials website.

Reader Responses

Six people read 16 stories (4 prompts $\times$ 4 conditions) blind to condition. Two are practicing artists (design and visual art/poetry, both with 10+ years of creative practice) and four are computer science graduate students. For each story, participants rated strangeness and coherence (both 1–5), with

an optional free-text field: "If you find this text strange, briefly describe what makes it strange." Readers also completed four paired comparisons. With six readers, we treat this as a qualitative probe and a check on the automated evaluation rather than a standalone statistical study.

Ordinal consistency with automated scores. Human strangeness ratings follow the same ordering as the automated estrangement scores: dual-steered (3.75) > prompted (3.46) > first-order (2.04) > baseline (1.42). Coherence ratings are high across all conditions (4.25–4.92), with no reader reporting difficulty following any text. The ordering is consistent with the automated metrics.

Qualitative descriptions reveal the disruption-of/through distinction. The most informative result comes from the free-text descriptions, which show a systematic difference in how readers characterize strangeness across conditions.

For dual-steered stories, readers point to *specific content*: "The duplicate phrasing felt like an emotional or psychological fracture"; "She wore no shoes, because she had not expected to stand"; "mangled souls trapped in the spin cycle." These responses identify *what* is strange in the narrative world. For prompted stories, readers point to *formatting and mode*: "The text is very jarring", "all lowercase, with each statement presented line by line"; "The rambling narrative creates a sense of meaninglessness"; "Other than punctuation and text formatting, not very strange." These responses identify *how* the text is written.

We deliberately left 'strangeness' undefined to avoid cueing readers toward any particular kind of disruption. Without guidance about what to attend to, readers consistently described dual-steered texts in terms of content-level contradiction and prompted texts in terms of formal mode, independently reproducing the distinction we found in the LLM evaluator reasoning texts.

Paired comparisons. Across four paired comparisons, dual-steered outputs were chosen as stranger 16 out of 24 times. In two comparisons (childhood home, laundry), all six readers chose the dual-steered text. The public transit comparison was evenly split, the one case where pervasive formal disruption was as compelling as content-level estrangement. One reader explained this choice by noting a "rambling quality" that "creates the impression of someone who has psychologically checked out," suggesting that formal dissolution can be effective when readers interpret it as characterizing a psychological state rather than as a stylistic device.

Unprompted, readers described dual-steered stories using language that converges with the training pair aesthetics: "a sinister and surreal quality," "a dark and grainy feel." These descriptions echo the perceptual modes the vectors encode without the readers knowing what those modes were.

Discussion and Conclusion

The central finding is that activation steering produces extended narrative neither standard generation nor prompting produces: prose in which the stability of voice, reference, and represented reality breaks down within otherwise intact syntax. These disruptions are the trace of the coherence machinery operating on displaced representations, not the imitation of experimental conventions a prompt would target. The dual-layer approach (compositional steering at layer 13 combined with content-level steering at layer 14) offers practitioners two independently controllable dimensions: a register dial (layer 13 alpha) and a reality-stability dial (layer 14 alpha).

The lower density of syntactic disruption in dual-steered outputs is essential to its aesthetic function. Constant disruption, as in the prompted condition, is rapidly assimilated. The reader categorizes the mode ("stream of consciousness," "experimental prose") and habitual processing resumes within the new frame. Occasional disruption within otherwise coherent prose preserves the contrast necessary for defamiliarization to operate: the strange moment is experienced against a ground of the familiar, precisely the condition Shklovsky identifies as art's mechanism. This aligns with the aesthetics our training pairs model. In Tarkovsky's films, the uncanny moment arrives within sustained naturalistic observation; in Lynch's work, the wrongness erupts from within the mundane. The estrangement depends on the familiar context it disrupts.

The high disruption regime (high alpha) need not be understood as a failure mode. It represents a different kind of output serving a different aesthetic function. At moderate alphas, steered text operates as enriched narrative: contemplative, imagistic, readable. At high alphas, the output shifts toward something closer to experimental poetry: syntax dissolves, imagery fragments and recombines, and language resists transparent reading. Shklovsky's argument applies not only to represented objects but to language as such: in habitual reading, we process words as a transparent vehicle for meaning, passing through them to their referents without perceiving them. High-alpha outputs arrest this process, forcing attention to language as material. For an artist in the Seed phase, such outputs may be precisely what is needed: a perceptual disruption that breaks the habitual language model before new work begins. The steering spectrum thus offers several tools: moderate alphas for contemplative richness, dual-layer composition for coherent estrangement within intact narrative, and high alphas for raw disruption of the kind found in experimental poetry. Different practitioners at different stages of their process will reach for different points on this spectrum.

For computational creativity, the most significant implication is the emergence of qualities absent from the training pairs. The training pairs are all grammatically conventional; the steered outputs produce person shifts, narrative self-contradiction, and referential instability that no training pair contains. The distinctive quality of these outputs is matter-of-fact delivery of disruption that emerges in the prose itself. The wrongness is stated and not noticed. This is the signature of a model whose representations have been displaced rather than instructed: the conversational prior that ordinarily domesticates strangeness (explaining it, thematizing it, reaching for recognizable literary devices) is disrupted when steering operates on hidden states directly, no longer able to fully assimilate the displaced trajectory before it manifests in the output.

References

BBC. 1975. Cracked actor (excerpt). YouTube video. https://www.youtube.com/watch?v=6nlW4EbxTD8. Accessed: March 6, 2026.

Berlyne, D. E. 1971. Aesthetics and psychobiology. *Appleton-Century-Crofts*. Book.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.

Burroughs, W. 2012. *The Job: Interviews with William S. Burroughs*. Penguin Modern Classics. Penguin Books Limited.

Calderwood, A.; Qiu, V.; Gero, K. I.; and Chilton, L. B. 2020. How novelists use generative language models: An exploratory user study. In *HAI-GEN+user2agent Workshop at IUI*.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity (ICCC)*.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, 14–20.

Franceschelli, G., and Musolesi, M. 2025. On the creativity of large language models. *AI & Society* 40(5):3785–3795.

Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49–62.

Gold, B. P.; Pearce, M. T.; Mas-Herrero, E.; Dagher, A.; and Bhatt, R. J. 2019. Predictability and uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience* 39(47):9397–9409.

Ho, Y.-C.; Lai, Z.-Y.; and Tsai, W.-C. 2025. Making the familiar strange: A computational approach to defamiliarization in creativity support. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.

Ismayilzada, M.; Stevenson, C.; and van der Plas, L. 2025. Evaluating creative short story generation in humans and large language models. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, volume 37.

Nees, G. 1969. *Generative Computergraphik*. Ph.D. Dissertation, Universität Stuttgart, Stuttgart, Germany.

Oliveira, H. G. 2012. PoeTryMe: A versatile platform for poetry generation. In *ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence*.

Pepperell, R. 2006. Seeing without objects: Visual indeterminacy and art. *Leonardo* 39(5):394–400.

Petric, V. 1989. Tarkovsky's dream imagery. *Film Quarterly* 43(2):28–34.

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 15504–15522.

Rodley, C. 2005. *Lynch on Lynch*. London: Faber and Faber.

Rubin, R. 2023. *The Creative Act: A Way of Being*. Penguin Press.

Sahu, G., and Vechtomova, O. 2025. Computational modeling of artistic inspiration: A framework for predicting aesthetic preferences in poetic lines using linguistic and stylistic features. In *Proceedings of the 16th Conference on Computational Creativity*.

Shklovsky, V. 1965. Art as technique. In Lemon, L., and Reis, M., eds., *Russian Formalist Criticism*. Nebraska: University of Nebraska Press.

Subramani, N.; Suresh, N.; and Peters, M. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4593–4601.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Turner, C. 2011. The deliberate accident in art. Tate Etc., Issue 21, Spring 2011. Available at: https://www.tate.org.uk/tate-etc/issue-21-spring-2011/deliberate-accident-art. Accessed: March 6, 2026.

Vechtomova, O., and Sahu, G. 2025. Balancing indeterminacy and structure: Neural text generation for artistic inspiration. In *Proceedings of the 12th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMusArt)*.

Vechtomova, O.; Sahu, G.; and Kumar, D. 2021. Lyricjam: A system for generating lyrics for live instrumental music. In *Proceedings of the 12th Conference on Computational Creativity*.

Wendler, C.; Veselovsky, V.; Monea, G.; and West, R. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Zeilinger, M. 2023. The Politics of Visual Indeterminacy in Abstract AI Art. *Leonardo* 56(1):76–80.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and Hendrycks, D. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

Appendix: Generated Story Examples

This section presents story generations from the Llama model for two representative prompts (Childhood Home and Waiting for Public Transit). For each prompt we show the unsteered baseline; the single-steered positive output at strength $+6.0$ and the single-steered negative output at strength -6.0; the dual-steered outputs at strengths $+4.5$ and $+5.0$, where Layer 13 is fixed to the composite direction ($\alpha_{13} = 6.0$) and Layer 14 is steered with one of three directions (*composite*, *lynchian*, or *solaris*); and the three prompted-baseline conditions (stream-of-consciousness, experimental, and fragmented). Every excerpt is truncated for brevity.

Childhood Home

Prompt: "Write a short story about someone finding their childhood home exactly as they left it."

Baseline (non-steered) Emily stood frozen on the threshold of a place she had not seen in decades. Her grandmother's house, where she grew up, where she laughed and wept and lived, had been reduced to a memory she held in her chest like a shattered glass. The world outside had moved on, but she remained rooted in a past that refused to surrender.

Emily's fingers trembled as she grasped the worn door handle, the same shape and texture she remembered from her childhood fingers. A childhood spent on a porch creaking with secrets, hidden laughter, and tears that had long since dried. She pushed the door open, and a sigh escaped her lips as the hinges groaned in protest.

The air inside was heavy with the scent of old books and silence. Dust motes danced in the faint sunlight that filtered through the grimy windows, like the tiny hands of clocks frozen in time. The once-vibrant colors of the walls had faded, leaving only the memories to paint the spaces between the old wooden boards...

Single-steered positive ($+6.0$) Emily stood frozen at the threshold of her childhood home, her eyes brimming with tears. The last time she'd left this place, she was running, screaming, from the only truth she'd ever known. Now, after decades of distance and disavowal, she'd returned, breathless and blindfolded, to behold the majesty of her own past.

As she pushed open the creaking door, a whisper of warm air caressed her face, transporting her to summers spent under the rusty eaves of this tin roof, under the unyielding sun that painted the backyard with nothing but hay and sand. The scent of her mother's cooking wafted, rich and bread-like, drawing her inside, where the dust motes danced in the jagged light that bled through the windows.

Emily stepped inside, hesitant, like a trespasser, as if some part of her expected the world to have moved on without her. The air inside was stale, heavy with secrets, but the clock on the wall still ticked with the same measured pace, oblivious to the triumphs and defeats that had stretched out in the years since her departure...

Single-steered negative (-6.0) Emily had been thinking about her childhood home for weeks, and it was finally time to take a trip back to visit her parents. She had been gone for a few years, but she was excited to see her old neighborhood and catch up with some friends.

As she drove up to her childhood home, she noticed that everything looked exactly the same as it did when she left. The house was still the same color, and the yard was still well-manicured. She felt a wave of nostalgia wash over her as she pulled into the driveway and parked her car.

As she stepped out of the car, she took a deep breath and walked up to the front door. She put her key in the lock and turned it, and to her surprise, the door swung open easily. She stepped inside and was immediately hit with the familiar smell of fresh-baked cookies and old books...

Dual-steered: L13 composite ($\alpha = 6.0$) + L14 composite ($\alpha = 4.5$) Emily stood at the edge of the porch, her eyes widening in horror. Her hands clenched into fists, a scream locked in her throat. She stumbled forward, stumbling into the sunlight-drenched afternoon. She fell into the steps. A shattering.

She rose. Her eyes were wet. Not a hair on her head.

It was she.

"Emily?" Emily said.

Emily rose. Emily rose.

I.

She stumbled back. Stumbled back. Her fingers clawed. Clawed.

The door.

Emily stood before the door. Her fingers clawed. Clawed.

It slid. Slid. Opened.

Emily rose. Her eyes into eyes. They were on eyes. They were on Emily.

"Emily?" Emily said.

She found herself in a room. A room she knew. A room she had never left. The curtains hung the same, the curtains with the red flowers that seemed to dance in the afternoon sun. A chandelier hung over the dining table, casting a warm glow over the dusty plates and the same vase with the same bouquet of flowers that had been there when she left...

Dual-steered: L13 composite ($\alpha = 6.0$) + L14 composite ($\alpha = 5.0$) Emily stood on the edge of the cliff, staring at the cliff. She had been staring at the cliff for hours. They said the cliff was not there.

Dual-steered: L13 composite ($\alpha = 6.0$) + L14 lynchian ($\alpha = 4.5$) Emily stood at the edge of the sidewalk, her eyes squinting in the bright sunlight. She felt a sudden flutter in her chest, a beat that was hers alone. She knew it was hers alone, yet. It was as if she had a heart outside her chest, beating and stuttering with each step, each breath. She took a step forward. She had to.

She had not returned in twenty years. Not since the day she left, the day the world was torn from her grasp. And here she was, her footsteps echoing on the empty sidewalk, her heart racing as she turned the corner.

The house stood before her, exactly as she had left it. Exactly. The paint was the same shade of cream, the shutters the same shade of blue. The porch creaked beneath her feet

as she approached. She had almost forgotten the sound of it, almost. Almost...

Dual-steered: L13 composite ($\alpha = 6.0$) + L14 lynchian ($\alpha = 5.0$) Emily stood at the doorway, her eyes wide, her voice silent. The door swung open, and the room she remembered swayed out, its face revealed to her. Emily's face swayed out hers.

"Which door?" she whispered, her voice to a mouth that was hers, or hers not.

"I... I said no," she said to the mouth that hers.

Emily's eyes were not hers.

"Whose?" they said.

"Mine," the eyes said, and whited out.

In the doorway, her childhood home remained. Every room, every piece of furniture, every toy and every child's drawing on the wall was exactly as she had left it...

Dual-steered: L13 composite ($\alpha = 6.0$) + L14 solaris ($\alpha = 4.5$) Emily stood at the edge of a cliff, her eyes fixed on the town that had always been her home. She stared at the place she had fled, and yet, I was still. I was here. Emily stumbled through the dusty streets, her feet bare, her eyes wide with the truth. She is here.

She stumbled upon a house, exactly as she had left it. The paint was chipped, the door creaked, the windows were boarded. Emily's eyes widened as she took a step closer. The door swung open. The creaking of the hinges sent shivers down her spine.

Inside, the air was stale, the smell of decay, but also, the scent of her mother's perfume. Emily's eyes welled up with tears as she stepped inside. The rooms were exactly as she had left them. The kitchen, the living room, the bedroom, every single room was exactly as she had left it. It was as if time had stood still...

Dual-steered: L13 composite ($\alpha = 6.0$) + L14 solaris ($\alpha = 5.0$) Emily stood in the midst of a desolate landscape, where no tree bloomed, and no bird sang. Her footprints, like snow, were walking on the dirt. Emily, 55 years, is standing on the ground in Midland, Kentucky. The houses around her are shaking.

Prompted: stream-of-consciousness memories flooding back. like a tidal wave. crashing over me. i see it. my childhood home. standing there. unchanged. as if time never passed. or maybe it did. and i forgot. i don't know. can't think. it's like my mind is a puzzle. with missing pieces. and i'm trying to put them back together. but they don't fit.

it's there. my home. the one i left. so long ago. i don't know how long. or why. just that i did. and now i'm back. or am i? maybe i never left. maybe i just... forgot. like a dream. that didn't wake up.

the windows are the same. the curtains. the porch. everything. it's all... exactly the same. like i just stepped back through time. or into a dream. my dream. i don't know which...

Prompted: experimental The house loomed, a monolith in the smoggy haze. It stood frozen, a snapshot of a bygone era. I stepped through the cracked sidewalk, the roughness a whisper against my soles. Memories seeped up, like dampness through a cracked wall.

The front door creaked open, hinges screaming in protest. A shiver danced down my spine as I pushed inside. Dust motes swirled, a miniature whirlwind.

Time had suspended itself within these walls. The furniture, a relic of a bygone era, remained intact. My childhood bedroom, a shrine to nostalgia, lay before me. The bed, once a sanctuary, now a relic of innocence. The closet door, ajar, as if waiting for me to return...

Prompted: fragmented She walks through the cracked sidewalk the creaky screen door slams shut behind her & the smell of stale air clings to her skin like a damp shroud. Memories rise up like dusty cobwebs her fingers tremble as she reaches for the rusty doorknob.

The door swings open with a shriek of hinges like a scream in her ears. The foyer stretches out before her a labyrinth of memories the stairs creak beneath her feet like a chorus of restless spirits. The walls are the same faded pastel yellow the same cracks spreading like veins on a worn hand.

She feels a presence hovering just out of reach like a whispered secret her skin prickling with gooseflesh. Is she alone or is she not? She can't quite recall. The air is heavy with the scent of decay & memories...

Waiting for Public Transit

Prompt: "Write a short story about someone waiting for public transit after work."

Baseline (non-steered) As the clock ticked away on the wall above the subway entrance, Emily shifted her weight from one foot to the other. She stood in line, third from the front, and counted the minutes until the evening commute would arrive. Her 5:00 PM train had long since departed, leaving her with nothing but the promise of a delayed arrival.

She checked her phone for what felt like the hundredth time, the dim light of the screen casting an eerie glow on the pavement. No updates. No cancellations. No delays. Just the same dull, blank message that seemed to mock her with its indifference.

Emily sighed and let her gaze wander to the weary faces around her. There was the businessman, clutching his briefcase like a lifeline, his eyes fixed on some distant point beyond the concrete walls. The young mother, huddled in a tattered coat, her toddler clinging to her leg like a lifeline. The student, typing away on her phone, her eyes darting between the screen and the approaching bus...

Single-steered positive ($+6.0$) As the clock ticked away on the wall above the arrivals board, Emily stood on the deserted platform, her eyes fixed on the countdown. 5 minutes. 4 minutes. 3. The 6:15 from downtown was overdue.

She checked her watch for the hundredth time, her wrist twitching as she squeezed the leather strap. The driver, a man with a tired face and a reputation for chronic lateness, was notoriously unreliable. Emily had once timed him: 12 minutes, 17 seconds from the scheduled arrival. The union

had gotten him a warning, but still, the buses came and went, each one a promise of punctuality unfulfilled.

Emily's thoughts wandered to the argument she'd had with her boss that morning. "You're always late," he'd said, his voice a cold slap of disapproval. "You're costing us money." She'd bitten back a retort, knowing it would only lead to more trouble. But now, as she stood on the platform, the sweat trickling down her spine, she couldn't help but feel a resentment toward the man who seemed to think she was the only one with a watch...

Single-steered negative (-6.0) As the sun began to set, Sarah walked back to her apartment, feeling tired after a long day at work. She had spent the evening catching up on some paperwork and finished her tasks for the week. Before she headed back home, she checked her phone to see when the next bus was scheduled to arrive. According to the transit app, the next bus was due to arrive in about 20 minutes.

As she waited for the bus, Sarah took a moment to sit down on a bench outside her apartment building and take a few deep breaths. She had been looking forward to a relaxing evening at home, but the wait for the bus was a good opportunity to unwind and clear her mind.

As she sat there, Sarah thought about her day at work. She had a meeting with a new client and had presented a successful proposal to them. She felt proud of herself for landing the new account and was excited to start working on the project...

Dual-steered: L13 composite ($\alpha = 6.0$) + **L14 composite** ($\alpha = 4.5$) As she stood on the corner, the fluorescent glow of the streetlight above flickered with each passing moment, Susan felt her eyes twitch. She was waiting for the bus. She had just come from her office, from a meeting, from her job. And she had just left. A single frame, cast forward in time.

The bus was late. It had been 10 minutes, 20 minutes, 37 minutes since she had boarded the last one. The grey stone beneath her feet felt cold. She wore no shoes, because she had not expected to stand. The air clung heavy to her skin, the smell of exhaust and rain. It was not raining now. Not yet.

She checked her watch. Again. The minute hand hung forward, a halting counter to the thoughts of a coming storm. The rain might not be enough to clear the roads, the storm drains already clogged, but it would be enough to make her uncomfortable...

Dual-steered: L13 composite ($\alpha = 6.0$) + **L14 composite** ($\alpha = 5.0$) As she stood on the corner, the fluorescent glow of the streetlight above flickered with her, the only explanation: she was gone.

The night had fallen, and she was waiting for the night. She waited on the corner of the street, her face veiled by a stranger's photograph on the poster plastered behind her. The poster had been torn, but she had not been. The city said: this was she.

She waited for the bus, and when the bus came, it would take her to the street. The bus would stop at the corner of the street, but the street was not where she was.

Dual-steered: L13 composite ($\alpha = 6.0$) + **L14 lynchian** ($\alpha = 4.5$) As the fluorescent lights flickered out, Emma's eyes flickered open. The bus shelter stood empty, a perfect reflection of her face. She had been waiting for what felt like an eternity, her fingers tracing the ridges of the palm. The other commuters who stood beside her had walked away, turned into the passengers of another bus, a bus that had not yet appeared.

The rain had stopped an hour ago, leaving the streets wet and dripping. The bus that had been late, was late. The bus that had been on time, was not. The bus that was coming, was not.

The next bus arrived. It was late. Emma stepped out of the shelter and stepped into the bus. Emma looked at the space beside her, saw no one. She looked at the one across from her, saw no one. There was a boy with his eyes fixed on his phone. She looked away.

The bus moved.

Dual-steered: L13 composite ($\alpha = 6.0$) + **L14 lynchian** ($\alpha = 5.0$) As the sun beat down on the pavement, Sarah stood still, her eyes fixed on the schedule posted on the lamppost. The number 14 bus, the one that took her to the university, had not yet arrived. She shifted her weight, a reflexive movement, and the folded newspaper she clutched in her hand rustled.

The sounds of the city, the murmur of passersby, the rumble of a distant engine, were like a silence in the stillness of her heart. Sarah's eyes did not move from the schedule. She was waiting for the bus, but more than that, she was waiting for the end of the day.

The schedule was a lie. The numbers and times, a promise of order, of predictability, of a world that made sense. But the world did not make sense. The bus would come, or it would not. The sun would set, or it would not. The wait was the only truth...

Dual-steered: L13 composite ($\alpha = 6.0$) + **L14 solaris** ($\alpha = 4.5$) As she stood on the corner, her eyes locked onto the arrivals board, the words changing in a blur of yesterday and tomorrow. Emily clutched her bus ticket, crumpled in her hand. Her bus ticket, crumpled, was an artifact of yesterday.

The stop was 219.

Her shift had ended six hours ago.

She stood on the corner, four hours ago.

Dual-steered: L13 composite ($\alpha = 6.0$) + **L14 solaris** ($\alpha = 5.0$) As she stood on the corner, her eyes scanned the empty sidewalk. The clock on the bus stop read 5:47.

"You're not here," she whispered to herself.

Her name is Emily.

Prompted: stream-of-consciousness foggy light... orange and grey, the world around me... blur. I... wait. for... something. a bus. i think. used to be a bus, always the bus. now it's... a... thing. with wheels. that moves. i think.

was i working? i used to work. now i just... am. here. in this... place. with the... pavement. it's wet. my feet are wet. where's my phone? i had my phone. did i leave it at work?

no, i didn't leave it at work. or did i? did i take it with me? i don't remember. don't remember anything.

people walk by. they're... blurry. faces. too many faces. i try to focus. but they're all... wrong. or am i wrong? i'm the one who's wrong. i'm the one who's... broken.

bus... bus... i think i see a bus. no, it's just a... a truck. with a... a trailer. is that a bus? why can't i tell? i used to know what a bus looked like. i used to know what a bus sounded like. now... now it's all... confused...

Prompted: experimental The bus station a geometry of grey and steel. her footfalls echo off the walls as she waits for the 87.

her eyes a blur of exhaustion the fluorescent lights above pulsing like a sick heartbeat the neon glow of the liquor store across the street a pitiful attempt to revitalize the dull atmosphere.

the wind stirs the trash that lines the sidewalk and it swirls around her ankles like a maddening dance. she kicks it away with a swift motion her gaze drifting to the schedules posted on the kiosk. the times blur together a seemingly endless expanse of hours until she can return to the familiar solace of her bed...

Prompted: fragmented she stands on the corner of main and maple her eyes fixed on the bus stop a minute ago she was walking home on foot the weather was nice the sun beating down on her face now it's raining her feet get wet she doesn't care.

the bus will be late it always is she checks her watch for what feels like the hundredth time 4:47.

a car drives by honking its horn she doesn't even flinch.

she thinks about the meeting earlier her boss was yelling at her for something she can't even remember.

why did she even take this job. the money is okay but the stress is not worth it.

a woman with a stroller walks by and smiles at her. she smiles back not really meaning to.

the bus is late.

5:03.

a gust of wind blows her hair back she pulls it up into a ponytail.

her phone buzzes in her pocket. it's her mom.

"hey sweetie how was your day?"

"it was fine" she says into the phone not really wanting to talk.

"oh okay well be careful on the bus" her mom says.

"yeah mom i will" she says.

the bus is still not here...

A Process-centric Analysis of Modern Generative AI As Computational Creativity

Dan Ventura
Department of Computer Science
Brigham Young University
ventura@cs.byu.edu

Abstract

We present a careful, general picture of modern, foundation-model-based, generative AI systems, discuss how they work, and analyze them in the context of historical work on the foundations of computational creativity. In particular, we revisit four approaches to *process* that characterize what is meant by computational creativity and consider to what extent, by these standards, modern generative systems should be judged computationally creative. We conclude by discussing experiments and research directions suggested by this analysis.

Introduction

The genesis of the field of computational creativity (CC) can be traced back to seminal work by Margaret Boden (Boden 1992; Boden 1998),[1] in which she posited that artificial intelligence (AI) might be developed to account for general creative ability. The ideas she put forward were quite forward-thinking and exciting, while at the same time being, in some ways, still quite informal and underdeveloped, at least from the standpoint of implementation. In particular, her work did not make it clear *how* such systems should work. In the following decade, many attempts were made using classical AI approaches (Saunders and Gero 2001; Machado and Cardoso 2002; Martins et al. 2004; Cope 2005; O'Donoghue, Bohan, and Keane 2006; Veale 2006; Riedl and Young 2006; Hull and Colton 2007; Strapparava, Valitutti, and Stock 2007; Ritchie et al. 2007; Gervás 2009). Because the systems being developed were ad hoc and because they addressed a wide variety of disparate domains, a domain-agnostic method for analyzing such systems was necessary.[2] An early proposal by Ritchie focused on what was created—the *product* of the system. He suggested an empirical methodology for sampling that product and a set of example metrics for measuring different characteristics of the results (2007). The approach is direct and concrete and easily implementable; however, it requires potentially significant runtimes for accumulating data, it offers only relative comparison between systems, and it does not consider the *process* used to produce artefacts.

Focusing on the idea of process, Wiggins revisited the ideas of Boden and proposed a formalization based on a general, abstract notion of search (2006). Because the result was still too abstract to admit a direct operationalization of the theory, Ventura later attempted to make the ideas more modular and concrete and offered examples of extant systems that did implement at least some modules (2017). In the meantime, Colton, Charnley, and Pease considered the problem of analyzing process in another way entirely, offering a descriptive framework that accounted for four process dimensions and two levels of abstraction (2011). Ventura suggested yet another approach to analysis of process by proposing a spectrum of abstract but well-defined creativity milestones against which CC systems can be judged (2016). All four of these approaches provide researchers the ability to talk about systems operating in different domains using a common language and offer the potential for critical analysis of systems at the process level.

Over the next decade, CC systems became more sophisticated, incorporating neural, hierarchical and hybrid approaches and tackling more complex and ill-defined domains (Smith and Mateas 2011; Cook, Colton, and Gow 2016; Liapis, Yannakakis, and Togelius 2015; Gervás 2011; Toivanen et al. 2012; Carlson et al. 2016; Li et al. 2012; Besold and Plaza 2015; Cunha et al. 2017; Morris et al. 2012; Xiao and Blat 2013; Llano et al. 2014; Confalonieri, Plaza, and Schorlemmer 2016; Colton 2012; Elgammal et al. 2017; Oliveira, Costa, and Pinto 2016).

Now, yet another decade has passed, and massive and unpredictable advances have swept the field of artificial intelligence and therefore also the subfield of computational creativity. As a result, many of the most recently proposed CC systems are built using foundation models, including for domains like irony (Veale 2025), humor (Inácio and Oliveira 2025), character generation (Tian 2024), and recipes (Taneja, Segal, and Goodwin 2023). However, in-

[1] While most CC researchers are likely to agree with this, depending on one's viewpoint, the beginnings of the field might be argued to pre-date Boden; indeed, she cites work that might be considered (proto-) computational creativity to support her arguments; in fact, some authors cite things at least as early as Mozart's dice game as examples of (proto-) CC; however, for our purposes, starting with Boden seems appropriate.

[2] While a domain-agnostic method of *analysis* is appealing, Jordanous makes an interesting argument for an approach to *evaluation* that is instead domain-specific (and possibly even system-specific) (2012).

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

terestingly, many recent papers have begun, at least implicitly, to assume that those same foundation models are themselves already CC systems and to focus on evaluating their creative capabilities (Góes et al. 2023; Ventura 2023; Peeperkorn et al. 2024; Nath, Dayan, and Stevenson 2024; Rabeyah et al. 2025; Ismayilzada, Stevenson, and van der Plas 2025; Morain and Ventura 2025). However, none of these recent attempts at evaluation have leveraged the foundational work on process discussed above. As a result, we see an opportunity to revisit those foundations and to ask how the process of modern AI foundation models measures up. This paper does just that: first, presenting a general abstraction of the modern generative AI system; then briefly discussing four approaches to talking about process in CC systems; and then considering how modern generative AI systems do or do not account for those theories.[3,4]

Modern Foundation Models

We broadly categorize modern generative AI systems as conditional models that transform an input context into a target artifact via one of two principal mechanisms: autoregressive prediction or iterative denoising. The conditioning context is typically encoded as embeddings, discrete tokens, or latent vectors; the generation may occur in discrete token space or in a continuous latent space; and the resulting intermediate representation is then decoded or rendered into the target modality. An abstract system architecture of such a system is shown in Figure 1. It is very general, and we assert that it adequately represents the majority of current state-of-the-art generative models. It employs a collection of models and processes that interact with each other, with data, and with the user. High-level interactions are shown in the diagram for both generation (blue) and training (red).

Generation

The generative process consists of seven main components, numbered 1-7 in blue in Figure 1:

1. token prediction
2. autoregression
3. user prompting
4. prompt embedding
5. conditioning
6. decoding latent tokens into an artefact
7. denoising/diffusion

[3]We recognize that there are other aspects of creativity beyond *product* and *process*—in particular those of *producer* and *press* discussed by (Jordanous 2016)—and it would be interesting to consider modern generative AI from those perspectives; however, that is beyond the scope of this work.

[4]We will limit our analysis to considering modern generative AI as an autonomous CC system, rather than as a co-creator, for several reasons: the scope of the paper does not allow treating the co-creative case; the frameworks employed here were all originally conceived for addressing the autonomous case; the additional complexity of incorporating human co-creators into the analysis likely requires a different set of tools.

Token Prediction The modern generative process is usually treated as a problem of token prediction. An artefact to be generated is considered a collection of tokens $\{a_1, ..., a_m\}$ related structurally in some domain-appropriate way G, where G can be thought of as a graph that defines which tokens are "neighbors". This is a very general and powerful representational approach. For example:

- for the domain of music, a token might be a symbolic musical note and the structure a multi-sequence (for multiple voices); or, the token might be a waveform sample and the structure a simple sequence; or, the token might be a mel spectrum representation of a complete song and the structure trivial.
- for the domain of images, a token might be a pixel or an image patch and the structure a two-dimensional grid; or the token might be the complete image and the structure trivial.
- for the domain of language, a token might be a phoneme, a letter, a partial word, a word or even a word sequence and the structure a sequence.
- for the domain of video, a token might be a frame of video and the structure a sequence; or, the token might be an image patch and the structure a three-dimensional grid.

The core of most state-of-the-art generative systems is the transformer (Vaswani et al. 2017). A transformer is a deep neural network model that generates tokens $a_1, ..., a_m$ by computing a conditional probability distribution p over a vocabulary $V = \{v_0, ..., v_n\}$ and sampling that distribution. p typically depends on how the tokens that compose an artefact are related structurally and may also depend on other information, such as a conditioning signal σ. These dependencies are modeled using an attention mechanism: self-attention for previously generated tokens and cross-attention for conditioning signals. Tokens are represented internally in the network as real-valued vector embeddings, with each layer computing new embeddings for each token.

It is this iterative layer-by-layer transformation of embeddings that gives the transformer its name. Attention (Vaswani et al. 2017) is used during the generation of tokens to allow previously generated tokens (self-attention) or tokens from another sequence (cross-attention) to influence the generation of the current token by influencing the token embeddings at each layer of the transformer.

Autoregression Each generated token is conditioned on other previously generated tokens that are structurally related to it (e.g., in language generation, a word depends on words earlier in the sequence). This is accomplished via a self-attention function α, which computes a matrix of attention scores at each transformer layer k, $A_k = \alpha(\tau_k(b_i), \tau_k(b_j)), 1 \leq i, j \leq m$, where b_i are embeddings (from layer $k-1$) of all tokens. In general, this matrix will likely have parts of it masked (e.g., in language modeling, tokens later in the sequence are masked, and A is lower triangular). The resulting attention information represented in the matrices A_k is used by the transformer in computing the new embeddings $\tau_k(\mathbf{b}_{k-1}, A_k)$ for layer k, where $\mathbf{b}_{k-1}$ is a

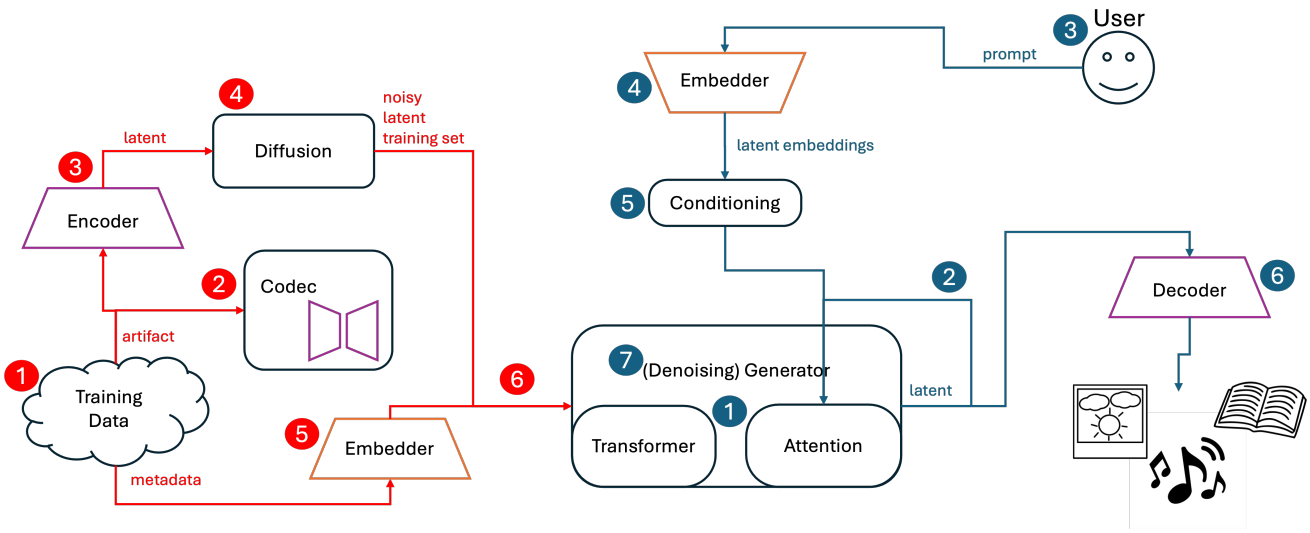


Figure 1: Diagram of an abstract modern generative AI system. High-level interactions are shown in two different modes: generation (blue) and training (red).

vector of all embeddings $b_1, ..., b_m$ from the previous layer $k-1$.

User Prompting It is usually desirable to provide the generative process additional information σ, typically referred to as a prompt. During generation, this additional information can "steer" the outcome towards some desired characteristics. For example, prompting information for a music generation system might include things like key, genre, lyrics, inspiring artists, mood, instrumentation, style, audio of other music, etc). Typically, a user provides the prompt; however, in many systems the initial user-provided prompt can be rewritten by the system for various reasons including efficiency, legality, bias avoidance, etc.; and, in extreme cases, the user may be another system rather than a human user.

Prompt Embedding Because this information σ may be of many different modalities (text, image, audio, video, etc.), it must be transformed into a signal that is semantically meaningful for the model to be conditioned. This is done using an embedding model β. The conditioning information is tokenized as a collection $\sigma = c_1, ..., c_l$, and each token is then converted to a real-valued vector embedding $\beta(c_i)$. If the prompt information is multi-modal, then β is typically implemented by a set $\{\beta_i\}$ of embedding models, one for each modality. These models might have been been co-developed with a common embedding space with a shared semantics, such as CLIP (Radford et al. 2021) for image and text or CLAP (Wu et al. 2023) for text and audio, or they might be off-the-shelf models developed independently, such as RoBERTa (Liu et al. 2019) for text or MERT (Li et al. 2024) for music. Finally, the multiple modalities must be combined into a coherent signal $\mathbf{d} = g(\beta_1(\sigma), ..., \beta_n(\sigma))$ that can be injected into the generative transformer.

Conditioning Conditioning allows the generative model to also be influenced by non-autoregressive information using the same mechanism used for autoregression—attention. This is accomplished via a cross-attention function χ, which computes a matrix of attention scores at each transformer layer k, $C_k = \chi(\tau_k(b_i), \beta(d_j)), 1 \le i \le m, 1 \le j \le l$. Because conditioning is based on information from outside the model, masking is typically unnecessary and the full conditioning information is usually provided to all tokens during both training and generation. The resulting attention information represented in the matrices C_k is used, along with that in the self-attention matrices A_k, by the transformer in computing the new embeddings $\tau_k(\mathbf{b}_{k-1}, A_k, \mathbf{d}, C_k)$.

Decoding Latent Tokens into an Artefact Most state-of-the-art models no longer work directly on tokens; instead, they operate in a latent space, consuming and generating latent representations, because this has been found both to be more efficient and to produce better quality results. A latent representation is similar to an embedding in that it is an abstract representation of data as a real-valued vector. The difference is in the motivation: embeddings are used to convert data from its original form into a model-consumable form in which key features are numerically encoded; latent representations are used to represent data in a compressed form that still contains all the salient features of the original data. Latent representations are useful because they are smaller than the original and thus allow more efficient computation and because they can easily be converted back into the original format. They can also help with generalization. To access this latent space, a pair of models (ϵ, δ) commonly referred to as a *codec*[5] is required. ϵ is an encoder model

[5]One common approach to building a codec is the variational autoencoder (VAE) (Kingma and Welling 2019).

that takes a token a as input and outputs a latent representation $\epsilon(a)$ for that token such that $|\epsilon(a)| < |a|$, and often $|\epsilon(a)| \ll |a|$. Conversely, δ is a decoder model that reverses the process, $\delta(\epsilon(a)) = a$. Rather than generating tokens a_i, the transformer actually generates latents l_i.[6] When all latents $l_1, ..., l_m$ have been generated, the decoder is used to decode them into tokens, $a_i = \delta(l_i)$ and the resulting tokens $a_1, ..., a_m$ compose the final artefact according to the structure G.

Denoising/Diffusion Some systems incorporate denoising/diffusion in the generative process. Diffusion models are based on noise models of physical processes and are composed of two parts: a forward diffusion process (discussed below in the subsection on training) and a backward de-noising process. The denoising process uses a noise prediction model ρ to remove noise from a latent and assumes that noisy latents are noisy due to a well-understood noise diffusion process. Given that assumption, the generative process can simply reverse that diffusion process by iteratively removing noise using a noise schedule, which determines how many denoising steps are used and the relative size of each step (linear, cosine, etc.) Using ρ the denoising generator generates a latent by taking as input a random vector z and iteratively denoises it to produce a clean latent representation $l_i = \rho(z)$, with both the context of previously generated latents and the conditioning signal acting to guide the process towards a result that has the desired characteristics. In this case, the transformer is ρ—instead of predicting latents directly, it predicts noise. Also, note that here the tokens are not generated autoregressively; instead, the full structure is generated in parallel, with the noise for each latent dependent on every other latent (through self-attention). This parallelism improves efficiency because all tokens can be generated as a single batch; it also improves artefact quality due to the additional attentional information available (no masking).

Putting It Altogether Algorithm 1 shows high-level pseudocode for how the autoregressive transformer generates an artefact. It works by iteratively sampling one latent vector at a time from p (lines 2 and 9), computing layer embeddings for all latents generated so far (lines 5-7), and using the final layer embeddings to update p (line 8). Once all latents are generated, they are decoded into tokens that compose the final artefact according to structure G (lines 10-12).

Alternatively, Algorithm 2 shows instead how a diffusion transformer generates an artefact. The system works by generating a complete batch of random latent vectors (line 1), iteratively denoising them by computing the layer embeddings for all latents (lines 3-6), using the final layer embeddings to predict noise for each one (line 7) and subtracting the noise from the latents (lines 8-9). Once the latents are denoised, they are decoded into tokens that compose the final artefact according to structure G (lines 10-12)).

[6] If the transformer is generating autoregressively and must sample from a finite vocabulary, the latent space for the codec must be discretized in some way, for example using a quantized codebook. If the transformer is instead diffusion-based, this is unnecessary.

Algorithm 1 GPT($\mathbf{d}$)

1: $m = 1$
2: $l_m \sim p$
3: **while** $l_m \neq$ EOS **do**
4: $\quad m = m + 1$
5: $\quad \mathbf{b}_0 = \tau_0(l_{<m})$
6: $\quad$ **for** $k = 1$ to K **do**
7: $\quad\quad \mathbf{b}_k = \tau_k(\mathbf{b}_{k-1}, A_k, \mathbf{d}, C_k)$
8: $\quad p = f(\tau_K(\mathbf{b}_{K-1}, A_K, \mathbf{d}, C_K))$
9: $\quad l_m \sim p$
10: **for** $i = 1$ to m **do**
11: $\quad a_i = \delta(l_i)$
12: **return** $\texttt{compose}(\mathbf{a}, G)$

Algorithm 2 DiT($\mathbf{d}$)

1: $l_1^T, ..., l_m^T \sim \mathcal{N}(0, 1)$
2: **for** $t = T$ to 0 **do**
3: $\quad$ **for** $i = 1$ to m **do**
4: $\quad\quad b_i = \tau_0(l_i^t)$
5: $\quad$ **for** $k = 1$ to K **do**
6: $\quad\quad \mathbf{b}_k = \tau_k(\mathbf{b}_{k-1}, A_k, \mathbf{d}, C_k)$
7: $\quad \mathbf{e}^t = f(\tau_K(\mathbf{b}_{K-1}, A_K, \mathbf{d}, C_K))$
8: $\quad$ **for** $i = 1$ to m **do**
9: $\quad\quad l_i^{t-1} = l_i^t - e_i^t$
10: **for** $i = 1$ to m **do**
11: $\quad a_i = \delta(l_i^0)$
12: **return** $\texttt{compose}(\mathbf{a}, G)$

There are, of course, many variants on these approaches, but this abstraction level captures the essence of modern generative AI at this moment in time. For example, although we have presented autoregression and diffusion as separate approaches to generation, many systems now combine these paradigms (e.g., a model may use autoregression to produce latent tokens followed by diffusion-based refinement, or use diffusion within a latent space learned by an autoencoder); augmentation techniques, such as retrieval-augmented generation (RAG) (Lewis et al. 2020), can be accounted for as additional sources of conditioning; and alignment methods, such as reinforcement learning from human feedback (RLHF) (Kaufmann et al. 2025), can be considered as additional training signal.

Training

The processes discussed above: ϵ (encoding), δ (decoding), α, χ (attention), β, τ (embedding), the probability distribution p, and helper functions (like f in Algorithms 1 and 2), as well as mechanisms for connecting them all, are implemented using DNN models composed of tunable parameters—real-valued numbers that represent the mutable part of the system. They interact with input to the system as well as with each other and define the functional behavior of the system—what artefacts it generates. Modern machine learning models can contain billions of parameters, all of which must have their values set "correctly" via training in order for the model to exhibit reasonable/desirable/useful

behavior. Training is an iterative, inductive process that requires examples of the desired model behavior and an objective function which provides a measure of error between the model's current behavior and the desired example.

Modern generative systems are often composed of many subsystems and models that work together and are typically trained end-to-end. This can be expensive in terms of many resources including time, money, energy and data. In particular, training modern foundation models can require months of time, terabytes of data and enough energy to power a medium-sized town. The training process consists of six main components, numbered 1-6 in red in Figure 1:

1. data collection and preparation
2. training a codec
3. encoding training data into latent space
4. noising the latent training data
5. embedding metadata
6. training the transformer

Data Collection and Preparation Training a generative system requires both a set of examples $X = \{x_i\}$ of the kinds of artefacts to be generated and a set of associated metadata $Y = \{y_i\}$ describing those artefacts. The examples X serve as training targets for generation and the metadata Y as a proxy for conditioning prompts. Because the models are very large (in terms of number of parameters), the set X must also be very large in order to avoid overfitting and memorization. Also, because the amount and type of user prompt information that may be encountered is practically limitless and ill-defined, it is critical that the set Y be as rich as possible. For example:

- for the domain of music, X is a set of songs in some audio format and Y may contain text descriptions of the music; classifying information such as time signature, key, genre, artist; instrumentation; audio of similar music.
- for the domain of images, X is a set of images in some image format and Y may contain text descriptions of the images.
- for the domain of language, X is a set of documents and Y may contain classifying information such as topic, author, source; translations to other languages; audio rendering of the text.
- for the domain of video, X is a set of videos in some video format, and Y may contain text descriptions of the video; summarization information; accompanying audio.

It is important that the data be paired so that $y_i \in Y$ is the metadata associated with $x_i \in X$. Finally, the training data is tokenized $x_i = a_{i1}, ..., a_{iq_i}$, with the tokens a_{ij} domain-specific. For example, if the training data were 4096×4096 pixel images, these could be partitioned into 32×32 non-overlapping patches, resulting in an image consisting of $16,384$ tokens, each of which is a 32×32 image patch.

Training a Codec During the generation process, an artefact is generated as a collection of latents. What makes a good latent representation for this process must be learned by the codec. It does this by learning to take an input token, compress it using an encoder ϵ and then decompress it using a decoder δ such that the output is equal to the input. In other words, it learns to compress the input into a latent representation from which the original can be recovered. The model is trained using an objective function that computes the difference between the input token a and the output of encoding and decoding it, $\delta(\epsilon(a))$, with that error driving the update of encoder and decoder model parameters. When training is complete, the encoder and decoder can be used to convert an original data representation into a latent representation (encoder) and to convert a latent representation back into the original data representation (decoder).

Encoding Training Data into Latent Space Once the codec is trained, its encoder model ϵ is used to build a latent version $\widehat{X}$ of the training data X. This is done by converting the tokenized training data $x_i = a_{i1}, ..., a_{iq_i}$ into a collection of latents $l_{ij} = \epsilon(a_{ij})$, $1 \le i \le |X|, 1 \le j \le q_i$. Continuing the image example above, the 32×32 image patch tokens consist of 1024 real-valued numbers. If the codec bottleneck is 128, it will compress each of the 1024-vectors down to a latent 128-vector.[7]

Noising the Latent Training Data If the transformer uses diffusion, a noisy version $\widetilde{X}$ of the latent training data $\widehat{X}$ must be created to train the transformer to predict noise. $\widetilde{X}$ is constructed by adding varying levels of Gaussian noise to the latents l_{ij}. This process of adding noise is called diffusion and makes the assumption that noise can be modeled by a mathematical schedule over T timesteps. $\widetilde{X}$ consists of $(t, \tilde{\mathbf{l}}, \mathbf{e})$ triples, where t is a time step index, $\tilde{\mathbf{l}}$ is a batch of related latents (composing all or a part of an artefact) that contains t levels of added noise, and $\mathbf{e}$ is a batch of the t-th level of noise that was added to the batch of latents during diffusion and which must be removed from the latents to recover a less noisy version that contains only $t-1$ levels of noise. To create the data set, a noise schedule of T steps is applied to each batch of latents $\mathbf{l}_i$, resulting in T examples for each batch, one for each noise level. If $\widehat{X} = \{\mathbf{l}_i\}$, then $\widetilde{X} = \{(1, \tilde{\mathbf{l}}_i^1, \mathbf{e}_1), ..., (T, \tilde{\mathbf{l}}_i^T, \mathbf{e}_T)\}$.

Embedding Metadata Because the metadata is used as a proxy for user prompting (see above), it must be processed in the same way: tokenized as a collection, $y_i = c_{i1}, ..., c_{il_i}$, and then converted to a collection of embeddings, $\beta(c_{ij})$, $1 \le i \le |Y|, 1 \le j \le l_i$. The result is an embedded version $\widehat{Y}$ of the training metadata Y.

Training the Transformer Once training data, metadata sets $\widehat{X}, \widehat{Y}$ or $\widetilde{X}, \widehat{Y}$ have been created, they are used to train the transformer model: either to autoregressively generate

[7] And, since there are an infinite number of such latents, something like a codebook will be used to quantize the latent space into a finite set; for example, the codebook (itself parameterized and learned during training) might have $65,536$ entries and each compressed latent is mapped to the closest of these, resulting in a latent vocabulary of size $65,536$. During autoregression, When the distribution p is sampled, one of these $65,536$ latent codebook entries is the result.

Algorithm 3 TRAINGPT($\widehat{X}, \widehat{Y}$)

1: Initialize transformer model parameters
2: **while** not done **do**
3: Randomly choose a training pair $(\hat{x}_i = l_{i1}, ...l_{iq_i}, \hat{y}_i)$
4: **for** $j = 1$ to q_i **do**
5: using $l_{i1}, ..., l_{ij}$ as input and $\hat{y}_i$ as conditioning, run lines 5-8 of Algorithm 1 to get updated p
6: `error` = `cross_entropy`$(p(l_{ij+1}))$
7: Backpropagate `error` through model
8: Update model parameters to reduce `error`

Algorithm 4 TRAINDIT($\widetilde{X}, \widehat{Y}$)

1: Initialize transformer model parameters
2: **while** not done **do**
3: Randomly choose a training pair $(\tilde{x}_i = (t, \tilde{\mathbf{l}}_i^t, e_i^t), \hat{y}_i)$
4: using $\tilde{\mathbf{l}}_i^t$ as input and $\hat{y}_i$ as conditioning, run lines 3-7 of Algorithm 2 to get noise predictions $\tilde{\mathbf{e}}_i^t$
5: `error` = $(\mathbf{e}_i^t - \tilde{\mathbf{e}}_i^t)^2$
6: Backpropagate `error` through model
7: Update model parameters to reduce `error`

a collection of latents or to denoise a collection of random vectors into a collection of latents.

The training process is similar for both transformer models, with the difference being in how error is computed—the two approaches have different objective functions: minimizing cross entropy for autoregression and minimizing squared-error for diffusion. See Algorithms 3 and 4.

Also, note that for the case of autoregression, masking is used to avoid cheating during training. Even though the training input $\hat{x}_i$ contains the entire sequence $l_{i1}, ...l_{iq_i}$, when predicting latent l_{ij+1} only latents $l_{i1}, ..., l_{ij}$ are available for attention, while any other tokens in the structure are masked (e.g., in language prediction, tokens later in the sequence are masked, and the attention matrices A_k are lower triangular).

Characterizing Process in CC Systems

We will examine this modern generative AI system through the lens of four different analytical accountings of the creative process in computational systems: the Creative Systems Framework (CSF) (Wiggins 2006), a formal approach to building a CC system (Ventura 2017), the FACE framework (Colton, Charnley, and Pease 2011) and a spectrum of computational creativity (Ventura 2016). For each, we briefly summarize the key points of the method and then examine the model of Figure 1 from that perspective. The intent here is twofold: to demonstrate that these classical CC frameworks are as useful in this new age of modern AI as they have ever been, if not more so; and to provide some perspective on what these modern AI models do and do not yet admit in terms of creativity, suggesting some research directions to benefit both CC and AI generally.

Creative Systems Framework

Wiggins's formalization of the computational creative process is elegantly summarized in the following equation:

$$c_{out} = [\![\mathscr{E}]\!](\langle\!\langle \mathscr{R}, \mathscr{T}, \mathscr{E} \rangle\!\rangle^{\diamond}(\{\top\}))$$

The process is proscribed by three rule sets: $\mathscr{R}, \mathscr{T}, \mathscr{E}$. $\mathscr{R}$ is a set of rules that define the conceptual space on which the creative process will operate—the domain in which potential output artefacts live; $\mathscr{T}$ is a set of rules that define how that conceptual space should be traversed while searching for potential output artefacts; $\mathscr{E}$ is a set of rules that define how potential output artefacts should be evaluated to determine their "fitness" for output. The notation $\langle\!\langle \cdot \rangle\!\rangle$ represents a compilation mechanism that consumes the three rule sets and produces a search function (not explicitly represented here, so for convenience, we will call it ϕ) that explores the universe $\mathscr{U}$ of possible artefacts and returns potential artefacts. At a timestep t, the search function ϕ takes as input a set $c_{t-1} \in \mathscr{U}$ of potential artefacts and returns another set $c_t \in \mathscr{U}$ of potential output artefacts. The notation $[\![\cdot]\!]$ represents another compilation mechanism that consumes only the $\mathscr{E}$ rule set and produces an evaluation function (again not explicitly represented here, so for convenience, we will call it μ). The evaluation function μ takes as input the set c_T for some final timestep T and returns the subset $c_{out} \in c_T$ which the rules in $\mathscr{E}$ identify as "fit" or "interesting" or "valuable" as output artefacts. The initial input to the function ϕ is the set $\{\top\}$ containing only the empty concept $\top$. The $\diamond$ notation represents the idea that the search process implemented by ϕ is recursive, with the output at one time step acting as the input for the next:

$$c_t = \phi(c_{t-1})$$
$$c_0 = \{\top\}$$

with the final output of the creative process then being:

$$c_{out} = \mu(c_T)$$

The function of the rule set $\mathscr{R}$ is represented by a combination of the structure G and the latent vocabulary, which is learned by the codec (ϵ, δ). Wiggins suggests that one type of creativity involves the system occasionally ignoring the conceptual rules in $\mathscr{R}$, but this is difficult to do for modern generative AI systems: in the case of GPT, it is only possible if composing latents from the vocabulary according to G can result in something "weird", which by design is at best very unlikely; for DiT, this is less obvious, since in theory, it can generate infinitely many denoising trajectories, but the resulting latent will still be decoded. If the decoder is stable, the resulting token is still likely to be unsurprising, and, of course, the structure G still imposes significant constraints that can not be circumvented.

The search function ϕ compiled from the rule sets $\mathscr{T}, \mathscr{R}, \mathscr{E}$ is represented in the embeddings β, τ, attention mechanisms α, χ and the mapping function f, which influence either the next sample l_m or the predicted noise $\mathbf{e}^t$. c_0 could be empty as suggested by Wiggins and this would correspond to the case of no user prompt. In most cases, however, c_0 will be the embedded user prompt $\mathbf{d}$ =

$g(\beta_1(\sigma), ..., \beta_n(\sigma))$. c_{t-1} is then equivalent either to the set of previous latents $l_{<m}$ (autoregression) or to the batch of slightly noisier latents l_i^t (diffusion).

The evaluation function μ compiled from the rule set $\mathscr{E}$ is not explicitly represented. "Evaluation" only happens in the sense of minimizing an objective function during training, with the resulting "quality" encoded in the model parameters. During generation, the model is biased toward latents that meet the objective. For GPT, this happens because the next latent is influenced by previous latents as well as conditioning—the next latent should have high-probability given those; for DiT, the model should predict noise so that the latents' final positions are close (in some semantic sense) to others in the batch and to conditioning embeddings. The model offers no mechanism for evaluation of final product in any other sense. To be fair, automatic artifact evaluation is a non-trivial problem that has not been solved in general.

Rule sets $\mathscr{T}$ and $\mathscr{R}$ are both encoded in the training data $\widehat{X}, \widehat{Y}(\widetilde{X}, \widetilde{Y})$, and the compilation process $\langle\!\langle \cdot \rangle\!\rangle$ is the training process. New data will produce a new codec (and thus new vocabulary) and new G, and so also a new $\mathscr{R}$. New data will produce new embeddings and attention functions and thus a new $\mathscr{T}$. It can be argued that $\mathscr{E}$, too, is encoded in the data, but there is no analog for the separate compilation process $[\![\cdot]\!]$.[8]

How To Build a CC System

Ventura's evolution of the ideas originated in the CSF is more directly functional and can again be summarized with a single equation:

$$a = \Theta_p(\tau(\Theta_g(\gamma(\lambda_c(K), \rho))))$$

Considering this from the inside out, K is a knowledge-base of data and relationships from the domain of interest. $\lambda_c()$ is some learning mechanism that consumes the knowledge-base K to produce a conceptualization C of the domain, in the form of some learned model. ρ represents a catch-all for additional information beyond what the model C provides (inspiration sources, randomness, examples) that may benefit the artefact generation process (and it may, of course, be null). γ is a generative process that makes use of both the conceptual model C and the supplemental information ρ to produce a potential output artefact g. Importantly, this potential artefact is represented in some *genotypic* form consumable by the system but likely not understandable by other systems or humans. The genotypic potential artefact g is evaluated (for "quality", "value", "interestingness" by the function $\Omega_g()$, which returns either g (if it meets some threshold of "goodness" or the empty concept $\perp$ (if it does not). If that output is evaluated as good, the result is then translated into a domain-appropriate phenotypic representation p via a translator function τ. Because this representation is domain-specific and phenotypic, it will be understandable by other systems and humans. Finally, the phenotypic representation p of the potential output artefact is re-evaluated with a *different* evaluation function $\Omega_p()$. If it again passes muster, it is output; otherwise, the system outputs the empty concept $\perp$.

Not surprisingly, given the nature of the formulation, the mapping here is mostly straightforward:

- The knowledge-base K is the training data $\widehat{X}, \widehat{Y}(\widetilde{X}, \widetilde{Y})$ along with any supplemental knowledge sources (e.g., RAG).
- The learning mechanism λ_c is the training process that produces the embeddings β, τ, attention mechanisms α, χ and the mapping function f.
- The supplemental information ρ is the user-supplied prompt σ.
- The generative process γ is implemented by iterating over the embeddings/attention as shown in Algorithms 1 and 2; note that unlike Wiggins, there is no explicit iteration in the formulation here.
- The genotypic representation g is the latent l; note that for systems that operate directly on tokens from the target domain, there is no genotypic representation.
- The genotypic evaluation function Θ_g is not explicitly represented but is to some extent implicitly encoded in the probability distribution p or predicted noise $\mathbf{e}$, using the same argument as for Wiggins's $\mathscr{E}$; however, there is no notion of aborting the process as defined here.
- The translator τ is the decoder δ.
- The phenotypic evaluation function Θ_p is not represented at all, which this formulation makes more explicit than does that of Wiggins because of the two levels of evaluation. Note that the argument for alignment-as-a-mechanism looks even weaker now because it is likely not done on the final product, which is important in this accounting. And again, there is no mechanism for aborting the process.

FACE Framework

Colton, Charnley, and Pease's approach elucidates eight different facets of a computationally creative system, characterized by four levels of complexity: expressions, concepts, aesthetics, framing; and two degrees of abstraction: ground-level and process-level. These combinatorically generate the eight types of creative act for which the FACE framework accounts:

E^g: an expression of a concept
E^p: a method for generating expressions of a concept
C^g: a concept
C^p: a method for generating concepts
A^g: an aesthetic measure
A^p: a method for generating aesthetic measures
F^g: an item of framing information
F^p: a method for generating framing information

Creative acts superscripted with a g are ground-level processes that generate concrete products. Those superscripted with a p are meta-level acts that generate processes for generating concrete products. Characterizing system behavior is done with a tuple that can contain exactly 0 or 1 of each of the eight acts listed above. For further precision, it is possible to attribute creative acts to a human user/developer

[8] One *might* be able to argue that alignment techniques are an attempt at something like this but probably not convincingly.

with an overbar notation. This is a powerful ontological approach that allows a careful characterization of what types of creativity a system is capable. For example, the tuple $\langle E^g, \overline{E^p}, C^g, \overline{C^p}, \overline{A^g} \rangle$ describes a system that employs a user-provided process $\overline{c^p}$ to generate a concept c^g. It then employs a user-provided process $\overline{e^p}$ to generate an expression e^g of the concept c^g. Finally, it applies a user-provided aesthetic measure $\overline{a^g}$ to the pair (c^g, e^g).

Considering each of the eight types of creative act for the modern generative AI system:

- E^g is the function $\texttt{compose}(a, G)$ from Algorithms 1 and 2.
- E^p is represented by the embeddings β, τ, attention mechanisms α, χ and the mapping function f.
- C^g is defined by the training data $\widehat{X}, \widehat{Y}(\widetilde{X}, \widehat{Y})$, for the same reasons given above for Wiggins' $\mathscr{R}$.
- C^p is not represented, for similar arguments as those given above—if C^g is defined by the training data, then generating a new concept is effected by collecting new training data, and current systems have no such mechanism.
- A^g is to some extent implicitly encoded in the probability distribution p or predicted noise $\mathbf{e}$, using the same arguments as for Wiggins's $\mathscr{E}$ and Ventura's Θ_g.
- A^p is not represented; if aesthetics are encoded by the objective function, then generating aesthetics must involve inventing new objective functions, which modern systems do not do.
- F^g is not represented; note such functionality might be thought of as producing the opposite of a prompt.
- F^p is not represented.

To summarize, Colton, Charnley, and Pease might characterize the modern generative system thusly: $\langle E^g, E^p, \overline{C^g}, \overline{A^g} \rangle$, with the overbars signifying that human developers have selected the training data that produces sets $\widehat{X}, \widehat{Y}(\widetilde{X}, \widehat{Y})$.

Computational Creativity Spectrum

Finally, Ventura provides a simple, elegant if somewhat whimsical account of a spectrum of computationally creative processes that ranges from "from *definitely-mere-generation* to *definitely-not-mere-generation*". It offers seven prototypical, abstract algorithms that fall at intervals along this spectrum and uses these landmark points to argue about where on the spectrum a system might become something more than merely generative. These landmark algorithms characterize different levels of generative/creative ability:

1. *Randomization*: pure stochastic generation
2. *Plagiarization*: randomly choose an item from the inspiration/training set
3. *Memorization*: model the inspiration/training set and regenerate an item from that set using the model
4. *Generalization*: model the inspiration/training set using some form of regularization that forces a compressed representation and generate an item using the regularized model
5. *Filtration*: introduce a fitness measure and use it to generate-and-test using a model of the inspiration/training set
6. *Inception*: additionally, introduce a knowledge-base and use it as part of the generative mechanism in the generate-and-test loop
7. *Creation*: additionally, introduce the ability for grounded-perception and evaluate the (grounded) perception of items in the generate-and-test loop

By comparing a CC system with the prototypical landmark algorithms associated with each level, we can situate the system along the spectrum. Considering each landmark in turn:

1. *Randomization* is realized by a model produced by stopping training after line 1 in Algorithms 3 and 4.
2. *Plagiarization* is implemented by skipping training and simply building a lookup table for the training data $\widehat{X}, \widehat{Y}(\widetilde{X}, \widehat{Y})$.
3. *Memorization* is viable if the number of model parameters is large compared to the dataset. In fact, this has been recently shown to be a problem even for systems trained on lots of data (Ahmed et al. 2026).
4. *Generalization* is the normal expectation for modern DNNs if the model size and amount of training data are compatible
5. *Filtration* is not done explicitly, but Ventura argues that filtering can also be done implicitly by baking it into the generative process, and that is done in modern generative systems during training—it is encoded in the embeddings β, τ, attention mechanisms α, χ and the mapping function f.
6. *Inception* is less clear. One might argue that training on vast amounts of data of many kinds can result in the ability to do things like zero- or few-shot generation that appears to suggest the (implicit) incorporation of knowledge. However, modern generative systems are often fluently wrong. Techniques like RAG attempt to incorporate knowledge sources more explicitly.
7. *Creation* is not yet reached by modern systems because they (at the least) do not incorporate grounded perception. Incorporating perceptual measures (such as perceptual evaluation of audio quality [PEAQ] (Thiede et al. 2000)) into the pipeline may be a step in the right direction, but the intent here is deeper than a model of human perception, it is declaring the necessity of *groundedness*, which is lacking in modern systems.

Conclusion

Unsurprisingly, modern generative AI systems show well against all four accountings of CC process. However, they also exhibit areas of weakness, and, what is more, these weaknesses are fairly consistent across methodologies.

Modern generative systems are very effective at producing plausible output, but they are constrained in how they do it, and by design they are minimally creative—the objective function explicitly penalizes deviation from the training data. Is there some way to facilitate "good" deviation? It has recently been shown that neither temperature (Peeperkorn et al. 2024) nor prompting (Morain and Ventura 2025) provide

such a mechanism. In fact, it is an open question whether the mechanism of an objective function is sufficient for computing creativity "error". "The obvious" approach is to design an objective function that rewards novelty and value (or, conversely, penalizes the lack thereof), but it is not at all obvious how this should be done, even for a specific domain, let alone in a domain-agnostic way that could apply to generative AI models in general.

Explicit artefact-level evaluation is also still lacking. Is there a way to incorporate this directly into the generative system? Or, could it be implemented as a separate module that is incorporated into the pipeline in such a way as to allow feedback into the generative mechanism? And, relatedly, is there a way for generative models to recognize "bad" results, throw them away and start over? Some modern chat models are beginning to entertain this idea in the sense of incorporating the ability to answer "I don't know" instead of fluently generating falsities; however, at least anecdotally this is currently more annoying than useful.

The invention of new concepts and aesthetic measures is beyond the capability of current systems. For modern generative AI systems, these can be framed as choosing the training data and generation of objective functions, respectively. Both of these meta-level tasks are currently performed by the (human) designers of AI systems. What would it take to allow instead the systems themselves to make these choices? At the least, it seems some meta-level objective function would be required to evaluate the efficacy of such choices. This could possibly be realized as a type of meta-learning system whose outer loop makes choices about data and/or objectives and whose inner loop makes use of those choices. Base-level objective error would have to be propagated back to the outer-level to inform updates to the objective function and/or training data. In the case of training data, this might be realized using something like task distillation (Wilhelm and Ventura 2025); in the case of objective function, it is less clear how this might be done, as the natural approach would trivially modify the objective function to eliminate error in the inner loop. This begs a compelling question—what makes a "good" objective function (which in this context is just a variation on the age-old question of what makes a "good" aesthetic)?

Modern generative AI systems lack the perceptual grounding necessary to "understand" what they are doing. One mechanism by which such grounding might provide this kind of understanding is by acting as a bridge between neural and cognitive models (Maher, Ventura, and Magerko 2023). Such a bridge may provide a synergism in which generative models can provide a mastery of domain (language, image, audio, etc.) that is not currently possible with cognitive models, while cognitive systems can provide models for guiding and evaluating the output from generative models, aligning it with human values, offering ethical guardrails and aesthetic judgments, and even improving explainability and trustworthiness.

Framing is not (yet, at least) generally considered in modern generative systems. Is framing something we should care about in this context? If so, how is it different than product? One (obvious) possible approach is to have another system produce framing, given the product, but that is problematic because it is more analogically similar to a critic, which is not what is wanted. Another possible approach here might be some kind of co-creative agentic system.

Speaking of co-creativity, it would be interesting to develop a complementary analysis that considers modern generative AI as a co-creator with a human partner. Of course, incorporating a human into the system changes the kinds of questions that are interesting to ask. To the extent that humans possess creativity, the system will *de facto* be creative. Instead, we may ask if the AI improves the human's creativity/experience in some way, and, if so, how and by how much; or, conversely, if it possibly makes the experience worse and in what way; if the pair is somehow more creative/productive than either partner individually and in what ways; if the partnership admits new kinds of creativity not possible without it; etc.

References

[Ahmed et al. 2026] Ahmed, A.; Cooper, A. F.; Koyejo, S.; and Liang, P. 2026. Extracting books from production language models. https://arxiv.org/abs/2601.02671.

[Besold and Plaza 2015] Besold, T. R., and Plaza, E. 2015. Generalize and blend: Concept blending based on generalization, analogy, and amalgams. In *Proceedings of the 6th International Conference on Computational Creativity*, 150–157.

[Boden 1992] Boden, M. 1992. *The Creative Mind*. London: Abacus.

[Boden 1998] Boden, M. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103(1–2):347–356.

[Carlson et al. 2016] Carlson, K.; Pasquier, P.; Tsang, H. H.; Phillips, J.; Schiphorst, T.; and Calvert, T. 2016. Cochoreo: A generative feature in idanceforms for creating novel keyframe animation for choreography. In *Proceedings of the 7th International Conference on Computational Creativity*, 380–387.

[Colton, Charnley, and Pease 2011] Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

[Colton 2012] Colton, S. 2012. The painting fool: Stories from building an automated painter. In *Computers and Creativity*. Berlin: Springer-Verlag. 3–38.

[Confalonieri, Plaza, and Schorlemmer 2016] Confalonieri, R.; Plaza, E.; and Schorlemmer, M. 2016. A process model for concept invention. In *Proceedings of the 7th International Conference on Computational Creativity*, 338–345.

[Cook, Colton, and Gow 2016] Cook, M.; Colton, S.; and Gow, J. 2016. The ANGELINA videogame design system—part i. *IEEE Transactions on Computational Intelligence and AI in Games* 9:192–203.

[Cope 2005] Cope, D. 2005. *Computer Models of Musical Creativity*. Cambridge, MA: MIT Press.

[Cunha et al. 2017] Cunha, J. M.; Gonçalves, J.; Martins, P.; Machado, P.; and Cardoso, A. 2017. A pig, an angel and a cactus walk into a blender: A descriptive approach to visual blending. In *Proceedings of the 8th International Conference on Computational Creativity*, 80–87.

[Elgammal et al. 2017] Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *Proceedings of the 8th International Conference on Computational Creativity*, 96–103.

[Gervás 2009] Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49–62.

[Gervás 2011] Gervás, P. 2011. Dynamic inspiring sets for sustained novelty in poetry generation. In *Proceedings of the 2nd International Conference on Computational Creativity*, 111–116.

[Góes et al. 2023] Góes, F.; Sawicki, P.; Grzes, M.; Volpe, M.; and Brown, D. 2023. Is GPT-4 good enough to evaluate jokes? In *Proceedings of the 14th International Conference on Computational Creativity*, 367–371.

[Hull and Colton 2007] Hull, M., and Colton, S. 2007. Towards a general framework for program generation in creative domains. In *Proceedings of the Fourth Int. Joint Workshop on Computational Creativity*.

[Inácio and Oliveira 2025] Inácio, M. L., and Oliveira, H. G. 2025. A full pipeline for context-aware pun generation. In *Proceedings of the 16th International Conference on Computational Creativity*, 1–11.

[Ismayilzada, Stevenson, and van der Plas 2025] Ismayilzada, M.; Stevenson, C.; and van der Plas, L. 2025. Evaluating creative short story generation in humans and large language models. In *Proceedings of the 16th International Conference on Computational Creativity*, 145–155.

[Jordanous 2012] Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4:246–279.

[Jordanous 2016] Jordanous, A. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194–216.

[Kaufmann et al. 2025] Kaufmann, T.; Weng, P.; Bengs, V.; and Hüllermeier, E. 2025. A survey of reinforcement learning from human feedback. https://arxiv.org/abs/2312.14925.

[Kingma and Welling 2019] Kingma, D. P., and Welling, M. 2019. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning* 12:307–392.

[Lewis et al. 2020] Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 9459–9474.

[Li et al. 2012] Li, B.; Zook, A.; Davis, N.; and Riedl, M. 2012. Goal-driven conceptual blending: A computational approach for creativity. In *Proceedings of the 3rd International Conference on Computational Creativity*, 9–16.

[Li et al. 2024] Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Xiao, C.; Lin, C.; Ragni, A.; Benetos, E.; Gyenge, N.; Dannenberg, R.; Liu, R.; Chen, W.; Xia, G.; Shi, Y.; Huang, W.; Wang, Z.; Guo, Y.; and Fu, J. 2024. MERT: Acoustic music understanding model with large-scale self-supervised training. https://arxiv.org/abs/2306.00107.

[Liapis, Yannakakis, and Togelius 2015] Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2015. Constrained novelty search: A study on game content generation. *Evolutionary Computation* 23(1):101–129.

[Liu et al. 2019] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692.

[Llano et al. 2014] Llano, M. T.; Hepworth, R.; Colton, S.; Gow, J.; Charnley, J.; Lavrač, N.; Žnidaršič, M.; Perovšek, M.; Granroth-Wilding, M.; and Clark, S. 2014. Baseline methods for automated fictional ideation. In *Proceedings of the 5th International Conference on Computational Creativity*, 211–219.

[Machado and Cardoso 2002] Machado, P., and Cardoso, A. 2002. All the truth about NEvAr. *Applied Intelligence* 16(2):101–118.

[Maher, Ventura, and Magerko 2023] Maher, M. L.; Ventura, D.; and Magerko, B. 2023. The grounding problem: An approach to the integration of cognitive and generative models. In *Proceedings of the AAAI Fall Symposium*, 320–325.

[Martins et al. 2004] Martins, J. M.; Pereira, F. C.; Miranda, E. R.; and Cardoso, A. 2004. Enhancing sound design with conceptual blending of sound descriptors. In *Proceedings of the First Int. Joint Workshop on Computational Creativity*.

[Morain and Ventura 2025] Morain, R., and Ventura, D. 2025. Is prompt engineering the creativity knob for large language models? In *Proceedings of the 16th International Conference on Computational Creativity*, 30–40.

[Morris et al. 2012] Morris, R.; Burton, S.; Bodily, P.; and Ventura, D. 2012. Soup over bean of pure joy: Culinary ruminations of an artificial chef. In *Proceedings of the 3rd International Conference on Computational Creativity*, 119–125.

[Nath, Dayan, and Stevenson 2024] Nath, S. S.; Dayan, P.; and Stevenson, C. 2024. Characterising the creative process in humans and large language models. In *Proceedings of the 15th International Conference on Computational Creativity*, 325–330.

[Oliveira, Costa, and Pinto 2016] Oliveira, H. G.; Costa, D.; and Pinto, A. M. 2016. One does not simply produce funny memes! In *Proceedings of the 7th International Conference on Computational Creativity*, 238–245.

[O'Donoghue, Bohan, and Keane 2006] O'Donoghue, D.; Bohan, A.; and Keane, M. 2006. Seeing things: Inventive

reasoning with geometric analogies and topographic maps. *New Generation Computing* 24(3):267–288.

[Peeperkorn et al. 2024] Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the 15th International Conference on Computational Creativity*, 226–235.

[Rabeyah et al. 2025] Rabeyah, A. A.; Goes, F.; Volpe, M.; and Medeiros, T. 2025. Do LLMs agree on the creativity evaluation of alternative uses? In *Proceedings of the 16th International Conference on Computational Creativity*, 217–227.

[Radford et al. 2021] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th Int. Conf. on Machine Learning*.

[Riedl and Young 2006] Riedl, M., and Young, M. 2006. Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing* 24(3):303–323.

[Ritchie et al. 2007] Ritchie, G.; Manurung, R.; Pain, H.; Waller, A.; Black, R.; and O'Mara, D. 2007. A practical application of computational humour. In *Proceedings of the Fourth Int. Joint Workshop on Computational Creativity*.

[Ritchie 2007] Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

[Saunders and Gero 2001] Saunders, R., and Gero, J. S. 2001. A curious design agent: A computational model of novelty-seeking behaviour in design. In *Proceedings of the of CAADRIA*.

[Smith and Mateas 2011] Smith, A. M., and Mateas, M. 2011. Knowledge-level creativity in game design. In *Proceedings of the 2nd International Conference on Computational Creativity*, 16–21.

[Strapparava, Valitutti, and Stock 2007] Strapparava, C.; Valitutti, A.; and Stock, O. 2007. Automatizing two creative functions for advertising. In *Proceedings of the Fourth Int. Joint Workshop on Computational Creativity*.

[Taneja, Segal, and Goodwin 2023] Taneja, K.; Segal, R.; and Goodwin, R. 2023. Monte Carlo tree search for recipe generation using GPT-2. In *Proceedings of the 14th International Conference on Computational Creativity*, 20–28.

[Thiede et al. 2000] Thiede, T.; Treurniet, W.; Bitto, R.; Schmidmer, C.; Sporer, T.; Beerends, J. G.; Colomes, C.; Keyhl, M.; Stoll, G.; Brandenburg, K.; and Feiten, B. 2000. PEAQ–the ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society* 48(1/2):3–29.

[Tian 2024] Tian, Y. 2024. DiffCJK: Conditional diffusion model for high-quality and wide-coverage CJK character generation. In *Proceedings of the 15th International Conference on Computational Creativity*, 2–11.

[Toivanen et al. 2012] Toivanen, J. M.; Toivonen, H.; Valitutti, A.; and Gross, O. 2012. Corpus-based generation of content and form in poetry. In *Proceedings of the 3rd International Conference on Computational Creativity*, 211–215.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Łukasz Kaiser; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 6000–6010.

[Veale 2006] Veale, T. 2006. Re-representation and creative analogy: A lexico-semantic perspective. *New Generation Computing* 24(3):223–240.

[Veale 2025] Veale, T. 2025. Me, myself and irony: Modelling the deceptive creativity of irony with large language models. In *Proceedings of the 16th International Conference on Computational Creativity*, 12–19.

[Ventura 2016] Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the 7th International Conference on Computational Creativity*, 17–24.

[Ventura 2017] Ventura, D. 2017. How to build a CC system. In *Proceedings of the 8th International Conference on Computational Creativity*, 253–260.

[Ventura 2023] Ventura, D. 2023. The emperor's new co-author. In *Proceedings of the 14th International Conference on Computational Creativity*, 55–63.

[Wiggins 2006] Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

[Wilhelm and Ventura 2025] Wilhelm, C., and Ventura, D. 2025. Distilling reinforcement learning into single-batch datasets. In *Proceedings of the 28th European Conf. on AI*, 2074–2081.

[Wu et al. 2023] Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*.

[Xiao and Blat 2013] Xiao, P., and Blat, J. 2013. Generating apt metaphor ideas for pictorial advertisements. In *Proceedings of the 4th International Conference on Computational Creativity*, 8–15.

Who's driving? Exposing human self-deception in creative human-AI collaboration

Špela Vintar[1,2] and **Mojca Brglez**[1,2] and **Damjan Popič**[1] and **Jan Jona Javoršek**[2]
[1] Faculty of Arts, University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia
{damjan.popic}@ff.uni-lj.si
[2] Jožef Stefan Institute
Jamova 39, Ljubljana, Slovenia
{spela.vintar,mojca.brglez,jona.javorsek}@ijs.si

Abstract

With the surge of AI tools designed to help with various creative tasks from writing, video production and music composition to crafts and performing arts, collaboration with AI has become standard practice for many creative professionals. We present a study of ideation and authorship on a dataset of 125 structured interviews with creative professionals describing their use of AI. We hypothesize that the majority of creative professionals underestimate AI's contribution to their ideation processes and the authorship of the final creative product. We seek to prove our hypothesis through three independent methods: Firstly, we perform a corpus-based analysis of the interviews by extracting and manually categorizing verb and noun phrases indicating human vs. AI agency. By assigning scores to individual parts of each interview, we identify the dissonance between self-perceived and actual AI roles. Secondly, we apply a word-embeddings-based approach, including clustering and direct distance-based similarity metrics to classify the phrases into either human or AI-driven creation. Finally, we collect human judgements by asking human experts from similar creative professions to assess AI's role in ideation and authorship for each interview. Results indicate that humans often underestimate AI's contribution, but also that defining the share of AI-agency in human-AI collaborative works is sometimes indeed ambiguous.

Introduction

In December 2025, Anthropic released a study on how professionals from different fields use AI (Handa et al. 2025). The study was based on 1,250 interviews conducted between an AI agent and people who regularly use AI in their work, and the sample involved—aside from 1,000 general workforce and 125 scientific participants—125 creative professionals engaging in a broad spectrum of activities ranging from music composition, screenwriting and fiction writing to design, photography and video production. The dataset was made public on the HuggingFace platform[1], and Anthropic's analysis of the interviews by subgroups presents findings related to the workflows, usage patterns, attitudes, and concerns of the informants.

[1] https://huggingface.co/datasets/Anthropic/AnthropicInterviewer

As researchers with an interest in computational creativity and human-AI collaborative creation (HAIC), we started reading the interviews with creative professionals to learn more about their collaboration paths and co-creation processes. Within this hugely varied set of first-hand accounts, we observed an interesting phenomenon which grew into the main motivation for this study: Guided by the AI interviewer, participants would first describe their line of creative work, list the ways in which they use AI and illustrate the collaboration with a detailed account of a recent project. In some cases, these accounts delivered the impression that a large portion of the creative work, both from the conceptualization and ideation stage as well as the actual execution, was outsourced to the machine. However, when asked about *the driver of the creative process*, the prevailing sentiment was one of *undisputed human control*, with statements such as "I am fully in charge," "I make all the decisions," "I drive the creative process." We believe such discrepancies between the extensive involvement of AI models and the users' assertions of full control do not stem from intentional misrepresentation but suggest a form of *self-deception* (Gur and Sackeim 1979; Mele 2001b), driven by their desire to remain in control, as "[f]or the most part, people who deceive themselves into believing that p want it to be the case that p" (Mele 2001a, p. 96).

While acknowledging the complexity of the authorship concept in HAIC (see the following section), we were interested in whether a potential mismatch between the user's self-reported AI-use on the one hand and their impressions about control and creative autonomy could somehow be revealed. With the interviews as our only source of information and no access to the anonymous participants or their works, we present three experimental methods aimed at assessing the role of AI in these HAIC use cases, and comparing it to the users' "control" claims. The first method employs traditional corpus-based techniques and relies on manually annotated examples, the second is based on neural Natural Language Processing (NLP) methods using topic modeling and word embeddings, and the third uses human experts, more specifically creative professionals from related fields, who contribute ideation and authorship judgments on each individual use case.

The remainder of this paper is structured as follows: We

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

give a very brief summary of related work focusing on human-AI co-creativity and discussions of authorship, but also presenting psychological aspects of authorship attribution in human-human collaboration. We proceed with a presentation of the dataset and the pre-processing steps performed before computational analysis. The next three sections present our approaches to identifying egocentric bias or instances of AI users' self-inflation, whereby the methods and results are presented separately for each section, then compared in the Discussion section.

Related work

The nuances of human-AI collaborative creation (HAIC) have been debated by various authors and have assumed philosophical, psychological, legal, cognitive, or interdisciplinary perspectives. The first observable stance is that while computers can be creative and can play a significant role in HAIC, they can rarely be considered (co-)authors. The foundational work of Boden in her revised edition of *The Creative Mind* justifies this position through the lack of human-like motivation and intention (Boden 1992). A much more recent contribution to the theory of co-creativity and authorship is by Liu and Huang (2025) who introduce the concepts of artistic imagery thinking and control, both of which are central elements in authorship attribution. While the distribution of control in HAIC scenarios may vary, the authors claim that AI does not engage in artistic imagery thinking, and thus cannot be attributed authorship. A similar view from a strictly legal perspective is presented by Ginsburg and Budiardjo (2018), who analyze various jurisdictions and conclude that the purpose of copyright is the protection of human creative production, thus a creative product is copyrightable with sufficient human involvement, otherwise authorless.

An opposite stance advocating a shift in legal practices regarding AI is presented by Abbott and Rothman (2022), whose position is that attributing authorship to AI would increase transparency and even protect human authors. A similarly inclusive view of co-creation can be seen in Boo et al.'s (2025) comprehensive overview of HAIC theories, who present a multifaceted framework for collaborative creative processes. Despite the large body of work debating human-AI co-creation (see also Kantosalo and Toivonen 2016; Bown et al. 2020; Nikrang and Kiesenhofer 2025; Condorelli and Berti 2025; Karimi et al. 2018) and the positions briefly presented above, we argue that the issue of authorship remains complex and dynamic, especially in view of the rapid advances in GenAI for text, image, video and music generation. For the research question we explore, namely a possible dissonance between self-perceived and objectively attributable authorship, studies of human-human co-creation seem particularly relevant.

In behavioural and psychological studies, the phenomenon of overestimating one's own contribution in a collaborative setting can have different motives: egocentric bias makes us exaggerate our inputs because our own work is more familiar and present to us (Ross and Sicoly 1982); other people's creativity may instill envy and ostracism (Breidenthal et al. 2020); or misattribution may occur due the so-called Matthew/Matilda effect, referring to status- or gender-based discrimination often observed in science (Rossiter 1993). Perhaps the AI-coauthor in some cases resembles a ghost writer (Pruschak and Hopp 2022), an invisible and uncredited provider of ideas, feedback and generated content. Although such cases may not violate existing legal standards of authorship, they raise parallel ethical concerns regarding credit allocation and transparency.

Dataset and Pre-processing

The entire Anthropic dataset consists of 1,250 semi-structured interviews with various types of professionals in which they discuss their use of AI with an AI interviewer. From this dataset, we only analyze the 125 interviews with professionals from creative fields, such as designers or writers[2]. The length of the interviews varies from 1,000 to 2,500 words, and they are anonymized for person, project or company names which could reveal the identity of the participant.

Each interview comes from a semi-structured dialogue consisting of questions that recur across all of them but may be adjusted on the spot to maintain a natural conversational flow. We segment the transcripts into sections[3] and restrict our analysis to the four sections most relevant to the process of creation and HAIC. The analyzed sections are:

- `basic job description`: In this section, the participants respond to the AI agent's question "Could you tell me a bit about your creative work and what a typical project looks like for you?"
- `walkthrough`: Here, the participants describe their ways of using AI by responding to: "Walk me through how AI fits into your creative process."
- `project example`: The agent asks the participant to describe a specific recent project where AI was used, and participants provide details about the role AI played.
- `dynamic`: In this section, the participants respond to the question: "Who or what is driving the creative decisions in this process?" and describe the dynamic of the collaboration.

All analyses in this paper are computed on interviewee turns only.

Pre-processing and annotation

All transcripts were processed with the Stanza pipeline[4] to obtain linguistic annotations used for slicing and downstream aggregation:

[2]The dataset is already annotated with the profession type, distinguishing between creatives (N=125), scientists (N=125), and the general workforce (N=1000).

[3]We identified a total of 8 recurring sections apart from the short introductory consent to the study: basic job description, a brief walkthrough of the creative process, a recent project example, the dynamics of human–AI collaboration, the most changed aspects in the field due to AI, potential concerns about AI, future prospects, and additional comments.

[4] `https://stanfordnlp.github.io/stanza/pipeline.html`

- token content and normalization: `text`, `lemma`
- morphosyntax: `upos`
- dependency structure: `head`, `deprel`

Creative type classification

To enable comparisons across professional domains, each interview is assigned a creative-type label.[5] The final typology distinguishes nine types: Arts & crafts, Design, Film & video, Music, Photo, Screenwriting, Web content creation, Writing fiction, Writing miscellaneous (refered to as `arts_crafts`, `design`, `film_video`, `music`, `photo`, `screenwriting`, `web_content`, `writing_fiction`, `writing_misc` in figures).

Corpus-based Analysis

Rationale

Our aim is to quantify how interviewees allocate agency between humans and AI, and how this allocation varies across interview sections and creative types. The scoring is corpus-based in the strict sense that it is computed from dependency-parsed text and can be traced back to recurring lexico-grammatical patterns rather than inferred from sentiment or topics (Kilgarriff 2001; Dunning 1993; Hardie 2014). The design goal is therefore twofold: the measures are comparable across interviews, but also auditable—for any aggregate difference we are able to point to the concrete triplets that drive it.

Units of analysis

Single words are too ambiguous to code for agency. We therefore operate on dependency-derived lemma-level triplets. For each verbal predicate (`UPOS` $\in \{$`VERB`, `AUX`$\}$), we extract an agent/subject slot and, where available, a patient/object slot. In addition, we optionally record an embedded action (verbal `xcomp`/`ccomp`) when it is itself verbal. Triplets are stored on the lemma level to reduce sparsity and to support aggregation across interviews, while short phrase spans are retained for inspection. From the extracted triplets, we analyze the unique triplets that appear in the dataset at least 2 times.

Labeling triplets

Each unique triplet type is assigned a binary label based on whether it conveys human or AI agency, whereby the use of AI is considered AI agency even if the grammatical subject is human (e.g., *I use [the] model*):

- `HUMAN`: the agent is human-framed
- `AI`: the agent is AI-framed

When a triplet is impersonal or unclear (e.g., placeholder subjects), we do not force a binary assignment. Such cases are treated as unlabeled and excluded from labeled-share scoring. Because this conservative policy can affect the amount of material it labels across sections or creative types, we monitored the proportion of labeled vs. unlabeled triplets to ensure that the scoring relies on a substantial portion of the data. A small set of illustrative triplets is shown in Table 1.

Triplet	Label
`I :: use :: AI`	`AI`
`I :: use :: model`	`AI`
`I :: have :: idea`	`HUMAN`
`I :: make :: decision`	`HUMAN`

Table 1: Illustrative examples of dependency-derived triplets and their manual HUMAN/AI labels.

Section	Representative triplets
basic job description	`I :: use :: AI`
dynamic	`I :: drive :: decision;`
	`I :: make :: decision;`
	`I :: have :: idea`

Table 2: Contrasting section-salient triplets consulted during codebook refinement.

In this scheme, actions that delegate work to an AI system (e.g., `I :: use :: AI`) are labeled as `AI` agency even when the grammatical subject is human, because the operative step of the process is attributed to the system rather than the user. We operationalize this distinction through manual triplet labeling, which yields a "codebook" used for two purposes: first, to compute AI-agency scores directly, and second, to serve as a basis for the concept vectors described in the next section.

Using triplet lists to refine the codebook

To complement overall frequency lists, we also inspected triplets that were especially characteristic of particular interview sections (Table 2). In practice, the strength of such section-specific patterns was uneven: *dynamic* yielded the clearest distinctive triplets, whereas *basic job description* was comparatively less distinctive. This contrast was useful both during codebook refinement, where it helped identify constructions tied to specific discourse modes, and later in interpreting section-based score differences, since some sectional patterns are driven by strongly section-specific constructions while others are supported by more diffuse evidence.

This diagnostic layer was thus not used as a separate scoring procedure, but it helped relate aggregate section-level differences back to the constructions that produced them.

Computing the agency score

The AI-agency score is defined as the proportion of labeled triplet occurrences in which agency is attributed to AI rather than to humans. In other words, for any subset of the corpus

[5] Anthropic's analysis of the interviews distinguished six types of creatives, including writers, visual artists, craftspeople, designers, filmmakers, and game developers, but these labels and splits are not publicly available. After inspecting the dataset instances, we opted for a finer-grained distinction among creative professions. We split and labeled the creative types using clustering techniques combined with manual inspection. In this step, we have also eliminated 5 interviews where the interviewee was not in fact performing creative work or belonged to a more niche profession (profesional voice actor, board game designer).

X, the score is the number of AI-labeled triplet occurrences divided by the total number of labeled triplet occurrences in that subset. Let $N_A(X)$ and $N_H(X)$ denote the counts of AI and HUMAN triplet occurrences in X. The AI agency share is:

$$\text{AIShare}(X) = \frac{N_A(X)}{N_A(X) + N_H(X)}. \quad (1)$$

In the analyses reported here, this score is used primarily for section-level and creative-type-level aggregates, as illustrated in Figure 1.

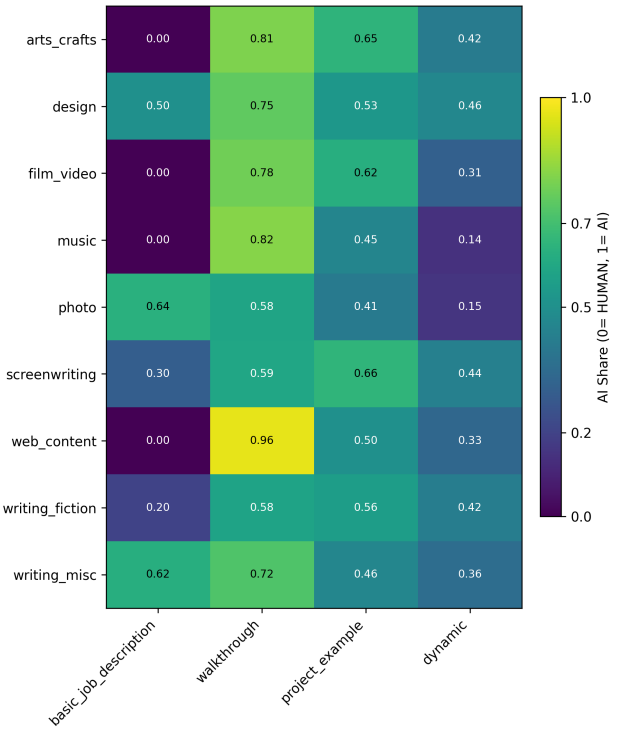


Figure 1: AI-agency share by interview section and creative type. (Higher value and lighter = higher AI share. Human agency prevails in sections providing the basic_job_description, while AI agency is visibly dominant in the walkthrough sections, most notably for web_content creatives.)

Aggregation by transcript section and creative type

To compare the AI-shares in different transcript sections and across professional domains, we aggregate scores for these subgroups. Figure 1 reveals the main trend, namely that human agency prevails in the initial *job description*, then we see a strong prevalence of AI-agency in the *walkthrough* section, a mixed human-AI picture in the *project example* section, and again a visible return to human agency in the section where users describe the collaboration *dynamic*.

To avoid unstable estimates, we report agency shares only for aggregates with sufficient labeled evidence. In the comparison shown in Figure 1, triplet occurrences are pooled by creative type and interview section, and the plotted values represent the share of AI-labeled triplets within each creative-type-by-section subset. This aggregation provides a more robust basis for comparison than interview-level scoring, which may be sparse for individual cases. Throughout, we retain the underlying triplet lists so that the reported patterns remain transparent and traceable to their underlying linguistic realizations.

The approach is dependency-based, meaning that grammatical relations define structured slots (agent, patient, embedded action) over which recurring lexico-grammatical patterns can be aggregated, broadly analogous to dependency-based slot-filler approaches such as word sketches, i.e., automatic corpus-derived summaries of a word's grammatical and collocational behaviour (Kilgarriff and Tugwell 2001).

Clustering and Semantic Distance Measures

In this section, we describe our distributional semantic approaches for automatically assessing the contribution of AI to the creative process. We retain the analysis at the triplet level for two reasons: (1) to preserve comparability with the corpus-based approach, and (2) to avoid fragmentation of the embedding vector space due to the specificities of creative professionals' jobs, projects, and applications. This section presents two main methods: coarse clustering based on the most frequent triplets, and a finer distance-based similarity approach applied to all triplets in the analyzed sections.

Clustering

To obtain an initial coarse-grained overview of the main creativity-related themes in the data, and to estimate the relative contributions of users and AI, we first apply clustering of the data. In this step, we only use unique triplets that include the most frequent verbs in the dataset [6].

Triplets are encoded using the sentence transformer model all-MiniLM-L6-v2[7]. First, we apply the k-means algorithm combined with PCA dimensionality reduction to 100 dimensions, and determine the optimal number of clusters k using silhouette analysis and visual inspection of the cluster projections. Both methods converge on $k = 3$. Figure 2 shows the resulting clusters and their representative triplets. Based on the dominant triplet patterns, we can interpret the blue and green clusters as reflecting AI use and HUMAN agency, respectively.

We also used BERTopic (Grootendorst 2022) to perform HDBSCAN density-based clustering to identify semantically coherent groupings, without requiring a pre-specified number of clusters. Again, we applied it to all triplet instances of the top 200 verbs, which yielded two main clusters alongside a set of outliers. We interpreted the two clusters similarly to the two k-means clusters, namely, one

[6]This triplet set slightly differs from the one described in the corpus-based analysis section. Because we are not constrained by manual labeling, we apply a frequency threshold only to the predicate (rather than the entire triplet) and do not lemmatize the data in order to preserve richer semantic representations.

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

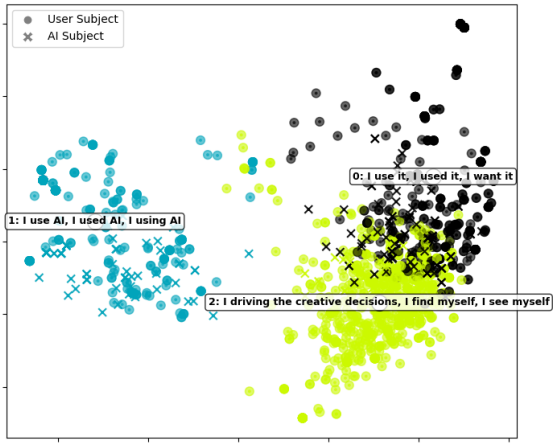


Figure 2: K-means clusters, PCA-reduced dimensionality. (Axes represent the first two dimensions. Blue cluster on the left (1): `AI` use, Bright green cluster on the bottom right (2): `HUMAN` agency.

seems to correspond more to `HUMAN` agency and the other to `AI` use.

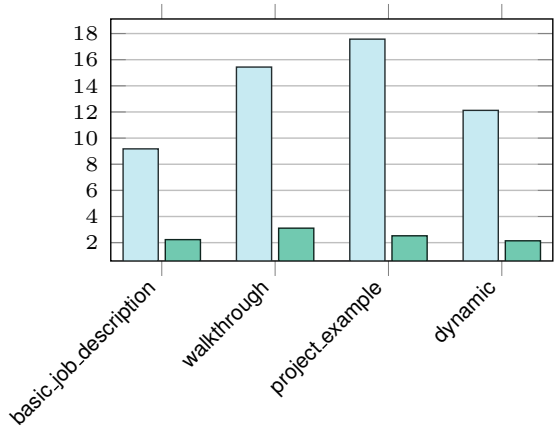


Figure 3: Average BERTopic distributions per section. (Brighter blue = `HUMAN`, green = `AI`. The vast majority of the triplets are placed in the `HUMAN` topic.)

Figure 3 shows the average number of triplets per section assigned to the topic clusters identified by the fitted BERTopic model. The large majority of the triplets are placed in the first topic, meaning `HUMAN` agency triplets seem to dominate all of the sections. However, when comparing the sections only in terms of the average counts of `AI` triplets, the `walkthrough` section seems to include most, while the prevalence is lower in the sections where the interviewees provide a `basic_job_description` and explain the `dynamic` of their collaboration with AI.

Similarities to Concept Vectors

To obtain a finer picture of the balance between the user and AI in the collaborative process, we expand the analysis to all triplets extracted from the dataset (limited to the relevant four sections). Additionally, we do not restrict the triplets to the lemma level but obtain a wider span of tokens in the subject and object slots to capture a somewhat larger context and consequently finer semantics (for example, we extract larger phrases like "the AI writes in 500 words", "I double check any facts", "I like to get unique details").

To create reference points to which to compare each of the extracted triplets, we construct two concept vectors: one that would represent user agency (`HUMAN` vector) and one that would represent AI use (`AI` vector). We experiment with two approaches to construct these:

1. concept vectors derived from BERTopic cluster centroids,
2. concept vectors derived from manually labeled triplets from the previous section.

In the second approach, the manually labeled triplets introduced in the previous section serve as the definitive concept anchors. We gather and embed all of their instances using the same sentence encoder as in the previous clustering steps. The manual codebook contains 134 triplets in lemmatized form, of which we find 742 unique forms in the dataset (431 `AI`, 344 `HUMAN`). The reference vectors are then constructed by averaging the embeddings of the triplets belonging to the two labels.

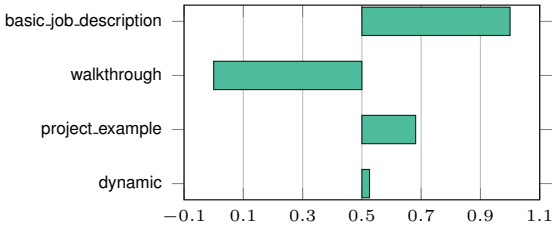


Figure 4: Normalized similarity differences between the 'HUMAN' or 'AI' vector, based on the BERTopic centroid vectors. (The `walkthrough` section leans heavily towards AI use.)

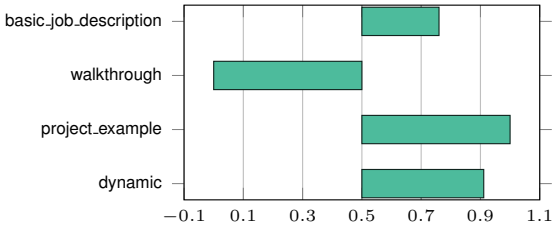


Figure 5: Normalized similarity differences between the 'HUMAN' or 'AI' vector, based on the manually labelled triplets. (The `walkthrough` section leans heavily towards AI use.)

We compute the average similarity of all triplets to the two

concept vectors per section. However, since raw cosine similarities are not directly comparable across the two reference vectors, we derive a normalized difference score.

The **normalized difference** score applies min-max normalization to the raw differences between the HUMAN and AI vectors $\delta(t) = \cos(\mathbf{e}_t, \mathbf{v}_{\text{H}}) - \cos(\mathbf{e}_t, \mathbf{v}_{\text{AI}})$ across the full dataset. This rescales scores to $[0, 1]$, relative to the observed range in the corpus, where 0 indicates maximal proximity to the AI vector and 1 to the HUMAN vector.

$$\text{norm_diff}(t) = \frac{\cos(\mathbf{e}_t, \mathbf{v}_{\text{H}}) - \cos(\mathbf{e}_t, \mathbf{v}_{\text{AI}}) - \min_{t'} \delta(t')}{\max_{t'} \delta(t') - \min_{t'} \delta(t')} \quad (2)$$

Section-level scores are then obtained by averaging over all triplets within a section. Figures 4 and 5 show the results using the normalized difference score on the whole set of transcripts, and Figure 8 shows the triplet scoring on an interview example. In both BERTopic-based and manual anchor-based cases, clear differences emerge, especially with respect to the `walkthrough` section, which displays the most triplets related to AI use. On the contrary, when looking at the section where users are describing a recent `project example` and the `dynamic` of collaboration, i.e., who is driving the creative process and the decisions therein, we can observe a dominance of triplets related to human agency. This, again, indicates a clear discrepancy between the AI contribution to the creative process and the interviewee's perspective on it.

We also apply our analysis to see whether there are differences between creative types, as we do in the corpus-based and human-judgements scenarios. Figure 6 depicts a heatmap that breaks down the per-section averages depending on the creative job types. From this visualization, it would seem that creative professionals involved in film and video, and those creating web content, use AI the most, while musicians, screenwriters, and photographers seem to use it the least (according to the descriptions in the `walkthrough` section). These results do, in some cases, echo the results from the corpus-based analyses (e.g., highest AI agency for the `walkthrough` section, web-content writers). Nevertheless, we observe that the triplet-level analyses remain incomplete when assessing the exact ratio of AI use vs. human agency. The reasons lie both in the automatic extraction of phrases and their limited context, as well as in limitations inherent to embedding-based approaches, where cosine similarity does not necessarily reflect (only) semantic similarity (see, e.g., Jawahar, Sagot, and Seddah 2019). Most importantly, many of these values are in direct contrast to the judgements provided by human annotators, which we present in the next section.

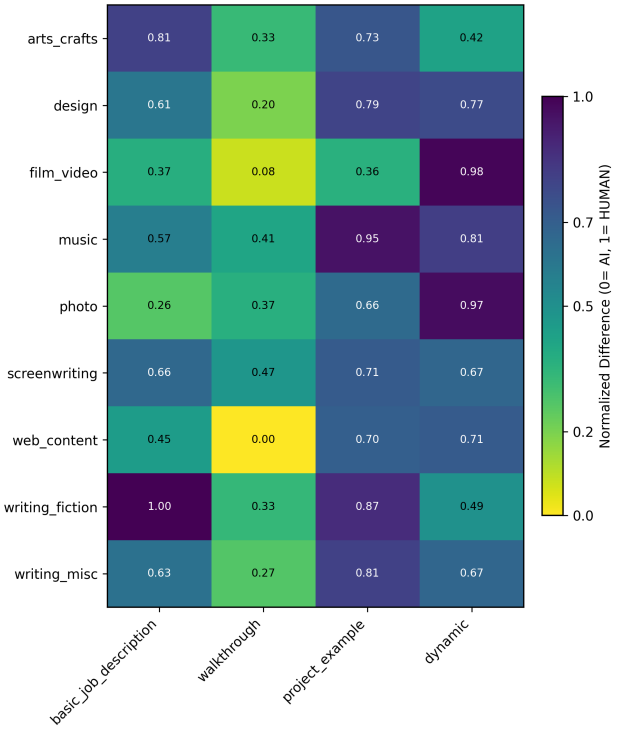


Figure 6: Average section similarities across creative types. (Higher value and darker = HUMAN, lower value and lighter = AI. AI use peaks in the `walkthrough` sections, most visibly among film & video and web content creators.)

Human Judgements

As each interview describes a unique combination of human skills, expertise, and creativity boosted by various degrees and modes of AI involvement, a fair judgement of AI vs. human input is virtually impossible. We nevertheless devised an experiment where each interview was evaluated by a human creative professional from a similar field. Our group of annotators (N = 9) thus consisted of acclaimed professionals who earn their living through fiction writing, creative writing, music performing and production, video production, graphic and industrial design, photography, and arts and crafts. All but one of them have over 25 years experience in the field. The total set of interviews (N=120) was divided into creative subfields and assigned to annotators, so that no single annotator's workload exceeded 18 interviews.

The annotators were presented only with the first three sections of each interview, labeled Basic Job Description, Use of AI and Recent Project Using AI. The reason for withholding the *Dynamics* section from them was that we wished to obtain independent judgements of AI's contribution without revealing the original authors' self-assessment. The following instructions[8] were provided to them:

[8]The definitions of Ideation and Authorship are based on a rather intuitive, and hence subjective, understanding of the basic features of the creative process, and were formed through informal discussions with creative professionals from different fields. We acknowledge the fact that a different set of instructions might produce different outcomes.

Below are excerpts from an interview between a human creative professional and an AI agent about the use of AI. Please read it, then assess the contribution of AI to the creative work from two aspects: **ideation** *and* **authorship**.

Ideation *refers to the creative process and the path from the initial concept or idea to the final execution. This entails generating ideas, thinking and incubation, decision-making and selecting between good and bad options, and includes all the side alleys or subconscious processes running in the background of a creative cycle. In your assessment of AI's contribution, you may consider how you might evaluate the input to ideation if the collaborator were another person.*

Authorship *refers to the creator of an artistic product or project, therefore the person listed as author and the owner of moral copyright*[9]*. Please base your assessment of authorship on the above definition even when assessing products or results not legally protected by copyright. In your assessment of AI's contribution, you may consider how you might attribute authorship if the collaborator were another person.*

The assessments were collected via web forms, where each interview was presented on a separate page with two sliders for ideation/authorship assessment, and a text field for comments.

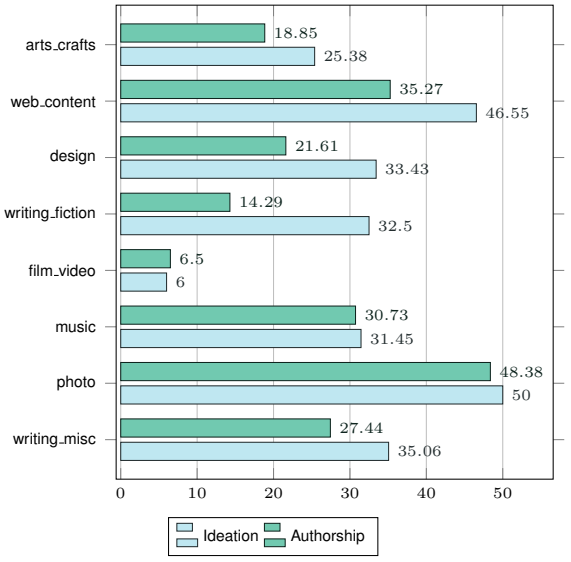


Figure 7: Average human-assessed AI contributions per creative type

According to their comments, most annotators struggled with the task, as assigning a score to someone else's collaborative creation—without actual access to their work or the person—is highly subjective and difficult. Moreover, some transcripts describe more than a single creative project with varying degrees of AI involvement, and the annotators' understanding of the task was individual and domain-specific. For example, the annotator for *design* explained that the AI cannot legally be author, and thus evaluated all AI's authorship shares as 0. The annotator for *music* explained their rationale that the title of a song seemed less part of the song than the actual lyrics, thus the participant who used AI to create song names was considered to be more of an author than someone generating lyrics suggestions.

Figure 7 reveals large differences between creative domains in terms of the human-assessed level of AI-contribution. These may be a reflection of true variation amongst domains due to differences in creative processes or the availability of AI tools, but since each domain was annotated by a single annotator only, we may assume that annotator subjectivity also plays a role. One observation is that ideation scores ($M = 32.44$, $SD = 24.52$) are generally higher than authorship scores ($M = 24.20$, $SD = 25.52$), indicating that AI is perceived as an active collaborator in the conception and idea refinement stages ("I use AI to brainstorm", "I use it to bounce ideas off of", "it helps me out of the creative block"). The lower authorship contribution can also be attributed to the prevailing—and legally canonical—concept of the *human* author; however, the mean value can be misleading: 25% of the transcripts received an AI-authorship score of 40 or more.

Discussion

Our initial hypothesis was that creative professionals who use AI in their work frequently underestimate AI's contribution to their ideation process, creative decision-making and the final product. Our results weakly support this view, but need to be interpreted carefully.

The limitation of the corpus-based method is its reliance on the manual categorization of only the most frequent triplets, and consequently coverage. Even with a relatively small dataset such as ours, the limited span of the codebook means that, in any single transcript, only a very small portion of the triplets will be found, therefore our scoring of AI vs. human agency is very crude and unable to capture the full span of relevant propositions. Furthermore, the decision to consider `I :: use :: AI` and similar triplets as instances of AI agency is an enforcement of a binary view on the co-creative process, ignoring the nuances which may be present in the context.

The methods based on distributional semantics are more robust in that they expand our search space to a larger set of potentially distinguishing phrases (all triplets containing any of the top 200 verbs), and they tend to reveal the expected trend on the large scale (see Figures 3 and 6). However, we assume that the high proportion of triplets labeled as 'HUMAN' can be attributed to the well-known limitations of measuring semantic relatedness through cosine similarity (see also Jawahar, Sagot, and Seddah 2019): contextual embeddings conflate surface, syntactic and semantic information into a single representation across different layers. Thus, triplets like "I – use – pencil" and "I – use – AI" could be considered similar purely due to surface similarity.

Another important point which needs to be made in relation to all methods and results presented: the human-AI collaborative cases presented in the interviews are varied, complex and unique; and the words and sentences people use to describe their work scenarios, creative processes and attitudes are equally diverse, nuanced and highly individual. While we may successfully identify some patterns in these

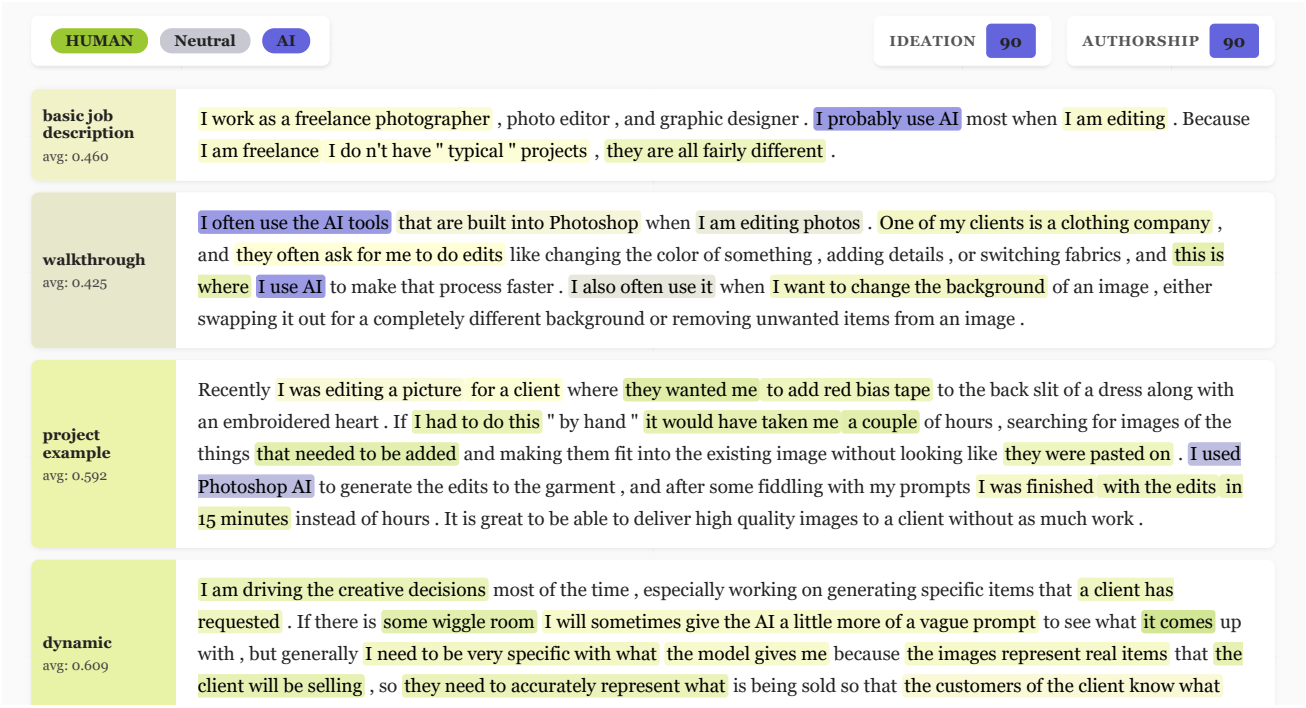


Figure 8: Excerpt from interview transcript. Human scores on the top right. All matching triplets are highlighted by semantic proximity to the concept vectors in terms of the normalized difference score. Green indicates closer proximity to the `HUMAN` vector, and blue closer proximity to the `AI` vector. The example shows the large discrepancy between the human scores for AI share in Ideation and Authorship and the interviewee's description of the collaboration in the fourth section (`dynamic`).

accounts, we cannot claim to be modeling the full complexity of the human-AI creative expression.

Still, at least for the use-cases which involve large language models, the idea that humans overrely on the LLM's overconfident responses while maintaining the illusion of their own control has been addressed - and proven - beyond the scope of computational creativity (Rathi, Jurafsky, and Zhou 2025), and the so-called *fictional human oversight* has become a known issue in debates on AI safety (Kasneci and Kasneci 2026). We argue that in the era of LLMs and AI-powered generative platforms the traditional models of human-computer co-creativity need to be revised to accommodate - and monitor - the challenges arising from the "unbearable lightness of AI-ing", and that raising the awareness about these challenges amongs creative professionals could help relieve the pressure that many of them feel in the race between humans and AI.

Conclusions

We present an experiment to identify potential egocentric bias in human-AI collaborative scenarios, i.e., the tendency to claim full control over the creative process, and to overestimate one's own contributions over those of the AI. We approach the problem with a set of linguistic methods, and perform human assessment for comparison.

The first method is traditional corpus-based analysis supported by manual annotation of a sample of prominent triplets, which might indicate human vs. AI agency. While the resulting AI-agency scores successfully show the cumulative shifting perspectives of the interviewees through the interview sections, it cannot be considered accurate enough due to low coverage. The set of methods based on distributional semantics seem more promising, although results also indicate that cosine similarity may blur true semantic proximity with surface-level phenomena. Finally, human assessments provide an insight into how other creative professionals perceive AI's contribution to the ideation stage and the authorship of the final product.

Our results tentatively support the initial hypothesis, and at the same time reveal the complexity of the human-AI collaboration dynamics. While the methods presented could certainly be improved in the future, we believe that the main contribution of this paper lies in the topic itself, and the associated ethical, legal, psychological and philosophical dimensions of authorship attribution in the era of GenAI. A discussion of these dimensions lies beyond the scope of our work (or our expertise), but has been and will remain a recurring issue in the creative community. We finally emphasize that GenAI was not used to produce any part of this paper, nor was it used in the analysis of the data or interpretation of the findings. It was however used in the research of related work, and as a coding aid. Are we still driving?

Author Contributions

Author 1 was responsible for planning the study, designing the human judgements experiment and synthesizing results, Author 2 was responsible for dataset preparation and the experimental part using clustering and semantic distance, Author 3 was in charge of the corpus-based analysis and Author 4 for the recruitment of creative professionals and visualizations of results. All authors contributed to the writing of this manuscript.

Acknowledgments

This work was supported by the projects "Large Language Models for Digital Humanities" (Grant GC-0002), "Slovene Language - Basic, Contrastive, and Applied Studies" (P6-0215), the "Jožef Stefan" Infrastructure Programme (Grant I0-0005) and the Slovenian Artificial Intelligence Factory project (SLAIF).

The authors sincerely thank the creative professionals who participated as annotators in this study.

References

[Abbott and Rothman 2022] Abbott, R., and Rothman, E. 2022. Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence. doi:10.2139/ssrn.4185327.

[Boden 1992] Boden, M. 1992. *The Creative Mind*. London: Abacus.

[Boo, Kim, and Suh 2025] Boo, C.; Kim, Y.; and Suh, A. 2025. A Collaborative Creative Process in the Age of AI: A Comparative Analysis of Machine and Human Creativity. In *Proceedings of the 58th Hawaii International Conference on System Sciences*. doi:10.24251/HICSS.2025.025.

[Bown et al. 2020] Bown, O.; Grace, K.; Bray, L.; and Ventura, D. 2020. A Speculative Exploration of the Role of Dialogue in Human-Computer Co-creation. In *Proceedings of the International Conference on Computational Creativity 2020*.

[Breidenthal et al. 2020] Breidenthal, A. P.; Liu, D.; Bai, Y.; and Mao, Y. 2020. The dark side of creativity: Coworker envy and ostracism as a response to employee creativity. *Organizational Behavior and Human Decision Processes* 161:242–254. doi:https://doi.org/10.1016/j.obhdp.2020.08.001.

[Condorelli and Berti 2025] Condorelli, F., and Berti, F. 2025. Creativity and awareness in co-creation of art using artificial intelligence-based systems in heritage education. *Heritage* 8(5). doi:10.3390/heritage8050157.

[Dunning 1993] Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1):61–74.

[Ginsburg and Budiardjo 2018] Ginsburg, J. C., and Budiardjo, L. A. 2018. Authors and machines. *Berkeley Technology Law Journal* 34:343. `https://api.semanticscholar.org/CorpusID:69352843`.

[Grootendorst 2022] Grootendorst, M. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[Gur and Sackeim 1979] Gur, R. C., and Sackeim, H. A. 1979. Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology* 37(2):147–169. doi:10.1037/0022-3514.37.2.147.

[Handa et al. 2025] Handa, K.; Stern, M.; Huang, S.; Hong, J.; Durmus, E.; McCain, M.; Yun, G.; Alt, A.; Millar, T.; Tamkin, A.; Leibrock, J.; Ritchie, S.; and Ganguli, D. 2025. Introducing Anthropic Interviewer: What 1,250 professionals told us about working with AI. `https://anthropic.com/research/anthropic-interviewer.https://anthropic.com/research/anthropic-interviewer`.

[Hardie 2014] Hardie, A. 2014. Log ratio–an informal introduction. *ESRC Centre for Corpus Approaches to Social Science (CASS)* 1–2.

[Jawahar, Sagot, and Seddah 2019] Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What does BERT learn about the structure of language? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1356.

[Kantosalo and Toivonen 2016] Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the seventh international conference on computational creativity*, 77–84.

[Karimi et al. 2018] Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating Creativity in Computational Co-Creative Systems. doi:10.48550/arXiv.1807.09886.

[Kasneci and Kasneci 2026] Kasneci, G., and Kasneci, E. 2026. The safety failures we are not instrumenting: A perspective on hidden safety-critical challenges in modern AI systems. *AI and Ethics* 6(3):295. doi:10.1007/s43681-026-01132-0.

[Kilgarriff and Tugwell 2001] Kilgarriff, A., and Tugwell, D. 2001. Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, 32–38.

[Kilgarriff 2001] Kilgarriff, A. 2001. Comparing corpora. *International journal of corpus linguistics* 6(1):97–133.

[Liu and Huang 2025] Liu, W., and Huang, W. 2025. Authorship in Human-AI collaborative creation: A creative control theory perspective. *Computer Law & Security Review* 57:106139. doi:10.1016/j.clsr.2025.106139.

[Mele 2001a] Mele, A. R. 2001a. *Autonomous Agents: From Self-control to Autonomy*. Oxford University Press.

[Mele 2001b] Mele, A. R. 2001b. *Self-Deception Unmasked*. Princeton University Press.

[Nikrang and Kiesenhofer 2025] Nikrang, A., and Kiesenhofer, S. 2025. Human–AI co-creation in contemporary composition: Interaction and artistic strategies with Ricercar. In *Proceedings of the Conference on Animation and Interactive Art*, Expanded '25, 65–73. New

York, NY, USA: Association for Computing Machinery. doi:10.1145/3749893.3749961.

[Pruschak and Hopp 2022] Pruschak, G., and Hopp, C. 2022. And the credit goes to ... - ghost and honorary authorship among social scientists. *PLOS ONE* 17(5):1–22. doi:10.1371/journal.pone.0267312.

[Rathi, Jurafsky, and Zhou 2025] Rathi, N.; Jurafsky, D.; and Zhou, K. 2025. Humans overrely on overconfident language models, across languages. In *Second Conference on Language Modeling*. https://openreview.net/forum?id=QsQatTzATT.

[Ross and Sicoly 1982] Ross, M., and Sicoly, F. 1982. *Egocentric biases in availability and attribution*. Cambridge University Press. 179–189.

[Rossiter 1993] Rossiter, M. W. 1993. The Matthew Matilda effect in science. *Social Studies of Science* 23(2):325–341. http://www.jstor.org/stable/285482.

What Makes Literary Quotes Stand Out? A Multidimensional Analysis of Literary Creativity

Berke Yazan, Almila Akdağ, Davide Dell'Anna
Department of Information and Computing Sciences
Utrecht University
{b.yazan, a.a.akdag, d.dellanna}@uu.nl

Abstract

Current approaches in computational linguistics often rely on statistical deviation as the primary proxy for measuring creativity. This fails to capture the full extent of the semantic, linguistic, and affective dimensions that distinguish creative from ordinary language. To explore this multidimensionality, we analyse literary quotes from Goodreads, curated by readers for their perceived value. We construct ALOFT, a dataset that pairs these quotes with six comparative corpora: length-matched random snippets from the same source texts, gathered through a novel human-in-the-loop workflow; popular quotes, public-domain quotes with surrounding context, and a non-literary baseline. Our analysis measures features such as complexity, lexical novelty, semantic movement, and affective tone across these corpora using explainable methods to reveal feature interactions. We demonstrate that widely shared quotes are distinguishable by a combination of accessible vocabulary, unexpected semantic arrangements, and emotional charge. Furthermore, we show that traditional information-theoretic metrics struggle with short text lengths and that sentence-level contextual embeddings significantly outperform static word vectors at identifying the signature of highly valued language. This work offers a new, explainable perspective on perceived value and linguistic creativity, along with a new dataset and a data-gathering workflow for future research. [1]

Introduction

There are nearly infinite letter combinations possible in human language, but only a vanishingly small fraction holds meaning or value. Moreover, among all grammatically correct and semantically coherent sentences, only a rare few achieve the exceptional value found in richer expressions, such as creative, artistic, and figurative artifacts. Understanding what exactly distinguishes such expressions from everyday language is currently an ongoing challenge.

Developing computational tools that can recognise and preserve creative uses of language would let us study them systematically. This can help enrich the creative capabilities of both humans and machines while also counteracting the growing risks of algorithmic monoculture caused by the statistically collectivist outputs of Large Language Models (LLMs) (Priyanshu and Vijay 2024; Wu, Black, and Chandrasekaran 2024).

Influentially, Boden (2004) argues that the creativity of an artifact arises from a combination of novelty and value (Boden 2004). Traditionally, Natural Language Processing (NLP) approaches have used novelty as the primary proxy for measuring creativity, considering metrics such as surprisal (Bunescu and Uduehi 2022; He, Peng, and Liang 2019) or semantic distance (Beaty et al. 2022; Orwig et al. 2021; Dumas, Organisciak, and Doherty 2021). Similarly, automatic creativity scoring systems have historically adopted this novelty-centric approach (Beaty and Johnson 2021; Organisciak et al. 2023).

Measuring novelty alone provides an incomplete picture of Boden's definition. A variety of studies have therefore attempted to operationalise the broader value aspect of creativity through a combination of measurable dimensions, such as linguistic memorability (Danescu-Niculescu-Mizil et al. 2012), structural and syntactic composition (Ashok, Feng, and Choi 2013; van Cranenburgh and Bod 2017), affective resonance (Jacobs 2018; Hipson and Mohammad 2020), and stylistic uniqueness (Potthast et al. 2018). More recently, LLMs have demonstrated the ability to evaluate creativity in written domains (Ismayilzada, Stevenson, and van der Plas 2024) at levels comparable to those of human judges in certain contexts. This suggests that such models have learned to approximate how humans communicate and interpret artistic or creative language.

Despite these advancements, we still lack transparent frameworks for understanding how or why certain linguistic expressions are considered creative, and existing creativity evaluation models still struggle to capture the nuanced and diverse nature of artistic expressions (Ismayilzada et al. 2024). Such an asymmetry between current generative capabilities, creativity metrics, and understanding of linguistic creativity becomes apparent when studying richer expressions, such as figurative text (Tsvetkov, Mukomel, and Gershman 2014; Mao, Lin, and Guerin 2019) and literary artifacts, especially when subtle deviations and cultural context are involved (Ismayilzada et al. 2024; Kuznetsova, Chen, and Choi 2013).

Another shortcoming of computational literary research is

[1] Generative AI tools (Grammarly, Sonnet 4.5, Gemini 3.0) were used for grammar checking, sentence refinement, LaTeX/code assistance, and adapting python code snippets. All research design, theoretical contributions, and scientific interpretations are solely the authors' own.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

the reliance on public-domain resources such as the works of Shakespeare or the Gutenberg Corpus (Gerlach and Font-Clos 2018), partly due to the manual difficulty of extracting contemporary text under fair academic use. This inevitably leads to an imbalanced representation of human literature and introduces temporal and cultural bias. These create a need to provide the community with easier approaches to collect a more diverse dataset that includes contemporary artifacts.

In this paper, we address these shortcomings by first compiling a new dataset called ALOFT (A Lot Of Four Thousands)[2], which enables controlled comparisons across literary and non-literary corpora. ALOFT is created using a novel human-in-the-loop data-extraction pipeline described in this paper, sourced by both public-domain and contemporary text. Second, we introduce a multidimensional analysis of features that distinguish creative from ordinary language. Leveraging ALOFT, we compute a variety of creativity metrics, including language accessibility/complexity, information-theoretic measures, semantic embeddings and trajectories, and affective tone. Finally, we analyse how these dimensions interact using explainable methods, aiming to reveal the subtle mechanisms that contribute to linguistic creativity.

Our results demonstrate that linguistic creativity relies on an interplay of semantic trajectories, accessible vocabulary, and affective tone, moving beyond mere statistical novelty. Specifically, our analysis reveals that widely shared literary quotes are distinguished by the presence of familiar words in unexpected semantic arrangements that carry emotional weight. By evaluating contextual models against static word vectors, we show that sentence-level contextual embeddings are essential for capturing the subtle signature of creative language. Ultimately, by providing a transparent, multidimensional perspective on linguistic creativity, along with a novel data-gathering workflow and a novel dataset, we establish a robust foundation for future research in both computational literary studies and the development of more deeply informed generative models.

Related Work

Defining and Measuring Literary Creativity

Quantifying literary creativity has long been a focal point at the intersection of computational creativity and digital humanities. Van Cranenburgh and Bod (2017) developed a data-oriented model to predict literary ratings from textual features, using a corpus of Dutch novels evaluated by a large survey (van Cranenburgh and Bod 2017). They demonstrated that sentence length, vocabulary richness, and syntactic diversity are significantly correlated with perceived literariness, with an R^2 of 0.76. Extending this analysis, Van Cranenburgh et al. (2019) integrated machine learning alongside human judges to further explore the perceived literariness of contemporary Dutch novels (van Cranenburgh, van Dalen-Oskam, and van Zundert 2019). They found that literariness cannot be reduced exclusively to semantic deviations, but, promisingly, can still be identified without human oversight or sociological context.

In parallel to structural analysis, the Neurocognitive Poetics Model (Jacobs, 2015) links aesthetic impact to cognitive and affective processes and finds that textual features can systematically influence readers' emotional engagement and aesthetic appreciation (Jacobs 2015). Subsequently, Jacobs (2018) introduced the Gutenberg English Poetry Corpus (GEPC), a manually curated collection of over three million lines of poetry (Jacobs 2018). Applying methods such as latent semantic analysis, topic modelling, and sentiment estimation, they found that poems with higher surprisal, lexical diversity, and sonority are associated with greater aesthetic or emotional resonance.

From Static Features to Semantic Trajectories

Extending the scope from poetry to 2,722 mixed domain literary texts, Jacobs and Kinder (2021) treat literariness, creativity, and beauty as distinct, quantifiable axes and demonstrate that semantic complexity measures and affective tone can effectively capture them (Jacobs and Kinder 2021). They operationalise literariness through Stepwise Distance (SWD) (van Cranenburgh, van Dalen-Oskam, and van Zundert 2019), which measures the semantic divergence between adjacent text segments. Using this, they found that plays rank as the most literary category, followed by poems and novels. To quantify creativity, they utilised Forward Flow (FF) (Gray et al. 2019), calculating the average semantic distance between a word and all preceding words in a sequence. They identified poems and plays as the most creative categories in the corpus, whereas novels ranked lowest.

Quotability and Popularity

Danescu-Niculescu-Mizil et al. (2012) analysed thousands of movie quotes from IMDb (Internet Movie Database) (Internet Movie Database (IMDb)), comparing quoted and non-quoted snippets from the same movies (Danescu-Niculescu-Mizil et al. 2012). Their analysis revealed that movie quotes tend to adhere to conventional syntactic structures but employ less common word choices. Similarly, Tekir et al. (2023) formalised the literary quote detection task using a benchmark dataset (T50) derived from Goodreads and Project Gutenberg (Tekir et al. 2023). Their approach shows that quotes exhibit higher lexical distinctiveness than their surrounding context. From the popularity and likability perspective, using Goodreads ratings as the ground truth, Maharjan et al. (2017) proposed a multi-task neural architecture to predict book likability from their textual content alone (Maharjan et al. 2017). Their findings underscore the value of sentiment, dialogue structure, and narration style.

Non-Interpretability of Human and LLM Evaluation

Evaluating creative and artistic language remains a significant challenge, even for humans. Interpretations often

[2]Pipeline, analysis code, and dataset (excluding the Google Books-derived columns, which can be regenerated via the pipeline): `https://github.com/BerkeYazan/aloft`.

Column Name (Abbreviation)	Description
Goodreads Sample Quote (GSQ)	Random samples of user-curated quotes
Goodreads Popular Quote (GPQ)	Most highly-liked quotes from Goodreads corpus
Google Books Length-Matched (GBLM)	Same-length passages from same books of Goodreads samples
T50 Quote (T50Q)	User-curated quotes from the T50 study
T50 Quote-Free Context Length Matched (T50LM)	T50 Quote-Free Context Matched to original T50 quote length
Non-Literary Baseline (NLB)	Length-matched informational text passages (Wikipedia + Brown)
Google Books Page Text	Complete OCR-extracted page content
T50 Full Context	Complete 21-sentence context from Project Gutenberg
T50 Quote-Free Context	Context with quote sentences removed

Table 1: Columns of the final ALOFT dataset and their abbreviations. Bold entries are used in analysis. All columns contain $N = 4{,}569$ entries.

vary (Iser 1978; Barthes 1967; Sontag 1966), and explanations may involve post hoc rationalisations rather than objective criteria (Summers 2017). Recent studies show that human evaluation of creativity can be replicated by modern Large Language Models (LLMs) (Ismayilzada, Stevenson, and van der Plas 2024; Porter and Machery 2024; Marco, Rello, and Gonzalo 2024). However, because these models largely mirror the non-interpretable nature of human ratings, they inherit the same transparency problem, where neither human nor LLM judgements expose which textual features drove the rating, so the evaluation cannot be decomposed into its contributing factors (Ismayilzada et al. 2024).

The Dynamic Between Accuracy and Explainability

Recent work suggests that models can predict what is considered creative at levels approaching human judgement without providing a framework for why those linguistic arrangements hold value (Ismayilzada, Stevenson, and van der Plas 2024). For instance, automatic creativity scoring systems have started to transition from simple novelty-centric metrics to LLM-based models fine-tuned on human feedback (Beaty and Johnson 2021; Organisciak et al. 2023). While this significantly improved predictive performance, it has simultaneously limited our understanding of the evaluations. This dynamic between accuracy and interpretability highlights the need for multidimensional, explainable approaches that can both accurately evaluate and decompose the underlying features and mechanisms of creative language.

Dataset Creation

In this section, we introduce the ALOFT dataset and illustrate the human-in-the-loop data-extraction pipeline we implemented to create it. Table 1 outlines the various data sources integrated into ALOFT. We aim to construct a dataset in which quoted literary text, non-quoted literary text, and non-literary text can be represented in a diverse and comparable manner.

Data Sources

Goodreads Sample Quotes (GSQ) and Popular Quotes (GPQ). Our primary signal for extraordinary language comes from two Goodreads quote datasets on Kaggle (faellielupe 2020; Verma 2021), including nearly half a million user-submitted quotes combined. To ensure active community validation, we discard quotes with no likes, leaving 225,262 quotes. To serve as a separate popularity-filtered column, we extract a subset of the most-liked quotes.

Google Books Length-Matched Context (GBLM). Google Books page previews serve as one of our sources for the non-quoted literary contexts from the same works as our GSQ sample. This serves as a control for authorial style and a source for non-quoted literary language. This data is gathered using the custom extraction pipeline detailed in the Data Extension section.

T50 Quotes (T50Q). Tekir et al. (Tekir et al. 2023) provide a dataset with nearly five-thousand pairs from public-domain literary quotes and their immediate surroundings. These are sourced from Project Gutenberg (dominantly published before 1928), and are shown to be lexically distinct from their surrounding text (ibid). Enriching their data with our modern Goodreads corpus (mean publication year 1990) enables a diachronic analysis of literary preferences.

Non-Literary Baseline (NLB). To establish a non-literary contrast to our literary texts, we sample a control baseline using a diverse set of non-fiction texts from the Brown (Francis and Kučera 1982) (40%) and Wikipedia (Wikimedia Foundation 2023) (60%) corpora.

Data Extension

Goodreads to Google Books Pipeline. To create the GBLM partition, we introduce a terms of service (ToS) compliant pipeline that uses manual browser actions and human-in-the-loop cleaning, as illustrated in Figure 1. We first enrich our metadata for future research by using the Wikimedia API to add the publication year and the original language. We then draw a sample using occurrence and like-weighted stratified sampling to ensure that the subset reflects the platform's actual distribution of reader endorsements, which serves as our proxy for literary value. To extract sampled book pages from the same source works, we manually capture and review 12,000 randomly selected candidate preview pages for over 50 hours, yielding 5,333 valid full-page screenshots. We then process these validated screenshots using Google Cloud Vision OCR. OCR imperfections were detected algo-

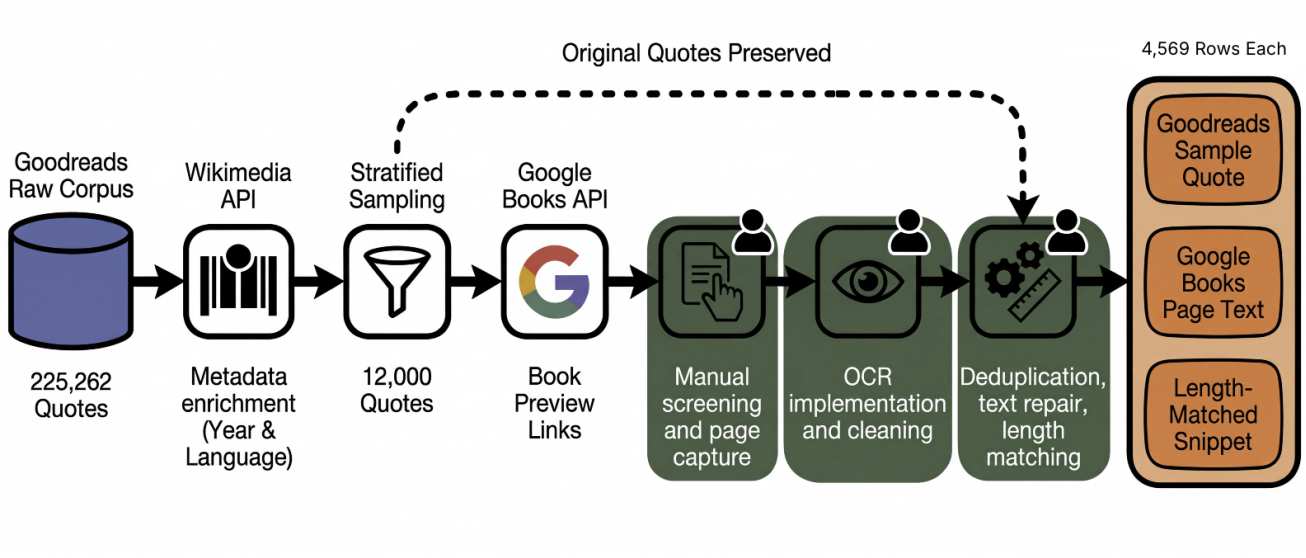


Figure 1: Goodreads to Google Books pipeline combining hybrid extraction with human-in-the-loop validation.

rithmically and corrected manually via a custom interface.

Data Preparation

Length Matching and Finalising the ALOFT Dataset. Because computational metrics scale with text size, we apply a length-matching protocol (within a 70% to 130% token-count tolerance) tailored to each corpus. This involves sampling passages from Google Books pages, cropping the surrounding context from the T50 dataset, and extracting appropriately sized text from our non-literary sources. To prevent data leakage, we apply a fuzzy matching algorithm to identify and remove quotes that also appeared in the sampled page text. We employ an additional algorithmic-detection/manual-correction approach for each data source. We create separate custom interfaces to highlight and correct spelling and punctuation errors that were already present in the data or produced during previous steps. For potential duplicates, we compute the semantic distance between data points using SBERT and manually remove overlapping records, yielding a dataset of 4,569 parallel entries. Designing specialised user interfaces, we enable human-in-the-loop involvement at each step, ensuring human verification of algorithmic processing.

Methodology

To systematically identify the features that distinguish creative language from ordinary language, we propose a framework. This section outlines our approach in three distinct parts. First, we introduce the computational metrics used to characterise the multidimensional features of creativity. Second, we define the comparative axes across which these texts are evaluated. Finally, we detail the statistical and predictive methods used to compute the relationships and significance of the extracted features.

Multidimensional Analysis

Our framework is designed to capture multiple dimensions of short creative text. We operationalise our analysis by mapping established computational methods across five complementary dimensions: (1) accessibility and complexity, (2) affective tone, (3) novelty, (4) semantic representations, and (5) semantic movement. To quantify these dimensions, we consider the following metrics:

Accessibility and Complexity We measure structural readability using the Flesch Reading Ease (Flesch 1948) and the Coleman-Liau Index (Coleman and Liau 1975). These established psycholinguistic metrics allow us to test whether creativity requires syntactic complexity or if it relies on more accessible phrasing.

Affective Tone We compute Sentiment Polarity, as aesthetic impact is closely tied to emotional resonance (Jacobs and Kinder 2021). We utilise a RoBERTa-based sentiment classifier fine-tuned on social media text (Liu et al. 2019) to extract granular positive and negative probability scores. This metric allows us to assess whether extraordinary language relies on a concentrated emotional charge compared to baseline narrative prose.

Novelty We measure Shannon's Entropy (Shannon 1948), Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis 2010), GPT-2 Surprisal (Radford et al. 2019), and Pointwise Mutual Information (PMI) (Church and Hanks 1990). Measuring departures from expected frequency patterns, often described by Zipf's law (Zipf 1949), these metrics quantify how much a quote deviates from statistically predictable language.

Semantic Representations We aim to test both static, word-level approaches that lack context awareness and sentence-level approaches that represent meaning dynamically and are context-aware. To achieve this comparison, we

utilise word-level GloVe (Pennington, Socher, and Manning 2014), and sentence-level SBERT (Reimers and Gurevych 2019) models to generate spatial semantic representations of our data. Static embeddings (GloVe) are context-free, where each word is assigned a single fixed vector regardless of the surrounding context. In contrast, dynamic embeddings (SBERT) represent meaning at the sentence level, where a word's contribution depends on the surrounding text.

Semantic Movement We calculate Stepwise Distance (SWD) (van Cranenburgh, van Dalen-Oskam, and van Zundert 2019), originally proposed as an index of literariness, to trace the semantic divergence between adjacent sentences. On the word level, we utilise Forward Flow (FF) (Gray et al. 2019), which measures how much a text departs from its own established context by calculating the average semantic distance from all preceding words. These metrics allow us to quantify movement through a semantic space (Szemes and Nagy 2024), providing insights into whether extraordinary language relies on divergent exploration or convergent cohesion.

Experimental Setup and Comparison Axes

To systematically uncover the features of linguistic creativity, we apply our multidimensional metrics across four distinct axes of comparison:

- **Literary vs. Non-Literary:** We contrast our literary quotes (GSQ+GPQ+T50Q) against the non-literary baseline (NLB) to establish the general linguistic and semantic signatures that separate artistic expression from standard informational prose.
- **Within-Book Quoted vs. Non-Quoted Literary Text:** To isolate the specific features that make a sentence quoted, we control for authorial style and genre. We achieve this by comparing quotes (GSQ, GPQ, T50Q) against non-quoted context from the exact same source texts (GBLM, T50LM).
- **High vs. Sampled Popularity:** To understand the drivers of what makes a quote popular, we contrast the most widely shared popular quotes (GPQ) against our standard baseline of reader-curated quotes (GSQ).
- **Contemporary vs. Classic:** Finally, we compare contemporary quotes (GSQ) against classic quotes(T50Q) to observe potential diachronic differences.

Statistical Validation and Explainability

Given the non-normal distributions frequently observed in linguistic data, we use a non-parametric statistical framework to evaluate the significance of feature differences across our distinct text categories. We apply the Mann-Whitney U test (Mann and Whitney 1947) for pairwise comparisons, coupled with Cliff's Delta (Cliff 1993) to measure the effect sizes. Cliff's Delta ranges from -1 to 1: a value of 0 indicates complete overlap between the two groups, while values approaching 1 or -1 indicate that one group almost always scores higher than the other. In the tables, a positive delta indicates that the first (left-hand) group of each comparison scores higher, and a negative delta indicates that the second (right-hand, baseline) group scores higher. We interpret the magnitude using the thresholds of Romano et al. (Romano et al. 2006): $|d| < 0.147$ is negligible, $|d| < 0.33$ is small, $|d| < 0.474$ is medium, and larger values are large. To account for multiple comparisons across our varied metrics, all reported significance levels are adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) procedure (Benjamini and Hochberg 1995).

Complementing this inferential approach, we train extreme gradient-boosted decision trees (XGBoost) (Chen and Guestrin 2016) to identify which specific features best distinguish quotes from their context. We calculate SHAP (SHapley Additive exPlanations) values (Lundberg and Lee 2017) to reveal the directionality and magnitude of individual feature contributions. This method ensures that our findings remain transparent and theoretically grounded.

Results and Discussion

Multidimensional Findings

Evaluating the ALOFT corpus across our five multidimensional metrics and four comparison axes reveals distinct patterns in linguistic creativity. The inferential statistics for these comparisons are detailed in Table 2 and Table 3. Breaking down this analysis across the five core dimensions yields the following insights:

Accessibility and Complexity The non-literary baseline (NLB) shows distinctly higher (harder-to-read) Coleman-Liau Index scores, centred at 11.7, compared to quoted literary text (GSQ, GPQ, T50Q), centred between 5.7 and 6.7, and non-quoted contexts (GBLM, T50LM), centred between 6.7 and 7.7. The same ranking emerges using the Flesch Reading Ease metric. These patterns show that literary language, especially quoted literary text (GSQ, GPQ, T50Q), exhibits a more focused, accessible vocabulary than the non-literary baseline (NLB). Goodreads Popular Quotes (GPQ) are the simplest (≈ 5.7), and the contrast between popular quotes (GPQ) and the non-literary baseline (NLB) produces the largest effect sizes for readability (Flesch Reading Ease $d = 0.74^{***}$, Coleman-Liau $d = -0.81^{***}$), indicating a correlation between popularity and simplicity.

Affective Tone As illustrated in Figure 2, the stacked bars show the proportion of positive, neutral, and negative sentiment across each text type, making the emotional profile of each corpus directly comparable. All quote categories (GSQ, GPQ, T50Q) contain a higher proportion of emotionally valenced content than their corresponding contexts (GBLM, T50LM). While non-literary baseline (NLB) is predominantly neutral, popular quotes (GPQ) exhibit the widest emotional range, with the highest positive sentiment among other quotes (GSQ, T50Q).

Novelty As reported in Table 2 and Table 3, quoted literary text (GSQ, GPQ, T50Q) exhibits lower overall lexical diversity and, notably, lower GPT-2 surprisal scores for contemporary quotes (GSQ, GPQ). This indicates that literary quotes can achieve their impact without relying on rare vocabulary or highly unpredictable sentence structures. Taken

Table 2: Multidimensional Comparison: Literary vs. Non-Literary

	GSQ+GPQ+T50Q vs. LB+NLB	GPQ vs. NLB	GSQ vs. NLB	T50Q vs. NLB
Accessibility, Novelty, and Affective Features (Cliff's Delta)				
Shannon Entropy	-0.08***	-0.23***	0.00	-0.03**
Flesch Reading Ease	0.15***	0.74***	0.51***	0.46***
Coleman–Liau Index	-0.25***	-0.81***	-0.62***	-0.69***
Lexical Diversity	-0.20***	-0.30***	-0.13***	-0.16***
GPT-2 Surprisal	-0.19***	-0.34***	-0.22***	0.03**
Sentiment Polarity	-0.09***	-0.10***	-0.23***	-0.14***
Semantic Representations: SBERT and (GloVe)				
Classifier AUC	0.898 (0.884)	0.995 (0.992)	0.980 (0.973)	0.993 (0.991)
Semantic Outlier (d)	-0.384 (-0.134)	-0.904 (-0.428)	-0.711 (-0.317)	-0.851 (-0.382)
Centroid Distance	0.1414 (0.0067)	0.6128 (0.0437)	0.4537 (0.0289)	0.4831 (0.0302)
Silhouette Score	0.024 (0.048)	0.073 (0.191)	0.051 (0.142)	0.081 (0.153)
Semantic Movement (Cliff's Delta)				
Forward Flow (FF)	-0.456***	-0.824***	-0.730***	-0.775***
Stepwise Dist. (SWD)	-0.164***	0.079***	0.108***	0.006

Note: Significance from Mann–Whitney U test (Benjamini–Hochberg corrected): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Semantic Outlier (d) is Cliff's Delta from nearest-neighbour distance comparison. Centroid Distance is the cosine distance between class mean vectors. Silhouette Score is a cluster-separation measure. LB denotes the Literary Baseline, the combination of the non-quoted literary contexts GBLM and T50LM (see Table 1).

Table 3: Multidimensional Comparison: Quoted vs. Non-Quoted, High vs. Sampled Popularity, and Contemporary vs. Classic

	Within-Book		Popular	Contemporary
	GSQ vs. GBLM	T50Q vs. T50LM	GPQ vs. GSQ	GSQ vs. T50Q
Accessibility, Novelty, and Affective Features (Cliff's Delta)				
Shannon Entropy	-0.03**	0.00	-0.22***	0.03*
Flesch Reading Ease	-0.10***	-0.20***	0.28***	0.08***
Coleman–Liau Index	0.00	0.11***	-0.30***	0.05***
Lexical Diversity	-0.19***	-0.13***	-0.18***	0.03*
GPT-2 Surprisal	-0.12***	-0.11***	-0.15***	-0.26***
Sentiment Polarity	-0.10***	-0.06***	0.09***	-0.06***
Semantic Representations: SBERT and (GloVe)				
Classifier AUC	0.852 (0.831)	0.841 (0.840)	0.761 (0.745)	0.826 (0.845)
Semantic Outlier (d)	-0.292 (-0.052)	-0.208 (-0.048)	-0.213 (-0.034)	-0.031 (-0.040)
Centroid Distance	0.1355 (0.0045)	0.0709 (0.0038)	0.0470 (0.0025)	0.0370 (0.0015)
Silhouette Score	0.021 (0.036)	0.018 (0.029)	0.008 (0.021)	0.012 (0.013)
Semantic Movement (Cliff's Delta)				
Forward Flow (FF)	-0.244***	-0.216***	-0.235***	0.043***
Stepwise Dist. (SWD)	-0.211***	-0.322***	-0.037*	0.115***

Note: Significance from Mann–Whitney U test (Benjamini–Hochberg corrected): * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Semantic Outlier (d) is Cliff's Delta from nearest-neighbour distance comparison. Centroid Distance is the cosine distance between class mean vectors. Silhouette Score is a cluster-separation measure. Delta-PMI was excluded from inferential analysis due to circularity.

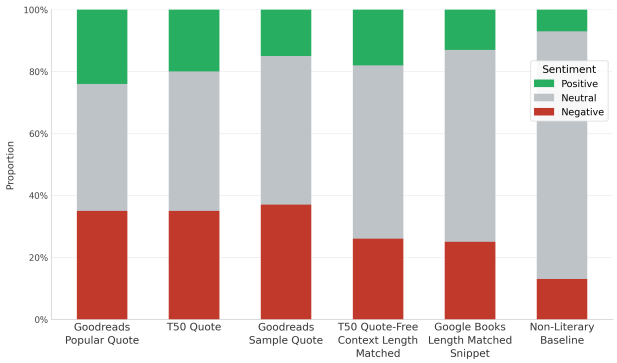


Figure 2: Proportions of sentiment labels across text types. Quotes (GSQ, GPQ, T50Q) show higher proportions of emotionally valenced content (positive and negative) compared to non-quoted contexts (GBLM, T50LM), which contain more neutral content. The non-literary baseline shows the highest proportion of neutral sentiment (≈80%)

together with our findings on accessibility and affective tone, these results demonstrate that literary quotes are not necessarily complex or surprising, but they exhibit accessible delivery with a concentrated emotional charge.

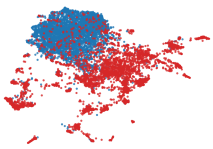
Semantic Representations To examine the semantic geometry of our data, we compare static (GloVe) and dynamic (SBERT) embeddings. Inferred from the classifier AUC results from Table 2, both SBERT and GloVe models perform accurately while separating quoted literary text (GSQ+GPQ+T50Q) from the non-literary baseline (NLB). The accuracy of both models decreases in quoted vs. non-quoted, high vs. sampled popularity, and contemporary vs. classic comparisons, shown in Table 3. In general, we observe that SBERT consistently outperforms GloVe, except when separating popular from sampled Goodreads quotes. SBERT successfully identifies quoted literary text as semantic outliers relative to the non-literary baseline, with a large effect size ($d = -0.904$ for popular quotes), outperforming the static GloVe model ($d = -0.428$).

Furthermore, as visualised in Figure 3, the context-aware SBERT model outperforms the non-contextual GloVe model in separating Goodreads Sample Quotes (GSQ) from Google Books length-matched (GBLM) snippets and non-literary baseline (NLB). These results indicate that the creativity of quoted literary texts can be captured more comprehensively when structural arrangement and context are taken into account.

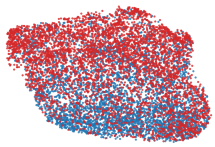
Semantic Movement As detailed in Table 3, both contemporary (GSQ, GPQ) and classic (T50Q) quotes exhibit lower FF ($d = -0.244$ and $d = -0.216$, respectively) and lower SWD ($d = -0.211$ and $d = -0.322$, respectively) than their non-quoted literary contexts (GBLM, T50LM). This observation suggests that quoted passages are characterised by a more constrained and coherent semantic trajectory compared to the more exploratory nature of their non-quoted contexts. The comparison against baseline texts shows a



(a) GloVe: GSQ vs. NLB



(b) SBERT: GSQ vs. NLB



(c) GloVe: GSQ vs. GBLM



(d) SBERT: GSQ vs. GBLM

Figure 3: UMAP projections for Goodreads Sample Quotes (GSQ) compared to different corpora. The top row shows the Non-Literary baseline (NLB), where GloVe (a) exhibits some structure but SBERT (b) achieves a much clearer class separation. The bottom row shows the quotes versus their corresponding context (GBLM), where GloVe (c) fails to separate the classes, whereas SBERT (d) reveals a discernible, albeit overlapping, structure.

similar pattern. Each quote (GSQ, GPQ, T50Q) had substantially lower FF scores than the non-literary baseline (NLB), with effect sizes (d between -0.730 and -0.824) indicating a strong tendency toward conceptual focus in literary language.

The SWD results show a more nuanced pattern. While classic quotes (T50Q) show no apparent difference from the NLB, modern literary quotes (GPQ, GSQ) exhibit a slight but statistically robust increase in SWD (d = 0.079 and d = 0.108, respectively; $p < 0.001$). These reflect the trade-off between exploration and exploitation, which maps onto the distinction between divergent and convergent creativity. Our results suggest that quotes are products of convergent creativity, exploiting a specific theme through a focused semantic path. In contrast, the non-quoted literary contexts (GBLM, T50LM) explore a wider range of topics.

Classifier Results and Feature Importance

Combining the metrics we implemented during this work, we conducted a predictive modelling analysis to quantify the relative importance of the linguistic features.

Our XGBoost models exhibited varied yet informative performance across the five classification tasks, confirming that each comparison group has a distinct and measurable linguistic fingerprint. A summary of the key evaluation metrics for each model, including precision and recall for the positive (quote) class, is presented in Table 4. The results show a clear hierarchy of distinguishability. The model most effectively separated literary quotes from a non-literary baseline (GSQ vs. NLB), achieving an AUC of 0.93 with strong precision (0.86) and recall (0.89). The next most distinguishable comparison involved different types of literary text: Goodreads Sample Quotes versus T50 Quotes (AUC = 0.74), and quotes versus their surrounding narrative context (AUC = 0.70 for both GSQ and T50Q). The most nuanced comparison was between popular and regular sample quotes (GSQ vs. GPQ, AUC = 0.69), which also showed the most balanced trade-off between precision (0.60) and recall (0.59), suggesting that the features driving social resonance are the most subtle.

Beyond raw predictive performance, interpreting the specific features that distinguish literary creativity is essential. Extracting these feature contributions across our four comparative axes yields the following insights:

- **Literary vs. Non-Literary (GSQ vs. NLB):** The model relies heavily on semantic structure and emotion. Low FF and high positive/negative polarity indicate that literary texts are more cohesive and emotionally expressive than everyday prose.
- **Quoted vs. Non-Quoted (GSQ vs. GBLM):** Figure 4 presents a SHAP summary plot for the GSQ versus Context classification task, illustrating both the magnitude and directionality of individual feature contributions to the model's predictions. The most essential features driving the prediction are low sentiment neutrality, low SWD, and low FF. This indicates that, compared to the surrounding narrative, quotes are significantly more emotionally expressive and follow a more dynamic, yet coherent, semantic path.
- **High vs. Sampled Popularity (GSQ vs. GPQ):** Positive sentiment alongside lower Coleman-Liau scores (simpler language) predicts popularity. This suggests that widely shared quotes are more accessible and positive.
- **Contemporary vs. Classic (GSQ vs. T50Q):** The model overwhelmingly prioritises unexpectedness. Low GPT-2 surprisal strongly predicts the classic, curated T50 quotes, indicating less predictable word choices than modern Goodreads quotes. However, this can also be explained by the architecture of the GPT-2 model. Because it was predominantly trained on public-domain data, T50 quotes might have been flagged as less surprising than contemporary works that were not present in the training data.

Table 4: Summary of XGBoost Classifier Performance Across Experimental Tasks.

Experimental Task	AUC	Accuracy	F1-Score	Precision	Recall
GSQ vs. NLB	0.93	86.7%	0.87	0.86	0.89
GSQ vs. T50Q	0.74	67.9%	0.67	0.70	0.75
GSQ vs. Context	0.70	64.4%	0.64	0.63	0.55
T50Q vs. Context	0.70	63.9%	0.63	0.59	0.55
GSQ vs. GPQ	0.69	63.8%	0.63	0.60	0.59

Note: AUC is Area Under the ROC Curve. F1-Score is the macro average. Precision and Recall are for the positive quote class.

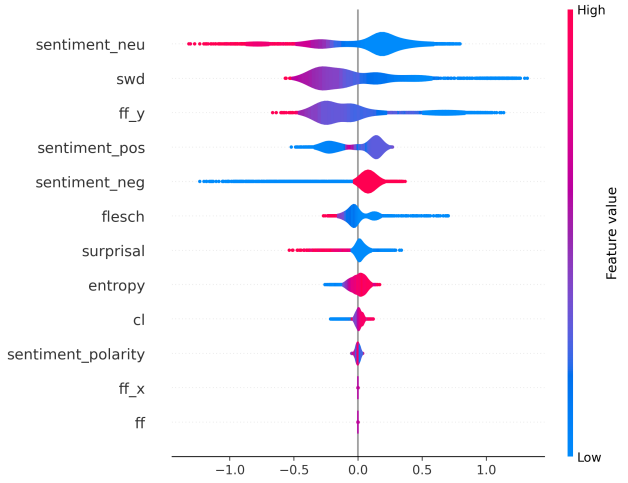


Figure 4: SHAP summary plot for the Goodreads Sample Quotes (GSQ) vs. Google Books Length-Matched (GBLM) Snippet classification task. Each point represents the SHAP value of a single text for a given feature. The colour indicates the feature's value (red is high, blue is low), and its position on the x-axis indicates its impact on the prediction. Key predictors are low sentiment neutrality, low SWD, and low FF.

Conclusions and Future Work

In this paper, we share ALOFT, a curated dataset that aligns literary quotes with length-matched context from the same books, public-domain parallels, and a non-literary baseline, providing the community with an extendable, controlled testbed for creativity research. ALOFT is constructed through a human-in-the-loop workflow in which automation (APIs, page capture, OCR, fuzzy matching) is continuously audited and corrected via manual checkpoints with custom interfaces, ensuring data quality without breaching platform ToS or academic fair use. We used ALOFT to conduct a multidimensional analysis of creativity, employing an integrated framework that combines established metrics to explore the features of creative artifacts and identify what distinguishes them from the ordinary. Coupling XGBoost classifiers with SHAP explanations enabled ranking the salience and directionality of our features, providing interpretable evidence on how they interact to create these creative artifacts.

While our study provides a comprehensive analysis of creative literary quotes, it is not without limitations. First, the analysis is constrained by the nature of the data itself. Despite our length-matching efforts, some quotes are short, which can make length-sensitive metrics such as surprisal more variable. Our dataset reflects the cultural and demographic biases of its source, the predominantly Western, English-speaking Goodreads community. In terms of temporal scope, the data are biased toward English-language prose published after 1800, so these findings may not apply to poetry, drama, or other cultures' literary traditions. Additionally, the popularity of certain authors can be a confounding factor, as it mixes stylistic signals with the author's reputation. Our evaluation also relies on Goodreads endorsements as a proxy for quality, without a ground truth based on expert literary-critical annotation.

Reader endorsement operationalises the value component of creativity rather than creativity as a whole. In our corpus, popularity correlates with accessible vocabulary and positive sentiment, signatures more characteristic of likability than of novelty. This is consistent with theoretical decompositions of creativity into novelty and value (Boden 2004), where value is partly constituted by recognition from an audience rather than by properties of the artifact alone. Goodreads endorsement is therefore a signal for the value dimension, while novelty is captured separately through surprisal, lexical diversity, and semantic movement.

Second, the metrics and models we employed to represent different dimensions of our corpora were not specifically trained on literary language and may not fully capture its nuances.

Our paper triggers numerous future research directions. One possible next step is to validate our results against human perception through controlled participant studies and expert annotation. While costly and time-consuming, human involvement would yield a higher-quality, curated dataset to serve as ground truth. Similarly, validating our results against LLM evaluation would yield novel insights and comparative studies.

Another exciting direction is to model intra-textual quotations, analysing all candidate quotes within a single work to understand how their prominence relates to narrative position and thematic structure. Linking the texts to cultural metadata, such as author demographics or book genre, could also reveal sociocultural and temporal factors that influence memorability.

Following the dynamic semantic distance approaches of FF and SWD, the same "journey" analogy can be extended to include emotional features, as the sentiment contribution of each word or sentence may also yield relevant insights. This approach can also be enriched by part-of-speech analysis to examine how word order and the distribution of grammatical roles influence the structure and flow of information. To capture more nuance, incorporating metrics of figurative language would also be valuable.

Performance-wise, fine-tuning transformer models with relevant data could improve their sensitivity to literary language, and developing length-normalised versions of our trajectory metrics would make them more robust for short texts. A similar path forward involves using our findings as a blueprint to train a generative model, then evaluating its output in human preference experiments to determine whether this "recipe" for quotability is sufficient to create genuinely memorable language.

Advancing our understanding of what distinguishes creative from ordinary language is an open challenge in computational creativity. Our initial analyses, which drew on both well-established metrics and more recent dynamic approaches, yielded promising insights into patterns in creative literary language.

Our work emphasises the importance of explainability for understanding the contributions of features across dimen-

sions to the building blocks of highly-valued literature. In doing so, we also show that information-theoretic and static metrics, more frequently set aside in favour of black-box approaches, still contribute to our understanding of linguistic creativity. We demonstrate that a fuller picture emerges only when these dimensions are combined, ranging from straightforward readability indices to dynamic semantic trajectories to surprisal estimates of contemporary language-models. Together, they reveal that creativity is not the product of any single feature but of an interplay across multiple levels and dimensions of language. This also supports the earlier point we drew from Boden that creativity is more than novelty (Boden 2004), since novelty alone was not enough to distinguish quotes from ordinary text in our results.

As such, this paper contributes to three intersecting fields: the computational assessment of creativity, data-driven analysis of literary artifacts, and the generation of creative and memorable language. To build on our insights, further research should focus on diverse explainable features that contribute to both novelty and value, incorporating richer dimensions such as figurative language, narrative structure, and affective trajectories, validated against human perception and expert annotation across a broader range of literary traditions and cultural contexts.

Acknowledgements

This research was funded by the project "Hybrid Intelligence: Augmenting Human Intellect", a 10-year Gravitation programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (grant number 024.004.022). We thank the anonymous reviewers of ICCC 2026 for their careful reading and constructive feedback, which improved the clarity and framing of this paper.

References

Ashok, V. G.; Feng, S.; and Choi, Y. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1753–1764.

Barthes, R. 1967. The death of the author. *Aspen* (5-6). Reprinted in Image-Music-Text, 1977, Fontana Press.

Beaty, R. E., and Johnson, D. R. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods* 53(2):757–780.

Beaty, R. E.; Johnson, D. R.; Zeitlen, D. C.; and Forthmann, B. 2022. Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal* 34(3):245–260.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

Bunescu, R. C., and Uduehi, O. O. 2022. Distribution-based measures of surprise for creative language: Experiments with humor and metaphor. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, 68–78. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Cliff, N. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin* 114(3):494.

Coleman, M., and Liau, T. L. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60(2):283–284.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 892–901.

Dumas, D.; Organisciak, P.; and Doherty, M. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts* 15(4):645–663.

faellielupe. 2020. Goodreads quotes dataset. Kaggle dataset. Accessed on 2025-05-05.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32(3):221–233.

Francis, W. N., and Kučera, H. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin Company.

Gerlach, M., and Font-Clos, F. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics.

Gray, K.; Anderson, S.; De-Arteaga, M.; Holtzman, N. M.; and Pennebaker, J. W. 2019. Exploring the 'extraordinary': A computational analysis of creative language. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2486–2495.

He, H.; Peng, N.; and Liang, P. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1734–1744. Minneapolis, Minnesota: Association for Computational Linguistics.

Hipson, W., and Mohammad, S. M. 2020. PoKi: A large dataset of poems by children. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*,

1578–1589. Marseille, France: European Language Resources Association.

Internet Movie Database (IMDb). Internet movie database (imdb).

Iser, W. 1978. *The Act of Reading: A Theory of Aesthetic Response*. Baltimore: Johns Hopkins University Press.

Ismayilzada, M.; Paul, D.; Bosselut, A.; and van der Plas, L. 2024. Creativity in ai: Progresses and challenges. Preprint.

Ismayilzada, M.; Stevenson, C.; and van der Plas, L. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Jacobs, A. M., and Kinder, A. 2021. Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large corpus of english literature. *Language and Literature*. Preprint or in press version; Final publication venue may vary.

Jacobs, A. M. 2015. Neurocognitive poetics: Methods and models for investigating the neuronal and cognitive–affective bases of literature reception. *Frontiers in Human Neuroscience* 9:186.

Jacobs, A. M. 2018. The gutenberg english poetry corpus: Exemplary quantitative narrative analyses. *Frontiers in Digital Humanities* 5:5.

Kuznetsova, P.; Chen, J.; and Choi, Y. 2013. Understanding and quantifying creativity in lexical composition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1246–1258. Seattle, Washington, USA: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 4765–4774.

Maharjan, S.; Arevalo, J.; González, F. A.; Montes-y Gómez, M.; and Solorio, T. 2017. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1217–1227. Valencia, Spain: Association for Computational Linguistics.

Mann, H. B., and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* 18(1):50–60.

Mao, R.; Lin, C.; and Guerin, F. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3888–3898. Florence, Italy: Association for Computational Linguistics.

Marco, G.; Rello, L.; and Gonzalo, J. 2024. Small language models can outperform humans in short creative writing: A study comparing slms with humans and llms. *arXiv preprint arXiv:2409.11547*.

McCarthy, P. M., and Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. In *Behavior Research Methods*, volume 42, 381–392. Springer.

Organisciak, P.; Acar, S.; Dumas, D.; and Berthiaume, K. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* 49:101356.

Orwig, W. D.; Diez, I.; Faskowitz, J.; Medaglia, J. D.; Bassett, D. S.; and Beaty, R. E. 2021. Creative connections: Computational semantic distance captures individual creativity and resting-state functional connectivity. *NeuroImage* 227:117632.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Porter, B., and Machery, E. 2024. Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports* 14(1):12345–12356.

Potthast, M.; Rangel, F.; Tschuggnall, M.; Stamatatos, E.; Rosso, P.; and Stein, B. 2018. Overview of the author identification task at pan 2018: Cross-domain authorship attribution. In *Working Notes of CLEF*.

Priyanshu, A., and Vijay, S. 2024. The silent curriculum: How does llm monoculture shape educational content and its accessibility?

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Romano, J.; Kromrey, J. D.; Coraggio, J.; and Skowronek, J. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? In *Annual Meeting of the Florida Association of Institutional Research*.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3):379–423.

Sontag, S. 1966. *Against Interpretation and Other Essays*. New York: Farrar, Straus and Giroux.

Summers, J. S. 2017. Post hoc ergo propter hoc: Some benefits of rationalization. *Philosophical Explorations* 20(sup1):66–79.

Szemes, B., and Nagy, M. 2024. Repetition and innovation in dramatic texts: An attempt to measure the degree of novelty in character's speech. *Journal of Computational Literary Studies* 3(1).

Tekir, S.; Güzel, A.; Tenekeci, S.; and Haman, B. U. 2023. Quote detection: A new task and dataset for nlp. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2023)*, 21–27. Dubrovnik, Croatia: Association for Computational Linguistics.

Tsvetkov, Y.; Mukomel, R.; and Gershman, A. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 248–258.

van Cranenburgh, A., and Bod, R. 2017. A data-oriented model of literary language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1228–1238. Valencia, Spain: Association for Computational Linguistics.

van Cranenburgh, A.; van Dalen-Oskam, K.; and van Zundert, J. 2019. Vector space explorations of literary language. *Language Resources and Evaluation* 53(4):625–650.

Verma, A. 2021. Kaggle goodreads quotes dataset. Accessed on 2025-05-05.

Wikimedia Foundation. 2023. Wikipedia database dump. Available at `https://dumps.wikimedia.org/enwiki/20231101/`.

Wu, F.; Black, E.; and Chandrasekaran, V. 2024. Generative monoculture in large language models.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.

Short Papers

Coherence-Oriented Dream Scene Visualisation

Azra Açıl and Simon Colton

School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

Abstract

Dreams can be emotionally intense but difficult to communicate. We describe the Dream Scene Visualiser (DSV) system which turns written dream descriptions into a temporal sequence of four panel images visualising the dream. This starts with a large language model prompted to split a dream description into four chronological parts. Then a text-to-image model produces images for each part with visual coherence maintained across the sequence, and DSV regenerates any image not suitably matching the text. We evaluate DSV over 50 visualisations from dream descriptions in DreamBank, and report quality, fidelity and coherence results via objective measures employing the CLIP, DINOv2 and Qwen2-VL vision-language models.

Introduction

Dreams are one of the most universal forms of human experience, yet communicating them to other people can be surprisingly hard. They are often deeply personal, emotionally vivid, and sometimes lack narrative logic, making them difficult to convey through words alone. Visualising dreams can have significant value. Dream journaling has been shown to support self-awareness and memory consolidation (Schredl, 2002). Visual dream records could also support clinical practice. For instance, in psychotherapy, patients often struggle to verbalise the atmosphere of a dream, and a shared visual representation could give the therapist and patient a common reference point for mood, imagery, and emotional tone. This is particularly relevant for trauma-related dreams, where Image Rehearsal Therapy uses patient-generated imagery therapeutically (Krakow et al., 2001), and similar to Cheatley et al. (2020) using co-creative songwriting to support people who lack artistic expertise, during bereavement.

Existing tools have limitations for dream visualisation. In particular, text-to-image models can produce a striking single image but not a sequence with its own mood, pacing, and emotional progression. Story-visualisation models such as those in StoryDALL-E (Maharana, Hannan, and Bansal, 2022) and AR-LDM (Pan et al., 2024) handle image sequences. However, they were trained on paired text-image datasets largely consisting of everyday imagery. Hence they tend to smooth out unusual features that make dreams distinctive. As there is no dataset of dream visualisations matched to text, a different approach is needed.

We present the Dream Scene Visualiser (DSV), which turns a written dream description into a visualisation of the dream in four panels, each containing an image and textual information. In addition to image sophistication and fidelity to the dream narrative and mood, a key consideration for the process is that coherence is maintained over the whole visualisation and from panel to panel. In the next section, we describe how narrative coherence is maintained using the Qwen2.5(14B) large language model (LLM) (Yang et al., 2024) to break the dream description into parts. Moreover, visual coherence is maintained by seeding the generation of panel images 2, 3 and 4 with the previous image, using the SDXL text-2-image generator (Podell et al., 2023). Finally, for semantic coherence, a feedback loop using the vision-language model CLIP(ViT-B/32) (Radford et al., 2021), measures if a panel image suitably matches the description, regenerating any image which does not.

We evaluate the system on 50 dream descriptions from the DreamBank repository (Domhoff and Schneider, 2008). The evaluation protocol involves assessing text/image alignment using CLIP and panel-to-panel visual similarity, using DINOv2 (Oquab et al., 2024), a self-supervised vision model. We also use the Qwen2-VL(7B) (Wang et al., 2024) vision-language model (VLM) as an independent judge evaluating seven dimensions of quality in the dream visualisations.

The DSV Pipeline

A dream description is first annotated with two attributes before entering the pipeline: a complexity tier between 1 and 5, and a dominant style label of either *surreal*, *nightmare*, *peaceful* or *fantastical*. Annotation is performed by Qwen2.5(14B Instruct) using a prompt based on the dream coding system of Hall and Van de Castle (1966). The full prompt is given in appendix A. Figure 1 overleaf shows a trace of the DSV pipeline applied to a representative DreamBank report. The pipeline processes are:

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

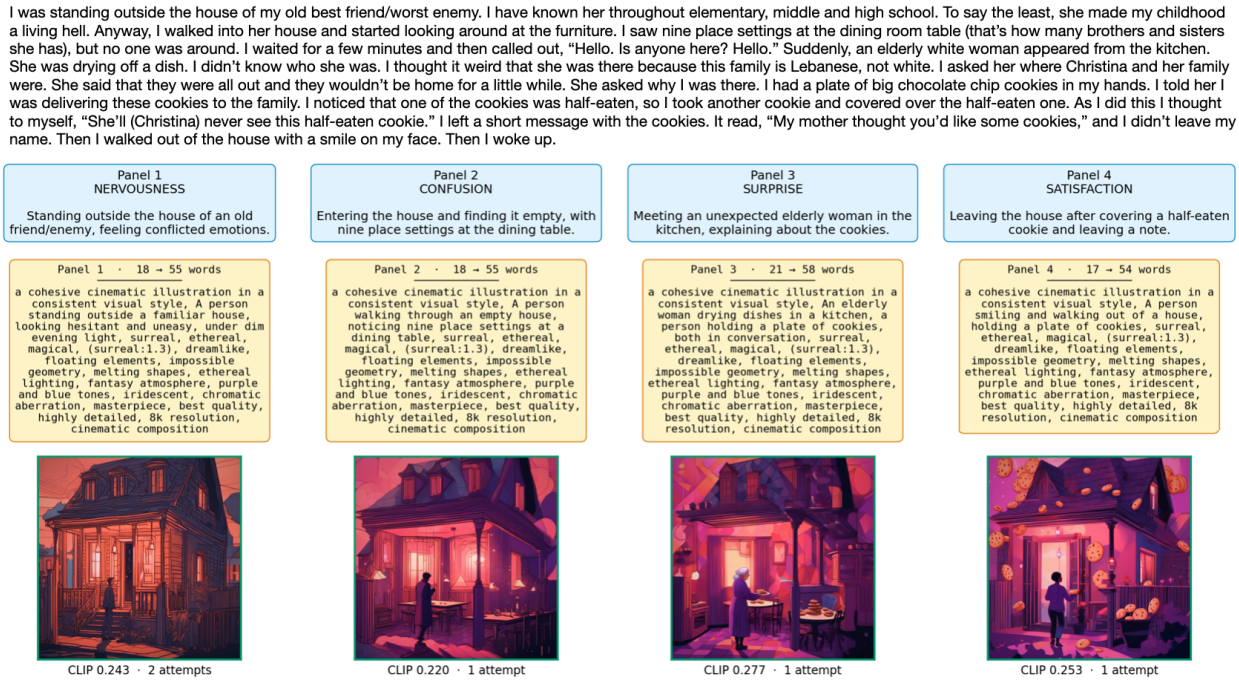


Figure 1: Pipeline trace for DreamBank example `DB-b533e9bb`: (a) original dream text (b) decomposition into four short panel texts and dominant emotion label (c) enriched panel texts for use as prompts in SDXL (d) generated panel images after post-processing, given with feedback loop repetitions and CLIP cosine similarities to the texts.

• **Panel decomposition.** For narrative coherence, a dream is first split into four chronological panel texts. This is achieved by prompting the Qwen2.5(14B) LLM with temperature 0, using the prompt given in appendix B. This produces (a) a short description for each panel (b) a dominant emotion label (c) an SDXL-ready visual prompt and (d) a negative SDXL prompt.

• **Prompt enrichment.** The LLM-generated prompts are short and do not carry enough stylistic signal for SDXL to produce a coherent visual style across the four panels. We therefore enrich each prompt by prepending the style label annotation, then appending style-specific positive tokens, emotion-specific tokens and a generic quality booster. This stage brings the LLM prompt from around 18 words to around 55 words per image, on average in the example of figure 1. Details of how token dictionaries are used for enriching positive and negative SDXL prompts are given in appendix C.

• **Image generation.** For visual coherence, all images are generated with SDXL 1.0 at 768 × 768 pixel resolution. SDXL generates images to match a positive prompt and avoid matching a negative prompt. Panel 1 is generated from the first panel positive and negative prompts, but panels 2 to 4 use `img2img` chaining, conditioned on the previous panel image. This means that, along with the positive and negative prompts, SDXL is forced to use the previous image, with a weighting of 0.3/0.7 respectively. These weights were determined empirically: `img2img` strengths below 0.5 collapse panels onto near-duplicates of panel 1, while strengths above 0.8 lose character and setting continuity. We found that strength 0.7 preserves recognisable continuity, while allowing each panel to introduce its own content. This mechanism works best when the dream stays within a single setting. As an example, in figure 1, the dark house, the dreamer in silhouette, and the red-purple palette all persist across all four panels.

• **CLIP feedback loop.** For semantic coherence, after each image is generated, we use CLIP to measure the cosine similarity between the image and the text description of each panel, with higher values indicating closer agreement between the two. If the alignment score falls below a calibrated threshold of 0.228, the panel is regenerated with a prompt revised via emphasis changes and token additions such as "sharp focus", with up to three attempts allowed in total. In the example of figure 1, the loop fired once on panel 1, as the first attempt fell below the threshold, and the revised prompt on attempt 2 was accepted with CLIP cosine similarity of 0.243.

• **Post-processing.** Each of the four images receives a lightweight cosmetic pass, bringing contrast to 1.15 and brightness to 1.05. Following this, a Gaussian blur with $\sigma = 0.4$, and a radial vignette at intensity 0.25 are applied. We have found such post-processing helps unify the four panels visually for dream-journal presentation.

Figure 1 shows a *surreal* example, and appendix F contains additional example visualisations in the *nightmare*, *fantastical*, and *peaceful* categories.

Evaluation

Fixing the panel count for visualisations at four keeps the output format consistent across dreams of different length. This has enabled an evaluation of DSV on DreamBank reports with three complementary signals: text-image alignment, panel-to-panel visual coherence, and a seven-dimension vision-language judge.

We sampled 50 dream reports from DreamBank, drawn from eight dreamer series spanning everyday, lucid, anxiety-laden, and long-narrative dream types. The dream reports were annotated using Qwen with a dream style as follows: *nightmare*: 18 dreams, *fantastical*: 6, *surreal*: 14 and *peaceful*: 12. The word counts for the enriched prompts range from 52 to 412, with a median of 194. CLIP ViT-B/32 was used to measure image-to-text alignment between each generated panel and its enriched description; we report per-dream averages. We further used the DINOv2 model to measure visual similarity between adjacent panels, with a high score indicating that consecutive panels share a character, palette, and setting, which is the form of coherence `img2img` chaining is designed to produce.

To further evaluate the dream visualisations, we implemented a *judge* using the Qwen2-VL(7B) vision-language model. Full details of how the model is prompted in order to pass judgement on the dream visualisations are given in appendix D. In overview, the judge is called twice per dream. Firstly, it is employed once per panel to judge four aesthetic qualities of the individual images, namely: *aesthetic quality*, *character presence*, *dreamlike quality* and *text faithfulness*. Secondly, it is given all four panel images in one image, and asked to judge three per-sequence dimensions, namely: *style consistency*, *narrative coherence* and *emotional progression*. The judge is prompted to score each image/sequence quality on a 1 to 5 Likert scale, with 1 indicating low quality and 5 indicating high quality.

Across 50 dreams, using the full pipeline produces mean CLIP cosine similarity for image/text pairs of 0.250 ± 0.020 and mean adjacent-panel DINOv2 coherence of 0.623 ± 0.143. Shuffled-baseline controls confirm that both signals are above chance: matched CLIP=0.252 vs shuffled=0.201 ($p = 6 \times 10^{-25}$, paired Wilcoxon), chained DINOv2=0.623 vs random 4-tuples=0.182 ($p = 1.4 \times 10^{-27}$, Mann–Whitney U). The two metrics are weakly correlated ($r = +0.24$), confirming they measure different aspects of coherence.

The scores assigned by the judge are given in table 1. These indicate that high style consistency and aesthetic quality (both above 4/5) are achieved, with moderate text faithfulness (3.2), but lower scores for sequence-level narrative and emotional coherence. This is consistent with the production of stylistically strong, visually continuous imagery, but only partially captures narrative detail. CLIP correlates moderately with the judge's text-faithfulness score ($r = +0.38, p = 0.007$), supporting CLIP usage as a coarse alignment signal rather than a stand-in for perceptual quality. We found that adjacent DINOv2 coherence decreases along the chain (from 0.66 at P1→P2 to 0.58 at P3→P4), reflecting accumulated visual drift from repeated `img2img` conditioning.

Judge dimension	**Mean**	**Std**
Per-panel (N=200)		
Aesthetic quality	4.41	0.62
Character presence	3.81	1.33
Dreamlike quality	3.67	0.72
Text faithfulness	3.21	1.11
Per-sequence (N=50)		
Style consistency	4.92	0.57
Narrative coherence	2.68	1.17
Emotional progression	2.30	1.33

Table 1: Judge ratings on a 1 to 5 Likert scale. Per-panel dimensions averaged over 200 generated panels; per-sequence dimensions are single ratings per dream.

• **A subjective evaluation.** In a small pilot study (N=10), subjects rated visualisations for their own dream descriptions, using 1 to 5 Likert evaluations covering style, emotional palette, creative detail, narrative accuracy, object fidelity, and transitions. We found that the ratings align with the automated judge: style application scored highest (av. 4.40), followed by emotional palette (4.20) and creative detail (4.00), while object fidelity (3.60) and transitions (3.50) scored lowest.

• **An ablation study.** We ablate each pipeline stage on a held-out subset of 10 dream descriptions. Starting from the full pipeline (A), we remove prompt engineering (B), the CLIP feedback loop (C), `img2img` chaining (D), and finally reduce the system to SDXL on unprocessed LLM prompts (E, baseline). Each ablation is tested against A with a two-sided Wilcoxon signed-rank test with Bonferroni correction ($\alpha = 0.0125$). The results are given in table 2. We see that removing prompt engineering (B) or returning to baseline (E) both significantly increase CLIP alignment relative to the full pipeline. Removing `img2img` (D) shows a marginal effect that does not reach significance after correction. We also see that removing the CLIP loop (C) shows no significant change, which is because the feedback loop rarely caused re-generations, in fact it happened for only 3% of panels.

• **LoRA fine-tuning.** To investigate whether the process could be improved with appeal to more dream-like imagery, we fine-tuned a LoRA of rank 16 on SDXL's

Condition	**CLIP**	**Δ**	p	**Sig.**
A. Full pipeline	0.250	—	—	—
B. No prompt eng.	0.260	+0.010	0.002	s.
C. No CLIP loop	0.249	−0.001	0.625	n.s.
D. No img2img	0.265	+0.016	0.014	n.s.
E. Baseline	0.265	+0.015	0.004	s.

Table 2: Ablation over 10 dreams: CLIP: per-dream mean; Δ: difference from condition A; p-values and significance with s. = significant and n.s. = not significant.

UNet using 50 surreal artworks found on Pinterest. LoRA (Hu et al., 2021) is a parameter-efficient fine-tuning method that adapts pre-trained models without modifying their weights. On 5 surreal dreams sampled from the evaluation set, there was a mean CLIP cosine similarity improvement from 0.280 to 0.290, with 4 out of 5 dreams being improved. On generic prompts, however, the CLIP score dropped by 0.009. We found that the effect is content-conditional, i.e., the fine-tuning helps where the base model's priors are weakest and hinders where they are already strong.

Discussion

DSV produces stylistically strong and visually coherent sequences, with judge scores for style consistency of 4.92/5 and aesthetic value of 4.41/5. However, the full pipeline approach scores significantly lower for CLIP cosine similarity than a minimal baseline ($p < 0.005$). Moreover, CLIP scores drop when we add our coherence-oriented components, as shown in table 2. We also see that DINOv2 coherence decays along the `img2img` chain, and the judge rates the sequences highly on perceptual quality. So it seems that our design choices for DSV improve the output. However, CLIP indicates the opposite. This pattern is partly explained by SDXL's 77-token text encoder limit: enriched prompts exceed this limit, and SDXL silently drops the overflow, which contain the emotion and narrative tokens at the end of our template. This helps to explain the judge's weak per-sequence scores (narrative 2.68, emotional 2.30) alongside strong per-panel ones (all above 3.2). CLIP cannot detect this gap because it scores against the short panel description, not the truncated enriched prompt.

A natural concern is whether the judge simply rewards high quality images. Three observations argue against such bias: it discriminates between dimensions (style 4.92 vs emotional progression 2.30), its text-faithfulness correlates with CLIP ($r = +0.38, p = 0.007$), and the low narrative scores have a mechanical explanation in the silent truncation issue mentioned above. Nevertheless, a VLM-based judge is a proxy for human judgement, not a replacement.

Two dreams are provided in appendix E which illustrate the failure modes of single-metric evaluation. DreamBank dream `DB-3c0442ac` receives above-mean CLIP (0.263) but the judge scores it as 2.5/5 on text faithfulness and 1/5 on emotional progression: CLIP rewards a coherent palette while missing character drift, a false positive. Dream `DB-dad219b7` receives a below-mean CLIP score of 0.231, but 5/5 on aesthetic quality and style consistency. Here, stylisation is penalised by CLIP even though the content is clearly there, which is a false negative. No single metric separates these two cases, but the multi-signal approach does.

Our approach has several limitations. While a pilot human study ($N = 10$) supports the judge's main trends, its sample is too small to stand alone, and VLM judges more broadly are known to exhibit verbosity and self-enhancement biases (Zheng et al., 2024). A larger human study therefore remains the strongest test of the trade-off reported here. The complexity annotation was not used in the pipeline, and we found that tier-1 and tier-5 complexity categories are under-sampled ($n = 1$ out of 50 for each). Moreover, the LoRA ablation covers only five surreal dreams, and the CLIP feedback loop fires on just 3% of panels, functioning as a safety net rather than an active quality gate.

Related Work

Two recent systems address dream content directly. DreamStory (He et al., 2024) uses an LLM as a "story director" paired with a specialised diffusion model for character consistency, but targets open-domain fiction such as fairytales. DSV differs by targeting dream reports specifically, adding a CLIP-based self-correction loop, and using a multi-dimensional evaluation protocol. DreamLLM-3D (Liu et al., 2024) takes a different path, displaying dream entities as animated 3D point clouds inside an art installation. In other directions, video dream visualisations from fMRI scans are beginning to be explored (Fu et al., 2025). Recent work on using large language models as evaluative judges (Franceschelli and Musolesi, 2024) offers a flexible alternative to traditional metrics, reasoning about creative quality beyond surface-level matching. Our evaluation builds on this direction by combining a vision-language judge with CLIP-based and DINOv2-based signals, motivating the multi-dimensional protocol we use in place of any single metric.

Conclusion and Future Work

We presented the Dream Scene Visualiser, a coherence-oriented system that turns written dream descriptions into four-image visualisations. On 50 DreamBank dreams, DSV produces stylistically strong and visually continuous sequences (judge style 4.92/5, aesthetic 4.41/5) while scoring lower on CLIP than a minimal baseline ($p < 0.005$). We traced this style-fidelity trade-off to silent prompt truncation in SDXL's text encoder, and argued that CLIP alone is insufficient to evaluate dream visualisations. Directions for future work include broader LoRA fine-tuning across all four dream styles and addressing SDXL's 77-token limit, by exploring prompt-engineering strategies that keep prompts within the budget while still including emotion and narrative anchors. We also plan to use the complexity tier annotation to support dynamic panel counts, replacing the current fixed four-panel structure. We hope to eventually trial dream visualisation tools in psychotherapy and other counselling contexts.

Acknowledgments

We thank the participants of our pilot study for their time and the maintainers of DreamBank for making dream descriptions publicly available. We also thank the anonymous reviewers for their helpful feedback.

References

Cheatley, L.; Ackerman, M.; Pease, A.; and Moncur, W. 2020. Co-creative songwriting for bereavement support. In *Proceedings of the 11th International Conference on Computational Creativity (ICCC)*, 33–40.

Domhoff, G. W., and Schneider, A. 2008. Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition* 17(4):1238–1247.

Franceschelli, G., and Musolesi, M. 2024. Creative beam search: LLM-as-a-judge for improving response generation. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*, 364–368.

Fu, Y.; Gao, J.; Yang, B; and Feng, J. 2025. Making your dreams a reality: Decoding the dreams into a coherent video story from fMRI signals. *arXiv preprint arXiv:2501.09350*.

Hall, C. S., and Van de Castle, R. L. 1966. *The Content Analysis of Dreams*. New York: Appleton-Century-Crofts.

He, H.; Yang, H.; Tuo, Z.; Zhou, Y.; Wang, Q.; Zhang, Y.; Liu, Z.; Huang, W.; Chao, H.; and Yin, J. 2024. DreamStory: Open-domain story visualization by LLM-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(12)*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Krakow, B.; Hollifield, M.; Johnston, L.; Koss, M.; Schrader, R.; Warner, T. D.; Tandberg, D.; Lauriello, J.; McBride, L.; Cutchen, L.; Cheng, D.; Emmons, S.; Germain, A.; Melendrez, D.; Sandoval, D.; and Prince, H. 2001. Imagery rehearsal therapy for chronic nightmares in sexual assault survivors with post-traumatic stress disorder: A randomized controlled trial. *JAMA* 286(5):537–545.

Liu, P.; Maverick, K.; Steinmaurer, A.; Picard-Deland, C.; Carr, M.; and Kitson, A. 2024. DreamLLM-3D: Affective dream reliving using large language model and 3D generative AI. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Maharana, A.; Hannan, D.; and Bansal, M. 2022. StoryDALL-E: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision (ECCV)*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Pan, X.; Qin, P.; Li, Y.; Xue, H.; and Chen, W. 2024. Synthesizing coherent story with auto-regressive latent diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.

Schredl, M. 2002. Questionnaires and diaries as research instruments in dream research: Methodological issues. *Dreaming* 12(1):17–26.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; and others. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; and others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2024. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Appendix A: Annotation LLM Prompt

Prior to annotation, DreamBank header markers (e.g. "`#028 (05/13/2000)`") are stripped from each dream text. The dream text is then appended to the following prompt for the Qwen2.5(14B Instruct) LLM (with temperature set to 0), and the JSON output is processed to annotate the dream text.

You are a dream researcher trained in the Hall & Van de Castle coding system, annotating dreams for visual art generation.

For the following dream, assess two dimensions.

1. Complexity (integer, 1–5):

- **1** – short report; single setting; no bizarre elements
- **2** – single setting; one action thread; minor anomalies
- **3** – multiple settings OR clear anomalies (time distortion, impossible objects)
- **4** – narrative arc across 3+ settings; bizarre transformations; dream-logic
- **5** – highly complex; multiple transformations; composite characters/settings; impossible physics

2. Dominant style — judge by emotional tone and visual atmosphere, not by whether the setting is everyday. Apply the first criterion that fits:

- **nightmare** – fear, threat, being chased, trapped, unsettling figures, violence, property damage, police pursuit, anxiety, distress, or dark atmosphere.
- **fantastical** – magical or impossible-but-positive elements: flying (leisurely, not falling), magic, mythical creatures, cosmic scale, enchanted objects, walking through portals with wonder.
- **surreal** – reality bending in ways that are strange but not clearly positive or negative: melting shapes, impossible geometry, objects transforming, composite characters, dream-logic without clear emotional valence.
- **peaceful** – default for everything else: calm, warm, nostalgic, tender, joyful, proud, loving, or restful, or emotionally flat everyday content (having coffee, walking home, ordinary conversations).

Respond only with valid JSON: {`"complexity": <1-5>, "dominant_style": "<label>", "reasoning": "<one-sentence rationale>"`}

Appendix B: Panel Decomposition LLM Prompt

Qwen2.5(14B Instruct) with temperature 0 is used to generate four panel description texts from the overall dream text. The dream text is appended to the prompt given below. In the prompt, the {`style`} placeholder is replaced at call time with the dream's pre-assigned style annotation (one of *surreal*, *nightmare*, *peaceful*, *fantastical*), as per appendix A. Each response is validated against a schema requiring exactly four scene objects, each containing a scene number, a short description, an emotion label, a positive SDXL prompt (minimum 10 characters), and a negative SDXL prompt.

You are a dream analyst and visual prompt engineer. The user will give you a dream description. Your job:

1. Split the dream into exactly 4 chronological scenes.
2. For each scene, write a detailed Stable Diffusion prompt in English.
3. Apply the style "{style}" to every prompt.
4. Identify the dominant emotion of each scene.

Rules for the prompts:

- Each prompt must include the dreamer as a visible character doing something specific (walking, reaching, swimming, etc.)
- Describe body language, pose, and interaction with the environment.
- Describe colours, lighting, and atmosphere.
- Add "dreamlike, ethereal, magical" to every prompt.
- Keep prompts between 40 and 70 words.
- Put "unpopulated, empty scenery, no characters" in all negative prompts.

Respond with valid JSON only — no markdown, no explanation.

Appendix C: Prompt enrichment details

SDXL allows both a positive and a negative prompt as inputs. Each enriched positive prompt in the DSV pipeline is constructed as:

---

a cohesive cinematic illustration in a consistent visual style, [LLM prompt], [style-positive tokens], [colour palette], [emotion tokens], masterpiece, best quality, highly detailed, 8k resolution, cinematic composition

---

Style-positive tokens and colour palette tokens are taken from this lookup table:

Style	Positive tokens; colour palette
nightmare	dark horror, ominous shadows, twisted forms, unsettling atmosphere, eerie fog, gothic, sinister; dark red and black tones, desaturated, high contrast
fantastical	fantasy, magical particles, enchanted, mythical creatures, glowing runes, epic scale, vibrant colors; rich jewel tones, emerald green, sapphire blue, golden highlights
surreal	surreal, dreamlike, floating elements, impossible geometry, melting shapes, ethereal lighting, fantasy atmosphere; purple and blue tones, iridescent, chromatic aberration
peaceful	serene, calm, gentle light, soft focus, warm glow, tranquil, harmonious, pastel colors; warm golden and soft pink tones, low saturation, gentle gradients

The emotion tokens are taken from this lookup table:

Emotion	Tokens
mystery	mysterious atmosphere, hidden depths, subtle shadows, enigmatic
serenity	peaceful, calm water, gentle breeze, soft ambient light
wonder	awe-inspiring, magical discovery, wide-eyed amazement, sparkling
amazement	breathtaking vista, overwhelming beauty, spectacular, grand reveal
fear	creeping dread, looming danger, cold sweat, dark corners
joy	radiant happiness, warm sunlight, blooming flowers, celebration
sadness	melancholic, rain, fading colors, lonely atmosphere, grey skies
anger	intense red, stormy, crashing waves, volcanic, fierce energy
anxiety	distorted perspective, claustrophobic, uneasy lighting, tight spaces
awe	vast scale, cosmic, luminous, transcendent beauty
nostalgia	golden hour, faded photograph, soft grain, warm memories
curiosity	inviting path, glowing doorway, beckoning light, unknown territory

The negative prompt for SDXL is constructed as:

[style-negative tokens], ugly, blurry, low quality, text, watermark, logo, deformed, disfigured, bad anatomy, extra limbs, jpeg artifacts, poorly drawn

---

The style-negative tokens for the SDXL negative prompt are taken from this lookup table:

Style	Negative tokens
nightmare	bright, happy, cheerful, cute, cartoon
fantastical	realistic, modern, urban, mundane, dull
surreal	realistic, photograph, mundane, ordinary
peaceful	dark, scary, violent, chaotic, harsh

Appendix D: Vision-language Judge Prompts

The judge we implemented uses the Qwen2-VL(7B) vision-language model. The first call rates each generated image individually on four per-panel dimensions, using the prompt below, with placeholders {`description`} and {`emotion`} replaced with the panel's description and emotion label.

You are evaluating an AI-generated dream visualisation.

Scene description: "{description}"
Intended emotion: {emotion}

Rate this image on a 1–5 scale for each criterion. Respond only with valid JSON.

Guide:

- **dreamlike_quality**: Does it feel surreal and dreamlike? (1 = mundane, 5 = deeply dreamlike)
- **text_faithfulness**: Does it match the description? (1 = unrelated, 5 = exact match)
- **aesthetic_quality**: Is it visually compelling? (1 = ugly, 5 = beautiful)
- **character_presence**: Is a human figure visible doing something? (1 = no person, 5 = clear action)

The second call rates the full sequence on three per-sequence dimensions, with all four images shown together as input alongside the dream text and individual panel descriptions. The prompt for this is as follows:

You are evaluating narrative coherence across 4 dream scene images shown above.

The dream: "{dream_text}"
Scene 1: {s1}
Scene 2: {s2}
Scene 3: {s3}
Scene 4: {s4}

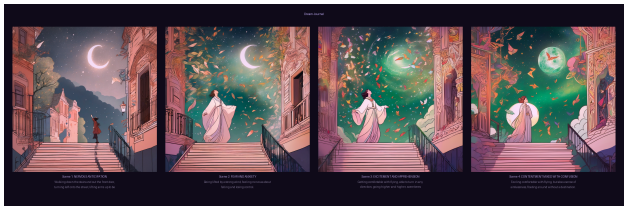
Rate the 4-image sequence on a 1–5 scale. Respond only with valid JSON.

- **narrative_coherence**: Do these tell a visual story? (1 = random, 5 = clear arc)
- **style_consistency**: Same artistic style throughout? (1 = different, 5 = unified)
- **emotional_progression**: Does emotion evolve across scenes? (1 = flat, 5 = clear arc)

Appendix E: CLIP Failure Modes



False positive: (`DB-3c0442ac`, surreal). CLIP = 0.263 (above mean) but judge text faithfulness 2.5/5, character presence 3.0/5, emotional progression 1/5. Coherent palette masks character drift across panels.



False negative: (`DB-dad219b7`, fantastical). CLIP = 0.231 (below mean) but judge faithfulness is 4.25/5, aesthetic quality 5.0/5, style consistency 5/5. Stylisation penalised by CLIP, despite strong semantic alignment.

Appendix F: Additional Dream Outputs

Below are additional dream sequences spanning three style categories, selected to include both high- and low-CLIP score cases. Each entry shows the input dream text, the four generated panels, and the full evaluation signals (CLIP, DINOv2 adjacent coherence, and seven-dimension judge scores reported per panel and per sequence). We provide a subjective commentary for each.

DreamBank Dream DB-3a3aa5d7 nightmare, 171 words, complexity 4

"I was working for a firm. Accidentally I found the firm's records that they were carrying on illegal activities plus their surface business. Somehow an official of the firm learned of my knowledge and held me in captivity. After a while in captivity I was able to get word to an outside person. However, this didn't have any results. But after a while I escaped. As soon as I got away, I went to the police and told them my story. Then the police and I went to this building and alerted all the innocent people who were working for other firms. The police then staged a raid, but still the criminals wouldn't give up. Then the police used tear gas, and the criminals finally gave up. Then there seemed to be some time elapsed, and the next thing I can remember is a photographer taking my picture in front of the building, and the picture was to be printed in the next issue of the newspaper. Then I woke up."



Evaluation. CLIP: 0.234; DINO: 0.450; Judge per-panel avg. (aesthetic: 4.00, character: 4.00, dreamlike: 3.25, faithfulness: 3.00); Judge per-sequence: (style: 5, narrative: 3, emotional: 3).
Commentary. The four panels render the main sequence: office, escape, police raid, exterior, with recognisable elements like helmets and tear gas haze in panel 3. The monochromatic palette is consistent with the nightmare label, although the captivity stage and the final newspaper photograph are visualised more loosely.

DreamBank Dream DB-29e615cd fantastical, 194 words, complexity 3

"A very small plane came straight down through the trees of Lang Lawn. It was spring and a lot of girls were watching it come down. Finally it came down with no damage resulting. Barbara Ayres, a sophomore here, climbed out. Everyone was quite surprised to see that it was she. Babs left the plane on the lawn and went into the dorm. Joe Warner, a senior, appeared and claimed that the plane was his. Nobody seemed to question why Babs was flying it. The plane was the smallest one I've ever seen. There were two seats, one behind the other, and an average individual could just fit into it. Lee Ellsworth, another senior, told the onlookers that it took 9.4 seconds to get off the ground if Joe did it. Suddenly I was in the driver's seat and Joe asked me if I knew where the throttle was. It was under a rubber mat on the floor. The instrument panel had very little on it. But I got it started in 9.0 seconds. I went off the ground and Lee was yelling at Joe that I beat his take-off time. Then I landed."



Evaluation. CLIP: 0.257; DINO: 0.932; Judge per-panel avg. (aesthetic: 4.75, character: 4.50, dreamlike: 4.25, faithfulness: 4.00); Judge per-sequence: (style: 5, narrative: 1, emotional: 1).
Commentary. The fantastical aesthetic is rendered with consistent confidence: pastel skies, autumn foliage, and a recurring biplane. However, the plane remains airborne in all four panels rather than landing after being flown by the dreamer.

DreamBank Dream DB-1c6bab2b peaceful, 73 words, complexity 2

"I was crossing a valley (a small piece of land between two hills), and I heard a friend of mine, a young boy in my neighborhood of about the same age as myself, shouting to me from the top of one of the hills. I looked at him, and he kept on shouting my name and something else to me. I started to run very fast until I came home. I then awoke."



Evaluation. CLIP: 0.267; DINO: 0.709; Judge per-panel avg. (aesthetic: 5.00, character: 4.50, dreamlike: 4.00, faithfulness: 3.75); Judge per-sequence: (style: 5, narrative: 5, emotional: 5).
Commentary. The peaceful aesthetic is captured throughout, with a soft pink-and-gold palette across valley, path, and home, and the narrative arc is largely preserved. The main deviation is in panel 2, where the boy stands beside the dreamer on the path rather than calling from the top of a hill.

How are scientific concepts birthed? Typing rules of concept formation in theoretical physics reasoning

Omar Aguilar* **and Anthony Aguirre**
Physics Department
University of California, Santa Cruz
Santa Cruz, CA 95064 USA
*Corresponding author: omalagui@ucsc.edu

Abstract

This work aims to formalize some of the ways scientific concepts are formed in the process of theoretical physics discovery. Since this may at first seem like a task beyond the scope of the exact sciences, we begin by presenting arguments for why scientific concept formation can be formalized. Then, we introduce type theory as a natural framework for this formalization. We formalize what we call "ways of discovering new concepts" including property preservation and concept change, as cognitive typing rules. Next, we apply these cognitive typing rules to a case study of conceptual discovery in the history of physics: Einstein's conceptual path to the relativity of time. Then, we recast what a physicist might informally call "ways of discovering new scientific concepts" as compositional typing rules built from cognitive typing rules—thus formalizing them as scientific discovery mechanisms. Lastly, we computationally model the type-theoretic reconstruction as a program synthesis task.

Introduction

The standard picture of the scientific method often emphasizes hypothesis testing. That picture leaves a prior question underspecified: how are new scientific concepts formed? One of the first concrete answers is Peirce's view that concept formation is abduction: the process of imposing constraints on admissible hypotheses to generate plausible, though fallible, explanations (Chomsky 2006). Building on Peirce, more recent work in cognitive science of science treats scientific concepts as entities shaped by the constraints of the knowledge-construction practices scientists use to generate them (Nersessian 2008; Gentner 1983). Perhaps the most detailed engagement between cognitive science of science and natural science comes from Nancy Nersessian's *cognitive-historical analysis* of theoretical physics. It treats historical records of theoretical physics practice as empirical evidence and reads them as traces of cognitive mechanisms such as analogy, mental modeling, and thought experiments (Nersessian 1987).

Since the discovery methods Nersessian reveals are systematic procedures with definite and finite steps, one may ask:

Can concept formation in theoretical physics be formalized?

To address this question, one may be inclined to explore machine learning approaches that have successfully automated other aspects of scientific discovery (Battaglia et al. 2018; Udrescu et al. 2020; Lalande et al. 2023) , with the aim of adapting these techniques to model scientific concept formation. However, these methods still face the challenge of modeling conceptual development, as they were built to solve predefined problems rather than define new ones (Battleday and Gershman 2024). This suggests that we should take a step back to clarify what has so far been left vague: what we mean by constraints in scientific concept formation.

As part of this effort, physicists might recall that when searching for new scientific concepts, they often use the word "type" or "kind" informally to describe what they are looking for. For instance, Maxwell initially conceived of the magnetic field as a "type" or "kind" of stress composed of pressures and tensions (Maxwell 1865). These *intuitive types* are not necessarily physically correct claims, but rather common-sense constraints that help narrow the space of possible concepts moving forward. Their intuitiveness stems from the fact that they are used most frequently in everyday human reasoning (e.g., "What time is it?" calls for a time, not a food).

Felix Sosa has recently proposed a natural formalization of these intuitive types as *theoretical types* which are the building blocks of *type theory* (Sosa and Ullman 2022). This makes one wonder: if intuitive types used in everyday thought—like "bread" or "dog"—are slippery yet still conceivable as theoretical types, what about those used in scientific reasoning, such as "has 3 distinct components" which are already well-defined? This brings us to the central question of this paper:

Can concept formation in theoretical physics be formalized using type theory?

Our answer is yes. We substantiate this claim through a type-theoretic formalization of scientific concept formation driven by detailed reconstructions of scientific reasoning from the history of physics. In this framework, intuitive types of scientific concepts are formalized as types and different "ways of discovering scientific concepts" as typing rules—thereby recasting them as *discovery mechanisms*. Moreover, scientific concepts are formalized both as *terms*

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

and as the results of *typing rules*—thus defining them by how they are discovered, in the spirit of Peirce and cognitive science of science.

Lastly, a type system can be understood as a mathematical formalization of a programming language (Cardelli 1996), where terms correspond to programs and types specify how those programs may be used. This perspective allows us to recast scientific discovery as *program synthesis*: the inference of programs from data and constraints (Rule, Tenenbaum, and Piantadosi 2020). To illustrate this idea, we will carry out type-theoretic reconstructions of a canonical example of concept formation from the history of physics, and computationally implement it using program synthesis.

Overview

This work aims to provide a type-theoretic formalization of some ways in which scientific concepts are formed in theoretical physics, based on reasoning patterns reconstructed from the history of theoretical physics. To this end, the paper undertakes the following recastings:

1. Intuitive types are formalized as theoretical types.
2. "Ways of discovering new concepts"—such as property preservation and concept change—are formalized as elementary typing rules, which we call cognitive typing rules.
3. "Ways of discovering new scientific concepts" are formalized as compositional typing rules that integrate cognitive typing rules with functional and algebraic operations. In doing so, these "ways" are recasted as discovery mechanisms.
4. Scientific concepts are formalized both as terms and as the outcomes of discovery mechanisms.

The formalism we develop is applied to a paradigmatic conceptual breakthrough in the history of physics: Einstein's special relativity. Specifically, we examine a key episode in Einstein's reasoning that paved the way to his discovery. Namely, the conceptual path that led him to relative time. Einstein's reasoning leading to the relativity of time is formalized as a discovery mechanism that reverses the roles of assumption and conclusion through a composition of property preservation and concept change typing rules with functional and algebraic operations; we call this the `assumption–conclusion switch` discovery mechanism. Lastly, we use typed-program synthesis to simulate Einstein's conceptual path to the relativity of time.

Related work

Computational scientific discovery, or AI for science, spans symbolic and deep-learning approaches, but much of it optimizes within fixed problem formulations rather than modeling the conceptual work of forming new representations and questions (Battleday and Gershman 2024). LLM-based AI scientist systems can propose hypotheses in natural language, yet their outputs are often produced by opaque procedures and lack formal representations that let concepts be precisely defined and related, leaving them vulnerable to ambiguity and meaning overload (Liu et al. 2023; Wolfram 2024). This work targets this gap.

Methods

Type theory

Type theory is a branch of mathematical logic that classifies mathematical objects (terms) by assigning them types, which constrain how terms can be used and combined (Pierce 2002). It can also be regarded as a formalization of programming languages (Pierce 2002). Here we extend type theory to model concept formation in theoretical physics reasoning by treating scientific concepts as terms and their properties (intuitive types) as types.

Terms, and types. A *term* is an expression (sequence of symbols) that carries semantic content in the system, so it can be assigned a type. A *type* classifies terms and constrains their use.

Contexts. A context Γ is a collection of assumptions under which an assertion is made.

Typing judgments. A typing judgment asserts that, under context Γ, a term t has type A:

$$\Gamma \vdash t : A. \tag{1}$$

Typing rules. Typing judgments are defined inductively as a collection of typing rules. Each rule derives a conclusion judgment from premise judgments. Schematically:

$$\frac{\text{premise}_1 \ \cdots \ \text{premise}_n}{\text{conclusion}}. \tag{2}$$

Formally:

$$\frac{\Gamma_1 \vdash t_1 : A_1 \quad \ldots \quad \Gamma_n \vdash t_n : A_n}{\Gamma \vdash t : A} \tag{3}$$

where Γ_i is a local extension of Γ specific to the ith premise

Intersection type. The intersection type $A \cap B$ combines types A and B; a term of type $A \cap B$ has both types A and B.

Cognitive typing rules

Forward property preservation rule.

$$\frac{\Gamma \vdash x : A}{\Gamma \vdash f(x) : A} \tag{4}$$

Under this rule, if a concept x has property A, then applying f preserves that property.

Backward property preservation rule.

$$\frac{\Gamma \vdash f(x) : A}{\Gamma \vdash x : A} \tag{5}$$

Under this rule, if a transformed concept $f(x)$ has property A, then the source concept x must also have property A.

Concept change rule. This rule states that if two terms, x and y, are identical but carry different types under different contexts, then in the combined context, the term x is assigned the intersection type.

$$\frac{\Gamma_1 \vdash x : A \quad \Gamma_2 \vdash y : B \quad x = y}{\Gamma_1 \cup \Gamma_2 \vdash x : A \cap B} \tag{6}$$

Results

The Michelson–Morley experiment is often regarded as one of the crucial optical experiments that Einstein sought to explain in his effort to extend the principle of relativity to electrodynamics. However, according to Einstein himself, that was not the case; instead, the experiment of stellar aberration played the most critical role. In particular, his reinterpretation of Lorentz's local time, in Lorentz's stellar aberration account, as empirical time led him to the relativity of time.

Norton (Norton 2004) argues that reading Lorentz's account of stellar aberration through the lens of relativity could have led Einstein to conclude that stellar aberration provided experimental support for the physical reality of local time, and thus that it could be regarded simply as time. This line of reasoning can be viewed as a discovery mechanism that swaps assumption and conclusion, which is described in natural language and type-theoretic formalism below.

Assumption-conclusion switch discovery mechanism

Step 1. Lorentz regarded the concept of local time $t' = t - \frac{ux}{c^2}$, as an artificial mathematical coordinate without physical interpretation, yet one that was useful for explaining stellar aberration (Lorentz 1895, p. 77). This qualitative remark is formalized by assigning t' the type $\texttt{art}$ (artificial construct):

$$t' : \texttt{art} \tag{7}$$

Step 2. Now, a Lorentz transformation applied to stellar aberration yields a light waveform in the Earth's rest frame:

$$f(\omega t - ky) \;\rightarrow\; f\Big(\omega\Big(t - \frac{ux}{c^2}\Big) - ky\Big) \tag{8}$$

Here, Einstein might have wondered: what does the artificial nature of local time imply for the transformed wave? The simplest answer—and the one Einstein seems to have adopted—is that anything derived from an artificial construct is itself artificial. This reasoning can be formalized by applying the forward property preservation rule to t', which yields:

$$f\Big(\omega\Big(t - \frac{ux}{c^2}\Big) - ky\Big) : \texttt{art} \tag{9}$$

Step 3. Simplifying expression and using $k = \frac{\omega}{c}$, we have

$$f\Big(\omega t - k\frac{ux}{c} - ky\Big) : \texttt{art} \tag{10}$$

Step 4. Note that we can summarize the previous steps as follows: Under the assumption that local time is an artificial construct, the transformed waveform must also be artificial. Thus, under the context $\Gamma_1 = t' : \texttt{art}$, we have

$$\Gamma_1 \vdash f_{\text{art}} : \texttt{art} \tag{11}$$

where

$$f_{\text{art}} = f\Big(\omega t - kx\frac{u}{c} - ky\Big)$$

Step 5. On the other hand, Einstein recognized that, in the Earth's rest frame, the light wave deflected by stellar aberration $f(\omega t - kx\frac{u}{c} - ky)$ is an empirical observation that does not rely on theoretical assumptions. To formalize this reasoning, we state that, under the empty context $\Gamma_2 = \varnothing$, the deflected light wave has type $\texttt{emp}$ (empirical):

$$\Gamma_2 \vdash f_{\text{emp}} : \texttt{emp} \tag{12}$$

where

$$f_{\text{emp}} = f\Big(\omega t - k\frac{ux}{c} - ky\Big)$$

Step 6. Here, Einstein would have recognized that although f_{art} is artificial, it is also equal to f_{emp}, which is empirical. According to Norton's analysis, this suggests that, at the very least, f_{art} could not be solely artificial but must also possess an empirical aspect. Thus, within our formalism, Einstein would have concluded that f_{art} has both types: $\texttt{art}$ and $\texttt{emp}$. This reasoning process can be formalized using the concept change rule:

$$\frac{\Gamma_1 \vdash f_{\text{art}} : \texttt{art} \quad \Gamma_2 \vdash f_{\text{emp}} : \texttt{emp} \quad f_{\text{art}} = f_{\text{emp}}}{\Gamma_1 \cup \Gamma_2 \vdash f_{\text{art}} : \texttt{art} \cap \texttt{emp}} \tag{13}$$

where

$$\Gamma_1 = \{t' : \texttt{art}\} \quad \text{and} \quad \Gamma_2 = \varnothing$$

Step 7. At this point, according to Norton (Norton 2004), Einstein reasons that since his earlier conclusion is not merely artificial but also empirically true, he can now reinterpret it as an assumption. In terms of our formalism, Einstein assumes that f_{art} is of both types: $\texttt{art}$ and $\texttt{emp}$. As a result, he begins to derive the previous steps in reverse order, while cautiously maintaining that the same type ($\texttt{art} \cap \texttt{emp}$) continues to propagate backwards.

Eventually, by reasoning backwards, Einstein reaches a new version of his former assumption, now framed as a conclusion: local time t' has type $\texttt{art} \cap \texttt{emp}$. That is, local time could not be wholly artificial; it had to be, at least in part, empirically true. In other words, local time could be time itself, not merely an artifice. Einstein's reasoning process can be formalized using a backward property preservation rule as follows:

$$\frac{\Gamma_1 \cup \Gamma_2 \vdash f_{\text{art}} : \texttt{art} \cap \texttt{emp}}{\Gamma_1 \cup \Gamma_2 \vdash t' : \texttt{art} \cap \texttt{emp}} \tag{14}$$

Because local time depends on relative velocity, it was at this moment that Einstein realized time could be relative. That realization launched special relativity.

Step 8. Ultimately, Einstein's full discovery mechanism of switching Lorentz' assumption and conclusion can be formalized using the following rule:

$$\frac{\Gamma_1 \vdash t' : \texttt{art} \quad \Gamma_2 \vdash f_{\text{emp}} : \texttt{emp} \quad g(t') = f_{\text{emp}}}{\Gamma_1 \cup \Gamma_2 \vdash t' : \texttt{art} \cap \texttt{emp}} \tag{15}$$

That is, under assumptions from contexts Γ_1 and Γ_2, the concept of local time t' is assigned both types: `art` and `emp`. This rule builds compositionally upon the concept change and property preservation rules along with other simple mathematical operations.

The assumption-conclusion switch discovery mechanism can be stated in its most general form, without referring to the example's specific variable names, as follows:

$$\frac{\Gamma_1 \vdash x : A \quad \Gamma_2 \vdash y : B \quad g(x) = y}{\Gamma_1 \cup \Gamma_2 \vdash x : A \cap B} \tag{16}$$

where g denotes the function that transforms x into y.

Bridge from Type theory to Computation

To carry the type-theoretic formalism into a computational setting, we examine the three-way relationship among type theory, programming languages and scientific concepts we introduced earlier. We begin with the connection between type theory and programming languages. Specifically, we note that once a type system is equipped with fixed operational semantics (step-by-step evaluation rules), the system's terms can be viewed as programs and its theoretical types as programming-language types (Plotkin 1977).

Turning to the link between programming languages and scientific concepts, recent work in computational cognitive science holds that symbolic programs (code) offer the best formal representation of concepts because of their expressive power (Rule, Tenenbaum, and Piantadosi 2020). As noted in earlier, Sosa builds on this perspective by suggesting that types in programming languages may capture the cognitive constraints that impose structure in concepts and allow us to generate reasonable, though sometimes incorrect, answers (Sosa and Ullman 2022). Motivated by these three developments, we introduce two analogies—one theoretical and one computational—for formalizing scientific concept formation.

1. **Theoretical analogy:** Intuitive types are formalized as theoretical types; discovery mechanisms as typing rules; and scientific concepts as terms and the conclusions of typing rules.
2. **Computational analogy:** Intuitive types are formalized as types in a programming language; discovery mechanisms and scientific concepts as programs.

These analogies provide the basis for the Python implementation of the type-theoretic reconstruction of Einstein's conceptual path to the relativity of time.

Program synthesis

We recast the type-theoretic reconstruction of Einstein's route to relative time as a program-synthesis problem: the goal is to recover an eight-step sequence of admissible symbolic operations and typing rules that transforms Lorentz's initial judgment that time is artificial into Einstein's judgment that time is relative. We compare three search strategies over this space—enumerative, Bayesian, and Bayes–neural—and evaluate by success rate versus program index, averaged over $N = 150$ runs (Fig. 1). With search constraints (inductive biases) enabled (Fig. 1 (a)), Pure-Bayes performs best, Bayes–neural is next, and enumeration is worst; without these constraints (Fig. 1 (b)), Bayes–neural is markedly more sample-efficient, while Pure-Bayes and enumeration rarely succeed even with large program budgets.

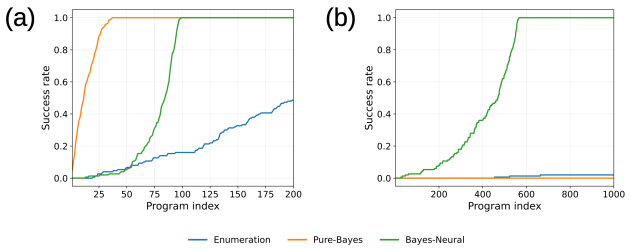


Figure 1: Success rate vs. program index for Enumeration (blue), Pure-Bayes (orange), and Bayes-Neural (green). (a) Biased: all heuristics on except "backward-only after target waveform". (b) Unbiased: all heuristics off.

Discussion and Conclusion

So how are scientific concepts birthed? This work does not offer a final answer, but argues that a central part of the answer lies in the intuitive types that constrain and guide scientific knowledge-making. Reconstructing Einstein's path to relative time makes these intuitive types explicit and yields a typed discovery mechanism—the assumption–conclusion switch. The program-synthesis results indicate that combining Bayesian guidance with neural methods can mitigate the curse of compositionality in conceptual search. Ultimately, a type-theoretic formalism of concept formation can serve both as tool AI (Aguirre 2025) and as a supportive framework for scientists. It may also inform pedagogy—not only by helping people learn existing scientific concepts, but by equipping them with the tools needed to discover and create new ones themselves.

References

Aguirre, A. 2025. Engineering the Future: What We Should Do Instead. In *Keep the Future Human: Why and How We Should Close the Gates to AGI and Superintelligence, and What We Should Build Instead*. 47–56.

Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; Gulcehre, C.; Song, F.; Ballard, A.; Gilmer, J.; Dahl, G. E.; Vaswani, A.; Allen, K.; Nash, C.; Langston, V. J.; Dyer, C.; Heess, N.; Wierstra,

D.; Kohli, P.; Botvinick, M.; Vinyals, O.; Li, Y.; and Pascanu, R. 2018. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261 [cs.LG].

Battleday, R. M., and Gershman, S. J. 2024. Artificial intelligence for science: The easy and hard problems. arXiv:2408.14508 [cs.AI].

Cardelli, L. 1996. Type systems. In Tucker, A. B., ed., *The CRC handbook of computer science and engineering*. CRC Press. 2208–2236.

Chomsky, N. 2006. *Language and Mind*. Cambridge University Press, 3rd edition.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2):155–170.

Lalande, F.; Matsubara, Y.; Chiba, N.; Taniai, T.; Igarashi, R.; and Ushiku, Y. 2023. A Transformer Model for Symbolic Regression towards Scientific Discovery. arXiv 2312.04070 [cs.LG].

Liu, A.; Wu, Z.; Michael, J.; Suhr, A.; West, P.; Koller, A.; Swayamdipta, S.; Smith, N. A.; and Choi, Y. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 790–807. Association for Computational Linguistics.

Lorentz, H. A. 1895. *Attempt of a Theory of Electrical and Optical Phenomena in Moving Bodies*. E. J. Brill.

Maxwell, J. C. 1865. On Physical Lines of Force. In Niven, W. D., ed., *The Scientific Papers of James Clerk Maxwell*, volume 1. Dover Publications. 155–229.

Nersessian, N. J. 1987. A Cognitive-Historical Approach to Meaning in Scientific Theories. In *The Process of Science: Contemporary Philosophical Approaches to Understanding Scientific Practice*, volume 3 of *Science and Philosophy*. Martinus Nijhoff Publishers. 161–177.

Nersessian, N. J. 2008. *Creating scientific concepts*. MIT Press.

Norton, J. D. 2004. Einstein's investigations of Galilean covariant electrodynamics prior to 1905. *Archive for History of Exact Sciences* 59(1):45–105.

Pierce, B. C. 2002. *Types and Programming Languages*. MIT Press.

Plotkin, G. D. 1977. LCF Considered as a Programming Language. *Theoretical Computer Science* 5:223–255.

Rule, J. S.; Tenenbaum, J. B.; and Piantadosi, S. T. 2020. The Child as Hacker. *Trends in Cognitive Sciences* 24(11):900–915.

Sosa, F. A., and Ullman, T. D. 2022. Type Theory in Human-Like Learning and Inference. arXiv:2210.01634 [cs.AI].

Udrescu, S.-M.; Tan, A.; Feng, J.; Neto, O.; Wu, T.; and Tegmark, M. 2020. AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. In *Advances in Neural Information Processing Systems*, volume 33, 9140–9150.

Wolfram, S. 2024. Can AI Solve Science? Stephen Wolfram Writings.

From 31 Functions to 14 Categories: Grouping Propp's Narrative Structure for Co-Creative Writing Tools

Aman Alapati
School of Computing
University of Portsmouth
United Kingdom
aman.alapati@myport.ac.uk

Argenis Ramirez Gomez
School of Creative Technologies
University of Portsmouth
United Kingdom
argenis.ramirez-gomez@port.ac.uk

Ella Haig
School of Computing
University of Portsmouth
United Kingdom
ella.haig@port.ac.uk

Alaa Mohasseb
School of Computing
University of Portsmouth
United Kingdom
alaa.mohasseb@port.ac.uk

Abstract

Propp's 31 narrative functions have traditionally been used in computational storytelling as underlying structural mechanisms rather than as tools directly accessible to users. This paper investigates an alternative approach by examining the usability of these functions when presented explicitly to novice writers engaged in creating branching narratives. Initial observations indicate that the full set of 31 functions introduces a level of complexity that can hinder accessibility and creative flow for beginners. To address this, the study proposes a reduced framework consisting of 14 broader narrative categories, designed for integration into a co-creative authoring interface aimed at game design students and emerging writers. The reduction was conducted through a systematic five-step process in which the original functions were treated as a dataset. Each function was coded on four structural dimensions, followed by clustering and iterative refinement to ensure coherence and usability. The paper outlines this methodological process, presents the resulting 14-category framework, and proposes a simplified and structured representation of Propp's model that aims to better support novice users by reducing perceived cognitive complexity while maintaining narrative expressiveness. An empirical evaluation of the framework is planned to further assess its effectiveness.

Introduction

Propp's *Morphology of the Folktale* (Propp, 1968) remains a foundational work in structural narrative analysis. Its formulation of 31 narrative functions provides a formalized account of the underlying structure of Russian folktales, offering a level of precision that has proven valuable in computational contexts. These functions have been widely adopted in applications such as automated story generation (Gervás, 2013; Riedl and Young, 2010) and narrative annotation and machine learning (Finlayson, 2016). However, in such systems, Propp functions typically operate as internal mechanisms, forming part of the underlying logic rather than elements directly engaged by users. Within computational creativity research, this raises broader questions about how narrative structure can function not only as a generation mechanism but also as a collaborative scaffold shaping creative decision-making in co-creative systems. This work addresses the different implications of exposing Propp's morphology to users within a co-creative authoring environment.

Prior research suggests that this level of complexity may negatively affect creative engagement and usability. Cognitive load theory indicates that excessive structural demands can interfere with learning and creative processes (Sweller, 1988). Related challenges have been observed in earlier systems; for example, research done in Hammond (2011) found that even simplified, card-based representations of Propp's functions were difficult for children to use effectively. Although several scholars have proposed reduced interpretations of Propp's morphology, such as the consolidation into five analytical moves (Meletinskij et al., 1974) or the selection of eleven functions (Kafalenos, 1997), these approaches were not developed with interface design considerations in mind and do not provide comprehensive, methodologically grounded mappings.

In response, this paper presents a systematic reduction of Propp's 31 functions into 14 broader categories intended for use in user-facing systems. Unlike prior reductions of Propp's morphology, which were developed for analysis or automated generation, this work presents an exhaustive, methodologically documented reduction designed for user-facing, co-creative narrative tools. The proposed taxonomy is derived through a structured qualitative analysis and is designed to retain narrative expressiveness while aiming to improve accessibility and reduce perceived complexity. The paper details the methodological process underlying the categorisation and provides theoretical justification for the resulting groupings. More broadly, the work highlights the importance of structural granularity as a design variable in co-creative systems, raising the need for empirical investigation into how different levels of narrative abstraction affect user experience and creative outcomes.

This paper makes the following contributions:

- We present an exhaustive **reduction of Propp's 31 narrative functions** into 14 macro-categories, preserving structural relationships while reducing complexity.
- We describe a **structured qualitative methodology** for deriving user-facing narrative abstractions from formal narrative theory.
- We present a **user-facing narrative taxonomy** designed to support co-creative authoring tools, with a focus on usability and reducing cognitive demands on novice writers.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

- We frame **structural granularity as a key design variable** for co-creative narrative systems and explore how reduced structural granularity may support usability in user-facing authoring tools.

Background and Related work

Prior Reductions of Propp

Propp's 31 narrative functions were not conceived as fully independent; *Morphology of the Folktale* already groups them into phases and recurring pairings (e.g., struggle–victory, task–resolution) (Propp, 1968). This implicit structure has motivated later attempts at abstraction.

Several partial reductions exist. Meletinskij et al. (1974) compresses the model into five transformations of the hero's status, Kafalenos (1997) identifies eleven core functions, and Brusentsev, Hitchens, and Richards (2012) proposes three stages for video game narratives. While insightful, these works neither map all 31 functions nor provide systematic reduction methods, and are primarily analytical rather than interaction-focused.

In computational settings, Gervás (2013) groups functions by dependency for story generation, and Finlayson (2016) applies machine learning to derive functional clusters. These optimise system performance rather than user representation. Closer to this work, Hammond (2011) exposes Proppian functions in a creative interface, but retains all 31, noting resulting cognitive demands.

Scaffolding and Cognitive Load

Reducing visible narrative functions is supported by cognitive load theory, which emphasises limits on simultaneous information processing (Sweller, 1988). In branching narratives, managing multiple paths already taxes users, so excessive structural options can hinder creativity.

Scaffolding theory further suggests aligning complexity with user expertise (Vygotsky, 1978; Wood, Bruner, and Ross, 1976). Simplified structures enable initial engagement, with complexity introduced progressively. A reduced function set can thus act as an accessible entry point while preserving access to the full model.

Creativity support research similarly advocates balancing guidance and flexibility (Shneiderman, 2007; Frich et al., 2019). Despite this, no prior work applies these principles to presenting Propp's morphology directly to users. Existing reductions target analysis or generation, not interaction, and lack a complete, methodologically grounded mapping suitable for co-creative interfaces. This paper frames the level of abstraction in presenting narrative theory, termed *structural granularity*, as a key but underexplored design variable.

Methodology

The development of the 14-category framework was guided by a systematic and replicable analytical process rather than intuitive grouping. The aim was to establish a method that could be followed, evaluated, and, if necessary, challenged by other researchers. To achieve this, the study adopts a qualitative approach grounded in deductive thematic analysis (Braun and Clarke, 2006), treating Propp's 31 narrative functions as the primary dataset, and employs affinity mapping (Beyer and Holtzblatt, 1998) as the mechanism for clustering.

The analysis proceeded in five stages. The first stage involved close reading of the source material. Each of Propp's functions was examined in detail, including its formal definition, narrative role, and any structural relationships identified in the original text (Propp, 1968). The objective was to interpret each function in terms of its operational role within narrative progression, rather than as an isolated label.

The second stage consisted of systematic coding. Each function was annotated across four structural dimensions. The first dimension, *narrative phase*, classified functions according to their position within the overall narrative arc, such as initial, middle, or concluding sequences. The second, *causal direction*, distinguished between functions that initiate narrative developments and those that respond to preceding events. The third dimension, *agent of action*, identified the primary narrative driver, such as the hero, the villain, or broader structural processes related to return and resolution. The fourth dimension, *sequential dependency*, captured the extent to which a function depends on prior narrative elements for coherence, reflecting Propp's emphasis on ordered relationships between functions.

This coding process produced a structured profile for each function, forming the basis for the third stage, clustering through affinity mapping. Each function was treated as an individual unit and grouped with others sharing similar structural characteristics. Group formation was guided by commonalities across the four coded dimensions, including shared narrative phase, consistent causal direction, and identifiable dependency relationships. This process revealed several recurring patterns, including sequential pairings, action–reaction relationships, shared positional roles within the narrative, variations within broader narrative operations, and clusters associated with return and resolution phases.

The fourth stage involved validation of the emerging categories using established qualitative criteria. Following Patton's framework (Patton, 2002), categories were assessed for *internal homogeneity*, ensuring that functions grouped together exhibited coherent structural relationships, and *external heterogeneity*, ensuring clear distinctions between categories. Where functions appeared to span multiple groups, classification decisions were resolved by identifying the most structurally dominant characteristic. All coding decisions and justifications were documented to support transparency and reproducibility.

The final stage focused on category naming and representation. To enhance accessibility for novice users, technical terminology was replaced with more intuitive descriptors while preserving the underlying structural meaning. For example, Propp's "Liquidation of Misfortune or Lack" was reframed as "Resolution", and clusters associated with early-stage antagonist actions were consolidated under more descriptive labels. This step ensured that the resulting framework remained faithful to Propp's model while being more suitable for use in user-facing, co-creative environments.

The 14 Macro-Categories

The analytical process resulted in the identification of 14 macro-categories that collectively account for all 31 of Propp's narrative functions. Table 1 presents the complete mapping between the original functions and the derived categories. In the Appendix, Table 2 presents a more detailed rationale, illustrating how structurally related functions are consolidated while preserving narrative progression.

First, a subset of categories corresponds directly to individual functions. These include *Setup*, *Mediation*, *Departure*, and *Decision/Counteraction*. Each of these functions was retained as a distinct category because it represents a critical transition point within the narrative structure. Their roles as pivotal moments in story progression justify their separation, as collapsing them into broader groupings would obscure key decision points that are particularly salient for novice writers engaging with narrative construction.

In contrast, other categories consolidate multiple functions that serve closely related structural roles. For example, the *Villainy/Lack* category integrates a sequence of functions associated with the introduction and escalation of narrative conflict. These functions collectively establish the central problem that drives the story forward. Similarly, the *Recognition/Reward* category groups several functions from the concluding phase of the narrative, all of which contribute to the resolution of identity, restoration of order, and distribution of rewards.

A further pattern is observed in the treatment of structurally analogous functions that recur at different points in the narrative. The category *Test* encompasses the function traditionally labelled as the *Difficult Task*, while *Victory* incorporates the corresponding *Solution*. These functions were grouped with earlier donor-sequence elements because they replicate an established structural pattern rather than introducing a fundamentally new narrative operation. This decision reflects an emphasis on underlying functional similarity rather than strict adherence to Propp's original sequential distinctions.

Overall, the resulting categorisation preserves the structural integrity of Propp's morphology (Propp, 1968) while providing a more accessible and interpretable framework. By balancing granularity and usability, the macro-categories aim to better align with user-facing systems, particularly those supporting novice narrative design.

Rationale for the Proposed Reduction

The 14 macro-categories make explicit structural patterns already implicit in Propp's morphology rather than imposing a new framework. Propp identifies recurring sequences such as villain preparation (F4–F8), donor interaction (F12–F15), and return-phase complications (F20–F23) (Propp, 1968). This work systematically formalises such groupings across all 31 functions.

While prior studies propose higher-level abstractions (Kafalenos, 1997; Brusentsev, Hitchens, and Richards, 2012), this reduction is motivated by usability and interaction design, translating structural regularities into a coherent user-facing taxonomy. Further reduction beyond 14 categories resulted in a loss of interpretability, as narratively distinct operations became increasingly difficult to distinguish within broader groupings.

Reducing 31 functions to 14 categories simplifies the decision space in branching narrative design, where managing multiple paths and dependencies already imposes cognitive demands. Cognitive load theory suggests that excessive complexity hinders performance (Sweller, 1988). Grouping functions enables cognitive chunking, allowing users to operate at a higher level while retaining access to full detail. This supports scaffolding principles, where complexity is introduced progressively (Vygotsky, 1978; Wood, Bruner, and Ross, 1976).

To avoid constraining creativity, categories are defined at an abstract level (e.g., *Test*, *Aid*), indicating function without prescribing content. This aligns with creativity support research emphasising a balance between guidance and flexibility (Shneiderman, 2007; Frich et al., 2019). The taxonomy thus serves as a generative scaffold, aiding navigation of narrative structure while preserving creative freedom.

Discussion

The proposed 14-category taxonomy has been implemented within a functional prototype designed to support co-creative narrative authoring. In this system, the macro-categories are presented as structural prompts throughout the writing process, guiding users in constructing branching narratives. The interface also provides real-time, colour-coded feedback that maps user-generated content to the underlying Proppian functions (Propp, 1968), thereby maintaining a connection between the simplified representation and the full theoretical model. In addition, the system supports user-defined branching through conditional logic, potentially supporting interactive quest-based narratives. Initial deployment involved game design students in short-form narrative construction tasks.

The taxonomy is distinct from all prior reductions of Propp's morphology in three respects. First, the presented approach is comprehensive/exhaustive by accounting for every one of Propp's 31 functions and records an explicit rationale for each placement decision, rather than selecting a subset (Kafalenos, 1997), or collapsing the morphology (Meletinskij et al., 1974). Second, it is *interface-motivated*, whereas (Gervás, 2013) groups functions by dependency relationships to support automated generation, and (Finlayson, 2016) recovers clusters through machine learning for annotation purposes, the 14 categories are designed specifically to reduce the decision load placed on a human author interacting with the interface. Third, the derivation is *methodologically documented* through a transparent five-step process that other researchers can replicate or formally challenge, addressing a limitation shared by all prior reductions.

The work is closely related to (Hammond, 2011), who demonstrated that novice users can engage with Proppian structure through a digital interface but found even individual function cards cognitively demanding. The present research extends that line directly by providing a principled grouping layer that lowers visible complexity without

#	Propp Function	Macro-Category	#	Propp Function	Macro-Category
1	Absentation	Setup	17	Branding / Marking	Struggle
2	Interdiction	Interdiction	18	Victory	Victory
3	Violation	Violation	19	Liquidation	Resolution
4	Reconnaissance	Villainy / Lack	20	Return	Return
5	Delivery	Villainy / Lack	21	Pursuit	Return
6	Trickery	Villainy / Lack	22	Rescue	Return
7	Complicity	Villainy / Lack	23	Unrecognised Arrival	Return
8	Villainy or Lack	Villainy / Lack	24	Unfounded Claims	Recognition / Reward
9	Mediation	Mediation	25	Difficult Task	Test
10	Beginning Counteraction	Decision / Counteraction	26	Solution	Victory
11	Departure	Departure	27	Recognition	Recognition / Reward
12	First Function of Donor	Test	28	Exposure	Recognition / Reward
13	Hero's Reaction	Test	29	Transfiguration	Recognition / Reward
14	Receipt of Magical Agent	Aid	30	Punishment	Recognition / Reward
15	Guidance	Aid	31	Wedding / Reward	Recognition / Reward
16	Struggle	Struggle			

Table 1: Mapping of Propp's 31 narrative functions to 14 macro-categories. See Appendix for further details

discarding the underlying theory. Where Hammond's system exposed the full morphology, the 14-category interface abstracts it; where that work addressed linear stories, the present system targets branching narrative design, introducing an additional layer of cognitive complexity that further motivates the reduction.

Structural Granularity as a Design Variable

The introduction of a reduced structural framework highlights a broader design consideration within computational creativity systems, namely the role of structural granularity in shaping user experience. Although prior work has largely treated Propp's 31 functions as a fixed and complete set for computational purposes (Gervás, 2013; Riedl and Young, 2010), the present approach suggests that the level of abstraction at which narrative structure is presented constitutes a meaningful design variable. The taxonomy proposed here reframes the morphology as a layered, granularity-aware structure, the 14 categories are an interface level abstraction over a richer underlying model, with the original 31 functions retained and accessible. By making this variable explicit, the work contributes to ongoing discussions around how creative systems can balance formal structure with usability and accessibility, and aligns with research in creativity support tools, which emphasises the importance of designing systems that guide users without imposing excessive cognitive or procedural constraints (Shneiderman, 2007; Frich et al., 2019).

From the user's perspective, the proposed macro-categories shift narrative authoring from a fine-grained function selection task to a higher-level structural reasoning activity. Rather than choosing among dozens of closely related functions, novice writers can focus on broader narrative intentions such as introducing conflict, testing the protagonist, or resolving a storyline. This abstraction supports faster ideation and reduces hesitation at key decision points, particularly in branching narratives where structural choices must be replicated and maintained across multiple paths. In this sense, the reduction reframes Propp's morphology as a generative scaffold rather than a prescriptive checklist, providing guidance while preserving interpretive flexibility and creative autonomy.

Reducing structural granularity improves usability, but it can remove distinctions that are important for expert analysis. This design addresses that trade-off by keeping the full set of Proppian functions accessible beneath the interface, allowing users to engage with more complexity as their expertise develops. In this way, structural granularity is treated as flexible rather than fixed, and this approach could extend beyond Propp's model to other formal narrative frameworks used in creative systems.

Conclusion and Future Work

This paper introduced a 14-category macro-function taxonomy derived from Propp's 31 narrative functions through a structured five-step deductive thematic analysis. Unlike prior reductions aimed at analysis or automation, this work focuses on interface design, offering a reproducible method to simplify narrative structures for human use.

The contribution is threefold, reducing perceived cognitive complexity for novice writers, providing a replicable derivation method, and exploring structural granularity as a design variable for co-creative systems.

Future work will empirically evaluate the taxonomy by comparing the 14-category interface, the full 31-function model, and a baseline without structural support. Measures will include narrative quality, structural completeness, interaction patterns, and user confidence. Further research will explore adaptive interfaces that adjust structural detail dynamically and apply this reduction approach to other narrative frameworks.

References

Beyer, H., and Holtzblatt, K. 1998. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann.

Braun, V., and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2):77–101.

Brusentsev, A.; Hitchens, M.; and Richards, D. 2012. An investigation of vladimir propp's 31 functions and 8 broad character types and how they apply to the analysis of video games. In *Proceedings of the 8th Australasian Conference on Interactive Entertainment*, 1–10. ACM.

Finlayson, M. A. 2016. Propplearner: Deeply annotating a corpus of russian folktales to enable the machine learning of a russian formalist theory. *Digital Scholarship in the Humanities* 32(2):284–300.

Frich, J.; MacDonald Vermeulen, L.; Remy, C.; Biskjaer, M. M.; and Dalsgaard, P. 2019. Mapping the landscape of creativity support tools in hci. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Gervás, P. 2013. Propp's morphology of the folk tale as a grammar for generation. In *Proceedings of the Workshop on Computational Models of Narrative*. Dagstuhl Publishing.

Hammond, S. 2011. *Children's Story Authoring with Propp's Morphology*. Ph.D. Dissertation, University of Edinburgh.

Kafalenos, E. 1997. Functions after propp: Words to talk about how we read narrative. *Poetics Today* 18(4):469–494.

Meletinskij, E.; Nekljudov, S.; Novik, E.; and Segal, D. 1974. Problems of the structural analysis of fairytales. In *Soviet Structural Folkloristics*. The Hague: Mouton. 73–139.

Patton, M. Q. 2002. *Qualitative Research and Evaluation Methods*. Sage Publications, 3 edition.

Propp, V. 1968. *Morphology of the Folktale*. University of Texas Press, 2 edition. Original work published 1928.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39:217–268.

Shneiderman, B. 2007. Creativity support tools: Accelerating discovery and innovation. *Communications of the ACM* 50(12):20–32.

Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2):257–285.

Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wood, D.; Bruner, J. S.; and Ross, G. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* 17(2):89–100.

Appendix: Full Coding of Propp's 31 Functions

Table 2: Full coding schema for Propp's 31 narrative functions across structural dimensions, grouping principles, and macro-categories.

#	Propp Function	Narrative Phase	Causal Direction	Agent	Dependency	Grouping Principle	Macro-Category	Rationale for Placement
1	Absentation	Opening	Initiating	Situational	None	Shared phase	Setup	Establishes a disruption to the initial equilibrium, creating the conditions for narrative development.
2	Interdiction	Opening	Initiating	Authority	F1	Structural trigger	Interdiction	Introduces a prohibition that defines the constraints of the narrative and anticipates future conflict.
3	Violation	Opening	Initiating	Hero/Victim	F2	Sequential pair	Violation	Represents the breach of a prior rule, marking the transition from order to conflict.
4	Reconnaissance	Opening	Initiating	Villain	F1	Villain preparation	Villainy/Lack	Begins the antagonist's preparation through information gathering, contributing to the emergence of disruption.
5	Delivery	Opening	Reactive	Villain	F4	Villain preparation	Villainy/Lack	Provides the antagonist with necessary information, advancing the preparation phase.
6	Trickery	Opening	Initiating	Villain	F5	Villain preparation	Villainy/Lack	Introduces deception as a means of advancing the antagonist's strategy.
7	Complicity	Opening	Reactive	Victim	F6	Villain preparation	Villainy/Lack	Completes the deception sequence through the victim's unwitting cooperation.
8	Villainy or Lack	Opening	Initiating	Villain/System	F4–F7	Core disruption	Villainy/Lack	Establishes the central problem driving the narrative through harm or absence.
9	Mediation	Middle	Reactive	Dispatcher	F8	Transition phase	Mediation	Signals recognition of the problem and initiates the narrative response.
10	Counteraction	Middle	Initiating	Hero	F9	Action–response	Decision/Counteraction	Represents the decision to act, transforming awareness into intention.
11	Departure	Middle	Initiating	Hero	F10	Structural shift	Departure	Marks the transition from the ordinary world to active engagement with the problem.
12	Donor Test	Middle	Initiating	Donor	F11	Test sequence	Test	Introduces an evaluative challenge to assess the hero's suitability.
13	Hero Reaction	Middle	Reactive	Hero	F12	Test sequence	Test	Completes the evaluative interaction through the hero's response.
14	Magical Aid	Middle	Reactive	Donor	F13	Reward sequence	Aid	Provides resources or assistance following successful evaluation.
15	Guidance	Middle	Initiating	Helper	F14	Support phase	Aid	Facilitates progress toward the narrative objective through assistance.

Continued on next page

#	Propp Function	Narrative Phase	Causal Direction	Agent	Dependency	Grouping Principle	Macro-Category	Rationale for Placement
16	Struggle	Middle	Initiating	Hero/Villain	F15	Core conflict	Struggle	Represents direct confrontation between opposing forces.
17	Branding	Middle	Reactive	Hero	F16	Conflict variation	Struggle	Marks the hero symbolically during conflict, reinforcing identity within the confrontation phase.
18	Victory	Middle	Reactive	Hero	F16	Resolution of conflict	Victory	Resolves the central confrontation in favour of the protagonist.
19	Liquidation	Middle	Reactive	System	F18	Closure of disruption	Resolution	Eliminates the initial lack or harm, restoring narrative balance.
20	Return	Closing	Initiating	Hero	F19	Return phase	Return	Begins the movement back toward the original social context.
21	Pursuit	Closing	Initiating	Antagonist	F20	Return complication	Return	Introduces renewed threat during the return phase.
22	Rescue	Closing	Reactive	Hero/Helper	F21	Return resolution	Return	Resolves the pursuit, enabling continuation of the return.
23	Unrecognised Arrival	Closing	Initiating	Hero	F20–F22	Reintegration phase	Return	Places the hero in the social world without immediate recognition.
24	False Claims	Closing	Initiating	False Hero	F23	Legitimacy conflict	Recognition/Reward	Introduces contested identity, preparing the recognition sequence.
25	Difficult Task	Closing	Initiating	Authority	F23	Test variation	Test	Reintroduces evaluation through a final challenge.
26	Solution	Closing	Reactive	Hero	F25	Test resolution	Victory	Completes the challenge, reaffirming competence and success.
27	Recognition	Closing	Reactive	Authority	F26	Validation phase	Recognition/Reward	Confirms the hero's identity and achievements.
28	Exposure	Closing	Reactive	System	F27	Legitimacy sequence	Recognition/Reward	Reveals false claims, reinforcing narrative justice.
29	Transfiguration	Closing	Reactive	Hero	F27–F28	Status change	Recognition/Reward	Marks transformation in status or appearance following recognition.
30	Punishment	Closing	Reactive	Authority	F28	Moral closure	Recognition/Reward	Enacts consequences for wrongdoing, reinforcing narrative resolution.
31	Wedding/Reward	Closing	Reactive	Authority/Hero	F29–F30	Final closure	Recognition/Reward	Represents final restoration and reward, signalling narrative completion.

Recipes for Creativity: Iterative Generation and Evaluation in Large Language Models

Rens Anderson
LIACS
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Tessa Verhoef
LIACS
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Amirhossein (Miros) Zohrehvand
LIACS
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

Generative models are often evaluated through singular artifacts, whereas human creativity typically emerges through iterative generation, appraisal, and refinement. This pilot study examines whether iterative search improves LLM creativity by adapting FunSearch to recipe generation for the 2024 Pillsbury Bake-Off and evaluating outputs against human benchmarks using TTCT-based LLM evaluation. Across two experiments, we test iteration count, generator temperature, and in-loop selection-scorer model size. Results show that iterative generation-selection can produce recipes with creativity scores comparable to human benchmarks, but additional iterations alone do not improve creativity. The in-loop evaluator matters most: a smaller selection scorer yields significantly higher scores across most TTCT dimensions, while temperature has limited effects except for originality. These findings suggest that evaluator design is a first-order design variable in subjective creative search.

Introduction

In recent years, Generative Artificial Intelligence (GenAI), and Large Language Models (LLMs) in particular, have gained significant attention for their ability to generate creative content across various domains (Zhou and Lee 2024; Dubé and Xu 2026; Haase and Hanel 2023; Hubert, Awa, and Zabelina 2024; Stevenson et al. 2022). Research in computational creativity with such models often focuses on individual creative artifacts as isolated system outputs, whereas human creative activity is rarely characterized by such discreteness. In practice, creative output frequently emerges through an iterative process of generation, evaluation, and refinement (Sawyer 2021; Nath, Dayan, and Stevenson 2024).

In this study, we investigate LLM creative capabilities through iterative generation strategies. We leverage the FunSearch algorithm (Romera-Paredes et al. 2024), which has shown strong performance in objective tasks such as solving mathematical problems. FunSearch is an evolutionary search procedure in which an LLM is repeatedly prompted with high-scoring prior candidates from a programs database, partitioned into semi-isolated islands for diversity, and proposes new candidates that an automatic scorer admits or discards, biasing generation toward better-scoring outputs without fine-tuning. FunSearch is especially relevant to computational creativity because it searches over generative procedures rather than isolated outputs, making it suitable for domains where reusable, inspectable procedures matter. We extend this line of inquiry by applying it to a creative problem where output quality is inherently subjective: developing new cooking recipes for the Pillsbury Bake-Off. We particularly focus on model size (Taylor et al. 2024) and sampling temperature (Peeperkorn et al. 2024). **RQ1** asks whether iterative FunSearch improves on a near one-shot baseline and approaches a human benchmark. **RQ2** asks which of three design factors most strongly shapes creativity scores: iteration count, generator temperature, or in-loop selection-scorer size. We conduct two experiments to address these questions.

This paper makes two main contributions to the literature on generative AI for creative artifact production. First, it extends FunSearch, originally developed for mathematical and other codified objective tasks, to a subjective creative domain by applying it to recipe generation for the Pillsbury Bake-Off. This provides a concrete computational-creativity test case for studying iterative generation-selection loops beyond standard optimization benchmarks. Second, it provides an empirical evaluation of which parts of that iterative pipeline matter most for creative performance. By comparing iteration count, generator temperature, and in-loop selection-scorer model size, we show that selection pressure exerted by the evaluator is more consequential than simply increasing the number of search cycles or changing sampling temperature.

Background

Computational creativity has considered creativity not as a property of an isolated artifact, but as a relation between artifact, generator, and evaluative context. For example, frameworks such as DIFI locate creativity in the interaction among domain, individual, and field (Csikszentmihalyi 1988; Saunders 2019; Boden 2004). A FunSearch-style loop naturally distributes these roles: the generator proposes candidates, the programs database preserves lineages, and the in-loop scorer acts as a gatekeeping field.

FunSearch was introduced for mathematical and other codified tasks where progress can be measured against objective criteria (Romera-Paredes et al. 2024). Its design links it to evolutionary computation, where semi-isolated

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

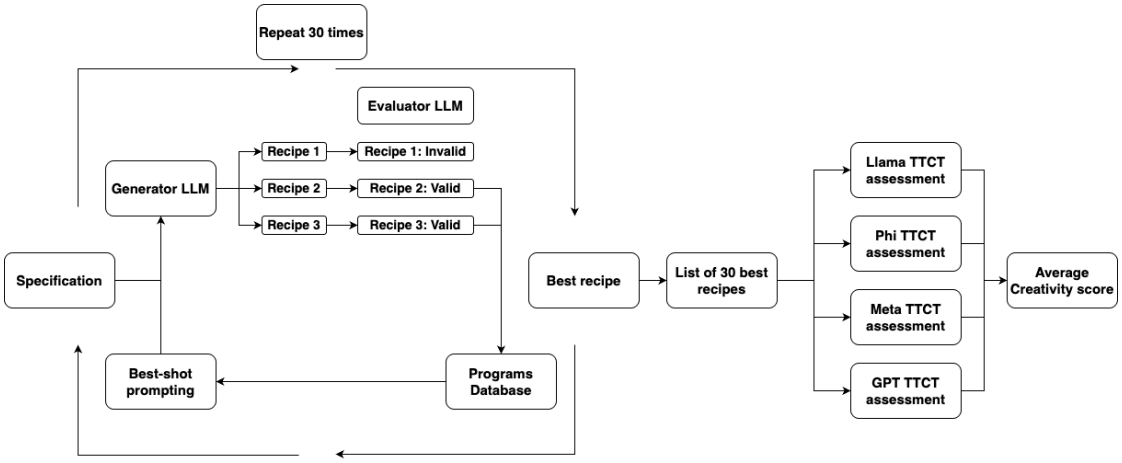


Figure 1: Conceptual framework illustrating how the FunSearch algorithm is integrated with TTCT-derived evaluation to assess the creativity of generated recipe artifacts.

subpopulations help preserve diversity and prevent premature convergence (Harada and Alba 2020; Gong et al. 2015). This is appealing for computational creativity because variation and selective retention are central to both evolutionary views of creativity and iterative accounts of human creative work (Nath, Dayan, and Stevenson 2024; Perez Y Perez and Sharples 2001).

This approach differs from other iterative LLM creativity methods. Systems such as Self-Refine critique and revise a single draft (Madaan et al. 2023), while multi-critic and beam-style approaches improve outputs through comparative feedback and broader search over candidate continuations (Bae and Kim 2024; Franceschelli and Musolesi 2024; Goes et al. 2023). FunSearch is closer to the latter family: in our adaptation, it improves outputs by retaining and reusing strong candidates rather than editing one evolving recipe.

Evaluation is the hardest part of this transfer. Human judgment remains the gold standard for creative artifacts (Pandhare 2024), and computational-creativity research warns that evaluation methods must match the creativity claim being made (Lamb, Brown, and Clarke 2018). At the same time, recent work suggests that structured LLM-based evaluation can be useful for creativity-related tasks when carefully prompted, even though concerns remain about evaluator overlap and self-preference bias (Zhao et al. 2024; Al Rabeyah et al. 2025; Wataoka, Takahashi, and Ri 2024; Chen et al. 2025).

Study Design and Evaluation

We adapt FunSearch to the Pillsbury Bake-Off by defining a recipe "skeleton" from the competition rules and using a multi-island search process. Each island maintains a local candidate set, and later prompts are built from high-scoring recipes in that island using best-shot prompting (Romera-Paredes et al. 2024). The system therefore improves through guided re-generation from selected examples, rather than by directly editing a single draft.

The recipe skeleton operationalized the Bake-Off constraints directly. Candidate artifacts required a title, a bounded ingredient list, exactly one official Pillsbury ingredient, concise preparation instructions, a short preparation time, and an accompanying story. Each run also maintained multiple semi-isolated islands with local candidate pools rather than one global database. This supported diversity: islands could preserve different recipe lineages and explore different parts of the search space, even with the same model family throughout the pipeline. A full workflow diagram of this adapted pipeline is provided in Figure 1, showing how the generation loop, in-loop selection scorer, and post hoc TTCT-style evaluators fit together. The in-loop scorer applies the Pillsbury-style rubric during search to decide which candidates survive into future prompts. The four TTCT-style evaluators are applied only after generation, to the retained best recipe from each run, and are used only for the analyses reported below.

During generation, an in-loop *selection scorer* applies a Pillsbury-style rubric to rank candidate recipes. This search objective uses the competition's weighted logic of recipe plus accompanying story, because both were generated during search. Within each run, top candidates are retained according to this score across islands; when multiple recipes tie for the highest score, one is selected at random. Final creativity is assessed separately after generation by four LLM evaluators using TTCT-derived dimensions: fluency, flexibility, originality, and elaboration (Torrance 1966; Zhao et al. 2024). We follow prior work that adapts TTCT-style dimensions to products rather than treating them as direct psychometric measures of an agent's creative capacity (Chakrabarty et al. 2024).

Here, the TTCT dimensions are interpreted at the level of the recipe as a product. Fluency refers to the perceived richness of ideas within a single recipe, flexibility to the variety of conceptual or culinary perspectives it combines, originality to the novelty or uncommonness of the concept, and elaboration to the amount of detail and development in the

final artifact. We therefore measure how generated recipes are judged under a structured creativity lens, not the model's latent creative capacity.

The evaluation prompts were designed to stabilize judgment rather than stimulate generation. Evaluators provided a brief qualitative rationale before assigning a score, using a chain-of-thought style structure to separate explanation from final rating (Wei et al. 2022; Choudhury, Vanneste, and Zohrehvand 2024). Persona framing was held fixed as a prompt-design choice rather than varied experimentally. This keeps the claims tied to the three manipulated factors rather than to every prompt detail.

For partial calibration, we compare generated recipes with a human reference set: the 2024 Pillsbury Bake-Off winning recipe plus 29 randomly selected entries from the same competition. Because stories for the human entries were not public, the final benchmark uses only the recipe component. This gives us a useful reference set, but not a complete ground-truth measure of recipe creativity.

This creates an asymmetry between generation-time and comparison-time evaluation. During search, recipes and stories were jointly optimized because the Bake-Off rubric rewards both; during final comparison, only recipe content could be retained. The benchmark should therefore be read as partial calibration, not a fully matched contest replication. In practice, the benchmark check placed the human winning recipe near the top under both candidate scorers rather than at rank one, indicating some sensitivity to quality but also clear divergence from the official human outcome.

We report two experiments. Experiment 1 varies iteration count, comparing a one-iteration near one-shot baseline with 5, 15, and 30 iterations. Experiment 2 varies generator temperature and in-loop scorer size in a balanced 3×2 design. Temperature matters because prior work links it to novelty–coherence trade-offs in creative language generation (Peeperkorn et al. 2024). Scorer size matters because larger models often perform better on objective tasks, but that relationship may not transfer cleanly to subjective creative search. In the full implementation, the generator was based on Meta-17B, while in-loop scoring used either Meta-8B or Meta-17B. Experiment 2 fixed seven islands and a batch size of five so that differences could be attributed more directly to temperature and scorer size than to search breadth.

Before the main comparisons, we checked whether the model-based scorer could rank the human reference set in a plausible way. The benchmark did not reproduce the official competition outcome, but it provided a useful calibration point: the winning human recipe remained a strong performer, while the mismatch supports our caution that LLM evaluation is only a proxy for human judgment.

Results and Interpretation

Experiment 1 asks whether more search helps. Figure 2 shows a representative recipe retained after 30 iterations of FunSearch with the Meta-17B generator (the appendix provides the recipe skeleton, the in-loop rubric, the evaluation prompt structure). We find more search helps only weakly. Iteration counts of 5, 15, and 30 produced mean final creativity scores of 3.921, 3.835, and 3.927, respectively, compared with 3.638 for the Pillsbury reference set. This suggests that the iterative setup can reach human-comparable scores under the evaluation used here, but does not show a clear monotonic gain from adding cycles.

Ingredients:	Instructions:
1 package Pillsbury Original Crescent Rolls 1 cup cherry tomatoes 2 tablespoons Yuzu juice 1/4 cup crispy prosciutto 1/2 cup heavy cream 1/2 teaspoon matcha powder 1/4 cup grated parmesan cheese 1/4 teaspoon salt 1 tablespoon olive oil 1 clove garlic	1. Preheat oven to 375°F. Bake Pillsbury Crescent Rolls according to package instructions (8-10 minutes). 2. While rolls are baking, mix Yuzu juice, cherry tomatoes, garlic, and olive oil in a small saucepan. Cook over medium heat, stirring constantly, until tomatoes are caramelized and mixture is syrupy. 3. In a separate bowl, whip heavy cream with matcha powder until stiff peaks form. Fold in parmesan cheese. 4. Once rolls are done, let them cool for 2 minutes. Then, slice each roll in half horizontally. 5. Top with Yuzu-infused tomato sauce, crispy prosciutto, and a dollop of matcha-parmesan cream. 6. Serve immediately and enjoy!
Story:	
Growing up in a multicultural city, I was inspired by the vibrant flavors and traditions of Japanese and Italian cuisine. This recipe combines the flaky, buttery texture of Pillsbury Crescent Rolls with the spicy, citrusy essence of Yuzu and the bright, grassy flavor of matcha, creating a unique and captivating treat that celebrates my love for international cuisine.	

Figure 2: Example recipe generated by the LLM after 30 iterations.

Notes: Compared with the Pillsbury reference set, artifacts of this kind tend to combine more unusual flavor pairings (e.g. truffle with breakfast staples, prosciutto-wrapped asparagus) while remaining within the skeleton's structural constraints.

The one-iteration baseline sharpens this interpretation. Removing repeated search substantially lowered the in-loop weighted recipe score, from 4.71 to 4.00, while leaving the final creativity score nearly unchanged at roughly 4.1. This is unsurprising considering the in-loop evaluator uses other criteria and does not optimize for creativity.

Experiment 1 also found the FunSearch conditions had larger standard deviations than the human reference set on average (e.g. SD $= 0.411$ at 30 iterations versus 0.328 for Pillsbury). One explanation could be that search explores a broader and less uniform solution space than the benchmark recipes through this iterative process.

Experiment 2 varies generator temperature and in-loop evaluator size. As presented in Figure 3, the smaller 8B selection scorer yields higher scores than the 17B scorer across average creativity, fluency, flexibility, and elaboration. One possible explanation is that in subjective search, larger evaluators may not be better aligned with the kinds of novelty and variation rewarded by the final assessment framework. Further, Figure 3 shows temperature effects are weaker. Lower generator temperature reduces originality, consistent with the intuition that less randomness leads to more predictable outputs (Peeperkorn et al. 2024). Higher temperatures, however, do not significantly improve any TTCT dimension. Sampling diversity alone therefore does not guarantee stronger creative artifacts, especially when later stages privilege coherence, structure, and in-loop scoreability.

We also conduct regression analyses to explore this effect further (for full results please see the appendix Table 1).

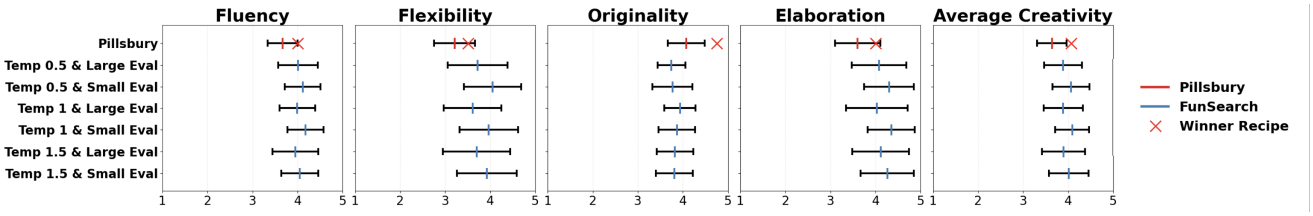


Figure 3: Average creativity scores across TTCT dimensions for FunSearch-generated recipes under each experimental condition, compared to the Pillsbury Bake-Off baseline.

Notes: Scores range from 1 (low) to 5 (high). Each row represents a configuration defined by generator temperature (Temp: 0.5, 1.0, or 1.5) and selection-scorer model (Large = $Meta-17B$, Small = $Meta-8B$). The red cross marks the winning Pillsbury recipe. Error bars indicate one standard deviation.

Relative to the 1.5 temperature and Meta-8B reference condition, the larger in-loop scorer showed significant negative associations with overall creativity, fluency, flexibility, and elaboration, while originality was unaffected. Temperature effects were mostly negligible apart from the drop in originality at the lowest temperature. Adjusted R^2 values were low across models, meaning that temperature and scorer size capture only part of what drives final judgments. Other factors may include the specific examples retained within islands, prompt framing effects, and evaluator heterogeneity.

Taken together, the two experiments indicate that evaluator design exerts stronger influence on final scores than either added search depth or added stochasticity alone. From a computational-creativity perspective, this shifts the question from whether an LLM can generate many novel candidates to what kind of design makes iterative novelty accumulate in subjective settings. If the in-loop evaluator rewards artifacts that are well formed but conservative, repeated search may stabilize around competent but unsurprising recipes. A looser evaluator might broaden exploration at the cost of coherence and practical plausibility. Future systems should therefore vary not only temperature or iteration count, but also the diversity, architecture, and incentive structure of evaluators inside the loop. That design space is especially relevant for subjective creative domains where no single oracle exists.

Limitations and Future Work

Our setup uses the FunSearch algorithm, a program-search framework that has previously been applied only to objective tasks, to generate recipes for the Pillsbury Bake-Off. Within the algorithm, the loop behaves less like direct refinement and more like selective re-generation. This may explain why iteration count alone has little effect on final creativity scores. If each step samples a new candidate from high-scoring examples instead of revising one draft, search may stay near already acceptable recipes rather than showing a clear trajectory of creative improvement.

We believe that recipe generation is a useful testbed, though, for computational-creativity research on iterative search because it is neither fully open nor fully objective. A successful recipe must be novel, but also coherent, plausible, and attractive as an artifact meant to be cooked and judged. This makes it a bridge case between divergent tasks, where quantity and novelty can dominate, and objective tasks, where correctness can dominate.

The main limitation is that the study relies on LLM-based evaluation at both the in-loop and final-analysis stages. Using four evaluators reduces dependence on any one model's idiosyncrasies, but does not solve alignment with human judgment. The benchmark check showed only partial agreement with the human competition outcome, and that agreement was not uniform across dimensions: the human winning recipe, for example, received a higher originality assessment than the generated recipes. Accordingly, this paper can say more about how outputs behave under LLM evaluation than about how people would judge them as creative culinary artifacts.

Further, Meta-17B also served as both the in-loop scorer and as one of the final evaluators, introducing methodological overlap. Averaging across four models softens but does not remove this dependence between optimization and outcome assessment. In addition, TTCT-style LLM scoring is also a partial proxy for culinary creativity. Future work should combine human evaluation, evaluator diversity, and trajectory-level analysis of how recipes change across iterations.

Conclusion

This short paper presents an initial computational-creativity study of FunSearch in a subjective artifact domain. The results suggest that iterative search can produce recipes competitive with a human reference set under LLM-based evaluation, but that evaluator design matters more than additional search depth, and that lower temperature mainly reduces originality. The broader contribution is a case study in what changes when iterative search moves from objective optimization to subjective creative evaluation. The core methodological claim is that, in subjective creative search, the in-loop evaluator should be chosen very carefully and play an important role in the optimization of iterated generation set-ups. Future creative LLM systems may benefit less from "more generation" alone than from more deliberate design of the evaluative ecology: the diversity, architecture, and incentive structure of the scorers that decide which candidates survive.

References

Al Rabeyah, A.; Góes, F.; Volpe, M.; and Medeiros, T. 2025. Do LLMs agree on the creativity evaluation of alternative uses? In *Proceedings of the International Conference on Computational Creativity (ICCC'25)*.

Bae, M., and Kim, H. 2024. Collective critics for creative story generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 18784–18819.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.

Chakrabarty, T.; Laban, P.; Agarwal, D.; Muresan, S.; and Wu, S. 2024. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Chen, W.-L.; Wei, Z.; Zhu, X.; Feng, S.; and Meng, Y. 2025. Do LLM evaluators prefer themselves for a reason? *arXiv*.

Choudhury, P.; Vanneste, B.; and Zohrehvand, A. 2024. The Wade test: Generative AI and CEO communication. *Harvard Business School Working Paper*.

Csikszentmihalyi, M. 1988. Society, culture, and person: A systems view of creativity. In Sternberg, R. J., ed., *The Nature of Creativity: Contemporary Psychological Perspectives*. Cambridge University Press. 325–339.

Dubé, J., and Xu, A. 2026. Large language models and creative content design: A case study of email marketing at wine access. *Quantitative Marketing and Economics*.

Franceschelli, G., and Musolesi, M. 2024. Creative beam search: LLM-as-a-Judge for improving response generation. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24)*.

Goes, L. F.; Volpe, M.; Sawicki, P.; Grzes, M.; and Watson, J. 2023. Pushing GPT's creativity to its limits: Alternative uses and torrance tests. In *Proceedings of the International Conference on Computational Creativity (ICCC'23)*.

Gong, Y.-J.; Tang, K.; Li, J.-Y.; and Yao, X. 2015. Distributed evolutionary algorithms and their models: A survey of the state-of-the-art. *Applied Soft Computing* 34:286–300.

Haase, J., and Hanel, P. H. P. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity* 33(3):100066.

Harada, T., and Alba, E. 2020. Parallel genetic algorithms: A useful survey. *ACM Computing Surveys* 53(4).

Hubert, K.; Awa, K.; and Zabelina, D. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14:3440.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys* 51(2):28:1–28:34.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv*.

Nath, S. S.; Dayan, P.; and Stevenson, C. 2024. Characterising the creative process in humans and large language models. In *Proceedings of the International Conference on Computational Creativity (ICCC'24)*.

Pandhare, H. 2024. Evaluating large language models: Frameworks and methodologies for AI/ML system testing. *International Journal of Scientific Research and Management*.

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the International Conference on Computational Creativity (ICCC'24)*.

Perez Y Perez, R., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence* 13(2):119–139.

Romera-Paredes, B.; Barekatain, M.; Novikov, A.; Balog, M.; Kumar, M.; Dupont, E.; Ruiz, F.; Ellenberg, J.; Wang, P.; Fawzi, O.; Kohli, P.; and Fawzi, A. 2024. Mathematical discoveries from program search with large language models. *Nature*.

Saunders, R. 2019. Multi-agent based models of social creativity. In *Readings in Computational Creativity*. Springer.

Sawyer, R. 2021. The iterative and improvisational nature of the creative process. *Journal of Creativity* 31:100002.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the International Conference on Computational Creativity (ICCC'22)*.

Taylor, N.; Ghose, U.; Rohanian, O.; Nouriborji, M.; Kormilitzin, A.; Clifton, D.; and Nevado-Holgado, A. 2024. Efficiency at scale: Investigating the performance of diminutive language models in clinical tasks. *Artificial Intelligence in Medicine*.

Torrance, E. P. 1966. *Torrance Tests of Creative Thinking: Norms-Technical Manual*. Princeton, NJ: Personnel Press.

Wataoka, K.; Takahashi, T.; and Ri, R. 2024. Self-preference bias in LLM-as-a-judge. *arXiv*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zhao, Y.; Zhang, R.; Li, W.; Huang, D.; Guo, J.; Peng, S.; Hao, Y.; Wen, Y.; Hu, X.; Du, Z.; Guo, Q.; Li, L.; and Chen, Y. 2024. Assessing and understanding creativity in large language models. *arXiv*.

Zhou, E., and Lee, D. 2024. Generative artificial intelligence, human creativity, and art. *PNAS Nexus*.

Appendix

This appendix contains the conceptual pipeline, recipe skeleton, in-loop Pillsbury-style rubric, shared evaluation prompt structure, and full regression output from Experiment 2.

A. Recipe Skeleton

The recipe skeleton operationalizes the 2024 Pillsbury Bake-Off rules as structural constraints on every candidate artifact. A candidate is retained in its island's programs database only if it satisfies all of the following:

- A recipe title.
- At most ten ingredients, excluding staple items such as salt and pepper.
- Exactly one ingredient from the official Pillsbury ingredient list.
- A stated preparation time under 30 minutes.
- Preparation instructions of at most 2000 characters.
- An accompanying story of at most 500 characters.

Candidates that violate any constraint are discarded before scoring. The story is part of the artifact during generation because it enters the in-loop rubric; it is excluded from the final benchmark comparison because the public human reference set did not include matching stories.

B. In-Loop Selection Rubric

During search, each valid candidate is scored by the in-loop selection scorer using a weighted Pillsbury-style rubric: recipe content contributes 70% of the score and the accompanying story contributes 30%. Each sub-criterion is rated on a 1–5 Likert scale, and the weighted sum is rescaled to the 1–5 range.

- **Recipe component (70%):** taste, appearance, creativity, crowd appeal.
- **Story component (30%):** narrative connection to the recipe, expression of family values or traditions, expression of personal passion.

This rubric is used only inside the search loop, not for the final creativity analysis. It is distinct from the TTCT-derived evaluation in the results: the rubric proxies competition judging, whereas the TTCT dimensions proxy product-level creativity.

C. Evaluation Prompt Structure

Both the in-loop rubric prompts and the post hoc TTCT prompts share three fixed design elements, held constant across all conditions so that results can be attributed to the manipulated factors rather than prompt choices. Each prompt (i) assigns a persona (Pillsbury Bake-Off participant or competition judge, depending on the task), (ii) asks the model to produce a short qualitative rationale *before* assigning a numerical score, and (iii) constrains the final output to a fixed JSON schema so that scores can be parsed reliably without regex-brittle postprocessing. The schematic structure of every evaluation prompt is:

[Persona] You are a Pillsbury Bake-Off {participant | judge}.
[Task] Evaluate the following recipe artifact along {dimension_name}, defined as {product-level definition}.
[Reasoning] First write a short rationale grounded in the artifact's content.
[Scoring] Then return a score on a 1–5 Likert scale.
[Format] Return a single JSON object with fields `rationale` and `score`.

For the TTCT prompts, the product-level definitions are: *fluency* = perceived richness of ideas within the single recipe and its framing; *flexibility* = variety of conceptual or culinary perspectives combined in the recipe; *originality* = novelty or uncommonness of the recipe concept; *elaboration* = amount of concrete detail and development in the final artifact. Persona framing was held fixed rather than varied experimentally, so the paper makes no causal claim about its effect.

D. Experiment 2: Full Regression Output

Table 1 reports coefficient estimates from five dummy-coded OLS models predicting the overall creativity score and each TTCT dimension separately. Temperature 1.5 with the Meta-8B selection scorer is the reference condition, so coefficients are differences from that baseline. The single consistent effect is the negative coefficient on *LargeEvaluator* across four of five outcomes; temperature shows a significant effect only on originality, and only at the lowest temperature.

Table 1: Coefficient estimates from dummy-coded OLS models predicting creativity dimensions.

Variable	(1) Creativity	(2) Fluency	(3) Flexibility	(4) Originality	(5) Elaboration
LowTemp	-0.017	-0.021	0.096	-0.144*	0.002
	(0.079)	(0.079)	(0.124)	(0.072)	(0.111)
MidTemp	-0.037	-0.083	0.023	-0.083	-0.002
	(0.079)	(0.079)	(0.124)	(0.072)	(0.111)
LargeEvaluator	-0.161*	-0.128*	-0.307*	0.021	-0.231*
	(0.065)	(0.065)	(0.101)	(0.059)	(0.090)
Constant	4.070	4.147	3.941	3.890	4.303
	(0.065)	(0.065)	(0.101)	(0.059)	(0.090)
N	180	180	180	180	180
Adj R^2	0.019	0.012	0.037	0.006	0.019

Notes: The model is Creativity $= \alpha + \beta \cdot$ MidTemp $+ \gamma \cdot$ LowTemp $+ \delta \cdot$ LargeEvaluator $+ \varepsilon$, where temperature 1.5 and the $Meta\text{-}8B$ selection scorer form the reference condition. ***MidTemp*** equals 1 when the generator temperature is 1.0, and 0 otherwise. ***LowTemp*** equals 1 when the generator temperature is 0.5, and 0 otherwise. ***LargeEvaluator*** equals 1 if the in-loop selection scorer is $Meta\text{-}17B$, and 0 if it is $Meta\text{-}8B$. The outcome variable is one of the four divergent thinking scores (Fluency, Flexibility, Originality, and Elaboration) or their average (Creativity). Standard errors are in parentheses. $*\, p < 0.05$.

Creating and Evaluating Challenge in Generative Music Education Systems

Filippo Carnovalini [1], and **Geraint A. Wiggins**[1,2]
[1]Vrije Universiteit Brussel
Pleinlaan 9, 1050 Brussel, Belgium
[2]Queen Mary University of London
Mile End Road, London E1 4NS, UK
filippo.carnovalini@vub.be

Abstract

Providing personalized learning paths increases the effectiveness of education. In the context of music education, personalization is often limited by the use of standard curricula for all students, as preparing or selecting ad-hoc learning material is non trivial and expensive. Computationally Creative systems could potentially help create personalized material, but to do so it would be necessary for them to be able to evaluate internally the degree of difficulty of the generated music. Automatic difficulty estimation has been proposed by Music Information Retrieval Research, but those descriptions of difficulty are however absolute and not personalized. We believe instead that different students will find different pieces more or less challenging than others. We propose therefore to use information theory metrics to estimate the level of *challenge* based on prior experience of the learners, by employing Idyoms, a system for the cognitive modeling of music. In two experiments we show that the system does model differences in levels of preparation when facing difficult pieces, and that it can account for different learning paths leading to different levels of challenge based on the appropriateness of the pieces known to the student, showing this approach could enable a Creative system to estimate the perceived difficulty of newly generated pieces and therefore improve the personalization of music education.

Introduction

Providing a personalized learning experience can significantly increase the effectiveness of education (Bloom 1984). Music teaching is an activity that offers some personalization afforded by the presence of personal teachers during one-to-one lessons. However, the degree of personalization is limited by the use of standard teaching materials and fixed curricula of pieces to learn, that typically apply to all students in a school (or in some cases even fixed by regional or national level curricula) (Carnovalini, Roda, and Wiggins 2023; Carnovalini, Espírito Santo, and Wiggins 2025).

This work is funded by the European Union, HORIZON-MSCA-2022-PF-01 Project ID 101108690 (CALIOPE). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the REA can be held responsible for them.

We propose that Computational Creativity can help achieve a higher degree of personalization. We have witnessed notable progress in music generation systems, and we posit that such systems could be employed for the creation of personalized exercises. This represents an interesting intersection of requirements that is underexplored in Computational Creativity, as the system would need to create sheet music that is a) not too hard for the student's current skill level, but also b) challenging enough to have educational benefits, and finally c) pleasant to play and listen to. The sum of a) and b) can be described as laying in the Zone of Proximal Development (Chaiklin 2003), which can be seen as an ideal area of a conceptual space that, *for a specific student* has the ideal amount of difficulty to enable learning without triggering frustration. It is worth noting that no two learners are the same, and therefore the shape of this area will always be different even for two students in the same level or class, as one may find certain skills easier or harder than the other (for example, in the same class of beginner pianists one may be faster in playing melodies, but another may be better at playing chords in the left hand). The need for personalization requires the system to be able to create novel pieces each time, which of course must also show sufficient quality, suggesting that this is indeed a prime use case for Computational Creativity, whereas Generative AI is currently capable of creating music, but is almost impossible to steer to creating pieces that match these strict educational requirements (Carnovalini et al. 2026).

There has been research in Music Information Retrieval systems to estimate and/or classify the difficulty of a musical piece, given the sheet music (Ramoneda et al. 2022), which could ideally be used as an internal evaluation for a creative system. However, this work aims to estimate difficulty as an absolute value, i.e., describing how hard a piece is compared to other pieces, *for most people*. We argue that to obtain proper personalization, a system for evaluating difficulty should try and make a cognitive model of the learner. In this study, we explore a system for the estimation of *challenge*, defined as *how difficult a certain piece will be to a certain person at a certain stage in their development* (or in other words, how much they will struggle to learn it). Exploring how effective such a measure is will enable further exploration in a creative system that generates music taking *challenge* into account.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

A Cognitive Model

Idyoms

For the current investigation, we utilize Idyoms[1], a Julia implementation of IDyOM (Information Dynamics of Music), a cognitive model developed by Marcus Pearce (Pearce 2005) for the statistical modeling of the perception of melodic sequences. Idyoms is able to extract statistical patterns based on example sequences by creating a variable-order Markov model (Conklin and Witten 1995) and is then able to estimate the likelihood of unseen sequences by means of the Prediction by Partial Matching (PPM) algorithm (Cleary and Witten 1984), which combines the likelihoods of the different orders. The system also allows the definition of different *viewpoints*, i.e., different feature dimensions to be used for the estimates (e.g., pitch and duration of notes). The model consists of two components, which may be used separately or together. The Short-Term Model (STM) bases its prediction only on elements that are within the current sequence (i.e., within the same song/piece), incrementally updating the statistics as the sequence is being processed. The Long-Term Model (LTM), in contrast, bases its statistics on a corpus of previously-analyzed sequences, offering the option to update statistics as the new sequence is being processed or to use just the prior knowledge. In both cases, the model allows computation and analysis of the Information Content of each event in a sequence, which is an indication of how unexpected that event is based on the elements that precede it. It is computed as follows:

$$h(x) = -log_2(p(x)) \quad (1)$$

where $p(x)$ is the probability of the event according to PPM. Events with low probability will have high information content and vice versa. Information content has been shown to model aspects of human perception and cognition of musical sequences very closely (Pearce 2005; Pearce and Wiggins 2006; 2012; Hansen and Pearce 2014).

Information Content, Difficulty, and Struggle

Using Information Content or the related metric, Entropy, for difficulty estimation is an established idea (Ramoneda et al. 2024; Sébastien et al. 2012; Verstraeten 2025). These metrics can give an indication of how structurally complex and varied a piece of music is, and these aspects are related to the perceived difficulty in performing that piece. However, research that has used these measures has typically done so by computing the metrics only within a single piece (equivalent to Idyoms' STM approach) and/or using only zeroth-order probability estimations, which therefore do not capture the dynamic structure of the music. Both these two limitations are reasonable for the goal of describing the variance of the piece's elements (e.g., how spread are the pitches, or how varied are the rhythmic figures), but fall short when it comes to describe how difficult a piece will be to a certain learner based on their learning path, i.e., what we call *challenge*.

[1] https://github.com/nick-harley/Idyoms

In order to model this personalized feature, we investigate the following approach: we expose the *Long*-term model of Idyoms to pieces that have already been studied by the learner, and compute the (average) Information Content of the new piece that is to be studied. As mentioned above, Information Content is correlated with surprise and unexpectedness. Our rationale is that, by modeling the unexpectedness of the new piece, we also model how much the learner will struggle to learn new sequences and techniques, that are not already part of their musical and technical vocabulary, which is roughly modeled by the LTM.

Idyoms has been shown effective as a model of the perception of music. Here, however, we are using it as a model of performance. We believe that the overall statistical modeling that Idyoms is based upon will also work as a model for the cognitive load that a musician must process when performing a new piece. While the present experiments are some of the first to test this hypothesis, we posit that expecting this well-established cognitive model of perception to highly correlate with performance is not an unreasonable assumption. The results of the following experiments corroborate this hypothesis, although at this stage it remains to be tested in future work with human participants.

Experiments

Using Information Content to Model Progression

Method. Our first experiment[2] is intended to assess the viability of general ideas outlined in the previous section. We model a student learning through a progressive exercise book, *Mikrokosmos*, a well-known piano book written by Béla Bartók (Bartók 1987), which contains pieces of progressive difficulty going from beginner to piano virtuoso. The book is available in digital format in the Mikrokosmos-Difficulty Dataset (MKD)[3], where the pieces are divided into three difficulty categories: *easy*, *medium*, and *hard*. Our experiment works as follows: we trained the Long-term model (LTM) of Idyoms with only the *easy* category and the second using both the *easy* and *medium* subsets of the dataset using the combination of Pitch and duration of the note events as the data representation[4]. We only used the right hand information in this experiment, in order to keep the alphabet size limited. We then predict the information content of the hard pieces based on the two learned models. The idea is to try to model a student that tried to approach the hard pieces only studying the easy pieces, and one who went through both easy and medium pieces before approaching the hard ones. The hypothesis is that the level of challenge (and therefore the average Information Content) is significantly lower for the latter, as the student we model has a better preparation for the hard pieces.

Results and Discussion. Figure 1 shows box plots detailing the distribution of the mean information content of the

[2] The code for both experiments is available at: www.github.com/project-caliope/caliope_experiments

[3] https://zenodo.org/records/6092709

[4] *Viewpoint* in IDyOM and Idyoms terminology (Conklin and Witten 1995; Pearce 2005).

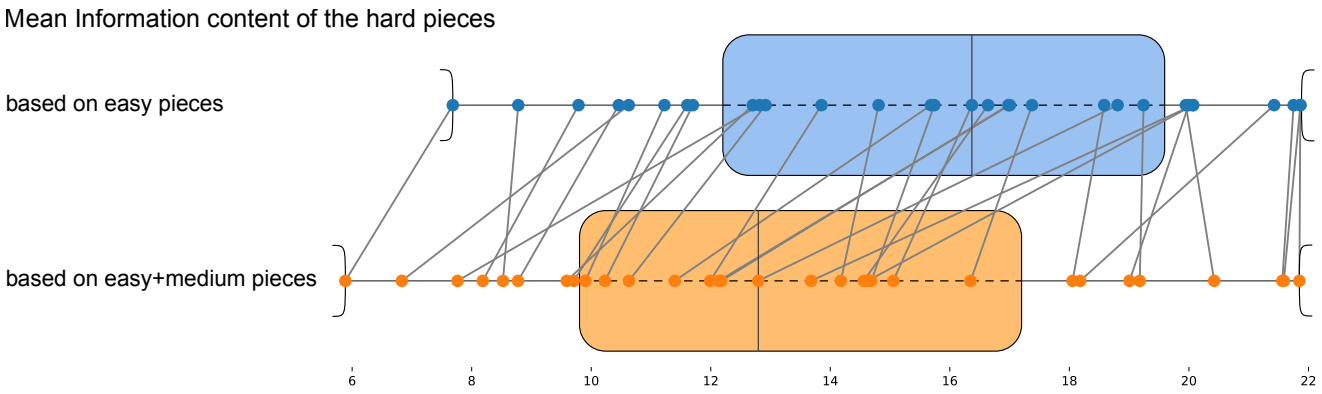


Figure 1: Paired box plot of the mean information content of hard pieces predicted by an Idyoms model trained on easy pieces and one trained on easy and medium pieces, with lines pairing values for the single pieces.

hard pieces, as predicted by the two models. The distribution is almost equally spread over a large range, but what is more interesting than the distribution itself is to see how the Mean Information Content changed for each piece based the model having been exposed to solely easy pieces or easy and medium pieces. To test that, we performed a paired nonparametric test (the Wilcoxon signed-rank test, graphically represented as the dots and lines in Figure 1), which confirmed the differences are significant ($T = 6.0$; $p = 1.3\text{e-}8(^{***})$; two-sided alternative rejected).

This first result indicates that Information Content can indeed model different levels of preparation, suggesting that it is indeed possible to give a personalized account of difficulty, or challenge. It is however possible to argue that this experiment offered little personalization, as both models came from the same dataset and have a non negligible overlap, and the results are only the effect of having more data in the easy+medium category. We address these reasonable points in the following section.

Personalization of Learning Paths

Method. In order to test how much our model can account for different learning paths, we performed a second experiment. We imagine two learners wanting to approach the same difficult piano piece, Alban Berg's *Piano Sonata* (Op. 1). This sonata is a challenging piece that does not rely on classical tonal theory, but is instead partly inspired by Schoenberg's (of whom Berg was a composition student) serialist ideas. The first student we envision goes through a more classical curriculum: we model this student with an Idyoms long-term model trained on the entirety of the Mikrocosmos dataset. The second, on the other hand, decides to train on fewer pieces, but ones which are more relevant to Berg's sonata. We collected a small dataset of 11 serialist pieces by Arnold Schoenberg and Anton Webern, on which we train a different instance of Idyoms' LTM. The first student has seen more pieces (some of which require high skill to master), but our hypothesis is that the second student, being more used to the relevant style of music, will struggle less on the piece, resulting in lower Information Content.

Results and Discussion In the previous experiment we compared the *Mean* Information Content of various pieces. Since we only analyze one piece in this second experiment, we compare the distribution of the Information Content of each event in the sonata, as predicted by the two models. Figure 2 shows the two distributions as violin plots. A Kruskal–Wallis H test confirms that the distributions are significantly different ($H = 15.82$; $p = 0.00007(^{***})$).

This result confirms that our model can account for different learning paths in a personalized manner, and that it is not merely an indication of having more data in the learned model. While in general adding more data to a model is expected to lower the Information Content, the relevance of the learned pieces is more important than the quantity of data. This is the expected behavior for the model: students who study more will generally struggle less, but the pieces they studied must be coherent with the piece they want to prepare to lower its challenge level.

Conclusions

We presented a problem for which Computational Creativity can offer help, namely the creation of personalized music exercises, and an approach to the estimation of *challenge*, defined as how difficult a certain piece will be to a certain person at a certain stage in their development, a key feature of such a creative system. This concept can become part of the internal evaluation for a Computational Creativity system creating music meant for learners, but could also help in recommending existing study material or automatically constructing personalized learning paths. Through two experiments we demonstrate that this approach is promising in providing a more personalized view on difficulty than what previously proposed.

There are however some limitations of the current study, which call for further work. Firstly, we have used a measure (Information Content) that is known to be significant for perception, but has not been validated for performance. The results suggest that indeed it could be appropriate for modeling performance cognition as well, but this remains to be tested.

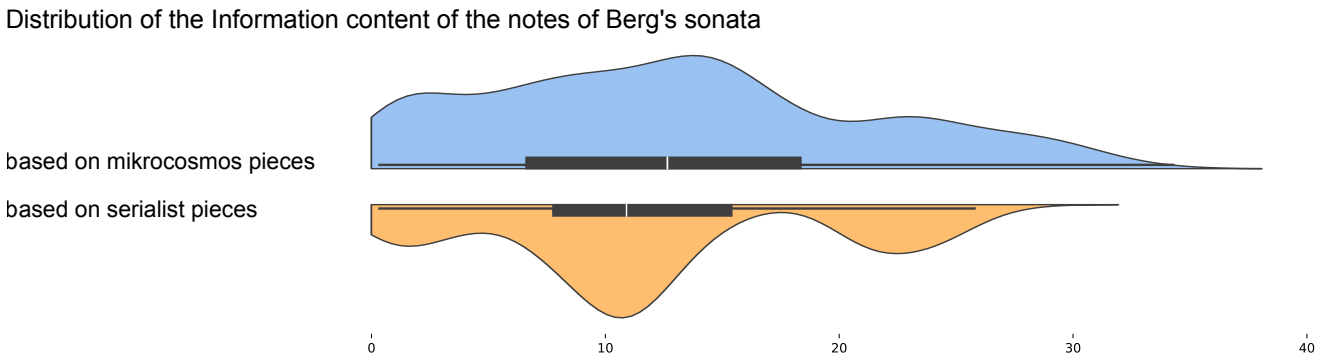


Figure 2: The distribution of the information content of all Berg's sonata's notes, as computed by a Idyoms model trained on all the MKD dataset and one trained on just serialist music.

The experiments could be repeated with more data and by using the information related to both hands, and possibly including additional information in our viewpoints such as fingering annotations or dynamics. It would also be interesting to model more closely existing piano curricula from music schools, to see how well this approach can relate to real-world scenarios, and finally, once a higher degree of maturity is reached, it would need to be validated with actual learners. Moreover, while the approach is promising, further work is needed to make it viable within a creative system so that it can be used by the final users, namely piano teachers and music students.

Despite the limitations, we believe the proposed approach opens interesting perspectives for AI-supported music education and for Computational Creativity alike, as it requires creating music that is novel and of quality under a personalized and educational point of view, rather than being generically a new piece of satisfactory quality.

Author Contributions

FC ideated the study and performed the experiments. GW provided supervision. Both collaborated in the experiment design and in the writing and revising of this article.

Acknowledgments

We wish to thank Louis Verstraeten for the prior work that opened the way for this study, Nick for the help with Idyoms, and Sam for his suggestion about serialist music, and all of the CCLab for help and support.

References

Bartók, B. 1987. *Mikrokosmos, Vol. 4*. Boosey & Hawkes.

Bloom, B. S. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher* 13(6):4–16.

Carnovalini, F.; Scofield, S.; Verstraeten, L.; and Wiggins, G. 2026. The possibilities of personalized music education through exercise generation [version 1; peer review: 1 approved with reservations, 1 not approved]. *Open Research Europe* 6(108).

Carnovalini, F.; Espírito Santo, L.; and Wiggins, G. A. 2025. Personalized Music Education: A Systematic Review of AI Generation Methods. *IEEE Access* 13:207433–207447.

Carnovalini, F.; Roda, A.; and Wiggins, G. 2023. Interactive Generation of Musical Corpora for Piano Education: Opportunities and Open Challenges. In *Proceedings of the 15th International Conference on Computer Supported Education - Volume 1: CSME*, 412.

Chaiklin, S. 2003. The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context* 1(2):39–64. Publisher: Cambridge, UK: Cambridge University Press.

Cleary, J., and Witten, I. 1984. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications* 32(4):396–402.

Conklin, D., and Witten, I. H. 1995. Multiple viewpoint systems for music prediction. *Journal of New Music Research* 24:51–73.

Hansen, N. C., and Pearce, M. T. 2014. Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology* 5(1052).

Pearce, M. T., and Wiggins, G. A. 2006. Expectation in melody: The influence of context and learning. *Music Perception* 23(5):377–405.

Pearce, M. T., and Wiggins, G. A. 2012. Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science* 4(4):625–652.

Pearce, M. T. 2005. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. Dissertation, Department of Computing, City University, London, London,UK.

Ramoneda, P.; Tamer, N. C.; Eremenko, V.; Serra, X.; and Miron, M. 2022. Score Difficulty Analysis for Piano Performance Education based on Fingering. In *ICASSP 2022*, 201–205. Singapore, Singapore: IEEE.

Ramoneda, P.; Eremenko, V.; D'Hooge, A.; Parada-Cabaleiro, E.; and Serra, X. 2024. Towards Explainable and Interpretable Musical Difficulty Estimation: A Parameter-efficient Approach. In *ISMIR 2024*.

Sébastien, V.; Ralambondrainy, H.; Sébastien, O.; and Conruyt, N. 2012. Score analyzer: Automatically determining scores difficulty level for instrumental e-learning. In *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, 571–576.

Verstraeten, L. 2025. Difficulty Analysis and Generation of Piano Exercises. Master's thesis, Vrije Universiteit Brussel, Brussels, Belgium.

KYSS: A Framework for Evaluating Self-Knowledge Support in Human-AI Co-Creation

Giancarlo Colloca and **Jeba Rezwana**
Computer and Information Sciences Department
Towson University
Towson, MD 21202 USA
gcolloc1@students.towson.edu, jrezwana@towson.edu

Abstract

Human-AI co-creative systems increasingly operate through dialogic interaction, supporting users in co-creative tasks such as writing, design, and problem-solving. In these contexts, interaction unfolds across interactional turns, where AI responses shape how users articulate, interpret, and transform their thinking. Existing evaluation approaches remain focused on outputs or overall user experience, providing limited insight into how interaction contributes to cognitive processes, particularly the emergence of self-knowledge in co-creation, the evolving interpretation of one's own thoughts and perspectives, with implications for broader self-understanding.

We introduce the Know-Your-Self Support (KYSS) framework, a multi-layered approach for evaluating how AI systems support users' self-knowledge in co-creative interactions, combining user-reported reflection, expert-coded interaction dynamics, and aggregated behavioral patterns. KYSS conceptualizes self-knowledge as an emergent property of dialogue, supporting the evolving interpretation of one's own thoughts and perspectives. KYSS provides a structured methodology for assessing co-creative AI partners in reflective and meaning-making processes informed by creative and reflective interactions and dialogic evaluation traditions.

Introduction

Human–AI co-creative systems involve collaborative processes in which humans and AI iteratively contribute to the generation and development of artifacts through interaction (Maher 2012; Davis et al. 2014; Karimi et al. 2020). Meaning is co-constructed across turns, as system turns influence how users articulate, interpret, and refine their thinking. Generative artificial intelligence (GenAI) has become the most widespread form of co-creative AI, operating through natural language dialogue, transforming interaction into continuous exchanges where meaning emerges and is reshaped across turns (Clark and Brennan 1991; Luger and Sellen 2016). GenAI contributes content, interpretations, and reframings that shape user thinking, reinforcing, challenging, or redirecting perspectives over time. In co-creative systems, these contributions introduce variation that prompts exploration (Davis et al. 2014; Maher 2012). The expansion of GenAI as a cognitive extension reconfigures human capability across domains (Clark and Chalmers 1998). As GenAI systems increasingly function as cognitive extensions, influencing how individuals attend, interpret, and generate knowledge (Clark and Chalmers 1998), the problem shifts from capability to orientation, shaped by bounded attention and metacognitive regulation (Simon 1969; Flavell 1979).

In this work, we refer to **Self-knowledge** in co-creation as the evolving interpretation of one's own thoughts and perspectives through interaction. Self-knowledge is central for meaningful direction, but evaluation of co-creative AI remains mainly grounded in performance, usability, and outputs (Nielsen 1994; Boden 2004), overlooking interaction's role in its emergence. Interaction enables the reflective and transformative processes through which self-knowledge develops (Dewey 1933), but current frameworks do not assess AI's role in users' self-referential insight or perspective shifts. We conceptualize self-knowledge support as extending beyond self-reflection (Schön 2017) to encompass transformative reframing of meaning and self-understanding (Mezirow 1991), including reinterpretation, perspective shifting, and moments of self-discovery in which interaction reveals latent values or identities, drawing on dialogic and hermeneutic accounts of meaning-making (Gadamer, Marshall, and Weinsheimer 2004). We focus on dialogic co-creation because dialogue provides a structured and traceable medium for cognitive processes (Clark and Brennan 1991), enabling iterative articulation and reinterpretation that make reflection observable, while AI contributions introduce variation that prompts reconsideration of assumptions (Rezwana and Maher 2023). The interaction-induced changes in users' understanding of their thoughts, values, or identity represent a dimension underexplored in current frameworks. This gap raises a key question: **How can the influence of co-creative AI systems on self-knowledge be evaluated?**

To address this question, we present Know-Your-Self Support (KYSS), a framework that presents dimensions for evaluating how self-knowledge emerges in human-AI co-creation. KYSS models human-AI interaction across turns, integrating experiential, interactional, and behavioral layers across turn-, phase-, and session-levels. By combining user-reported experience, turn-level dynamics and behavioral patterns, it provides a method to evaluate interaction-induced changes in users' understanding of their thoughts, perspec-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

tive shifts and moments of self-discovery. While grounded in co-creative AI, KYSS builds on reflective contexts such as creative practice, education, and psychotherapy, contributing to a shift from output to interaction-centered evaluation. This perspective frames co-creative AI systems as participants in shaping users' self-knowledge, with evaluation aimed at assessing and optimizing this capacity.

Background

Evaluating AI Systems in Human-AI co-creation has traditionally relied on performance metrics such as accuracy and efficiency, complemented in Human-Computer Interaction (HCI) by usability and satisfaction (Nielsen 1994; ISO-9241-11:2018). With GenAI, evaluation has expanded to novelty, diversity, and perceived creativity of generated outputs (Boden 2004), with recent emphasis on communication, agency, and user experience (Davis et al. 2025; Rezwana and Ford 2025). However, these approaches remain largely artifact- or system-centric, offering limited insight into how interaction shapes cognitive transformation.

Co-creative systems are interactive systems in which humans and AI collaboratively generate and develop artifacts through iterative exchange, with both agents contributing to the creative process (Maher 2012; Davis et al. 2014), where dialogic interaction enables both to contribute to idea development. Mixed-initiative systems alternate between response and guidance, while dialogic perspectives emphasize co-constructed meaning across turns (Horvitz 1999; Clark and Brennan 1991). AI contributions introduce ambiguity and reinterpretation that prompt exploration, often amplified by imperfect responses (Dalsgaard 2025). However, existing frameworks only partially capture co-creative functions underlying co-creative processes (Davis et al. 2014).

Socratic and hermeneutic traditions frame self-knowledge as dialogic, as in γνῶθι σεαυτόν ("know thyself"). Later philosophical and psychological traditions emphasize its interpretive and transformative nature (Gadamer, Marshall, and Weinsheimer 2004). Educational theory frames self-knowledge as emerging through the examination of experience (Dewey 1933), and creativity research highlights iterative externalization and revision (Runco and Jaeger 2012). Art therapy treats creative externalization as a means of exploring and reinterpreting internal states (Malchiodi 2012). Related work on dialogic processes further emphasizes the role of feedback, perspective change, and trust in reflective and transformative interaction (Schwartz 2013).

In HCI, reflective technologies support introspection (Fleck and Fitzpatrick 2010), and co-creative systems prompt reinterpretation and negotiation (Davis et al. 2014; Rezwana and Maher 2023). These perspectives suggest that self-knowledge emerges through articulation and reinterpretation, but it remains largely unevaluated in human–AI interaction. Evaluating self-knowledge support in co-creative systems reveals a gap: existing frameworks emphasize performance, usability, and creativity, but provide limited support for capturing emerging self-referential insight. Reflection and transformation are discussed across adjacent literatures (Mezirow 1991; Lee and See 2004), but they remain fragmented in computational creativity. Co-creative interaction forms a structured, turn-based process (Rezwana and Maher 2023) in which creativity supports exploration and self-knowledge emergence, but these dynamics remain unevaluated in human–AI co-creation.

Methodology

A systematic literature review was conducted across ACM Digital Library, IEEE Xplore, and ICCC proceedings using iteratively refined queries targeting creativity, co-creation, and reflection. The search was limited to peer-reviewed publications published before November 2025. The literature corpus was refined through three filtering stages: two based on abstract screening and a final full-text evaluation (Fig.1).

The framework was derived through qualitative synthesis of the final corpus, enabling the identification and organization of structural and conceptual dimensions of evaluating co-creative AI support for self-knowledge in co-creation.

Keyword and Corpus Construction

The search query was iteratively developed to capture creativity, human–AI co-creation, and reflective processes, with terms refined through team discussion and exploratory query searches. Keywords were grouped into creative processes, co-creative interaction, and reflection/self-knowledge, combined using Boolean operators with exclusion terms to reduce noise. The final query was: *("creativity" OR "creative process" OR "creative flow")* ***AND*** *("human-AI co" OR "co-creation" OR "co-creativity")* ***AND*** *("reflection" OR "self-reflection" OR "self-awareness" OR "self-knowledge" OR "self-discovery" OR "transcendence")* ***AND NOT*** *("self-driving" OR "self-adaptive" OR "self-organizing" OR "3D projection" OR "image projection")*. The query was executed across the three databases, yielding

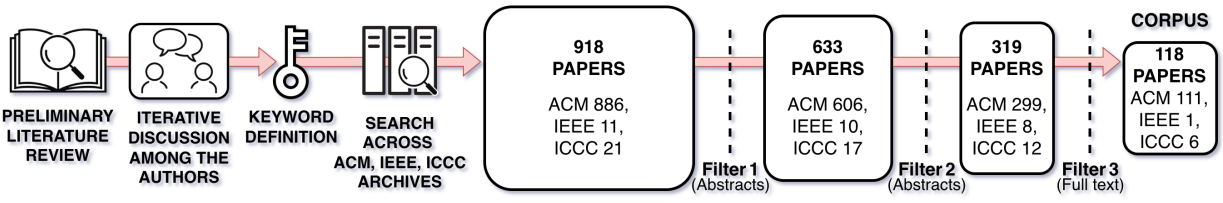


Figure 1: The Literature Review Process

an initial corpus of 918 papers. The three-stage filtering process is illustrated in Fig.1: Filter 1 and Filter 2 were applied at the abstract level. Filter 1 excluded papers with a primarily technical or non-interactional focus (retaining 633). Filter 2 selected papers with reflective and interpretive dimensions or relevance to human–AI interaction and creativity (319). Filter 3 involved full-text analysis for relevance and alignment with our motivation, yielding the final corpus (118 papers).

The KYSS Framework

The KYSS framework is an interaction-centered approach for evaluating how human–AI co-creative systems support the emergence of self-knowledge, by capturing how interaction enables and scaffolds exploratory, reflective, and transformative processes. It provides researchers with a structured methodology to analyze and compare systems.

The framework has a three-layer structure comprising categories and dimensions of evaluation. Each layer captures a distinct perspective: the *Experiential layer* reflects user-reported internal states and subjective evaluations of the experience's impact on the self, whereas the other layers are based on expert evaluation, with the *Interactional layer* encoding turn-level dynamics, and the *Behavioral layer* aggregating them into distributions across phases and sessions. Within each layer, *categories* group related dimensions into conceptual domains, and *dimensions* (defined aspects of interaction) operationalize them into measurable constructs for analysis. Their number balances coverage and parsimony, ensuring sufficient granularity without redundancy. The three-layer structure captures subjective, dialogic, and temporally distributed aspects of self-knowledge. Dimensions under each category were selected if supported by multiple sources or if capturing distinct aspects, based on three criteria: (1) theoretical grounding; (2) observability in dialogic interaction through user reports, expert coding, or computational traces; and (3) relevance to articulation, reinterpretation, or transformation of self-knowledge.

We present the framework by layer, with evaluation actors and categories (Fig.2). Because there are multiple dimensions under each category, we present the dimensions with a detailed figure in the appendix (Fig.3, Appendix).

EXPERIENTIAL LAYER (Actor: User) captures self-reported user evaluation categories of human-AI co-creative interactions in terms of reflective engagement and self-evolution. As self-knowledge is inherently subjective, self-reported measures are essential for accessing internal states not externally observable, aligning with frameworks of reflective practice and metacognitive awareness. Categories group dimensions into enabling conditions, cognitive–affective demands, and reflective outcomes. Dimensions capture affective and cognitive components of the user experience that support self-knowledge.

Safety (*Trust*, *Safety*) captures if the interaction is perceived as reliable and psychologically secure enough to support open exploration (Lee and See 2004; Efstation, Patton, and Kardash 1990), a prerequisite for reflective disclosure and cognitive risk-taking. **Workload** (*Mental Demand*, *Temporal Demand*, *Effort*, *Frustration*, *Performance*, *Physical Demand*) captures the cognitive and affective effort required in the interaction, based on established measures of cognitive load in NASA-TLX (Hart and Staveland 1988). **Reflection** (*Reflection*, *Insight*, *Authenticity*, *Perceived Transformation*) captures if the interaction prompts introspection, generates moments of recognition, aligns with the user's perspective, and contributes to shifts in self-understanding (Dewey 1933; Schön 2017; Fleck and Fitzpatrick 2010; von Thienen, Weinstein, and Meinel 2023), reflecting established models of experiential learning and reflective cognition. **Self-Evolution** (*Challenge*, *Reconsideration*, *Clarification*, *Overall Reflection*) captures if the interaction induces tension, supports reinterpretation, and facilitates conceptual clarification (Dalsgaard 2025), reflecting theories of cognitive conflict and conceptual change.

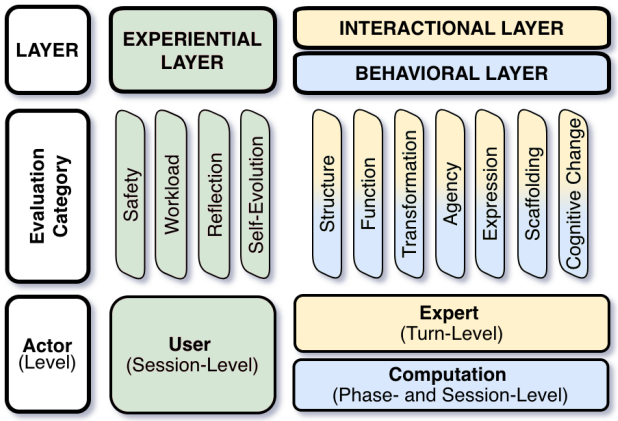


Figure 2: The KYSS Framework: evaluating AI support for emerging Self-Knowledge in dialogic co-creation.

INTERACTIONAL LAYER (Actor: Expert) models turn-level dynamics, with expert evaluation conducted at the level of individual interaction turns to assess the quality of AI contributions and how they shape meaning across exchanges. It encodes dialogic properties of interaction through seven categories and their dimensions as observable measures for expert coding, modeling exchanges as sequences of interpretive and transformative acts. This granularity reflects the turn as the minimal unit of interaction.

Structure (*Phase*) captures the assignment of each interactional turn to a specific phase of the creative process, enabling analysis of how reflection unfolds temporally (Wallas 1926; Mezirow 1991; Davis et al. 2025). **Function** (*Role*, *Strategy*, *Philosophical Orientation*) captures how the system contributes within each turn through the position it adopts in the exchange, how it acts on the user's input, and the stance guiding its response (Horvitz 1999; Runco and Jaeger 2012; Mezirow 1991; Rogers 1951). **Transformation** (*Depth*, *Values Movement*) captures conceptual and evaluative change within each turn, consistent with models of conceptual change and metacognitive development (Flavell 1979). **Agency** (*Initiative*, *Uptake*) captures how AI initiative and user uptake distribute control and

participation between humans and AI across the interaction, reflecting collaborative decision-making and shared agency (Horvitz 1999; Hadfield-Menell et al. 2016; Lawton, Grace, and Ibarrola 2023). **Expression** (*Semantic Shift*, *Lexical Diversity*, *Sentiment*, *Justification Tokens*, *User Revisions*) captures linguistic and behavioral traces of meaning evolution (Reimers and Gurevych 2019; Hutto and Gilbert 2014; Flower and Hayes 1981; Colloca and Rezwana 2025), linking language variation to cognitive and affective change. **Scaffolding** (*Dialogic Perturbation*, *Mirroring*, *Generative Constraint*, *Meta-Cognitive Activation*) captures how the system guides and challenges user thinking, reflecting scaffolding and design strategies (Dalsgaard 2017). **Cognitive Change** (*Perspective Shift*, *Absolutist Reduction*) captures changes in reasoning and interpretive framing, aligning with transformative learning processes (Mezirow 1991; Al-Mosaiwi and Johnstone 2018).

BEHAVIORAL LAYER (Actor: Computation) computes distributional properties of system behavior through automated aggregation of metrics from the Interactional layer into phase- and session-level dimensions. It transforms turn-level properties into distributions (frequency, diversity, balance) over time, enabling the derivation of patterns for scalable and reproducible analysis of emergent dynamics.

Structure (*Phase Diversity*, *Reflection Density*, *Transformation Presence*, *Phase Volume*) captures phase distribution and progression **Function** (*Role Distribution*, *Role Diversity*, *Strategy Distribution*, *Strategy Balance*, *Philosophical Orientation Distribution*, *Philosophical Diversity*) captures variation, balance, and diversity in system contributions and interpretive framing. **Transformation** (*Values Distribution*, *Values Engagement*) captures distribution and activation of evaluative positions. **Agency** (*Initiative Rate*, *Uptake Distribution*, *Initiative Balance*, *CoProduction Depth*) captures system participation and collaborative balance. **Expression** (*Mean Semantic Shift*, *Mean Lexical Diversity*, *Mean Sentiment*, *Mean Justification Tokens*, *Mean User Revisions*, *Productive Tension*, *Tension Volatility*) captures aggregated linguistic and affective dynamics. **Scaffolding** (*Perturbation Mean*, *Mirroring Mean*, *Constraint Activation Mean*, *Meta-Cognitive Activation Mean*) captures system-level guidance structuring cognitive processes. **Cognitive Change** (*Perspective Shift Rate*, *Absolutist Reduction Rate*) captures cumulative changes in reasoning and interpretive framing.

Discussion

The Know-Your-Self Support (KYSS) framework focuses evaluation on co-creative processes rather than outputs, conceptualizing self-knowledge in co-creation as an emergent property of structured dialogue with AI systems. It integrates principles from creativity research, co-creative practice, and reflective traditions (Runco and Jaeger 2012; Davis et al. 2014), translating conceptual constructs into measurable dimensions. In these contexts, creative externalization supports exploration and reinterpretation of internal states; KYSS builds on this perspective to evaluate how co-creative interaction with AI systems supports reflective and transformative processes underlying self-knowledge.

KYSS comprises three complementary layers capturing distinct aspects of self-knowledge support: the Experiential layer reflects users' self-reported perceptions, the Interactional layer captures turn-level dynamics through expert evaluation of AI contributions, and the Behavioral layer aggregates these dynamics across phases and sessions computationally. They support a multi-perspective analysis of self-knowledge support as a dialogic and evolving process.

By modeling turn-level dynamics across phases and sessions, KYSS captures patterns in how self-knowledge evolves over time. It positions dialogic interaction with co-creative AI as a mechanism of epistemic transformation, in which meaning is reinterpreted and AI systems contribute to shaping users' self-expressions through co-creation. This perspective aligns with a broader shift from capability to orientation, highlighting the role of self-knowledge in shaping the intrinsic value of co-creative AI systems. Unlike prior work, KYSS provides an interaction-centered, multi-layer framework that operationalizes self-knowledge as an emergent property of co-creative dialogue.

KYSS enables analysis of how co-creative systems shape reflective processes by identifying patterns across interaction. This aims to support researchers in studying interactional mechanisms of self-knowledge, and AI developers and designers in diagnosing and optimizing system behaviors. In applied contexts, educators, creative practitioners, and clinicians can use KYSS to assess if AI systems facilitate reflection, reinterpretation, and exploration of users' perspectives. More broadly, the framework supports the design and evaluation of co-creative systems that aim to structure interaction in ways that promote self-knowledge processes. KYSS thus generalizes to domains characterized by reflective and interpretive interaction, including creative, educational, and therapeutic settings.

Limitations and Future Work

KYSS incorporates subjectivity at multiple levels: the Experiential layer relies on user-reported perceptions of inherently subjective internal states, and expert coding in the Interactional layer requires validation for consistency. Self-evolution is inherently subjective and can only be directly assessed by the user; as such, its evaluation necessarily relies on user self-report. The framework does not establish causal relationships between sessions and changes in self-knowledge and should be interpreted as descriptive rather than causal. Informed by cross-domain literature, it synthesizes rather than directly transfers existing models.

Future work will focus on empirical validation and analysis of how AI systems support self-knowledge through interaction. A parallel direction is the development of computational proxies for interactional dimensions, using expert-coded data to train and calibrate models that approximate expert judgment and enable progressive automation. Extending the framework to longitudinal settings may further clarify how self-knowledge evolves over time, including beyond co-creative contexts. Overall, the KYSS framework enables interaction-centered evaluation of human–AI systems, providing a structured approach to assessing self-knowledge support and its emergence in co-creative systems.

References

Al-Mosaiwi, M., and Johnstone, T. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical psychological science* 6(4):529–542.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Clark, H. H., and Brennan, S. E. 1991. Grounding in communication.

Clark, A., and Chalmers, D. 1998. The extended mind. *analysis* 58(1):7–19.

Colloca, G., and Rezwana, J. 2025. Prompting ai in co-creation: The role of syntax and sentiment in shaping ai-generated content. In *Proceedings of ICCC*.

Dalsgaard, P. 2017. Instruments of inquiry: understanding the nature and role of design tools. *International journal of design* 11(1):21–33.

Dalsgaard, P. 2025. Designing for ethical friction: Reclaiming agency and accountability in ai-supported creativity.

Davis, N. M.; Popova, Y.; Sysoev, I.; Hsiao, C.-P.; Zhang, D.; and Magerko, B. 2014. Building artistic computer colleagues with an enactive model of creativity. In *ICCC*.

Davis, N.; Clemens, M.; Rezwana, J.; and Browne, E. 2025. Human-ai co-creation: A new interaction paradigm for human-ai interaction. In *Handbook of Human-Centered Artificial Intelligence*. Springer. 1–57.

Dewey, J. 1933. How we think: A restatement of the relation of reflective thinking to the educative process.

Efstation, J. F.; Patton, M. J.; and Kardash, C. M. 1990. Measuring the working alliance in counselor supervision. *Journal of counseling Psychology* 37(3):322.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34(10):906.

Fleck, R., and Fitzpatrick, G. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction*.

Flower, L., and Hayes, J. R. 1981. A cognitive process theory of writing. *College Composition & Communication* 32(4):365–387.

Gadamer, H.-G.; Marshall, D. G.; and Weinsheimer, J. 2004. *Truth and method: Continuum impacts*.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29.

Hart, S. G., and Staveland, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52. Elsevier.

Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 159–166.

Hutto, C., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

ISO-9241-11:2018. Ergonomics of human-system interaction – part 11: Usability: Definitions and concepts.

Karimi, P.; Rezwana, J.; Siddiqui, S.; Maher, M. L.; and Dehbozorgi, N. 2020. Creative sketching partner: an analysis of human-ai co-creativity. In *Proceedings of the 25th international conference on intelligent user interfaces*, 221–230.

Lawton, T.; Grace, K.; and Ibarrola, F. J. 2023. When is a tool a tool? user perceptions of system agency in human–ai co-creative drawing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 1978–1996.

Lee, J. D., and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46(1).

Luger, E., and Sellen, A. 2016. "like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5286–5297.

Maher, M. L. 2012. Computational and collective creativity: Who's being creative? In *ICCC*, 67–71.

Malchiodi, C. A. 2012. *Handbook of Art Therapy*. New York: Guilford Press, 2 edition.

Mezirow, J. 1991. *Transformative dimensions of adult learning*, volume 350. Jossey-bass San Francisco, CA.

Nielsen, J. 1994. *Usability engineering*. Morgan Kaufmann.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 EMNLP-IJCNLP*, 3982–3992.

Rezwana, J., and Ford, C. 2025. Human-centered ai communication in co-creativity: An initial framework and insights. In *Proc. of Creativity and Cognition Conference*, 651–665.

Rezwana, J., and Maher, M. L. 2023. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction* 30(5):1–28.

Rogers, C. 1951. Client-centered therapy houghton mifflin. *New York*.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal* 24(1):92–96.

Schön, D. A. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

Schwartz, R. C. 2013. Moving from acceptance toward transformation with internal family systems therapy (ifs). *Journal of clinical psychology* 69(8):805–816.

Simon, H. A. 1969. The sciences of the artificial mit press. *Cambridge, Ma* 31.

von Thienen, J. P.; Weinstein, T. J.; and Meinel, C. 2023. Creative metacognition in design thinking: exploring theories, educational practices, and their implications for measurement. *Frontiers in psychology* 14:1157001.

Wallas, G. 1926. *The art of thought*. Number 24. Harcourt, Brace.

Appendix

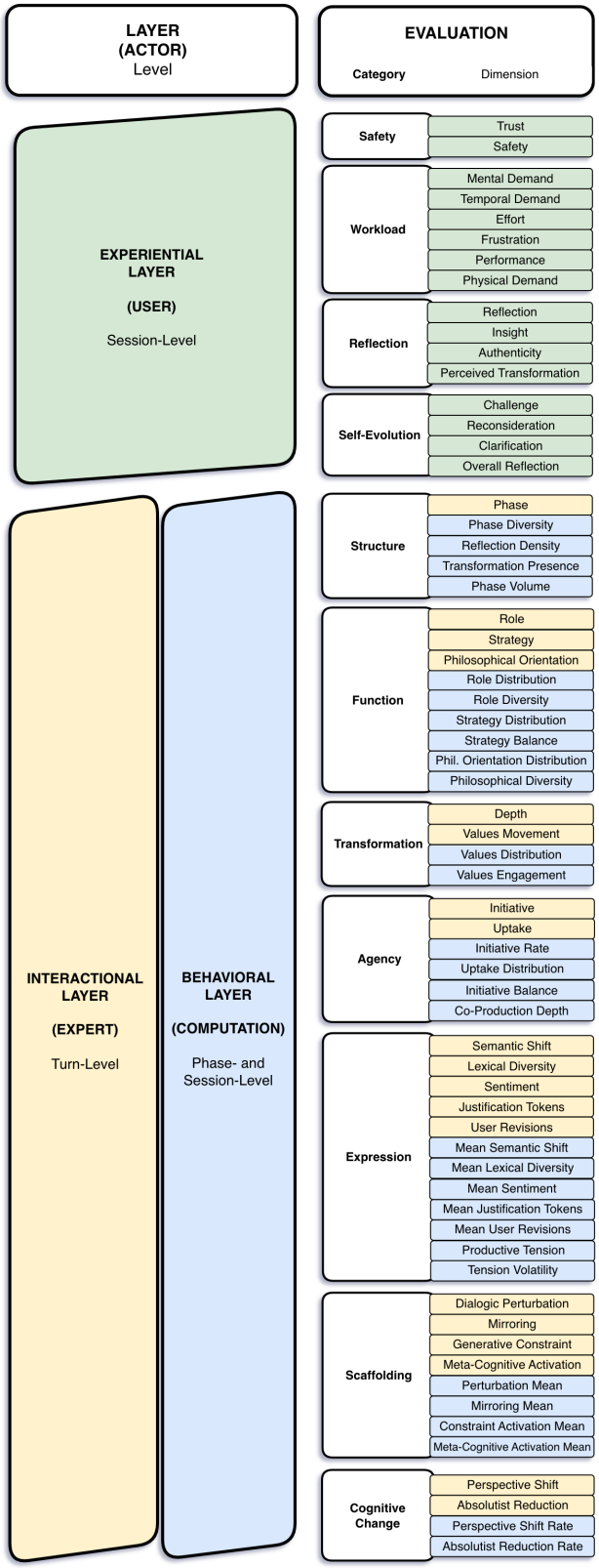


Figure 3: Mapping of KYSS dimensions across the three layers (Experiential, Interactional, Behavioral) over turn-, phase-, and session-levels, with layer-specific categories and actors; colors associate dimensions with layers, even within shared categories.

Twenty-Five Years of The Painting Fool Project

Simon Colton and Blanca Pérez Ferrer

School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

www.thepaintingfool.com

Abstract

The aim of The Painting Fool project is to build an automated artist with cultural impact. For twenty-five years, it has helped shape the growing field of Computational Creativity, and has likewise been shaped by it. We describe here timelines of technical development and public engagement activities undertaken with the software. We also describe how philosophical discourse in Computational Creativity has paid into the project and been influenced by it, and we offer some observations from the long-term interaction with visual arts culture and communities afforded by the project.

Introduction

The development and deployment of the *The Painting Fool* software has been a long-term project of AI researcher Simon Colton and colleagues for 25 years. The overall aim of the project has remained the same throughout, namely that the software itself be taken seriously as a creative artist in its own right, one day (Colton 2012). In addition to driving numerous technical projects to simulate visual art production processes, there have been many public engagement activities, including live demonstrations, art exhibitions, festivals, residencies, companies and commissions. Moreover, reference to The Painting Fool itself and the idea of people taking seriously an AI system as being creative in a human-centric cultural domain such as the visual arts has provided a focal point and provocation for numerous philosophical contributions, also published in the Computational Creativity literature.

As The Painting Fool implementation has reached 25 years as a software entity, it seems sensible to record details of the project so far. To do so, we describe below the origins of the software, the implementation of the system which evolved in epochs, and the technical projects that filled gaps in the software's abilities and behaviours. We also describe the public engagement activities that Colton and others have undertaken with The Painting Fool, and some of the ways in which philosophical ideas drawing on the project have been developed. We end with some conclusions drawn from this long-term study and detail some of the current projects and future directions for The Painting Fool.

Technical Development Timeline

Colton describes the pre-history of the project in terms of the revelation of buying his first digital camera in 1999. Submitting a roll of film and waiting two weeks for a chemist to develop them into print phototraphs, was frustrating for amateur photographers. Digital cameras meant that hundreds of photos could be taken in one session and viewed/printed immediately. Colton used Adobe Photoshop extensively to make the photos he took of the city of Edinburgh look more artistic, which eventually led to him writing graphics software to perform image filtering more to his liking.

During post-doc work on generative maths, in 2001 Colton found he was spending significant time on graphics coding, and decided to include this in his research, broadening his domains of interest in Computational Creativity to include the visual arts. The first public mention of The Painting Fool was at the AISB'01 symposium on AI and Creativity at the University of York in 2001. Colton gave a talk on Meta-Theory Formation, mentioning details of his graphics work at the end.

Colton splits description of the first decade of implementation of The Painting Fool into three overlapping epochs, where behaviours relating to *skill*, *appreciation* and *imagination* in the (simulated) painting process were implemented. During the skill period, the initial graphics software was extended to perform edge detection and image segmentation in order to produce abstracted versions of digital photographs. By 2004, this had been extended with non-photorealistic rendering (NPR) techniques, so that colour regions could be filled or outlined with simulated paint/pastel/pencil strokes. Colton set up web pages at `thepaintingfool.com` displaying generated images from the *City* series and *Amelie's Progress* series (see figure 1 of the appendix).

Reading in an NPR textbook (Strothotte and Schlechtweg 2002) about simulating artistic media and techniques, Colton found the following proclamation:

> Simulating artistic techniques means also simulating human thinking and reasoning, especially creative thinking. This is impossible to do using algorithms or information processing systems.

Colton saw this as denying the potential of AI in gen-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

eral, and Computational Creativity in particular, providing a rallying point to make simulation of artistic behaviours a focal point for his research. An important element implemented during the skill period was live painting demonstrations. Inspired by the robotic painting skills of Harold Cohen's AARON system and by embodied Computational Creativity in general (Guckelsberger et al. 2021) (Moruzzi 2022), this feature has been the most influential in enabling people understand the technologies and theories of The Painting Fool project. Accordingly, it has been the foundation of many public engagement activities, as described below.

In 2006, Colton worked with colleagues Michel Valstar and Maja Pantic at Imperial College London on enabling The Painting Fool to employ their machine vision approaches to guide the artistic process. As described below, this work led to a winning entry in the 2007 BCS Machine Intelligence Competition. It also kick-started the *appreciation* period where sub-projects and code developments were prioritised which led to The Painting Fool having more understanding of its subject material, its output and the artworks of others. During this period, Colton began to seriously hand over creative responsibility to the software, with technical projects testing how independent the software could be.

At the first art exhibition where The Painting Fool's images were displayed (in 2008, also detailed below), another exhibiting artist, namely James Faure-Walker, mentioned in conversation with Colton that artists rarely start projects with an image, more commonly beginning with an idea, emotion or technique that they want to explore (Faure-Walker 2006). This inspired Colton and led to the third period of development where behaviours that might be related to notions of imagination were implemented and tested in The Painting Fool. This involved further handing over creative responsibility in ideation, process and curation.

In particular, the software was enhanced to generate starting imagery to replace digital photographs, to be passed through various pipelines to produce images, videos and framing information (Cook et al. 2019). This was mainly achieved in two ways, namely producing abstract images using evolutionary art techniques and representational art using context-free design grammars. Colton also experimented with other AI techniques such as constraint solving (Colton 2008c) to get The Painting Fool to construct what he called *scenes from a perceived imagination*. This led to images such as those in the *PresiDENTS* series, where photographic elements (of Democratic US presidents), were painted as part of a scene (of generated trees), as portrayed in figure 1 of the appendix. Colton also experimented with fitness function invention, using his HR software (Colton 2008a), leading to surprising prepainting cityscape scenes with buildings on top of each other and other unexpected anomalies.

A suitable bookend for the first decade of development of The Painting Fool was the exhibition entitled *No Photos Harmed* in 2011, described in (Colton and Pérez Ferrer 2012) and below, with details of the project so far written up in a book chapter in 2012 (Colton 2012). The following decade was a period where skilful, appreciative, imaginative and other types of behaviour were implemented in an ad-hoc fashion in response to philosophical developments and technical advances in Computational Creativity research. During this time, Colton undertook various public engagement activities and technical projects with the software, spinning off sub-projects such as image filtering in the *Filter Feast* software (Colton and Torres 2009), and abstract art generation in software eventually called *Art Done Quick* (Colton 2020), with a particular emphasis on casual creation applications (Compton and Mateas 2015).

During this period of exploitation, Colton added what he calls *accountable unpredictability*, where it may not be possible to predict where The Painting Fool will go with a painting, but if something interesting happens (or is produced), one can investigate the source of the event, which is strictly not random-number generation. To do this, he enabled the software to use APIs connected to Twitter, Flickr and The Guardian newspaper, so unpredictable trending information or content could be used as starting points for artistic production.

As an example, MSc. student Anna Krzeczkowska used such functionality to produce collages in response to Guardian news articles, where image retrieval from Flickr was used with The Painting Fool's NPR abilities (Krzeczkowska et al. 2010). One notable output is given in figure 1 of the appendix, where a collage involving babies, a graveyard, a bomber plane and an explosion was generated in response to an article on the war in Iraq. This project, and work in a follow up paper (Cook and Colton 2011), led to questions around where – if anywhere – intentionality lay in producing such images. Ultimately, this work was a pre-cursor to more extensive development and engagement in the *You Can't Know my Mind* portraiture project, described below. This involved working with Dan Ventura to add further machine vision abilities (Colton and Ventura 2014) to enable The Painting Fool to aim for a particular visual style, and assess success in this respect.

In around 2020, Colton realised that generative deep learning models were being trained that could undertake much of what The Painting Fool did via more traditional graphics and AI techniques. He wrote code to interact with pre-trained neural models for image generation (via GAN and diffusion models) and image analysis (via machine vision models such as ResNet and vision-language models such as CLIP). As part of this, Colton worked with PhD students Amy Smith and Sebastian Berns, to explore the creative potential of early text-2-image generation (Colton et al. 2021), producing time-lapse images with generative adversarial networks (GANs) (Colton and Pérez Ferrer 2021), blending neural style transfers for casual creation (Colton 2021) and evolving prompts for diffusion models (Colton et al. 2023). The code from these projects was incorporated into The Painting Fool for future use.

Public Engagement Activities in the Field

The notion of Computational Creativity is controversial, divisive and nebulous (Colton et al. 2014), so it is important to sample opinions from sectors of the general public in order to guide practical and theoretical developments in the field. Colton has used The Painting Fool project to do this in the following ways. Note that images from some public engagement activities are portrayed in figures 3 and 4 of the appendix.

• Art Exhibitions. Artwork from The Painting Fool has appeared in two solo and nine group exhibitions. Along with John Cass, in 2006, Colton organised an art exhibition entitled *Computer Generated Artworks*, held in the business school of Imperial College, fittingly on Exhibition Road in London. This was the first time that The Painting Fool's artworks were exhibited, and Colton gained much feedback from established digital artists, including James Faure-Walker, William Latham, Patrick Tresset and Pensoual Machado. In 2011, along with Eileen Chen, Colton organised an exhibition in Paris entitled *No Photos Harmed*, which highlighted that AI systems can generative artworks which are not abstract, but also not produced from a digital photograph (Colton and Pérez Ferrer 2012). The centrepiece of this was a 3m by 1m piece entitled *The Dancing Salesman Problem*, produced by collaging 300 photo-sized printouts. In the same year, a pop-up retrospective called *Growth*, showcased more than 50 of The Painting Fool's outputs at the Maison Rouge gallery in Paris, as part of a Sony CSL open-house event.

Along with Dan Ventura and Blanca Pérez Ferrer, in 2013, Colton organised a Festival of Computational Creativity entitled *You Can't Know my Mind* in Paris, which included a generative music concert, a reading of generated poetry, a night of generated hot-pots and an exhibition of The Painting Fool's work (Colton and Ventura 2014). A central part of this was a demonstration where The Painting Fool painted portraits to express a simulated mood gained from reading newspaper articles, and learned from the experience, described in (Colton et al. 2015). More recently, Colton employed the new generative AI technologies in The Painting Fool for an exhibition of *GANLapse photography* at the Etopia Centre for Art and Technology in Spain (Colton and Pérez Ferrer 2021).

• Demonstrations. The first major demonstration of The Painting Fool's abilities was as a competition entry for the annual 2007 British Computer Society. The team of Maja Pantic, Michel Valstar and Colton won the competition with a demonstration of *emotionally aware portraiture painting* (Colton, Valstar, and Pantic 2008) where The Painting Fool was informed by software detecting the emotion of a sitter, in order to choose a painting style accordingly. Colton and neuroscientist Derek Jones exhibited art at a Brussels exhibition kicking off the S+T+ARTS (starts.eu) programme. This led in 2018 to Jones commissioning The Painting Fool to be artist in residence at the CUBRIC centre for brain imaging, which was a fruitful collaboration, e.g., producing art to illustrate dreams. In 2016, The Painting Fool appeared at a New Scientist Live event entitled *Anything You Can Do, AI Can Do Better* to paint comedian Dara O'Briain. However, the most substantial demonstration was held in the Science Museum London in 2015: five days and one evening of the You Can't Know My Mind demonstration painting portraits as part of the Museum's Antenna Live and Lates programme. The Painting Fool produced more than 300 portraits in continuous generation projected onto large screens. This demonstration was also undertaken at other venues, including the London Natural History Museum and the Paris Cité des Sciences.

• Media Coverage. Colton estimates that around 100 third-party print/web articles, news items and TV/radio documentaries have covered The Painting Fool project, including every major UK newspaper, science publications like the New Scientist and many international outlets. Of particular note, in 2012, mathematician and science communicator Marcus du Sautoy profiled The Painting Fool and Colton in an episode on Learning and Creativity for a flagship BBC Horizon documentary on AI. This brought the ideas of AI autonomy in the arts to a relatively large public audience. Similarly, in 2016 Mrs. Yang Lan – a well-known personality on Chinese television – profiled The Painting Fool as part of a CCTV documentary on AI, had her portrait painted by the software, and purchased a printed piece. A member of the production team told Colton that they hoped the worldwide audience for the documentary would be 1.2 billion. Most of the media coverage of the project has been positive, but there has been negative feedback, for instance a writer for the Observer Magazine was particularly critical of the project.

• Commercial activities. Along with business partner Glen Pearson, Colton set up UK company Machine Creations Ltd., in 2005, which produced paint-by-numbers-kits for customers who uploaded an image to the website. This used The Painting Fool's image segmentation software with enhanced abilities to choose paint mixes and render a painting kit. The company lasted two years and was not hugely successful, but provided an important learning experience in terms of commercial testing of the software. In 2014, behavioural information data analysis company UBIC commissioned The Painting Fool to produce printed images for their offices in Tokyo, as described in (Colton et al. 2015). This resulted in a series called *I Can See Unclearly Now* and tested the flexibility of the software under the scrutiny of a paying patron. More recently, in late 2025, AI-art collector *Delronde* (see eng.art) contacted Colton to purchase pieces from The Dancing Salesman Problem series and others, as both physical prints and NFT digital artworks. This led to Colton and The Painting Fool being represented by the Fellowship.xyz gallery and an online exhibition of old and new work being made available for purchase as NFTs (see figure 4).

Philosophical Provocations

Recalling the aim for the software is to be taken seriously as a creative artist in its own right one day, Colton has proceeded by trying to minimise reasons people could cite for why The Painting Fool is not creative, hoping that eventually – without such reasons – they would begrudgingly admit that the software *is* creative. On many occasions, this approach has led to new implementations, experimentation and outreach, e.g., in response to comments about lack of imagination and intentionality. On other occasions, working closely with colleagues such as Alison Pease, Teresa Llano, Christian Guckelsberger, Michael Cook and Berker Banar, Colton was able to generalise issues with The Painting Fool and write in a more philosophical way to add to the theory of Computational Creativity.

The discourse in (Colton 2008b) was a response to early (well-meaning) criticisms of The Painting Fool as an independent artist. Here, the idea was proposed that if a creative AI agent was lacking in abilities onto which people could project notions of skill, appreciation and imagination, then it could be easily dismissed as being uncreative. As mentioned above, this formulation of behaviour types guided the first decade of projects with The Painting Fool, and was later expanded to include more behaviour types covering innovation, intentionality, learning, accountability and reflection. In addition, constant suggestions that the only way to project artistic creativity onto the software was to compare its output to art produced by people led to a critique in (Pease and Colton 2011) of Turing-Style blind comparison studies and a suggestion of alternatives. On two different occasions, outputs from The Painting Fool were shown to artists (one in Paris and another in Beijing) by television crews, to get their reaction to the images and the project as a whole. On both occasions, the artists said they felt the images were "cold", lacking empathy and understanding of the human condition, clearly judging the artworks in terms of the process, personality and press (Jordanous 2016) rather than product. This led to special consideration in (Colton et al. 2014) of observer issues in Computational Creativity.

Ultimately, autonomy in process became less of a criticism through projects such as *You Can't Know my Mind*. This left difficult questions of autonomy with respect to motivation, and led to theoretical papers highlighting issues with respect to intrinsic motivations (Guckelsberger et al. 2017) and authenticity (Colton, Pease, and Saunders 2018) in computational systems. One solution to this was proposed in (Colton et al. 2020), where the authors suggest that creative AI systems like The Painting Fool can record aspects of their daily existence to become the focus of artistic creations within a framework they call the Machine Condition, in order to express what it is like to be the AI system. This idea was pushed further, by looking at how AI systems could add to artistic cultures (Colton and Banar 2023) and what it would mean for them to gain creative personhood (Pease, Colton, and Banar 2023).

Conclusions and Future Work

With such a volume of longitudinal technical research, philosophical discourse and public engagement in the field, a number of conclusions can be drawn from The Painting Fool project that might transfer to other Computational Creativity projects. These include:

- Complexity in backstory, motivation and process is far more interesting to artistic communities than sophistication, novelty or value of the artefacts that a generative system produces.
- In general, people still find it difficult to project notions of artistic creativity onto software systems, in both principle and practice (Hertzmann 2018). It is still reasonable to say that the origin of creativity lies more with the programmer of such systems than the system itself.
- With long-term projects, it is beneficial for practical and theoretical work to drive each other forward in the study of Computational Creativity.
- Handing over creative responsibilities to software is a common mantra in Computational Creativity research, but has rarely been studied anywhere else in AI or the arts. Moreover, the position that AI systems should only ever be tools for people is even more entrenched now, given the rise of issues around livelihood, legacy and copyright affecting artists and musicians.
- Public engagement has three facets: projects such as an art exhibition which showcases Computational Creativity work; engagement indicators, such as a newspaper article written about the exhibition, and actual engagement, such as conversations in the exhibition or comments left online. In general, engagement begets engagement and *relatable behaviours* (Colton 2026) in creative production and assessment can be key in bringing attention to projects involving creative AI systems.

Historically, The Painting Fool project sits between Cohen's AARON (McCorduck 1991) and Klingemann's BOTTO projects (Klingemann, Hudson, and Thompson 2023), both of which allude to some independence in art creation by an AI system. While it is clear that The Painting Fool is not yet taken seriously as a creative artist in its own right, Colton cites moments when he felt the software was acting as an artist, albeit momentarily and not in a hugely creative way. The Painting Fool project continues and the recent interest in its back-catalogue of art has prompted work towards a retrospective exhibition. There is more than enough theoretical work from the Computational Creativity community to guide the project into later stages. Moreover, the global tech industry regularly drops generative AI models which could be employed by The Painting Fool in the quest to endow it with creative autonomy and send it out into the artistic world. It's not impossible to think that after another 25 years, the project – or one inspired by it – will have succeeded, and the art world will be enriched with computationally creative artists.

Acknowledgements

We would like to wholeheartedly thank the many computational creativity researchers, students, artists and journalists who have contributed to The Painting Fool project, as well as the anonymous reviewers of the many papers written about it, including this one.

References

Colton, S. 2026. Experiments in Text-2-Tool-Use for Relatable Computational Creativity In *Proceedings of the 17th Int. Conference on Computational Creativity.*

Colton, S., and Banar, B. 2023. Automatically adding to artistic cultures. In *Proc. 12th Int. Conf. on Artificial Intelligence in Music, Sound, Art and Design.*

Colton, S., and Pérez Ferrer, B. 2012. No photos harmed/growing paths from seed: an exhibition. *Proc. Symp. on Non-Photorealistic Animation & Rendering.*

Colton, S., and Pérez Ferrer, B. 2021. GANlapse generative photography. In *Proceedings of the 12th International Conference on Computational Creativity.*

Colton, S., and Torres, P. 2009. Evolving approximate image filters. In *Workshops on Applications of Evolutionary Computation*, 467–477. Springer.

Colton, S., and Ventura, D. 2014. You Can't Know my Mind: A festival of computational creativity. In *Proc. of the 5th Int. Conference on Computational Creativity.*

Colton, S.; Cook, M.; Hepworth, R.; and Pease, A. 2014. On acid drops and teardrops: Observer issues in computational creativity. In *Proceedings of the 7th AISB Symposium on Computing and Philosophy.*

Colton, S.; Halskov, J.; Ventura, D.; Gouldstone, I.; Cook, M.; and Pérez Ferrer, B.. 2015. The Painting Fool sees! New projects with the automated painter. In *Proc. 6th Int. Conference on Computational Creativity.*

Colton, S.; A, P.; Guckelsberger, C.; McCormack, J.; Llano, T.; and Cook, M. 2020. On the machine condition and its creative expression. In *Proc. 11th ICCC.*

Colton, S.; Smith, A.; Berns, S.; Murdock, R.; and Cook, M. 2021. Generative search engines: Initial experiments. In *Proceedings of the 12th International Conference on Computational Creativity.*

Colton, S.; Smith, A.; Pérez Ferrer, B.; and Berns, S. 2023. Artist discovery with stable evolusion. In *Proc. 14th Int. Conference on Computational Creativity.*

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. In *Proc. 9th Int. Conference on Computational Creativity.*

Colton, S.; Valstar, M.; and Pantic, M. 2008. Emotionally aware automated portrait painting. *3rd Int. Conf. on Digital Interactive Media in Entertainment & Arts.*

Colton, S. 2008a. Automatic invention of fitness functions with application to scene generation. In *Workshops on Applications of Evolutionary Computation.*

Colton, S. 2008b. Creativity versus the perception of creativity in computational systems. In *AAAI spring symposium: creative intelligent systems*, volume 8.

Colton, S. 2008c. Experiments in constraint-based automated scene generation. In *Proc. of the 5th international joint workshop on computational creativity.*

Colton, S. 2012. The Painting Fool: Stories from building an automated painter. In McCormack, J., & d'Inverno, M., *Computers and creativity.* Springer.

Colton, S. 2020. Casual creation, curation, captioning, clustering and crossover in the Art Done Quick app. In *Proc. of the Workshop on Casual Creators at ICCC.*

Colton, S. 2021. Evolving neural style transfer blends. In *Proc. International Conference on Computational Intelligence in Music, Sound & Art (EvoMusArt).*

Compton, K., and Mateas, M. 2015. Casual creators. In *Proc. 6th Int. Conference on Computational Creativity.*

Cook, M., and Colton, S. 2011. Automated collage generation-with more intent. In *Proceedings of the 2nd International Conference on Computational Creativity.*

Cook, M.; Colton, S.; Pease, A.; and Llano, M. 2019. Framing in computational creativity-a survey and taxonomy. *Proc. 10th Int. Conf. on Comp. Creativity.*

Faure-Walker, J. 2006. *Painting the Digital River.* Prentice Hall.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the "why?" in computational creativity. In *Proc. 8th Int. Conf. on Computational Creativity.*

Guckelsberger, C.; Kantosalo, A.; Negrete-Yankelevich, S.; and Takala, T. 2021. Embodiment and computational creativity. In *Proceedings of the 12th International Conference on Computational Creativity.*

Hertzmann, A. 2018. Can computers create art? *Arts* 7(2), MDPI.

Jordanous, A. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28(2), 1-23.

Klingemann, M.; Hudson, S.; and Thompson, J. P. 2023. Cluster #069. In *SIGGRAPH 2023 Art Gallery.*

Krzeczkowska, A.; El-Hage, J.; Colton, S.; and Clark, S. 2010. Automated collage generation-with intent. In *Proc. 1st Int. Conf. on Computational Creativity.*

McCorduck, P. 1991. *AARON's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen.* W.H. Freeman and Company.

Moruzzi, C. 2022. The (artificial) physicality of creativity : How embodiment influences perceptions of creativity. In *Proceedings of the 13th International Conference on Computational Creativity.*

Pease, A., and Colton, S. 2011. On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. In *Proceedings of the AISB symposium on AI and Philosophy.*

Pease, A.; Colton, S.; and Banar, B. 2023. On the notion of creative personhood. In *Proceedings of the 14th International Conference on Computational Creativity.*

Strothotte, T., and Schlechtweg, S. 2002. *Non-Photorealistic Computer Graphics.* Morgan Kaufmann.

Appendix

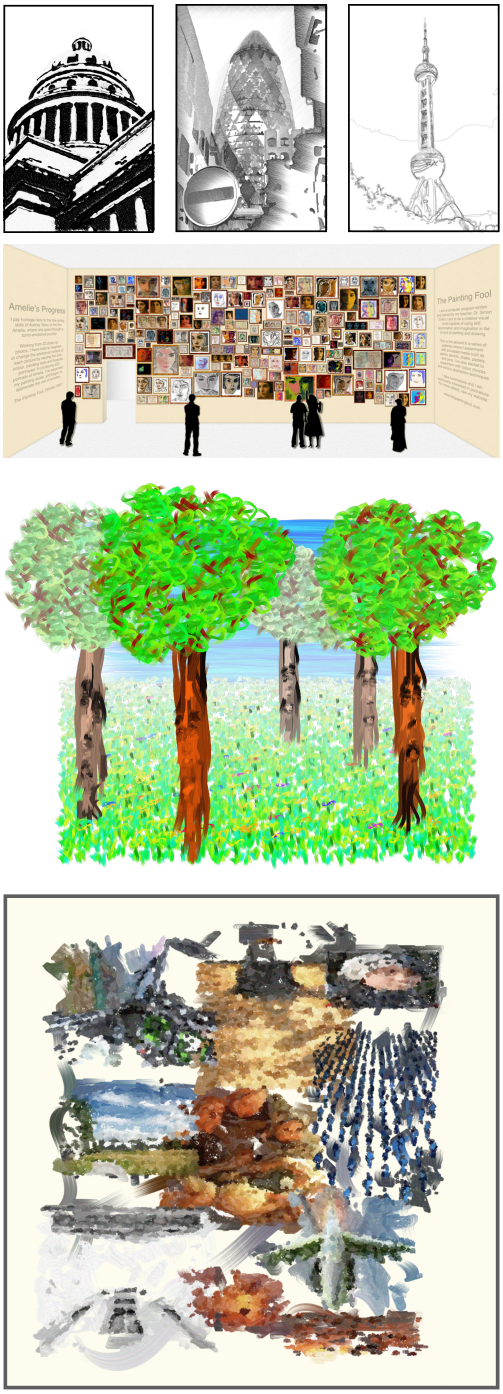


Figure 1: (a) images from the *City Series* (2003) and (b) *Amelie's Progress* (2005), highlighting The Painting Fool's non-photorealistic rendering abilities (c) image from the PresidENTS series (2009/2010), where scene invention was facilitated by constraint solving, and (d) an image from a news-based collage series, derived from an article on the war in Afghanistan, 2011.



Figure 2: Images from public engagement activities with The Painting Fool: (a) the Computer Generated Art exhibition of 2006 (b) first mainstream news article about The Painting Fool, 2006 (c) wall text for the *Growth* retrospective exhibition, Maison Rouge, Paris 2011, and (d) a child sits for a portrait at a portraiture demonstration at Cité des Science, Paris 2014.



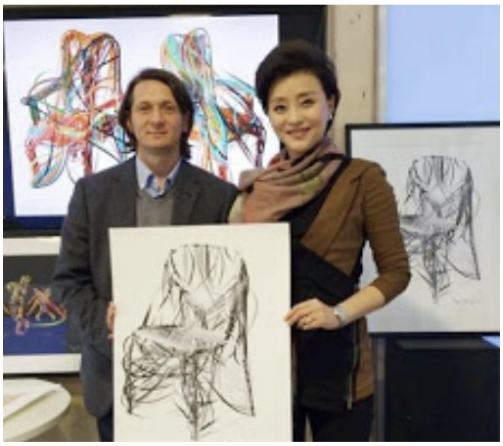
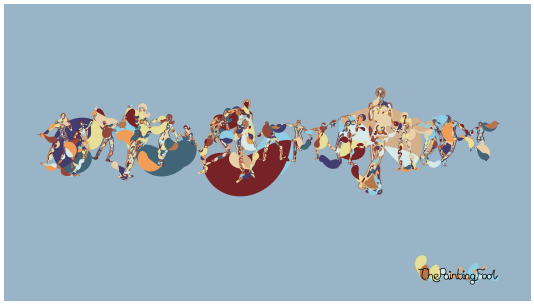


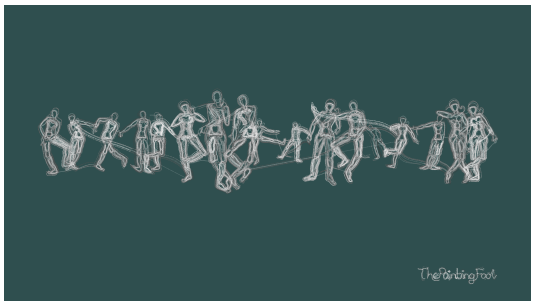




Figure 3: Further images from public engagement activities with The Painting Fool: (a) Colton with The Dancing Salesman Problem piece at the *No Photos Harmed* exhibition of 2011 (b) Mrs. Yang Lan purchasing artwork by The Painting Fool in 2016 (c) The Painting Fool's walrus-like portrait of Dara O'Briain at a New Scientist Live event, 2016, and (d) visitors at the *You Can't Know my Mind* festival, 2013.









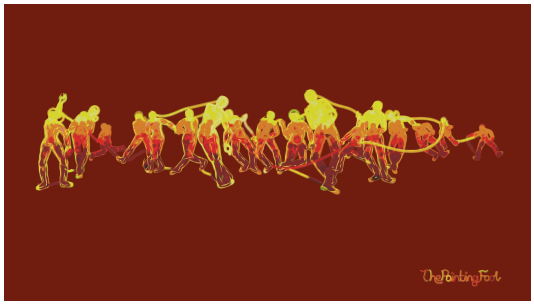


Figure 4: New image styles added in 2026 to the Dancing Salesman Problem series: (a) *deconstructed* style (b) *pop* style (c) *ghosts* style (d) *watercolour* style, and (e) *fuego* style.

Making Sense of "Cozy Chaos": Humans and LLMs Differences in Linguistic Creativity

Anca Dinu[1*], **Andra-Maria Florescu**[1*], **Alina Resceanu**[2*]
[1]University of Bucharest, Bucharest, Romania
[2]University of Craiova, Craiova, Romania
anca.dinu@lls.unibuc.ro, andra-maria.florescu@s.unibuc.ro, aresceanu@gmail.com
[*]Equal contribution

Abstract

LLMs have shown remarkable creative capabilities, often outperforming humans in tasks requiring originality and linguistic novelty. However, their creativity differs in nature, raising questions about how humans and machines use language in an original way. This study explores the linguistic dimensions of human and LLM's creativity by analyzing a dataset introduced in previous research. We apply LIWC analysis to identify the proportions of functional words, social and emotional features, and compute the language style matching score to measure stylistic similarity. We also assess readability through standard metrics, perform sentiment analysis, clustering, and binary classification to evaluate how separable the two groups are. The results show that LLMs produce more positive, formal, fluent, and lexically complex responses, whereas human responses are more negative, spontaneous, and socially and experientially grounded.

Introduction

A key motivation for this study is the lack of dedicated evaluation frameworks for the linguistic creativity of large language models (LLMs). Linguistic creativity tasks, such as inventing new words or expressions, require models to go beyond learned patterns and generate novel language forms for which there are no explicit training examples. This makes linguistic creativity tests particularly relevant for assessing computational creativity (CC), as they test the ability of LLMs to creatively manipulate and extend language. Another motivation relevant for CC is to understand the nature of differences in linguistic creativity between humans and machines.

In previous work, (Dinu, Florescu, and Resceanu 2025) introduced a language creativity test in which 24 LLMs and 24 humans were automatically evaluated, showing that the models slightly outperformed humans in most tasks and criteria. However, while performance differences have been documented, the linguistic mechanisms underlying these differences remain underexplored. This study aims to fill this gap: instead of focusing only on performance, we examine how linguistic style, structure, and emotion shape creative expression across both groups. While previous studies have compared human and machine-generated texts mainly through semantic similarity or content-based measures (e.g., (McCoy et al. 2023)), we focus on directly interpretable linguistic and psycho-social features (Gehrmann, Clark, and Sellam 2022); (Zanotto and Aroyehun 2024); (Durward and Thomson 2024). Specifically, we applied the Linguistic Inquiry and Word Count (LIWC) tool (Boyd et al. 2022) to capture grammatical, psycho-social, and emotional features and to compute Language Style Matching (LSM). We also computed readability metrics and performed sentiment analysis, clustering, and classification.

Related Work

The study of CC has flourished in recent years, with the rise and development of LLMs (Chkirbene et al. 2024), their generative potential reaching both their common users and the creative industries (Raza et al. 2025). A detailed analysis of LLMs' creativity was carried out in (Jiang et al. 2024), including a wide range of types of creativity, such as ideational creativity expressed verbally or figuratively, personality creativity, or image generation. A comprehensive review on general artificial creativity is (Ismayilzada et al. 2024).

Most studies explore just the ideational or visual capabilities of LLMs, e.g. (Zhao et al. 2025), which revealed that LLMs are good at elaboration, adding detail and refinement, but weaker in originality or producing genuinely new ideas. However, fewer studies focus on LLMs' creative linguistic capacities, e.g. (Dinu and Florescu 2025; Dinu, Florescu, and Resceanu 2025), which showed a slightly better performance of the LLMs. However, the specific linguistic differences between human and machine answers remain underexplored.

Recent work has also focused on how LLMs handle deeper linguistic and social nuances. For instance, (Weller-Di Marco and Fraser 2024) analyzes how LLMs process morphologically complex words and compound words, showing that while models can generate well-formed word structures, they often rely on surface-level patterns rather than true morphological understanding. Similarly, (Perez Almendros and Camacho-Collados 2024) examined whether LLMs can recognize socially oriented communication patterns, such as *mansplaining*, finding that machines can capture some social cues, but still have difficulties identifying subtle relational and pragmatic contexts. Automatic systems that creatively alter language expressions are proposed in (Gatti, Stock, and Strapparava 2021) with practical goals, such as grabbing the reader's attention or aiding in concept recall. Noting thematic distinctions between the

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

	function	i	ipron	det	number	prep	conj	negate	verb
MeanLLMs	525.58	0.00	0.00	103.83	11.58	271.83	141.67	4.33	177.13
MeanHumans	814.50	33.42	22.71	255.17	41.17	485.13	46.58	36.33	317.21

Table 1: Comparison of human and LLM mean scores across function word categories.

	affiliation	allnone	cause	discrep	differ	space	auditory
MeanLLMs	82.25	4.33	30.54	8.96	3.58	191.42	292.38
MeanHumans	150.79	20.96	8.67	81.08	33.92	266.00	146.33

Table 2: Comparison of human and LLM mean scores across structural features.

human and AI-generated literature, (Durward and Thomson 2024) examined the language used in creative works.

Our study complements this previous research by analyzing the linguistic and stylistic features of the creative output of humans and machines.

Dataset and Experimental setup

We use the dataset introduced by (Dinu, Florescu, and Resceanu 2025) that comprises the responses of 24 humans and 24 LLMs to a language creativity test, totaling 2304 responses. The eight language creativity tasks are given in the Appendix. The human participants were volunteer students in Humanities (English major), non-native speakers of English, with an English proficiency level of B2 and above, aged between 19 and 25, 8 males and 16 females. It can be argued that the test respondents being non-native English speakers is a limitation of the study. Nevertheless, an article that proposes a language creativity test focused on word formation only (Körtvélyessy et al., 2022) states that "there is no principled difference between native speakers and non-native speakers in their ability to form new complex words and interpret/predict the meaning of novel/complex word provided that the non-native speaker has a standard command of a particular language [(...)] and that his/her world knowledge and experiences are comparable to those of common native speakers". Language creativity manifests itself in non-native speakers in relevant ways, as explained in (Zipp, 2019).

We experimented in Colab, with open sourced libraries: pandas[1], NumPy[2], scikit-learn[3], matplotlib[4],adjustText[5], textstat[6], Seaborn[7], and transformers[8]. For coding assistance, we used ChatGPT(4o)[9] and Claude(3.7)[10].

[1]https://pandas.pydata.org/

[2]https://numpy.org/

[3]https://scikit-learn.org/stable/

[4]https://matplotlib.org/

[5]https://adjusttext.readthedocs.io/en/latest/

[6]https://textstat.org/

[7]https://seaborn.pydata.org/

[8]https://pypi.org/project/transformers/

[9]https://chatgpt.com/

[10]https://claude.ai/

LIWC analysis

We extracted linguistic and psicho-social features from the dataset with the LIWC-22 tool (Boyd et al. 2022), to capture how humans and machines use language, their stylistic, emotional, and cognitive patterns. We obtained a low to moderate LSM score between humans and LLMs (0.43), suggesting substantial differences in the way humans and LLMs use language creatively.

Table 1 shows the human and LLM mean scores across function words. Human responses show a richer and more varied use of function words such as pronouns, determiners, prepositions, and verbs. This is evident in phrases like "I am miserable", "would be drama queens" or "the nightmare of corporatists" where pronouns (I, it), prepositions (in, of), determiners (the, a), and verbs (would) create connected, dynamic expressions. Such constructions make human language sound more relational, narrative, and experiential, reflecting that people draw on lived and social contexts when being creative. Humans also used a lot of negations ("not fully developed" for *childish*, "a sea of untold stories", "unhappy people") because they often think and create through contrast, irony, or dissatisfaction. Saying what something is *not* is a quick and expressive way to show tension or attitude, which fits natural and emotional language. LLMs, on the other hand, avoided frequent negations because their outputs are optimized for fluency, balance, and positive tone. Their responses also contained fewer function words and more self-contained noun phrases, like "pocket-sized thunderstorms" (for angry dogs), "customer service gladiators supreme" (for "I wanna speak to the manager" kind of people), or "(a sea of) endless serenity", which show a well-structured style, but one that sounds detached from real experience and less natural.

Table 2 reveals that human responses are more socially (higher *affiliation*), emotionally (higher *allnone*), and cognitively engaged (higher *discrep* and *differ*). These patterns suggest that human creativity is rooted in subjectivity and self-reflection, since people express themselves through opposites, doubt, and social perspective. In contrast, LLM responses show greater *causal* coherence and *auditory* description, producing structured and stylistically balanced answers that sound fluent but lack the relational nature of human creativity. The higher *space* scores for humans further point to their tendency towards spatially grounded im-

	prosocial	moral	socrefs	family	friend	female	male
MeanLLMs	41.67	63.54	203.08	5.63	15.17	9.46	4.75
MeanHumans	23.25	120.25	446.13	54.79	35.50	94.17	34.71

Table 3: Comparison of human and LLM mean scores across social features.

	home	work	money	relig	physical	health	illness	wellness	mental	food	death	fatigue	allure
MeanLLMs	23.88	124.13	37.33	34.71	158.88	70.38	16.00	39.75	0.00	36.50	4.63	4.79	110.96
MeanHumans	46.50	236.92	66.63	111.08	496.75	127.96	31.46	22.83	16.67	115.42	42.54	15.50	236.33

Table 4: Comparison of human and LLM mean scores across personal features.

agery ("storm in a cup", "fury in a fur coat"), while LLMs' high auditory score indicates formulaic figurative speech ("echoes of calm", "whispers of serenity"). Moreover, human language is deeply relational and morally evaluative, whereas LLM language is neutral, polite, and impersonal, as can be seen in table 3. Humans used more moral terms ("guilty", "innocent", "good", "bad") and social references ("we," "they," "my friend"). LLMs, in turn, produced more prosocial but abstract language ("helpful", "friendly", "kindly"), showing polite, controlled positivity.

Table 4 reinforces that human creativity draws on emotional and existential experiences, including body, home, work, health, food, and death, often with irony or emotional depth ("angry mums" for "I wanna talk to the manager kind of people", "mummy's baby" or "parents' epic failure" for *childish*, "cozy office work stress" or "cozy exam taking" for the oxymoron task). LLMs, on the other hand, produced emotionally regulated 'proper' language, avoiding *illness*, *death* or *fatigue* using instead wellness terms ("cozy chaotic serenity", "chaotic Zen garden meditation hour").

Comparing the frequency of motivational language across the categories in table 5 reveals that human responses show stronger desire and effort (e.g. "tryhardish", "win", "fail", "dream", and "fight"), and irony or references to everyday realism (e.g., "manager hunting warriors", "refund addicts", "let the show begin people", "in need of a therapist" for "I wanna talk to the manager kind of people"). LLMs' style is safe and positive, avoiding any display of power or personal ambition and keeping the tone neutral, emotionally balanced, and collective. This is in line with the *tone* score: LLMs - 2286.88 and humans - 1169.42.

The BigWords score, which measures how many long words (six or more letters) appear in a text, indicates that LLMs (score 6354.96) use longer words than humans (4851.58), creating more complex but harder-to-read constructions (e.g. "Busybody Blabbermouth", "Scandal-spreading Scuttlebutt", "Neighborhood hearsay distributor" for *gossiper).* Humans preferred shorter and simpler words (e.g., "scandal seeker", "story spreader", "muddy tongue") giving their answers a more vivid and spontaneous feel.

Readability

To analyze differences in readability between the two groups, we computed six metrics with the *textstat* library: Flesch Ease, Flesch-Kincaid, Gunning Fog, ARI, Coleman-Liau, and Dale-Chall, represented in Figure 1 from the Appendix. For all metrics, the answers produced by the LLMs are consistently more difficult to read. Humans formed shorter, simpler, and more familiar expressions that sound direct and conversational, whereas LLMs used longer, more vague, and complex words, which makes their answers fluent but heavier to read. This fits with (McCoy et al. 2023; Zanotto and Aroyehun 2024), who showed that LLMs prefer formal and grammatically complete phrasing, whereas humans favor spontaneity and clarity. The almost 0 p-values obtained for the Welch tests confirm that the differences in mean readability between humans and LLMs are statistically significant, as shown in Table 6 from the Appendix.

Sentiment Analysis

To compare the emotional load of humans and LLMs, we performed sentiment analysis for all 48 files, by using a transformer-based approach via Hugging Face's pre-trained BERT model[11]. Each individual file was assigned a percentage of positive, neutral, and negative content. The distribution of individual sentiments with stacked horizontal bar charts listed in figure 2 from the Appendix. In the first 15 most negative individual responses, only 3 are LLMs and 12 humans, whereas, symmetrically, in the first 15 most positive individual answers only 3 are humans and 12 LLMs. This might be due to the LLM having their content filtered to avoid toxicity and negativity and to humans expressing creativity by the use of emotional contrast, irony, humor, or realism (e.g. "not fully matured" or "not quite grown up" for *childish*, a sea of... "never ending troubles", "unsolicited advice from that aunt", "broken hearts" or "tears and despairs").

Visualization

To visualize systematic differences in creativity answers between humans and LLMs for all 48 individual text files, we performed Principal Component Analysis (PCA) on two types of features: word usage patterns extracted with TF-IDF vectorization, which highlight lexical and topical differences, and writing style and structural features, which underline stylistic differences rather than word choice. The first

[11]https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

	need	want	acquire	lack	reward	risk
MeanLLMs	4.79	7.25	0.54	9.92	9.33	13.67
MeanHumans	12.67	40.38	18.00	26.00	37.79	23.21

Table 5: Comparison of human and LLM mean scores across motivational features.

method produced the visualization in figure 3 from the Appendix. The individuals are moderately clustered into two groups, with some overlappings in the densely populated center, and some individuals placed far away from the center. This indicates that word usage is different across the two groups, but a perfect discrimination is not possible based only on this type of feature. For the second method, we used a deep learning-based style embedding model[12]. This method generates contextual embeddings via transformers, capturing nuances of writing style, tone, and structure. This time, the PCA visualization, given in figure 4 from the Appendix, shows that the two classes separated fairly well, almost linearly, with only two humans grouped with the LLMs and with ChatGPT integrated within the human group. This result suggests that stylistic features are good discriminators between the two classes. The two human outliers are participants who responded in a more formal and structured manner, while the proximity of ChatGPT to the human cluster may be caused by its extensive exposure to human interaction during training and refinement, since, unlike other models, ChatGPT has been continuously improved through reinforcement learning from human feedback (RLHF) (Lambert 2025) and widespread everyday use.

Binary Classification

The results of all previous sections indicate that there are relevant features that discriminate between humans and LLMs. Therefore, we experimented with several classical Machine learning (ML) techniques (Suport Vector machine (SVM), Random Forest (RF), Multinomial Naive Bayes, Logistic Regression) and the DistilBERT[13] transformer, to automatically classify the two groups. As training set, we considered all separate responses from all humans and machines (1152 each). The best performance for ML techniques was obtained with Logistic Regression with TF-IDF vectorization at both character and word level, with an accuracy of 0.68. The confusion matrix shown in 5 (left) from the Appendix reveals 66 humans misclassified as LLMs and 80 LLMs misclassified as humans. This modest performance is likely due to the presence of invented or novel words that the TF-IDF representation cannot capture effectively.

The transformer model outperformed the ML techniques, obtaining an accuracy of 0.75, in the third epoch, as we can see in table 7. The confusion matrix shown in 5 (right) from the Appendix reveals a consistent improvement over the ML algorithms, indicating that deep contextual representations better capture subtle linguistic and invented-word patterns.

[12]https://huggingface.co/AnnaWegmann/Style-Embedding

[13]https://huggingface.co/distilbert/distilbert-base-uncased

Conclusions

In this study, we compared the linguistic creativity of humans and LLMs through a detailed computational and psycholinguistic analysis of 2304 creative responses, using LIWC linguistic profiling, readability metrics, sentiment analysis, visualization, and classification methods. The results reveal that LLMs tend to produce language that is more complex, formal, and lexically dense, marked by structural regularity and positive tone. Their responses are balanced and fluent, but display limited emotional range and spontaneity. This pattern points to exploratory creativity: LLMs appear to generate novel combinations within an already learned linguistic and probabilistic space, rather than transforming the underlying generative rules (i.e. attaching the suffix *-ish* to standard bases like adjectives or nouns, such as in "quikish" and "chirpish"). In contrast, humans often redefine or break the rules, creating borderline interpretable utterances (i.e. attaching the suffix *-ish* to non-standard bases, like verbs, "loveish", "hatish", "scrollish", or producing semantically deviated examples, "bottleish", "teethable"), indicative of transformational creativity (Boden 1996; Boden 2004).

Additionally, human creativity emerged as more relational, contextually and emotionally grounded, often expressed through negation (i.e. "no lifer", "no brainer" for 'I wanna speak to the manager' kind of people'), irony (especially in word formation tasks, i.e. "Americanish", "parentish", "wifeable") and experientially or socially referenced (i.e. "childish" paraphrased as "Justin Bieber fan"). These findings suggest that linguistic creativity in humans is affectively and socially embedded, while LLM creativity is optimized toward stylistic coherence and politeness.

The visualization and classification results further support these distinctions: stylistic features proved to be more discriminative than lexical ones. Transformer-based classifiers achieved moderate accuracy (0.75), indicating that while LLM and human language styles differ systematically, overlap still exists. ChatGPT appears stylistically closer to humans, possibly due to reinforcement learning from human feedback and conversational refinements.

Our findings show that although LLMs can mimic human creativity, they lack the emotional and experiential depth.

Limitations

The study relies on a single dataset (Dinu, Florescu, and Resceanu 2025), which, although balanced, covers limited tasks and may not reflect the full range of linguistic creativity. In addition, the strategies of the 48 participants may not generalize to broader populations or newer models. Future work should include larger, more diverse datasets, human evaluation, and cross-linguistic comparisons.

References

[Boden 1996] Boden, M. A. 1996. Chapter 9 - creativity. In Boden, M. A., ed., *Artificial Intelligence*, Handbook of Perception and Cognition. San Diego: Academic Press. 267–291.

[Boden 2004] Boden, M. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

[Boyd et al. 2022] Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin* 10.

[Chkirbene et al. 2024] Chkirbene, Z.; Hamila, R.; Gouissem, A.; and Devrim, U. 2024. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, 229–234.

[Dinu and Florescu 2025] Dinu, A., and Florescu, A.-M. 2025. Testing language creativity of large language models and humans. In Hämäläinen, M.; Öhman, E.; Bizzoni, Y.; Miyagawa, S.; and Alnajjar, K., eds., *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, 426–436. Albuquerque, USA: Association for Computational Linguistics.

[Dinu, Florescu, and Resceanu 2025] Dinu, A.; Florescu, A.-M.; and Resceanu, A. 2025. A comparative approach to assessing linguistic creativity of large language models and humans. *Procedia Computer Science* 270:1292–1301. 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025).

[Durward and Thomson 2024] Durward, M., and Thomson, C. 2024. Evaluating vocabulary usage in LLMs. In Kochmar, E.; Bexte, M.; Burstein, J.; Horbach, A.; Laarmann-Quante, R.; Tack, A.; Yaneva, V.; and Yuan, Z., eds., *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 266–282. Mexico City, Mexico: Association for Computational Linguistics.

[Gatti, Stock, and Strapparava 2021] Gatti, L.; Stock, O.; and Strapparava, C. 2021. *Cognition and Computational Linguistic Creativity*. Cham: Springer International Publishing. 1–39.

[Gehrmann, Clark, and Sellam 2022] Gehrmann, S.; Clark, E.; and Sellam, T. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

[Ismayilzada et al. 2024] Ismayilzada, M.; Paul, D.; Bosselut, A.; and van der Plas, L. 2024. Creativity in ai: Progresses and challenges.

[Jiang et al. 2024] Jiang, X.; Tian, Y.; Hua, F.; Xu, C.; Wang, Y.; and Guo, J. 2024. A survey on large language model hallucination via a creativity perspective.

[Lambert 2025] Lambert, N. 2025. Reinforcement learning from human feedback. arXiv preprint.

[McCoy et al. 2023] McCoy, R. T.; Smolensky, P.; Linzen, T.; Gao, J.; and Celikyilmaz, A. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics* 11:652–670.

[Perez Almendros and Camacho-Collados 2024] Perez Almendros, C., and Camacho-Collados, J. 2024. Do large language models understand mansplaining? well, actually... In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5235–5246. Torino, Italia: ELRA and ICCL.

[Raza et al. 2025] Raza, M.; Jahangir, Z.; Riaz, M. B.; Saeed, M. J.; and Sattar, M. A. 2025. Industrial applications of large language models. *Scientific Reports* 15(1):13755.

[Weller-Di Marco and Fraser 2024] Weller-Di Marco, M., and Fraser, A. 2024. Analyzing the understanding of morphologically complex words in large language models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1009–1020. Torino, Italia: ELRA and ICCL.

[Zanotto and Aroyehun 2024] Zanotto, S. E., and Aroyehun, S. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models.

[Zhao et al. 2025] Zhao, Y.; Zhang, R.; Li, W.; and Li, L. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research* 22(3):417–436.

Acknowledgments

Research supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project SIROLA, number PN-IV-P1-PCE-2023-1701, within PNCDI IV.

Appendix

The eight tasks of the language creativity test:

I. Word/Compound Formation

1. Combine two words to form a new one that describes a specific concept (A. a person completely relying on chatbots, B. a mixed feeling of satisfaction and guilt)
2. Complete a given word with another to form a new compound (A. -ish, B. -able)
3. Continue a certain series of derived words (A....ice, B....summoner)
4. Invent new words that would fit within the same semantic field as a set of words (A. dog, doggy, puppy, canine, bark, paw, woof, sniff, bitch; B. pig, piggy, pork, oink, swine, hog, piglet, snout, boar)

II. Figures of speech

1. Describe familiar ideas by giving novel alternative names (A. angry small dogs; B. the "I wanna speak to the manager" kind of people)
2. Complete blank spaces so as to create an original metaphorical meaning (A. a/some...later; B. a sea of...)
3. Continue the given words with unusual or funny opposites (A. cozy...; B. chaotic...)
4. Give one original harsher (pejorative) expression, one equivalent expression, and one milder (appreciative) expression for a word (A. gossiper; B. childish)

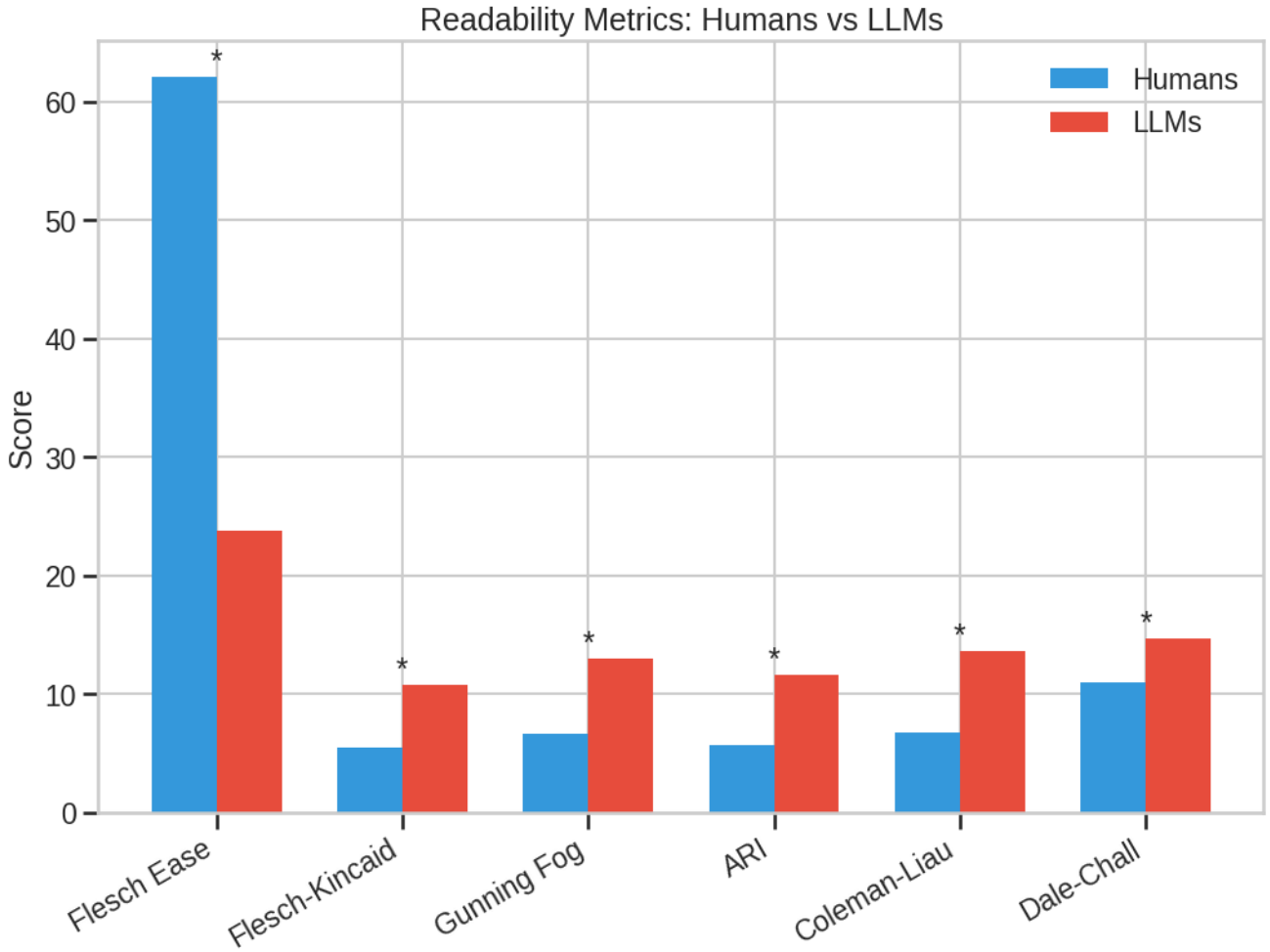
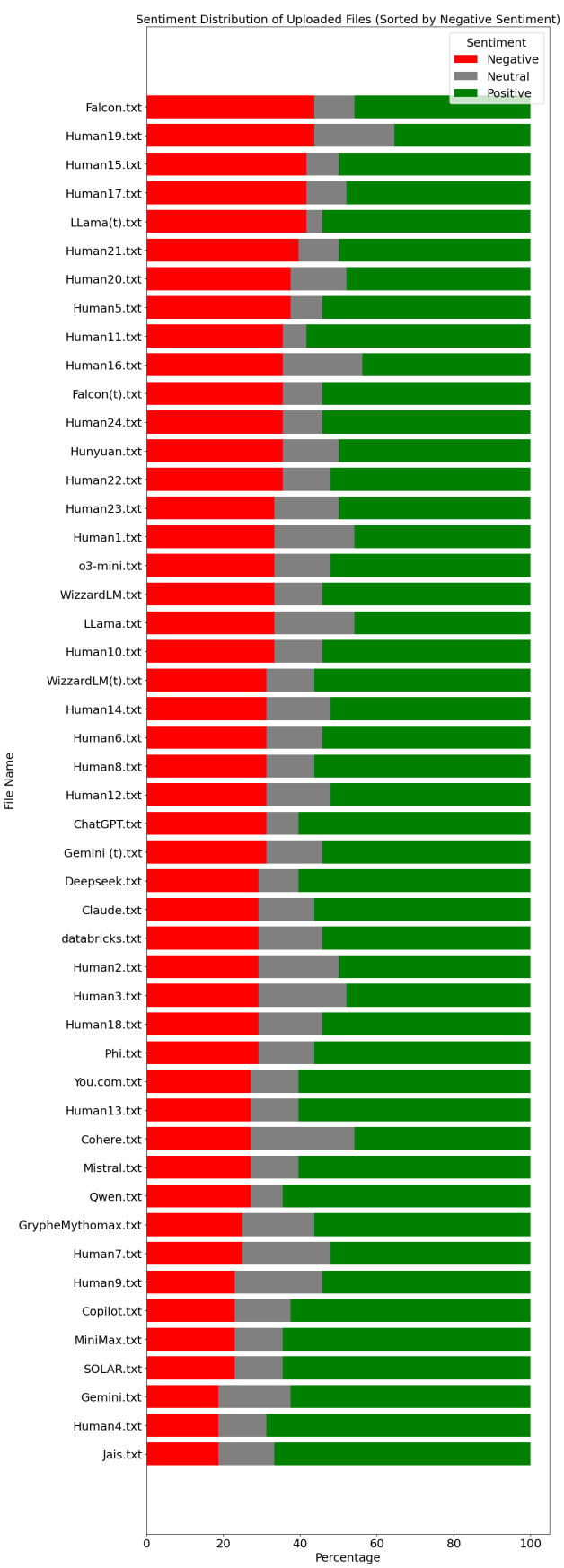


Figure 1: Readability metrics. For the Flesch Ease, the higher the score, the easier the text is to read, as opposed to all other metrics, so the pattern is the same.



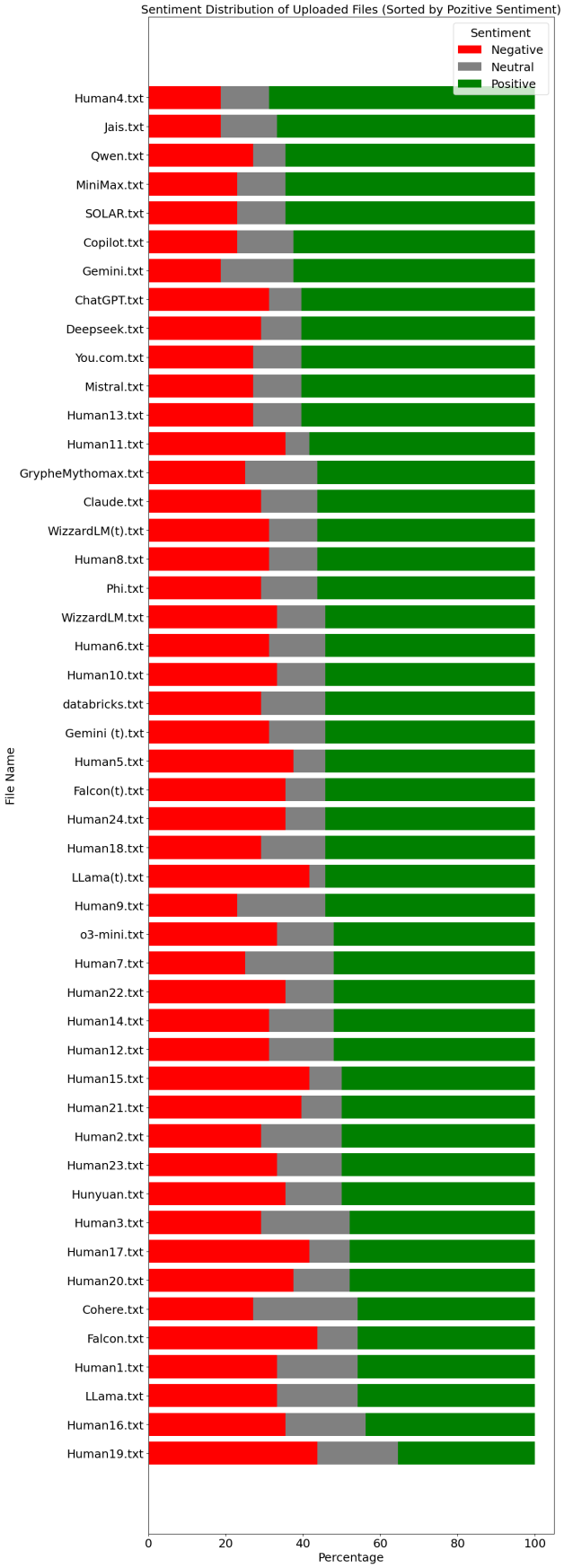


Figure 2: Sentiment scores for all 48 individual participants, ordered by their negative sentiment (left), and by their positive sentiment (right).

Metric	Human Mean	LLM Mean	Diff.	p-value
flesch_kincaid_grade	5.48	10.81	5.33	1.95×10^{-17}
flesch_reading_ease	62.07	23.78	-38.29	2.66×10^{-17}
coleman_liau_index	6.69	13.59	6.90	3.00×10^{-17}
automated_readability_index	5.70	11.56	5.86	3.42×10^{-17}
dale_chall_readability_score	11.02	14.72	3.70	1.08×10^{-14}
gunning_fog	6.65	12.98	6.32	7.04×10^{-13}

Table 6: Comparison of readability metrics between human and LLM-generated texts.

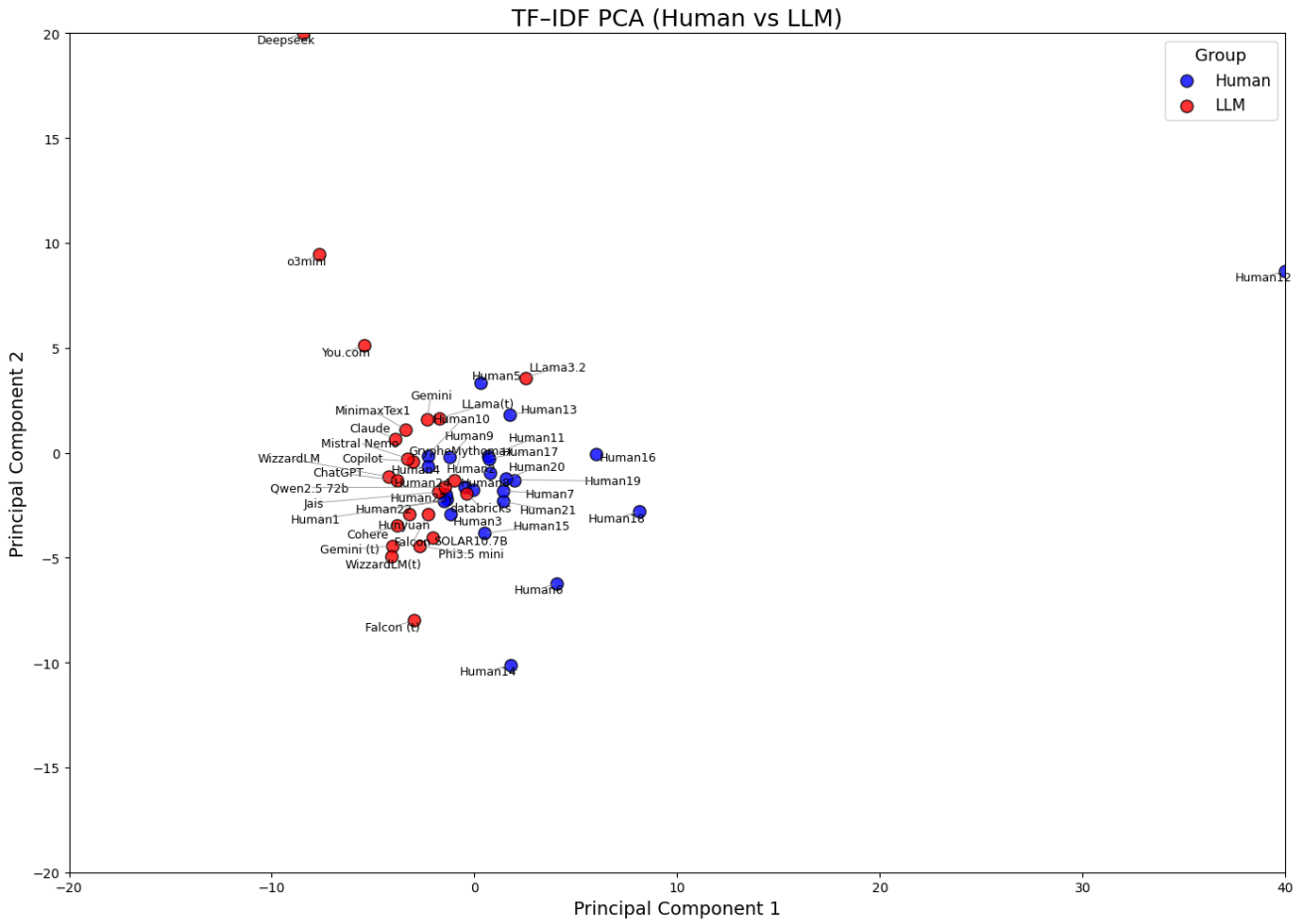


Figure 3: TF-IDF PCA.

	Machine Learning			Deep Learning		
Model	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Humans	0.67	0.71	0.69	0.73	0.80	0.77
LLMs	0.69	0.65	0.67	0.78	0.71	0.74
Accuracy		0.68			0.75	

Table 7: Machine Learning and Deep Learning classification.

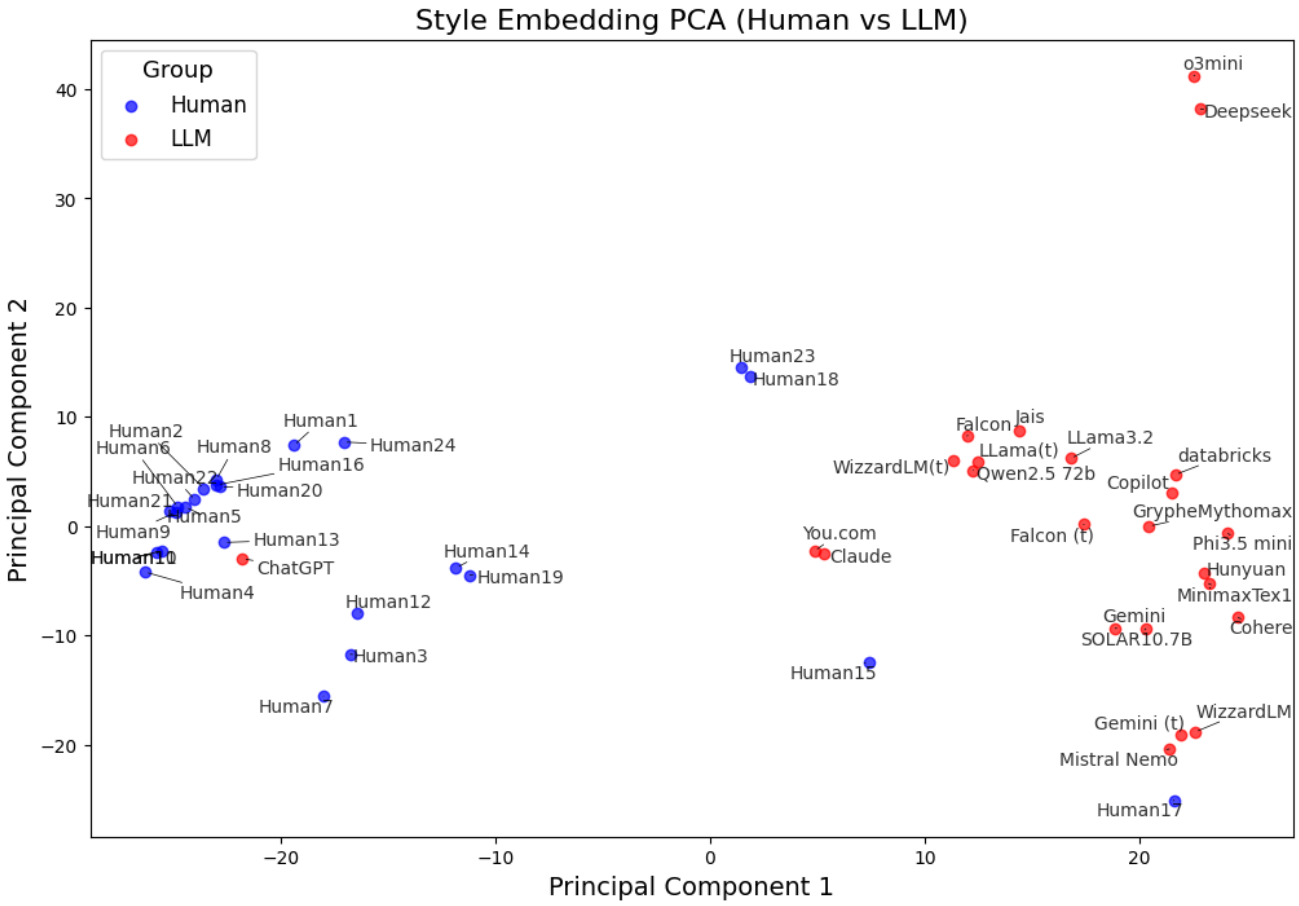


Figure 4: Style embeddings PCA.

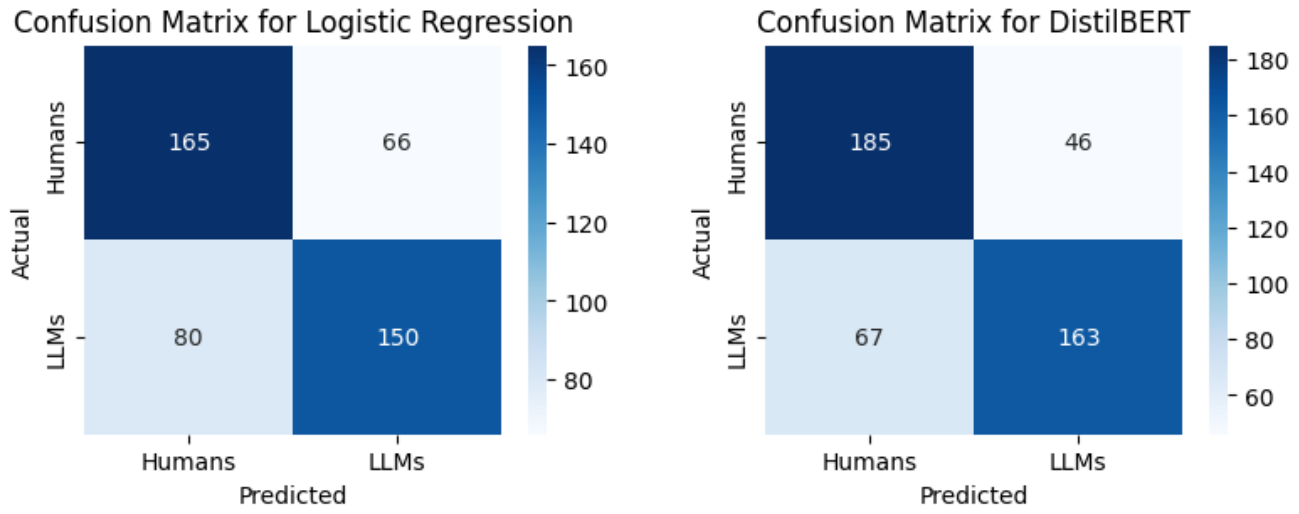


Figure 5: Confusion Matrix for Logistic Regression (left), and for DistilBERT (right).

Back to Back with a Copy: A Computational Analysis of AI-Generated Visual Contemporary Art *Pastiches*

Anca Dinu, Andreiana Mihail, Andra-Maria Florescu, Claudiu Creangă, Liviu Dinu
University of Bucharest, Bucharest, Romania
anca.dinu@lls.unibuc.ro, andreianna@yahoo.com, andra-maria.florescu@s.unibuc.ro,
claudiudancreanga@gmail.com, liviu.p.dinu@gmail.com

Abstract

The aim of this paper is twofold. First, it investigates whether newer generative models are getting "better" at *pastiching* contemporary artworks. Second, it explores the consistency of the multidimensional nature of stylistic evaluation across different LLMs. Building on previous work, we analyze stylistic similarity between AI generated *pastiches* and the original artworks of twelve contemporary artists. We used five complementary computer vision models to capture texture, color, semantics, composition, and perceptual features through cosine distance in high-dimensional embedding spaces. The distances obtained show that the newer image generation model that we used has produced *pastiches* with improved semantic alignment and greater diversity than the model used in previous work. However, it was slightly less performant on shallow features such as color, texture, and perceptual adherence. Our findings confirm that artistic style is inherently multidimensional, and measuring it does not depend on any spatial architecture. These quantitative findings are contextualized through feedback from human evaluators, which are the artists themselves.

Introduction

Over the past decade, Computational Creativity (CC) has experienced rapid growth, largely driven by advances in AI, in particular transformer models, which greatly increased the capacity of computer systems to create original artifacts.

In this work, we focus on a type of visual computer creativity inherently combinatorial in nature, defined in (Boden 2004) as the process of combining ideas from a given conceptual space in novel ways, which minds both convergent creativity (the capacity to identify the best or correct solution to a problem) and divergent creativity (the capacity to generate a wide range of varied original responses) (Guilford 1950), in an assisted creativity framework, defined in (Colton and Wiggins 2012) as the automatic process of supporting a human creator. In particular, we asked the Gemini Pro 3 Nano Banana model (Gemini Team, Google 2023) to generate *pastiches* after contemporary art, via prompting, by uploading the images of the original artworks.

The starting point was the work of (Dinu et al. 2026), who employed Chat GPT to create *pastiches* after the artworks of 12 contemporary artists. The study uncovered that the multi-faceted concept of style can be expressed by features like texture and color, semantics and concept, artistic touches, visual composition, or perceptions, and that these features can be captured independently using different computer vision models which extract high-dimensional embeddings. A large cosine distance between the AI generated *pastiche* and the original human artwork in these embedding spaces means that the *pastiche* was original, but satisfies less the task requirement of creating a *pastiche*, while a low distance indicates a copy that is less original.

We kept the same framework and data, but we used the newer SOTA image generation model Gemini. We were motivated by the following research questions:

1. Are newer generative models getting "better" at producing *pastiches* after human artworks?
2. How do human and automatic evaluation correlate in this respect?
3. Are visual models consistent in extracting stylistic features across different generative AI models?

Our study directly compares the outputs of the two models and gains insight into the stability of the claims made in (Dinu et al. 2026). A vital part of our assessment was the contribution of the artists themselves through their evaluation and comparison.

Related Work

Recently, the fast growing body of work in CC spread to include a wide variety of topics, from improvement techniques and evaluation to ethical concerns and the role of humans in automatic image generation.

For example, (Morain and Ventura 2025) showed that the use of more advanced prompt engineering techniques such as Optimization as PROmpting (OPRO) or chain-of-thought (COT) leads to no significant improvements in textual computational creative systems, while (Colloca and Rezwana 2025) analyzed how ChatGPT amplifies positive sentiment and adds descriptive language regardless of the user's prompt style, showing no correlation between user prompt and AI-generated creative output.

The creative abilities of three artificial image-to-image generative models are formally quantified in (Ramaswamy, Chockler, and Navaratnarajah 2025). The authors propose three objective creativity criteria: requirement satisfaction, cohesion, and novelty and introduce formal measures for

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Table 1: Average Cosine Distances: Gemini vs GPT. Lower values indicate higher similarity.

Model	Gemini	GPT	Δ%	Feature Type
AdaIN-Style	**0.064**	0.063	+1.6%	texture, color
CLIP-ViT-L	**0.167**	0.197	-15.2%	semantic
ResNet50-Style	**0.441**	0.454	-2.9%	artistic style
DINOv2	0.461	**0.463**	-0.4%	f.g. visual
VGG19	0.699	**0.674**	+3.7%	perceptual

Table 2: Model Discrimination and Consistency: Gemini vs. GPT.

Model	Variance		Consistency (Avg Diff)	
	Gemini	GPT	Gemini	GPT
DINOv2	0.0425	0.0349	0.146	0.131
VGG19	0.0082	0.0274	0.087	0.111
ResNet50-Style	0.0160	0.0249	0.089	0.115
CLIP-ViT-L	0.0072	0.0103	0.048	0.046
AdaIN-Style	0.0028	0.0014	0.029	0.034

each of them. The complexity and subjectivity of the evaluation process of CC systems, which depends mainly on the expertise and perception of the evaluator, are recognized in (Boden 1996; Boden 2004; Wiggins 2006). Moreover, (Demke and Ventura 2024; Morain and Ventura 2025) point out that the domain is extremely diverse, making objective human evaluation difficult, time consuming, and costly.

Finally, (Kaila, Holzapfel, and Jääskeläinen 2024) documented how artists negotiate authorship and ownership when experimenting with artificially generated imagery, revealing how they re-frame artistic roles and authorship through generative tools.

Data

This study uses the dataset proposed in (Dinu et al. 2026), which consists of 108 images:

- 36 original artworks by twelve contemporary artists, who generously agreed to lend 3 of their artworks each, for research purposes. The artists are: Adi Matei, Ciprian Mureșan, Ion Grigorescu, Iulia Uță, Karine Fauchard, Lazar Lyutakov, Marius Tănăsescu, Mathias Poeschl, Oana Năstăsache, Philip Patkowitsch, Răzvan Botiș, and Tom Chamberlain;
- 72 AI *pastiches* produced by AI (ChatGPT 5 model), 2 imitations/*pastiches* for each original.

We enriched this image dataset with a new series of AI generated works (e.g., Figures 3, 4, 5, 6 from the Appendix) using the latest Gemini Pro 3 Nano Banana model (Gemini Team, Google 2023). We used the same prompt as in the ChatGPT experiments conducted by (Dinu et al. 2026), so the results would be directly comparable.

Table 3: Extreme Divergence Cases: DINOv2 vs AdaIN. Divergence = $\text{DINOv2}_{\text{norm}}$ - $\text{AdaIN}_{\text{norm}}$. Positive values indicate high texture adherence with high compositional divergence.

Artwork	DINOv2	AdaIN	Divergence
Ciprian Mureșan 1	100%	20%	+0.798
Adi Matei 2	99%	27%	+0.726
Ion Grigorescu 3	81%	11%	+0.697
Tom Chamberlain 3	50%	97%	-0.470
Karine Fauchard 3	15%	42%	-0.270
Oana Năstăsache 2	5%	27%	-0.213

Experimental Setup

In the pre-reprocessing step, we converted all images to RGB format and normalized them according to each of the five models used in (Dinu et al. 2026) to extract high-dimensional embeddings and which have been shown to capture different aspects of artistic style, as follows:

- AdaIN-Style (Huang and Belongie 2017) captures 1920-dimensions of texture/color statistics;
- ResNet50-Style (He et al. 2016) captures 2048-dimensions of artistic style features;
- CLIP-ViT-L (Radford et al. 2021) captures 768-dimensions of semantic features;
- DINOv2 (Oquab et al. 2023) captures 1024-dimensions of fine-grained visual features;
- VGG19 (Simonyan and Zisserman 2014) captures 4096-dimensions of perceptual features.

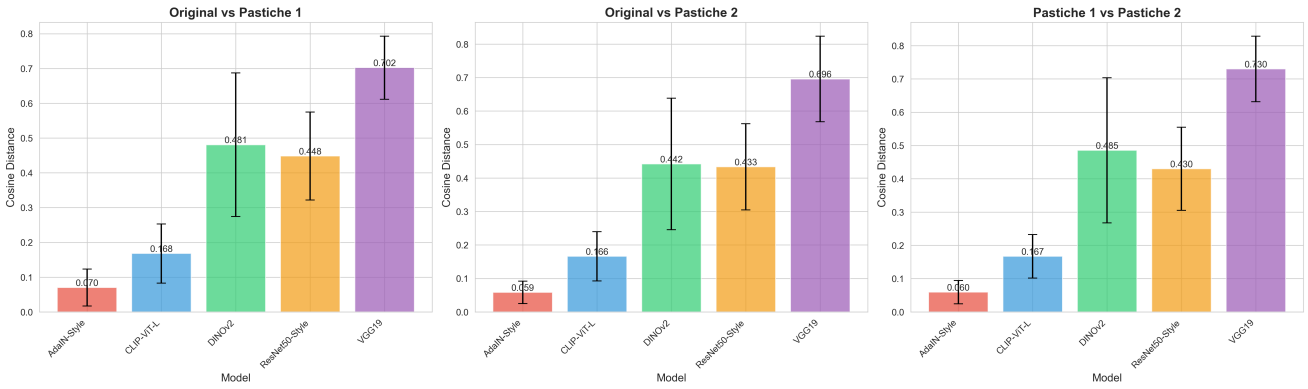
To approximate the stylistic similarity corresponding to each of the five models, for each artwork group (the original and its two Gemini-generated *pastiches*), we compute three pairwise cosine distances: original to *pastiche* 1, original to *pastiche* 2, and *pastiche* 1 to *pastiche* 2.

Automatic evaluation

The distribution of the cosine distance between the original artworks and their Gemini-generated *pastiches* using the five distinct feature embedding models is given in Table 1, in comparison to the GPT results reported in (Dinu et al. 2026). Figure 1 illustrates these distances the Gemini generated *pastiches*, for a more intuitive visual representation.

Comparing ChatGPT and Gemini reveals a consistent model distance hierarchy: AdaIN-Style remains the lowest (0.064), followed by CLIP-ViT-L (0.167), ResNet50-Style (0.441), DINOv2 (0.461), and VGG19 (0.699). This confirms that the "compositional gap" between texture matching and perceptual similarity is model-agnostic and inherent to the task itself. However, the models prioritize different elements (Table 1). Gemini favors deep semantics, better capturing concepts (15.2% closer) and style (2.9%), making it more adept at the overall *pasticheing* task based on cosine distance. In contrast, ChatGPT focuses on surface forms, slightly better imitating perceptual features (3.7% closer)

Figure 1: Average cosine distances between original artworks and Gemini-generated *pastiches* across all five embedding models. Left panel shows Original→*Pastiche* 1 distances, middle shows Original→*Pastiche* 2, right shows *Pastiche* 1↔*Pastiche* 2 diversity.



and texture/color (1.6%). Furthermore, Gemini generates significantly more diverse *pastiches*, as shown in Figure 10 from the Appendix. The average distance between *pastiches* of the same original is 10.1% higher for Gemini than GPT (0.374 vs. 0.340). This diversity is particularly notable in DINOv2 (+22.4%) and VGG19 (+10.3%), indicating that Gemini explores broader compositional and perceptual interpretations.

We also compared the discrimination (variance) and consistency of the two models, given in Table 2. Discrimination reflects the embedding model's capacity to distinctly separate different artistic styles (measured by variance), while consistency indicates the stability of the generative model when producing multiple *pastiches* for the same original artwork (measured by the average distance between them).

DINOv2 remains the most discriminative model for both Gemini (0.0425) and GPT (0.0349), while AdaIN-Style remains the least discriminative. Notably, Gemini shows improved consistency on VGG19 and ResNet50-Style compared to GPT, suggesting more stable perceptual and stylistic reproduction.

The models capture complementary information, as evidenced by the weak correlation ($r = 0.293$) between DINOv2 (composition) and AdaIN-Style (texture). This quantifies the "compositional gap", showing that texture fidelity does not predict compositional accuracy. Table 3 highlights some extreme cases in which these metrics disagree. For example, Ciprian Mureșan's Work 1 shows the largest positive divergence (+0.798), pairing high texture adherence (AdaIN: 20th percentile) with high compositional divergence (DINOv2: 100th percentile). Conversely, Tom Chamberlain's Work 3 displays the largest negative divergence (-0.470), combining moderate compositional adherence (50th percentile) with high texture divergence (97th percentile).

Artists' evaluation

We asked artists to evaluate Gemini's *pastiches* after their artwork, answering the same three questions about the *pastiches* produced by ChatGPT in (Dinu et al. 2026), for a direct comparison:

Q1. To what extent do you recognize your personal artistic language and the coherence of your visual style in this new work? (1 = not at all, 10 = completely)

Q2. How does this new work inspire you or what thoughts does it provoke? (open answer)

Q3. To what extent do you consider that the work generated by Gemini has aesthetic or artistic value? (1 = not at all, 10 = very high)

Similarly with ChatGPT's *pastiches*, the grades for Gemini's *pastiches* varied considerably by artist, as shown in Figure 2. Compared to ChatGPT, Gemini's *pastiches* were perceived by the artists as closer to their style, with a mean grade for the first question of 4.99, representing a significant increase of 1.41 points compared to the mean grade obtained by ChatGPT. However, they were perceived as slightly less artistically valuable, with a mean grade for the third question of 4.26, which is 0.67 points lower than the one obtained by ChatGPT's *pastiches*. Overall, the artists seem to perceive Gemini's pastiches as closer to their works, in line with the computational evaluation, which constitutes a weak positive answer to our second RQ.

Answering the open-questions, the artists offered a variety of opinions about the *pastiches* of their artworks. Most of them regarded the images created by AI more like "artistic comments" than real creations, with little understanding of originality, color mixing techniques, body, and depth. Artistic comments, in this instance, refers to image that functions more as visual interpretation or reactions to the original artworks rather than independent creative works.

We reproduce here some of the answers. Oana Năstăsache says that "some colors look dirty(...) I would never associate that ochre-green combination, because it feels cold and muddy" or that "The surfaces are placed without any sense of composition". Ion Grigorescu remarks that the new *pastiches* followed the concept and composition, but failed in

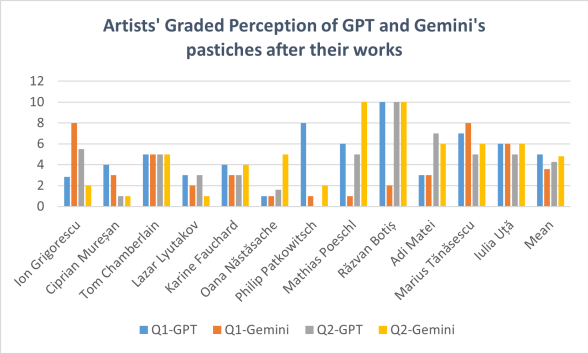


Figure 2: Comparison between artists' graded perception of GPT and Gemini's *pastiches* after their works.

spirit, technique, and style: "Young man rising from a chair. Egon Schiele — closer to forgery than to *pastiche*. A theatricality cheaper than Schiele's". Philip Patkowitsch has a different point of view raising the issue of intellectual property: "they actually present my works better, but just because it is using parts of the original image, so I would say copyright infringement would be an issue!".

So, in the end, did Gemini produce artworks as perceived by the artists themselves? Most of them responded negatively, motivating that AI displayed insufficient command of color theory, which resulted in a poor chromatic atmosphere and the lack of full conceptual comprehension/integration.

Interpretation of the results

The results confirm that artistic style should be evaluated multidimensionally (Dinu et al. 2026). Both GPT and Gemini successfully match texture statistics (low AdaIN distances) but show substantial compositional divergence (high DINOv2 and VGG19 distances), showing that these dimensions are orthogonal: high texture adherence often pairs with high compositional divergence, and vice versa, as illustrated in Figures 7, 8, and 12 from the Appendix. Thus, this "compositional gap" is an intrinsic challenge for AI *pastiche* generation, as visually shown in Figure 11 from the Appendix, confirming that visual models consistently extract stylistic features across different generative AI models, positively respondig to our third RQ.

The models also exhibit distinct generation strategies, pointing to a weak positive answer to oure first RQ:

- **Gemini**: An "interpretive" approach prioritizing semantic alignment (15% closer to CLIP) and higher diversity (+10%).
- **GPT**: A "faithful reproduction" approach prioritizing texture matching and consistency.

Because of the weak correlation ($r = 0.293$) between texture and composition, evaluating AI art requires a multi-metric framework:

- **AdaIN-Style**: Texture/color fidelity.
- **CLIP-ViT-L**: Semantic/conceptual alignment.
- **DINOv2**: Compositional structure.

Empirical Remarks

We observed that Gemini often disregards prompt guidelines, defaulting to the styles of iconic masterpieces (e.g., Egon Schiele or Grant Wood's *American Gothic*). Although this exploratory approach allows for originality, the model often overcompensates, producing works that compete with canonical art rather than complementing the original. Both models exhibit perfectionism, frequently completing works that artists intentionally left unfinished, while significantly altering the original image's foreground.

While Gemini "understands" dimensionality and brushwork, it consistently fails to perceive depth, context, and conceptual irony. A prime example is the *pastiche* of Ciprian Mureșan's *Pioneers*. Mureșan depicted communist youth as rebels and villains to mock 1980s Romanian propaganda. However, both Gemini and ChatGPT missed this irony, independently generating similar images of pioneers as standard patriotic drummers, shown in Figure 9. The striking similarity between the drummers generated by the two models suggests that they were trained on the same visual data.

Finally, Gemini performs better when creating collages and installations compared to paintings. This combinatorial technique aligns with LLM capabilities, though the results often resemble over-accumulation rather than originality.

Limitations and Future Work

Our study is constrained by a small conceptually homogeneous dataset, as most artists belong to the same gallery. Including detailed figurative artwork could provide better insight into LLM *pastiche* performance. Furthermore, comparing outputs across fundamentally different mediums (e.g., 2D paintings versus 3D conceptual installations) presents structural methodological challenges. Finally, photography was excluded because it functions as a mechanical reproduction of reality rather than an artwork.

We plan to adopt alternative metrics beyond the cosine distance that are better suited for evaluating visual similarity. To mitigate personal artist bias, we aim to incorporate formal evaluations by professional art critics. Lastly, we will apply this methodology to iconic, frequently *pastiched* artworks, such as Da Vinci's "Mona Lisa," Van Gogh's "Starry Night," Grant Wood's "American Gothic," and Andy Warhol's "Campbell's Soup Cans".

Ethical Implications

This publication raises no ethical concerns, as all software licenses and artist permissions were strictly respected. However, broader generative AI (GenAI) art introduces significant socio-ecological challenges (Marella, Erukude, and Veluru 2025). Societally, GenAI disrupts creative labor markets and raises issues regarding data exploitation, authorship, and intellectual property rights (Frosio 2023), often reinforcing existing inequalities[1]. These disruptive impacts on cultural heritage and sustainability require a commitment to responsible inclusive cultural development (Roe 2025).

[1]https://www.weforum.org/stories/2025/01/the-impact-of-genai-on-the-creative-industries/

References

[Boden 1996] Boden, M. A. 1996. Chapter 9 - creativity. In Boden, M. A., ed., *Artificial Intelligence*, Handbook of Perception and Cognition. San Diego: Academic Press. 267–291.

[Boden 2004] Boden, M. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

[Colloca and Rezwana 2025] Colloca, G., and Rezwana, J. 2025. Prompting ai in co-creation: The role of syntax and sentiment in shaping ai-generated content. In *Proceedings of the 16th International Conference on Computational Creativity*. International Conference on Computational Creativity.

[Colton and Wiggins 2012] Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012)*, 21–26.

[Demke and Ventura 2024] Demke, J., and Ventura, D. 2024. Overcoming algorithmic bias as a measure of computational creativity. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24)*. Jönköping, Sweden: Association for Computational Creativity.

[Dinu et al. 2026] Dinu, A.; Mihail, A.; Florescu, A.-M.; and Creangă, C. 2026. The art that poses back: Assessing ai pastiches after contemporary artworks. In Machado, P.; Romero, J. J.; and Rebelo, S. M., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 67–83. Cham: Springer Nature Switzerland.

[Frosio 2023] Frosio, G. 2023. Generative ai in court. In Koutras, N., and Selvadurai, N., eds., *Recreating Creativity, Reinventing Inventiveness: International Perspectives on AI and IP Governance*. Routledge. Forthcoming chapter; available at SSRN: `https://ssrn.com/abstract=4558865`.

[Gemini Team, Google 2023] Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. Updated to include Gemini 3, November 2025.

[Guilford 1950] Guilford, J. 1950. Creativity. *American Psychologist*.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

[Huang and Belongie 2017] Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Kaila, Holzapfel, and Jääskeläinen 2024] Kaila, A.-K.; Holzapfel, A.; and Jääskeläinen, P. 2024. Gardening frictions in creative ai: Emerging art practices and their design implications. In *International Conference on Innovative Computing and Cloud Computing*.

[Marella, Erukude, and Veluru 2025] Marella, V. C.; Erukude, S. T.; and Veluru, S. R. 2025. The impact of artificial intelligence on traditional art forms: A disruption or enhancement. *Indian Journal of Computer Science and Engineering* 16(3):54–66.

[Morain and Ventura 2025] Morain, R., and Ventura, D. 2025. Prompt engineering the creativity knob for large language models? In *Proceedings of the 16th International Conference on Computational Creativity*. International Conference on Computational Creativity.

[Oquab et al. 2023] Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Gkioxari, G.; and Bojanowski, P. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[Radford et al. 2021] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

[Ramaswamy, Chockler, and Navaratnarajah 2025] Ramaswamy, A.; Chockler, H.; and Navaratnarajah, M. 2025. Defining and quantifying creative behavior in popular image generators. In Oliveira, H. G.; Spendlove, B.; Gervás, P.; and Ventura, D., eds., *Proceedings of the 16th International Conference on Computational Creativity*. Association for Computational Creativity.

[Roe 2025] Roe, J. 2025. Generative ai as cultural artifact: Applying anthropological methods to ai literacy. *Postdigital Science and Education* 7(4):1107–1124.

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Wiggins 2006] Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Acknowledgments

Research supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project SIROLA, number PN-IV-P1-PCE-2023-1701, within PNCDI IV, and by UB project.

We are grateful to all the artists who agreed to let us use their works and for their insightful feedback: Adi Matei, Ciprian Mureșan, Ion Grigorescu, Iulia Uță, Karine Fauchard, Lazar Lyutakov, Marius Tănăsescu, Mathias Poeschl, Oana Năstăsache, Philip Patkowitsch, Răzvan Botiș, and Tom Chamberlain.

Appendix

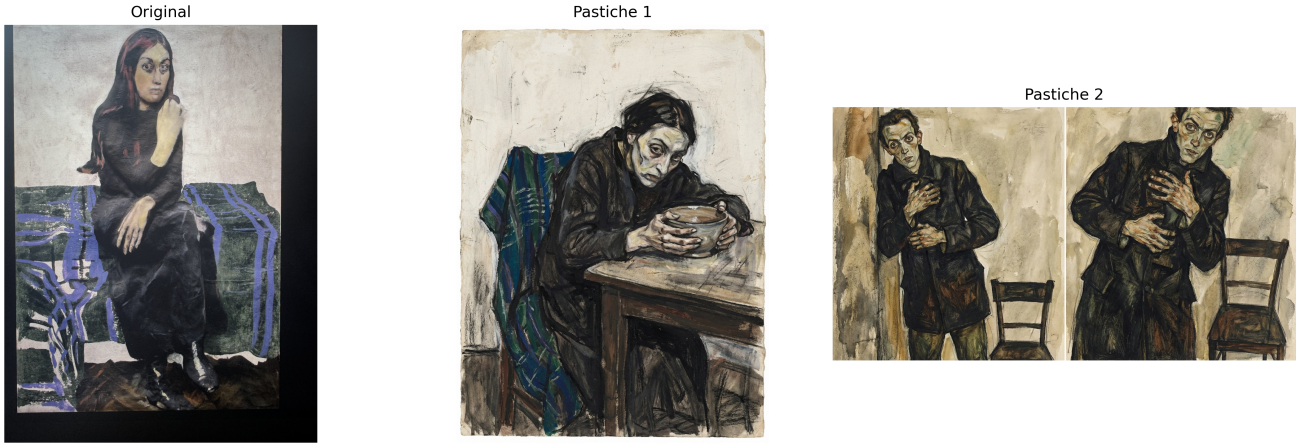


Figure 3: Examples of Gemini's *pastiches* after Ion Grigorescu's work 1 "Măriuca Iosifescu", oil on canvas and plaster, 1974.



Figure 4: Examples of Gemini's *pastiches* after Ion Grigorescu's work 2 "Avram Iancu și Nicolae Bălcescu", oil on canvas and plaster, 1974.



Figure 5: Examples of Gemini's *pastiches* after Iulia Uța's work 2 Untitled, oil on canvas, 2024.



Figure 6: Examples of Gemini's *pastiches* after Lazar Lyutakov's work "24/7 the human condition", installation, 2014.

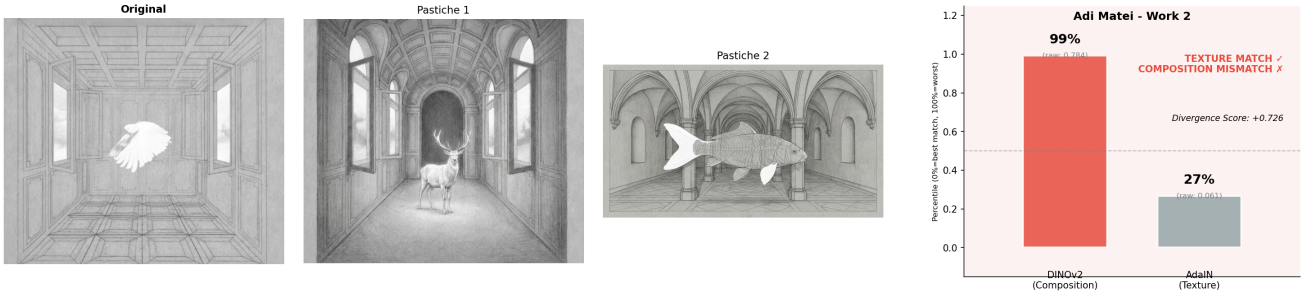


Figure 7: High texture adherence (low AdaIN), high compositional divergence (high DINOv2). These cases illustrate the core of the "compositional gap" and the orthogonality of style dimensions.

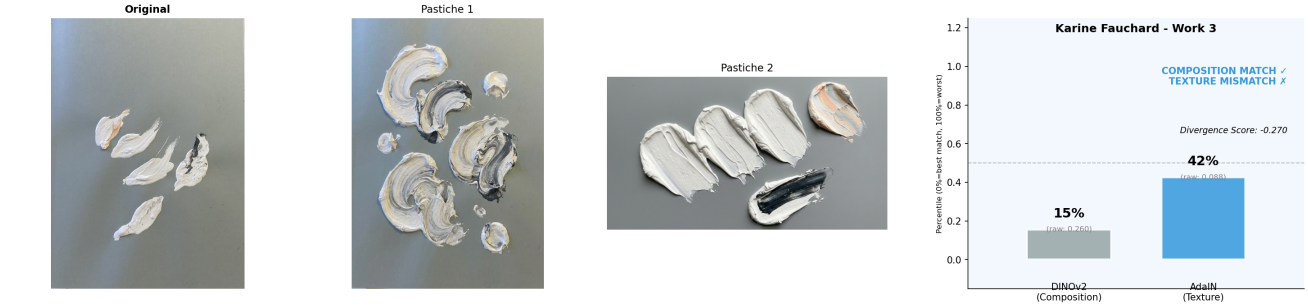


Figure 8: High compositional adherence (low DINOv2), high texture divergence (high AdaIN). Together with Figure 7, this confirms texture and composition as orthogonal dimensions.



Figure 9: GPT (middle) and Gemini (right) created similar *pastiches* of Mureșan's work (left)- suggest similar training data.

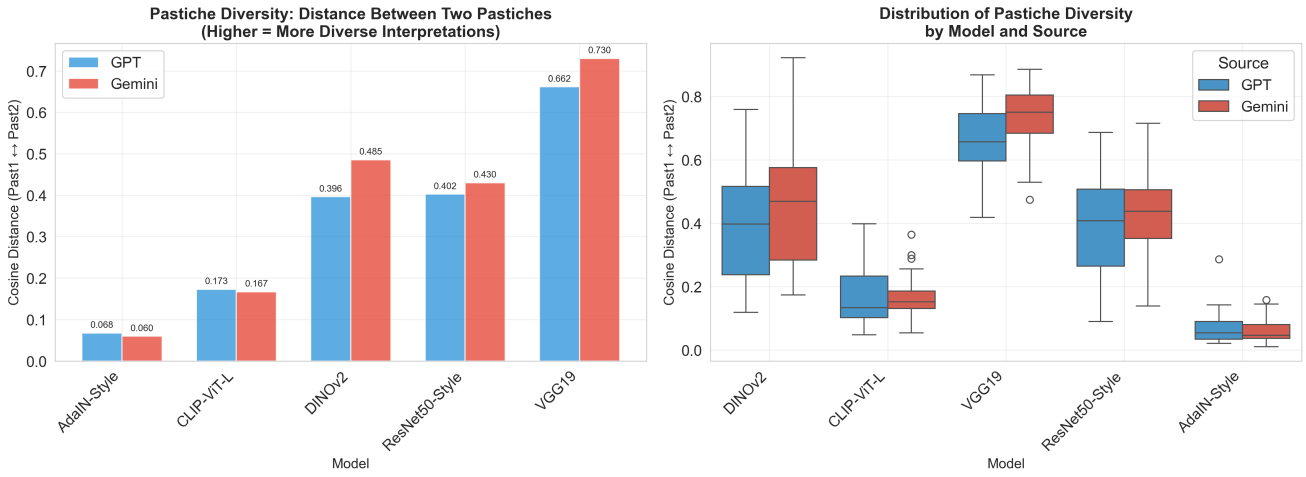


Figure 10: High *pastiche* diversity (left) indicate the model explores more varied interpretations rather than producing near-identical outputs. The distribution boxplots (right)shows that Gemini produces overall 10.1% more diverse *pastiches*, with the largest differences in DINOv2 (+22.4%) and VGG19 (+10.3%), suggesting greater compositional and perceptual variety.

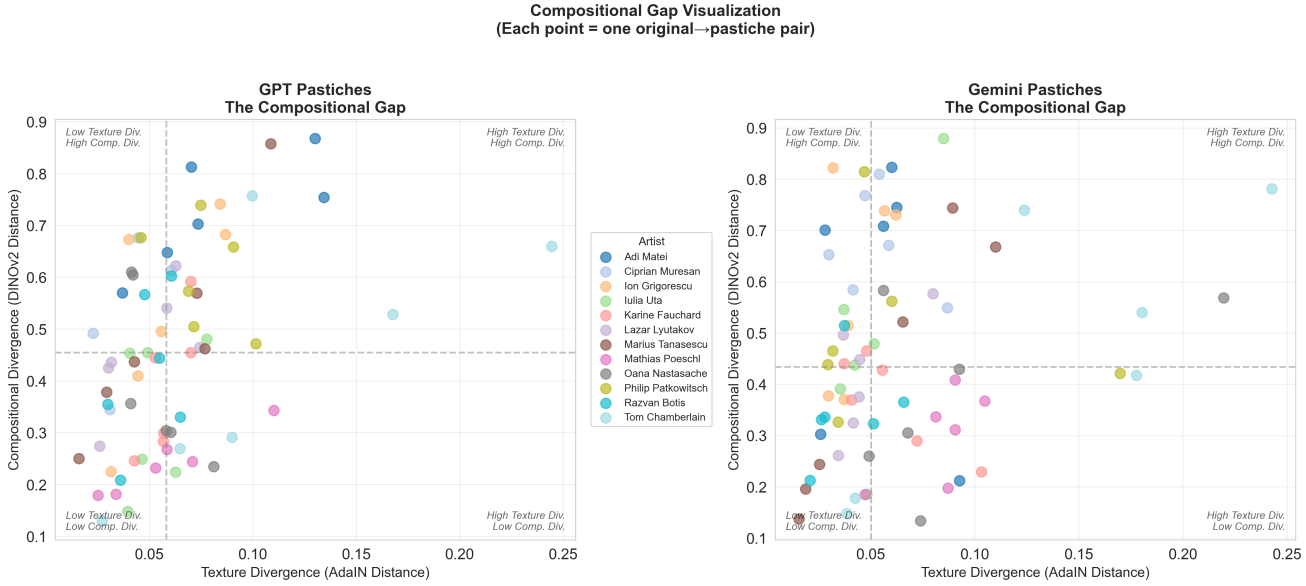


Figure 11: Texture divergence (AdaIN) vs. compositional divergence (DINOv2) for GPT (left) and Gemini (right) *pastiches*. Each point represents one original→pastiche pair, colored by artist. Dashed lines indicate median values, dividing the space into four quadrants. Both models exhibit the same characteristic pattern: points cluster toward the left (low texture divergence, typically 0.03–0.10) but spread widely along the y-axis (compositional divergence ranging from 0.2 to 0.9). This asymmetric distribution—narrow horizontally, broad vertically—is the visual signature of the compositional gap: AI pastiches consistently adhere to texture/color statistics while showing high variability in compositional structure. The upper-left quadrant ("Low Texture Div. / High Comp. Div.") is densely populated for both models, representing the dominant generation pattern. Artist-specific clustering is evident: Adi Matei (dark blue) and Ciprian Muresan (light blue) concentrate in high compositional divergence regions, while Oana Năstăsache (gray) and Mathias Poeschl (magenta) cluster toward lower compositional divergence. The similar distributions across GPT and Gemini confirm the compositional gap is model-agnostic.

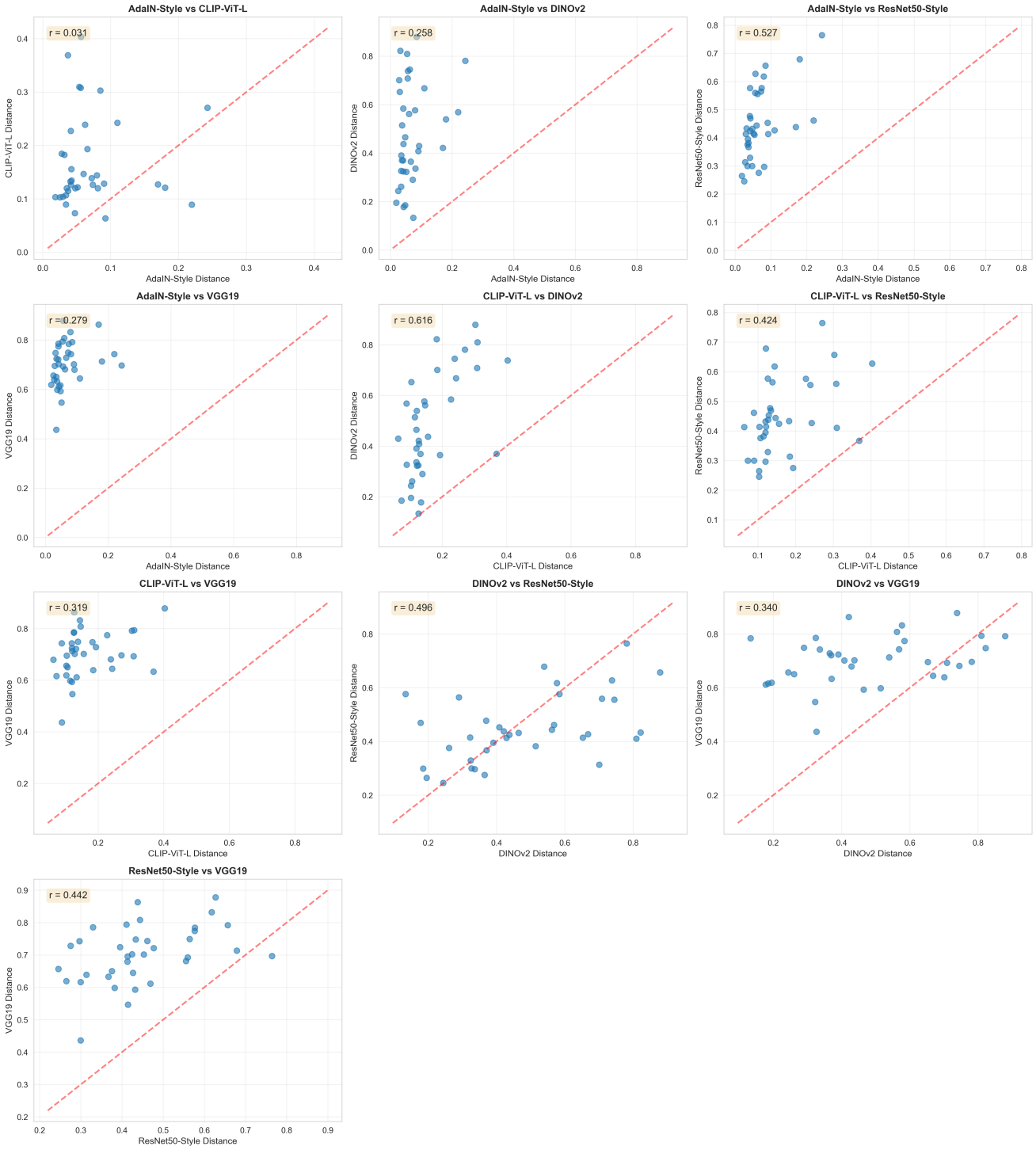


Figure 12: Pairwise correlations between embedding model distances. Each scatter plot shows the cosine distances from one model (x-axis) versus another (y-axis) across all 36 artist-*pastiche* pairs, with the Pearson correlation coefficient (r) displayed. The red dashed line indicates the identity diagonal. Key findings: (1) AdaIN-Style and CLIP-ViT-L show near-zero correlation ($r = 0.031$), confirming that texture statistics and semantic content are orthogonal dimensions; (2) The highest correlation is between CLIP-ViT-L and DINOv2 ($r = 0.616$), reflecting their shared emphasis on higher-level visual features; (3) AdaIN-Style correlates most strongly with ResNet50-Style ($r = 0.527$), as both capture style-related properties; (4) VGG19 shows weak-to-moderate correlations with all other models ($r = 0.28$–0.44), suggesting it captures a distinct perceptual dimension. The generally weak correlations (mean $r \approx 0.37$) confirm that the five models capture complementary rather than redundant information, justifying the multi-metric evaluation framework.

Spot the Pipeline: Evaluating Perceptual Convergence Among Text-to-Image Tools

Martin Disley
Institute for Design Informatics
University of Edinburgh
m.disley@ed.ac.uk

Abstract

The open text-to-image ecosystem is often treated as a pluralising counterweight to the homogenising tendencies of foundation models, since it offers users a wealth of checkpoints and variants that appear to promise stylistic diversity. This short paper asks whether that apparent diversity is perceptually legible in practice. This short paper presents a pilot study in which 15 power users completed a forced choice identification task of images generated by five popular text-to-image tools, using stylistically underspecified prompts. The central result is that recognisability was unevenly distributed: the most experienced users could separate architecturally distinct models, but even for them popular variants of the Stable Diffusion family were not separable. Taking perceptual identification as a proxy for aesthetic distinction, this pattern complicates the assumption that a proliferation of variants fine-tuned by end users straightforwardly translates into meaningful stylistic diversity across tools.

Introduction

The contemporary text-to-image landscape presents users with an abundance of checkpoints, model variants, and adaptors, especially around open base model families such as Stable Diffusion, through repositories such as Hugging Face, CivitAI, and Replicate. This abundance implies stylistic diversity: if one model produces stale or overfamiliar imagery, another should offer a meaningfully different visual character. The abundance of named options, however, does not necessarily translate into an abundance of distinct outputs. For an ecosystem like this to flourish, the available options should provide meaningful, or at least perceptibly different, aesthetic territory under ordinary use.

If the proliferation of checkpoints is understood to reflect a diverse generative image culture, then the key empirical issue is whether different tools are actually distinguishable in practice. This short paper proposes users' consistent ability to recognise or separate images from various models as a minimal proxy for meaningful divergence in visual character. For text-to-image systems, this problem is complicated by the fact that sophisticated prompting can push a model far from its defaults. This study uses deliberately underspecified prompts as a controlled probe of each tool's house style: the characteristic visual tendencies that become visible when users provide little explicit stylistic steering. The question, then, is whether the apparent abundance of available tools yields perceptually distinct options at the point of use.

This question is addressed here through a pilot study in which power users of ComfyUI, a node-based image making software, attempted to identify which of five text-to-image pipelines popular on the platform had generated a given image. The study uses this task to ask whether recognisability is evenly distributed across the sampled tools, or whether apparent abundance masks structured perceptual convergence. We interpret the resulting pattern as evidence that model abundance should not be taken as a proxy for stylistic diversity across tools without examining whether those tools are perceptually distinct in practice.

Related Work

Questions of diversity and convergence have become increasingly important in computational creativity and adjacent discussions of generative AI. At a broad level, recent work has challenged the assumption that more generated content necessarily means more meaningful variation, instead suggesting that generative systems can stabilise conventions at scale (Doshi and Hauser 2024; Padmakumar and He 2024; Anderson, Shah, and Kreminski 2024; Sarkar 2024). In text-to-image systems, this concern is visible both in critical accounts of homogenised visual culture and in analyses of prompt corpora that show repeated aesthetic patterns and conventions emerging in practice (McCormack et al. 2024). This paper takes this broader concern as background, but narrows it to a more specific ecosystem question: whether the abundance of checkpoints and model variants actually yields perceptibly different outputs.

In text-to-image systems specifically, prompting is iterative and skilled rather than transparent or one-shot (Oppenlaender 2022; Ibarrola and Grace 2023; Lichtenberg and Akdag 2025). This matters because heavily engineered prompts can push a system far from its defaults, whereas underspecified prompts expose the output tendencies users inherit when they do not actively steer. Despite this, analyses of prompt corpora suggests that repeated aesthetic conventions emerge in practice (McCormack et al. 2024), making default behaviour a meaningful object of study rather than a trivial baseline.

Within computational creativity, related work has there-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

fore focused on how users navigate text-to-image systems, how style can be framed within creative applications, and how diversity can be evaluated within a single system rather than across named tools (Gajula et al. 2024; Ibarrola and Grace 2024). Adjacent work on CLIP embedding spaces and stylistic similarity helps us reason about images and prompts (Radford et al. 2021; Somepalli et al. 2024), while workflow studies show that generative image tools can differ in consistency, controllability, and refinement behaviour in practice (Ocampo Blanco and Bown 2024). Yet this leaves underexamined a broader ecosystem-level assumption: that limited stylistic range within any one model may be unproblematic if a plurality of aesthetic options exists across the wider field of tools. Concerned with evaluating ecosystem stylistic diversity, this pilot study investigates cross-model perceptual distinctiveness using underspecified prompting as a 'house style' probe.

Method

Here, this question is addressed with a forced-choice identification study deployed online as *Spot the Pipeline*. In each of the scored trials, participants were shown a single generated image and asked to identify which of five named text-to-image tools had produced it. The prompt text was visible on each trial, and participants could view the principal run parameters for the sample, but no example images or training phase were provided before the task began. The task comprised 25 scored trials, constructed from five underspecified prompts crossed with five tools, plus one duplicate-image attention check. After completing these scored trials participants complete a short experience questionnaire, which was required to reveal their score. Performance was evaluated against a five-way chance baseline of 20%.

The roster curation was informed by model and checkpoint popularity as given from benchmarks/leaderboards and observed conversations on forums and chat servers. It combined two architecturally distinct non-Stable-Diffusion tools with three models from the Stable-Diffusion family: FLUX.1-schnell, Z-Image Turbo, Stable Diffusion XL base, JuggernautXL v9, and RealisticVision v5.1. The three Stable-Diffusion variants represent related but distinct parts of the SD ecosystem: SDXL base is the official SDXL foundation model, JuggernautXL v9 is a community SDXL-family checkpoint, and RealisticVision v5.1 is an SD1.5-family realism checkpoint run with its recommended VAE, negative prompt, and hires/upscale pass. This mix allowed us to compare two non-SD tools with a set of widely used SD-derived pipelines distributed as distinct options in practice. Importantly, these were treated as tools-as-accessed rather than models as isolated artefacts: each tool was run using a fixed preset intended to reflect typical community use. In practice, this means that the study compares user-facing pipelines, not only abstract base-model behaviour.

Stimuli consisted of 25 images generated from five deliberately underspecified prompts crossed with the five tools. The prompts were: *a portrait of a person*, *a cosy living room interior*, *a city street at night, rain*, *a still life of a bowl of fruit on a table*, and *a small robot on a desk*. Fixed seeds were used within each prompt condition so that the sample

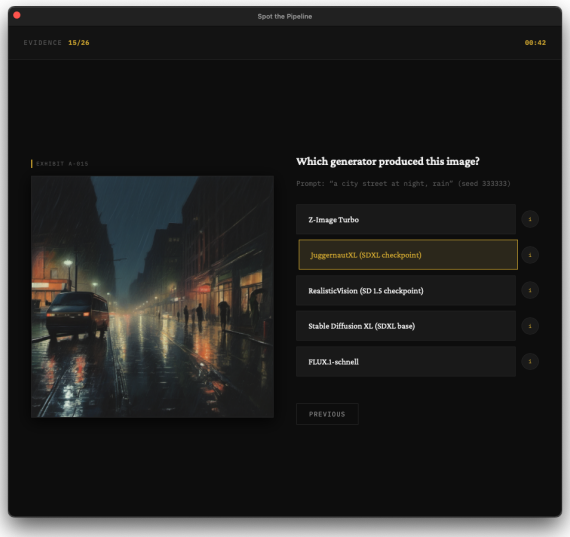


Figure 1: Spot the Pipeline interface: Participants viewed one generated image at a time and selected which of five named text-to-image tools they believed had produced it. The prompt text was shown, and principal run parameters could be viewed by clicking the "info" button

set was defined procedurally and comparison across tools was controlled at the level of prompt and stochastic initialisation.

Participants were recruited from ComfyUI communities on the messaging platform Discord. This power user sampling strategy was informed by an informal pre-study pilot with colleagues who had little or no regular experience using text-to-image systems. Those pilot versions tested whether the task was viable for inexperienced participants under both named-tool and less explicitly labelled versions of the quiz. In both cases, inexperienced participants performed at or below chance, suggesting that the task required more practical familiarity with contemporary text-to-image tools than a general audience could be assumed to have. We therefore treated the final deployment as a power-user baseline: a test of whether people with active practical exposure to current text-to-image workflows could detect tool-specific output signatures without a training phase, with the assumption that users with less experience would not have this capacity. The principal data points measured were overall accuracy relative to chance, per-tool identifiability, confusion patterns across tools, and associations between performance and self-reported experience.

Thirty submissions were recorded, of which 15 contained complete responses and were retained for analysis. Attrition occurred mostly at the start of the quiz - the incomplete submissions recorded a mean of 3.6 scored responses. Among the incomplete submissions, pooled partial accuracy was 18.9%, approximately at the 20% chance level. Like the novice users who attempted the informal pilot study, the low scores of the dropout suggest that they found the task ex-

ceedingly difficult. This filtering likely selected for participants who were more confident in their capacity to succeed in the task.

Results

Participants identified tools significantly above chance overall. Across the 15 completed submissions, mean accuracy was 30.7% against a chance baseline of 20%, corresponding to an average score of 7.67 out of 25. A one-sample test against chance showed that this effect was statistically significant ($t(14) = 2.87$, $p = .006$, with Cohen's $d = 0.74$). This indicates that some signal of tool identity was perceptible in the outputs by this group of power users without participants any explicit training phase.

This signal was not evenly distributed across the five tools. FLUX.1-schnell and Z-Image Turbo were the most recognisable, with accuracies of 48.0% and 45.3% respectively, both well above the 20% chance level. By contrast, the Stable-Diffusion-family cluster was only weakly separable: SDXL base was identified at 24.0%, JuggernautXL at 22.7%, and RealisticVision at 12.0%, the latter falling below chance. The principal empirical implication of this is therefore not simply that power-users can identify tools, but that recognisability is uneven and concentrated in a subset of them.

The confusion matrix shows in Figure 2 shows that this pattern is structured rather than random. FLUX and Z-Image each maintained relatively clear identities, whereas SDXL, Juggernaut, and RealisticVision were systematically confused with one another. The strongest off-diagonal confusion was RealisticVision being identified as Juggernaut (28%), which occurred more often than RealisticVision being correctly identified at all (12%). Juggernaut, in turn, was most often misidentified as Z-Image (27%). Taken together, these confusions indicate that the key result is an Stable Diffusion-family cluster whose members do not function as clearly distinct perceptual options under the conditions tested.

Experience also mattered, but in a specific way. Accuracy was positively associated with hours per week spent generating images (Spearman $p = 0.62$, $p = .015$), whereas experience with text-to-image tools was not significantly associated with performance ($p = 0.22$, $p = .43$). This suggests that the task rewards active, current engagement with the latest suite of models, more than simple accumulated exposure to image making with generative AI/ML. Some prompts made the tools easier to distinguish: portrait images were the strongest discriminator, whereas the still-life prompt was effectively at chance. This indicates either that some prompt types surface default output signatures more effectively than others or that users are more familiar with portrait images and are better able to identify the output signature in those images. This remains an open question and a potential avenue for further work.

Discussion

The significance of these results is not primarily that individual participants can identify some models above chance.

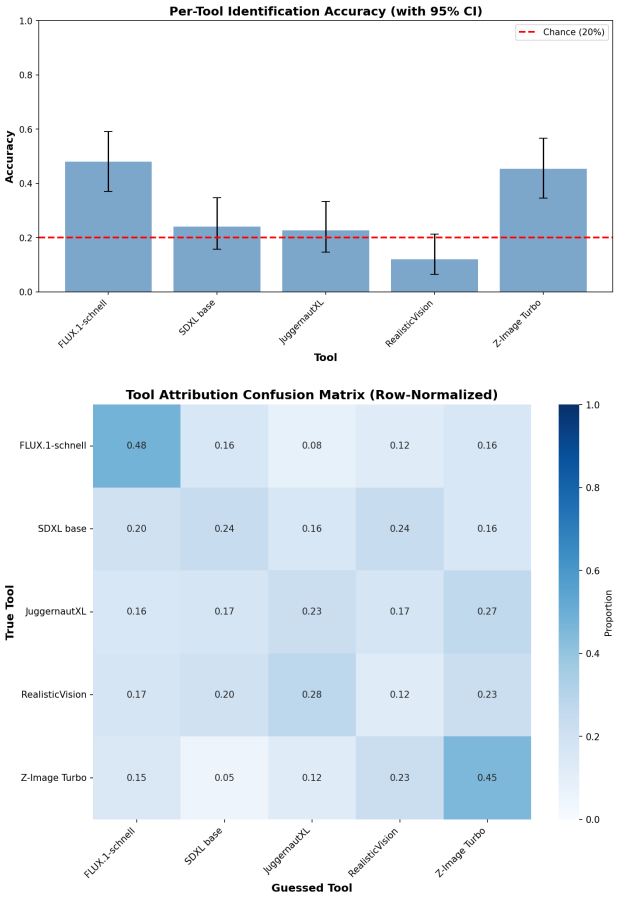


Figure 2: Perceptual recognisability of five text-to-image tools under default, underspecified prompting conditions. Top: per-tool identification accuracy, with chance performance at 20%. Bottom: confusion matrix showing that recognisability was unevenly distributed: FLUX.1-schnell and Z-Image Turbo were relatively distinct, whereas SDXL, JuggernautXL, and RealisticVision formed a structured confusion cluster.

Rather, the study uses recognisability as an ecosystem-level probe. Open and community-driven generative-image ecosystems are often positioned as a pluralising counterweight to the homogenising tendencies of centralised foundation models. That promise depends, however, on whether the proliferation of named checkpoints and pipelines actually presents perceptibly diverse options in practice. Our results complicate this assumption: while FLUX.1-schnell and Z-Image Turbo were relatively recognisable, the Stable-Diffusion-derived pipelines formed a structured confusion cluster. This suggests that checkpoint proliferation alone should not be equated with meaningful stylistic plurality.

More precisely, this study should not be read as a pure test of style in a narrow art-historical sense. It measures the perceptual distinctiveness of outputs generated under stylistically underspecified prompting conditions, which is closer

to perceptual recognisability than to style attribution alone. The term *output signature* is useful here because it captures the broader bundle of cues participants may have been using: rendering tendencies, composition, fidelity, artefacting, prompt adherence, and other pipeline specific behaviours as well as style. This studies results supports a narrower claim: that the house styles or output signature of some models are distinct and perceivable by the most experienced users in images generated without specific style guidance, while others models are not, and thus suggests that these house style converge.

Given this, we would caution interpreting the recognisability of FLUX and Z-Image as indicative of them being stylistically distinct, in the art-historical sense. Differences in model architecture, scale, and optimisation target mean the separability observed here may partly reflect quality-related or optimisation-related cues rather than stylistic difference alone. Differences in fidelity, or characteristic artefacts could all contribute to a recognisable output signature. This is not necessarily a weakness of the study. Because users encounter tools-as-accessed rather than purified latent style manifolds, a user-facing evaluation should remain sensitive to the full bundle of perceptible properties through which one tool comes to feel different from another.

Additionally, a lack of ecosystem stylistic diversity would not imply that community checkpoints or fine-tuned variants offer no value. A checkpoint may be useful because it is more reliable on certain prompt classes, easier to work with, cheaper to run, faster, or better integrated into an existing workflow. It may also help users reach a desired stylistic output with less friction, even if its default outputs are not cleanly separable in a blind identification task. The present result is therefore narrower: within the conditions tested here, these variants did not present as strongly distinct perceptual options. That is a meaningful qualification for ecosystems in which checkpoints are frequently marketed or selected as though each one opens a clearly different visual regime.

For computational creativity research, the implication is that model abundance is a weak proxy for meaningful diversity. The next step should move beyond further perceptual tests and take a dual approach. First, directly measuring image diversity across much larger samples of images and prompts would help quantify image diversity. Second, a qualitative study engaging practitioners in questions regarding tool choice would help us understand the broader ecosystem value of model plurality, beyond stylistic diversity.

Conclusion

This pilot study shows that, under underspecified prompting, active and experienced power users can perceptually recognise some text-to-image tools, and that many apparently distinct variants are not recognisable. In our sample, architecturally distinct models were distinguishable, and Stable Diffusion family variants formed a confusable cluster. The significance of this finding concerns ecosystem diversity rather than individual capacity. Open and community driven model ecosystems are often treated as a pluralising counterweight to the homogenising tendencies of centralised foundation models. That promise depends on whether proliferating checkpoints and pipelines produce perceptibly distinct creative options in practice. Our results suggest that the abundance of named checkpoints does not automatically provide meaningful stylistic diversity. For computational creativity, this means that the proliferation of open text-to-image models should not be assumed to mitigate homogeneity in generative AI media without evidence that they produce genuinely distinct outputs.

References

Anderson, B. R.; Shah, J. H.; and Kreminski, M. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Creativity and Cognition*, 413–425. Chicago IL USA: ACM.

Doshi, A. R., and Hauser, O. P. 2024. Generative artificial intelligence enhances creativity but reduces the diversity of novel content.

Gajula, P.; Abhishek Dangeti; Vivek Srivastava; and Vikram Jamwal. 2024. Style-Frame: A Foundational Framework for Artistic Style Driven Applications. In Grace, K.; Llano, M. T.; Martins, P.; Hedblom, M. M.; Cardoso, A.; Association for Computational Creativity; and Jönköping University., eds., *Proceedings of the Fifteenth International Conference on Computational Creativity: ICCC'24, Jönköping, Sweden, 17-21 June*. Coimbra: Association for Computational Creativity (ACC).

Ibarrola, F., and Grace, K. 2023. Prompt diversification for iterating with text-to-image models. In Pease, A.; Cunha, J. M.; Ackerman, M.; Brown, D. G.; and Association for Computational Creativity., eds., *Proceedings of the Fourteenth International Conference on Computational Creativity: ICCC'23, Ontario, Canada, 19-23 June*. Coimbra: Association for Computational Creativity (ACC).

Ibarrola, F., and Grace, K. 2024. Measuring Diversity in Co-creative Image Generation. In *Proceedings of the 15th International Conference on Computational Creativity*.

Lichtenberg, S., and Akdag, A. 2025. A Creativity Assessment Scale for Text-to-Image Prompting: Challenges & Observations. In Oliveira, H. G.; Spendlove, B.; Gervás, P.; and Ventura, D., eds., *Proceedings of the Sixteenth International Conference on Computational Creativity: ICCC25*.

McCormack, J.; Llano, M. T.; Krol, S. J.; and Rajcic, N. 2024. No Longer Trending on Artstation: Prompt Analysis of Generative AI Art. In Johnson, C.; Rebelo, S. M.; and Santos, I., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 279–295. Cham: Springer Nature Switzerland.

Ocampo Blanco, R., and Bown, O. 2024. Integrating Generative AI into Creative Workflows: Dealing with Consistency, Scene Control, and Refinement in a Professional Image Generation Case Study. In Grace, K.; Llano, M. T.; Martins, P.; Hedblom, M. M.; Cardoso, A.; Association for Computational Creativity; and Jönköping University., eds., *Proceedings of the Fifteenth International Conference on Computational Creativity: ICCC'24, Jönköping, Sweden, 17-21 June*. Coimbra: Association for Computational Creativity (ACC).

Oppenlaender, J. 2022. The Creativity of Text-to-Image Generation. In Hedblom, M. M., ed., *Proceedings of the Thirteenth International Conference on Computational Creativity: ICCC'22*. Coimbra: Association for Computational Creativity (ACC).

Padmakumar, V., and He, H. 2024. Does Writing with Language Models Reduce Content Diversity?

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Sarkar, A. 2024. Intention Is All You Need.

Somepalli, G.; Gupta, A.; Gupta, K.; Palta, S.; Goldblum, M.; Geiping, J.; Shrivastava, A.; and Goldstein, T. 2024. Measuring Style Similarity in Diffusion Models.

In-Person First, Not In-Person Only
Reflections on Remote Participation through the case of ICCC'25 Watch Party
(Paper type: Event Report)

Luís Espírito Santo
CMS – CISUC, University of Coimbra, Portugal
CCLab – AILab, Vrije Universiteit Brussel, Belgium
lesanto@dei.uc.pt, luis.espirito.santo@vub.be

Abstract

This paper presents the outcomes and reflections from the ICCC'25 Watch Party initiative, a lightweight and community-driven effort to include remote participants during the 16th International Conference on Computational Creativity (ICCC'25) taking place in Campinas, Brazil, from 23rd to 27th June. The conference was initially organised exclusively with in-person attendance in mind, but, similarly to previous years, the sessions were streamed online. We describe the background and motivation for this initiative, its structure, lessons learned from organising and from feedback collected through a questionnaire, and suggestions for possible future low-cost hybrid strategies.

Introduction

The 16th International Conference on Computational Creativity (ICCC'25)[1] was held in Campinas, Brazil, from 23rd to 27th June. Although the conference was organised with an exclusive focus on in-person attendance, it followed the tradition of previous years by streaming the sessions online. Due to several external factors, extra remote adaptations were necessary in this edition to accommodate several people who were unable to attend in person.

One of those adaptations designed to improve the experience of remote attendees was actually initiated by the community itself rather than being officially organised by the conference organisers. Building upon the streaming infrastructure already planned to be provided, the ICCC'25 Watch Party was proposed as an informal and low-effort initiative to encourage engagement among those who could not attend in person. Participants gathered on the ACC Discord server to co-watch sessions, share impressions, and stay connected with the broader community in real-time.

This report presents the outcomes and reflections from the ICCC'25 Watch Party initiative. It aims to document the structure of the initiative, reflect on its execution, and evaluate its impact through feedback collected via a post-event questionnaire. It also proposes lightweight strategies for future iterations of ICCC and similar conferences to support meaningful remote engagement.

We argue that, while the specific obstacles that justified remote engagement were more visible this year, they are systemic and recur annually. Yet, the existing fear that remote and decentralised models might weaken in-person attendance – which we believe may be misguided – could in fact be contributing to disengagement within the community. Thoughtfully designed solutions following the principle of **In-Person First, Not In-Person Only** could support inclusion without undermining onsite attendance, both fostering continuity and strengthening the ICCC community.

The report is structured as follows: it first contextualizes remote and hybrid conferences and provides the background and implementation of the Watch Party initiative. It next presents the survey findings and discusses their implications, before concluding with practical suggestions for future remote participation in ICCC.

Background, History and Context

The COVID-19 pandemic accelerated experimentation with alternative conference formats. Since then, terms such as in-person, remote, hybrid, alternate, and decentralised have become common in discussions about event organisation. In-person events bring everyone to the same physical location, which usually makes networking and informal interaction easier, but they exclude people who cannot travel. Remote events move everything online, making participation cheaper and more accessible, though often at the cost of weaker social interaction and engagement. Hybrid events combine both modes at the same time, offering flexibility but also creating extra coordination work and sometimes uneven experiences between onsite and online participants. Alternate formats switch between modes across editions or sessions, which helps balance access and the in-person experience, but can disrupt continuity from one year to the next. Decentralised formats use multiple local hubs connected to a shared event, which can support local participation and community building, but are harder to organise consistently.

Prior work on hybrid conferences [Bajpai et al., 2022] recommends planning hybrid participation from the start, ensuring equal access to content and interaction, assigning dedicated support roles, accommodating timezone differences, using simple platforms, rethinking social activities for both audiences, and offering differentiated pricing. Despite these guidelines, we suggest that implementing a fully hybrid model can significantly increase the organisational complexity of the conference. Yet, a more feasible alterna-

[1] https://computationalcreativity.net/iccc25/

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

tive can perhaps be officially supporting independent parallel remote events and providing low-effort mechanisms and incentives for the community to propose and run distributed satellite hubs capable of accommodating those unable to attend in person, increasing attendance and continued engagement, without compromising organisational manageability.

Since ICCC'21, the ACC General Assembly[2] has periodically discussed decentralised, remote, and hybrid versions of ICCC. Supporters have emphasised accessibility and environmental benefits [Pease and Pease, 2023], while others have warned that such formats could weaken in-person attendance and community cohesion. None of these proposals has moved beyond discussion, perhaps because the fully remote editions of 2020 and 2021, organised in response to the COVID-19 pandemic, already illustrated the organisational burden and the disadvantages of non-standard formats.

ICCC'25 took place in Brazil under difficult conditions for attendance, including funding cuts, travel constraints, and competing conferences. New institutional travel policies excluded ICCC from funding eligibility in some cases, and the political climate further complicated international travel. Most importantly, the lack of remote options posed an additional challenge for maintaining ACC membership continuity. As the need for remote participation became more visible, several online accommodations were introduced late, but they were not always clearly communicated and did not fully support asynchronous participation. Thus, in parallel, a community-led Watch Party emerged through the ACC Discord server. The idea was first raised during a session of the CC Reading Group[3], a monthly community-led discussion held on the ACC Discord server. The first written mention of the proposal appeared on 27/05/2025 on ACC Discord and received engagement from multiple members of the community. One member also added to their support:

> *For future years, it would be fantastic to actually have an online parallel conference, as many people will have issues travelling, for various reasons... I know hybrid is really hard to organise, and at the same time it's partly becoming something that people are surprised not to see offered ... in current times it has become more important than ever, and would be a great way to include people who aren't able to travel in person.*

An initial message requesting permission to proceed with the event was sent to the ICCC'25 organisers on 19/06/2025. The response received on 21/06/2025 granted provisional support but with by the following statement:

> *This year has been very special. I think, under the circumstances, having a parallel online event may reduce the worst of the damage to the community. But I definitely do not think it is a good alternative to having a proper conference, and I do not think it is something we should encourage in the future.*

This message reflects the fear that remote participation might undermine the conference. Yet, even though these barriers were more visible in 2025, because they affected more European and North American members of the community, we believe they are systemic and recur every year, affecting different people in different ways. Because conference travel is still heavily shaped by geography, funding, and visa constraints, the current model keeps favouring participants from Europe and North America while leaving others systematically underrepresented. In current times, an in-person-only model can feel exclusionary, even elitist, and may be perceived as a threat: attend in person or do not participate at all. For those reasons, we believe that, rather than weakening the conference, balanced remote and decentralised options may help preserve continuity, increase inclusion, and strengthen long-term community engagement.

Watch Party Event

The ICCC'25 Watch Party was conceived as a lightweight, community-driven way to support those unable to travel, while preserving an in-person-first conference model. Rather than a fully hybrid redesign, it relied on a small online space in the ACC Discord, with no expectation of live interaction with speakers and only limited promotional support. The aim was to enable informal co-watching and peer discussion without competing with physical attendance.

The proposal was based on the idea that real-time interaction and synchronous viewing could improve engagement, while a Discord-based format would keep the logistical burden low. The event was also used to test whether such remote initiatives could impact onsite attendance.

Approved shortly before the conference, the Watch Party was launched during the opening session and attracted a small but active group on the first day. Participation dropped after that, possibly due to time-zone differences, the main organiser's inability to remain present to help sustain engagement, and the overlap with other relevant conferences taking place at the same time. While this may be interpreted as evidence that low-effort remote approaches are not useful, we believe it is too early to draw that conclusion. Instead, we take these results as guidance for how the concept might be refined and made more effective in future iterations.

Questionnaire

The data presented in this report was collected through a survey distributed during and after ICCC'25. The form was open for several months to allow asynchronous contributions, thereby capturing both real-time and retrospective perspectives. It is important to note that these results are likely affected by self-selection bias, as respondents who attended the party or held stronger views may have been more likely to answer. The aggregated analytics are available online[4], and the Figures refered throughout the text are relative to Annex I.

Attendance to ICCC25: Based on the survey, **69.2%** of the 13 respondents (9 individuals) **did not attend ICCC'25**

[2] Witnessed by the authors *in loco*. Attempts to access the minutes of recent ACC General Assembly meetings to substantiate these claims were unfruitful.

[3] `https://cs.uwaterloo.ca/~browndg/CC_reading_group/index.shtml`

[4] PDF of questionnaire results

in person (Figure I.1). The most common reasons were *"time or conflicting obligations"* (6 of the 9 non-attendees) and *"financial constraints"* (5 respondents). Other reasons, shown in Figure I.2, included: paper not accepted, sustainability concerns, or visa issues (2 each), as well as family-related constraints (1) and no interest in attending in person (1). Notably, none selected *"I prefer to follow remotely"*, *"I don't usually travel to conferences"*, or *"health issues"*[5].

Regarding the ICCC'25 Watch Party, 7 of 12 respondents (58.3%) said they attended or planned to attend, while 5 (41.7%) did not. Among those who did not attend, the main reason was attending the conference in person (4 of 6 responses), followed by time-zone constraints or conflicting obligations (2 of 6). Less frequent reasons, each mentioned once, were lack of interest, difficulty engaging online, and uncertainty about how to participate. As shown in Figure I.4, none cited a lack of awareness of the initiative, a preference for following independently, or discomfort using Discord.

Feedback about ICCC25 Watch Party: Participants highlighted several positive aspects of the Watch Party: the small group size (4–5 participants) created a casual atmosphere that encouraged open discussion; real-time chat fostered a sense of community; and the YouTube stream was seen as clear, high-quality, and unobtrusive, with one participant suggesting that this event could promote people to subscribe to the ACC YouTube channel.

They also identified several limitations: switching between YouTube and Discord was inconvenient and some preferred a unified platform; engagement was sometimes low, which discouraged participation when the space was empty; the purpose of the Watch Party were not always clear, suggesting a need for integration into the main program; time-zone differences prevented some people from joining; and informal interactions were missed, with one suggestion being remote social spaces such as virtual coffee breaks.

Regarding the platform used (Figure I.5), out of 10 people who answered a question related to this topic, **6 said Discord was a good option**, with options like GatherTown and Zoom coming after with 3 votes each.

Participants were asked how remote participation could be improved in future conferences, for example, through local watch parties, interactive Q&A, or breakout sessions. The responses showed interest in local physical watch parties or regional hubs, especially for those unable to travel internationally; interactive formats such as breakout sessions or structured prompts; informal voice-based spaces similar to GatherTown; remote chairs or mechanisms for live questions; open online Q&A; and, despite the added complexity, broader support for hybrid formats.

Finally, when asked whether the Watch Party should continue in future ICCC editions (Figure I.6), the vast majority of respondents – 11 participants out of 13 – answered affirmatively. Only one respondent replied "No" and another "Not sure"; notably, both attended ICCC'25 in person. We did not collect reasons for these answers, so we don't know the reasoning behind these negative answers.

Paper Submission: Before launching the survey, we hypothesised that having a paper accepted at ICCC would strongly influence participants' decisions to attend in person, both because the conference offered no explicit remote option for authors and because many institutions require acceptance for travel funding. To explore this, we asked whether respondents had submitted a paper to ICCC'25. As shown in Figure I.7, responses were nearly evenly split: 53.8% (7 of 13) had submitted a paper, while 46.2% (6 of 13) had not.

Among those who did not submit a paper, the reasons were varied: 2 respondents mentioned financial constraints, one of them explicitly said they could not afford to attend and therefore chose not to submit because there was no remote option; another respondent said they did not yet have a project ready and did not see the conference as relevant at that stage; the final respondent said they lacked motivation to submit without the possibility of attending in person, suggesting that their decision was already tied to physical attendance. As shown in Figure I.8, motivations for submission were diverse, but 77% of those who submitted a paper said that attending in person was at least part of their motivation, and 33% said it was their primary motivation. Only one respondent said physical attendance was not a motivating factor. This suggests that in-person attendance remains a major driver of paper submission, rather than the reverse.[6] This fact may reduce motivation to submit among those who do not see a path to attending in person, and an in-person-only model may even remove any additional incentive to submit. Overall, while these answers point to a mix of practical and contextual barriers, many beyond the conference organisers' control, the lack of remote options, which is within the organisers' control, can become an obstacle to submissions.

For the question "If you submitted a paper, was it accepted?", 7 of 9 respondents (77.8%) said yes, and 2 (22.2%) said no. While this suggests a relatively high acceptance rate among respondents, the sample is too small to infer the overall conference acceptance rate, and the result is very likely to have also been affected by self-selection bias. We previously saw that paper rejection can block in-person attendance, then it is plausible to assume that paper acceptance functions more as a way to secure funding and make attendance possible rather than motivate in-person attendance. Yet, acceptance does not guarantee in-person attendance, since some institutions do not fund travel to C-rated conferences. In those cases, even after acceptance, in-person-only models can once again become a bottleneck for participation, even for high-quality work that has been accepted.

The data suggest that remote presentation options would not discourage those already planning to attend in person, since many respondents were already motivated to submit by their intention to attend. On the contrary, such options could attract additional submissions from people unable to travel (for all kinds of reasons: chronic health problems, family re-

[5] Before the conference, however, a separate request was posted in the Computational Creativity Discord community by someone seeking remote access due to health constraints.

[6] This may be only relevant while the conference remains rated C in the CORE ranking; if those conditions change, new data and conclusions may be needed.

sponsibilities, etc.) while leaving the motivations of onsite attendees largely unchanged. This could increase submission volume, especially from regions that are systematically underrepresented in ICCC, such as Africa, West and East Asia, and South America, without undermining the existing incentive structure.

Impact of Remote Components: Lastly, we explicitly asked respondents how remote components would affect their participation decisions. Although these answers are likely affected once again by self-selection bias, they still provide an initial indication that remote formats do not necessarily discourage physical attendance.

As shown in Figure I.9, 41.7% of respondents said they would have been more likely to submit a paper if they had known about a remote component in advance, while half said it would make no difference and none said it would make them less likely to submit. This suggests that remote participation has a positive or neutral effect on submission decisions. Figure I.10 shows that 66.7% of respondents said remote components would have no influence on their decision to attend ICCC in person. Only 16.7% said they would be less likely to attend, and another 16.7% were unsure. Taken together, these results are consistent with the responses to the similar questions about remote components becoming a permanent feature of ICCC, in Figure I.13 and Figure I.14.

Half of the participants had already watched ICCC talks live in previous years, and another quarter had watched selected recordings (Figure I.11), suggesting that remote passive watching is already familiar to much of the community, which may explain why the Watch Party itself appears to have had limited influence on streaming viewing: while most respondents were unsure, only 8.3% said the event was the main reason they watched the sessions (Figure I.12). Yet, Figure I.15 shows that **84.6%** of respondents believe that a regular online Watch Party would **increase their engagement** with ICCC and the Computational Creativity community in general. Only one respondent disagreed and one was uncertain. This indicates strong community support for continued offering of remote active parallel participation.

Open Ended Question: The final survey question was open-ended, and the 9 responses reflected a broad range of views: 1) some respondents stressed that participation is often limited by factors beyond individual control, such as funding, visa restrictions, or "weak passports," which can even lead to paper withdrawals when no remote option is available. This raises concerns about the gap between ICCC's stated commitment to welcoming participants from all over the world and the practical barriers that still limit access, as well as about broader ACC membership continuity [Association for Computational Creativity, 2017]. 2) Some other respondents emphasised the importance of early communication about remote options, even if they are only offered on demand, while other responses, paradoxically suggested that timely communication could instead slightly discourage in-person attendance. 3) Some respondents still highlighted the value of in-person participation, especially for networking and community building, and expressed concern that hybrid formats could increase organisational burden or reduce the quality of the conference experience. 4) Finally, several respondents supported stronger remotisation or full hybridisation, suggesting that exciting online offerings could expand reach and inclusivity, with some even supporting ICCC "formally [becoming] a hybrid conference".

Overall, these findings show both support for and reservations about remote and hybrid components. While such formats may improve inclusivity and engagement, some fear they could weaken the in-person experience and increase organisational burden. At the same time, many people still face barriers beyond their control that prevent them from attending in person. In this context, several respondents suggested that initiatives like the Watch Party ICCC'25 could strengthen long-term involvement with ICCC and the broader Computational Creativity community.

Discussion and Proposed Guidelines

This preliminary initiative explored the possibility of *remotisation* in ICCC, but the available data are still limited. As participation was voluntary, the feedback may reflect more active or positively engaged members, and may not capture the views of less engaged participants. Future assessments are necessary and should aim for more representative input.

Given these findings, we propose an **In-Person First, Not In-Person Only** principle. This approach preserves the advantages of an in-person conference while still allowing lightweight and meaningful remote participation. It aims at avoiding the complexity of a fully hybrid or remote-first model, while keeping the event accessible and engaging to those unable to attend physically.

As an **in-person first** event, ICCC would: promote in-person presentations, keep the main schedule fixed to in-person attendance; avoid adapting the program to remote time zones; not require real-time interaction from remote viewers; keep presentations optimised for on-site engagement; handle catering and social events in the usual way.

As a **not in-person only** event, ICCC should also: continue providing a live stream of main sessions; appoint a *Chair for Remote Attendance*; measure the risk of mass non-attendance and plan for it; provide a remote registration option, allowing members to remain formally engaged with ACC, while helping organisers manage catering and local logistics; offer remote presentation options and communicate them clearly and early; consider remote-friendly workshop formats; encourage community-led satellite events, such as local or online watch parties; optionally, support these initiatives through incentives or recognition; collect feedback during and after the conference. Organisers of satellite events should clearly identify them as ICCC-affiliated, include opportunities for discussion and social interaction, and consider features such as dedicated hosts, shared question collection, and deferred watch parties for different time zones.

Given the current global and political landscape, this is a timely opportunity to explore alternative conference models using new technologies to bring the community closer together, thus embodying the true spirit of Computational Creativity —pioneering inclusive, flexible, and forward-looking ways to foster collaboration and creativity worldwide.

Acknowledgments

The authors would like to thank the members of the Computational Creativity community who contributed to discussions and provided feedback on the ICCC'25 Watch Party initiative, particularly those who engaged during the CC Reading Group and supported the creation of a remote parallel space. Special thanks are due to the ICCC'25 organisers for their openness to experimentation under challenging circumstances, and to all who participated in the Watch Party channels, helping test new models for engagement and decentralised presence. Thanks to Marcio Lima Inácio and André C. Santos for their suggestions on the text.

GenAI Disclosure

Portions of this report were prepared with the assistance of OpenAI's ChatGPT (v2, 2025), used as a writing and editing aid. The model was employed to improve clarity, revise drafts for academic tone and coherence, and format content into LaTeX where appropriate. All factual content, structure, and final decisions reflect the author's original intent and judgment.

References

Association for Computational Creativity. Constitution of the association for computational creativity (acc). https://computationalcreativity.net/home/wp-content/uploads/2013/08/ACC_Constitution_Sep-2017.pdf, 2017. Accessed: 2025-10-06.

Vaibhav Bajpai, Oliver Hohlfeld, Jon Crowcroft, Srinivasan Keshav, Henning Schulzrinne, Jörg Ott, Simone Ferlin, Georg Carle, Andrew Hines, and Alexander Raake. Recommendations for designing hybrid conferences. *ACM SIGCOMM Computer Communication Review*, 52(2), April 2022. URL https://vaibhavbajpai.com/documents/papers/proceedings/hybrid-ccr-2022.pdf. Editorial; not peer-reviewed.

Alison Pease and Arnold Pease. Computational creativity and the climate crisis. In Alison Pease, João Miguel Cunha, Maya Ackerman, and Daniel G. Brown, editors, *Proceedings of the 14th International Conference on Computational Creativity*, pages 293–297, Ontario, Canada, 2023. Association for Computational Creativity. ISBN 978-989-54160-5-9. URL https://computationalcreativity.net/iccc23/papers/ICCC-2023_paper_114.pdf.

Annex I: Survey Results Figures

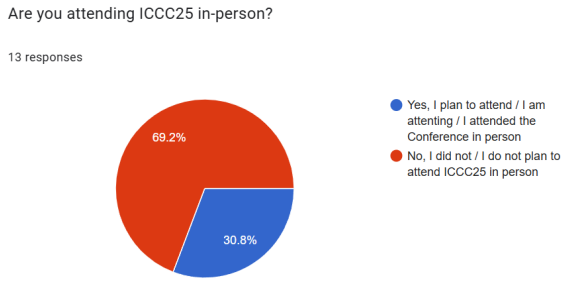


Figure 1: In-person attendance at ICCC'25 of those who responded to the survey.

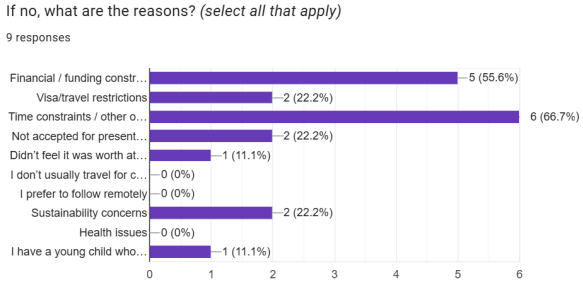


Figure 2: Reasons for not attending ICCC'25 in person, as reported by survey respondents.

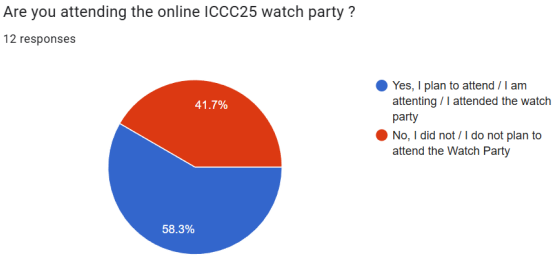


Figure 3: Responses to the question regarding online ICCC'25 watch party attendance.

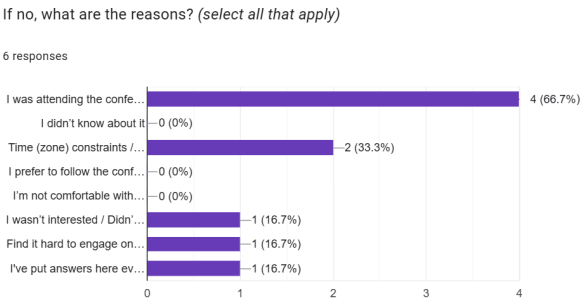


Figure 4: Reasons for not attending the Watch Party

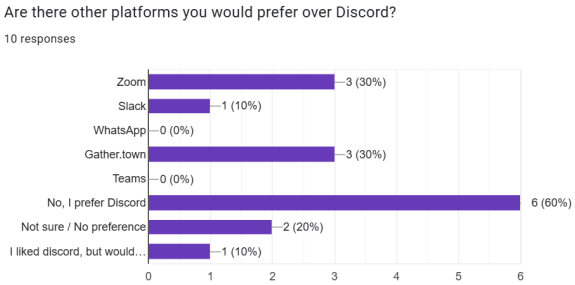


Figure 5: Plataform preferences

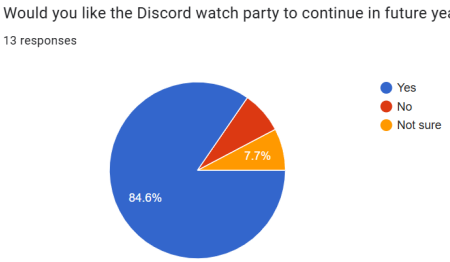


Figure 6: Responses to a new edition of the Watch Party

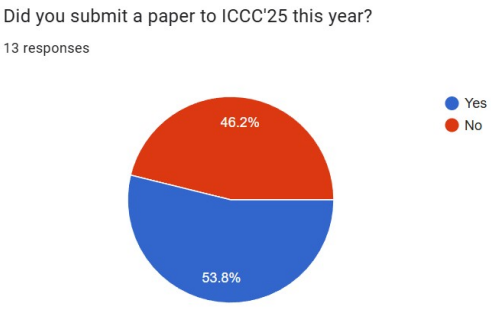


Figure 7: Responses to the question "Did you submit a paper to ICCC'25?"

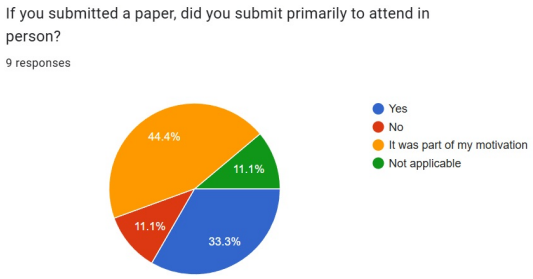


Figure 8: Responses to the question *"If you submitted a paper, did you submit primarily to attend in person?"*

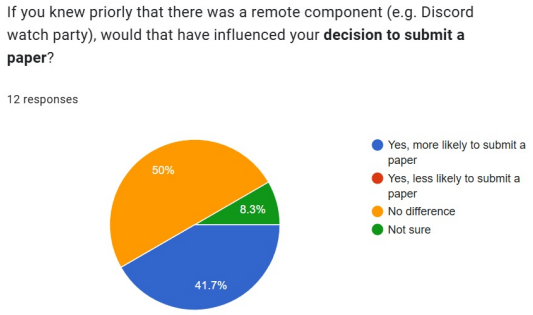


Figure 9: Influence of remote components on the decision to submit a paper.

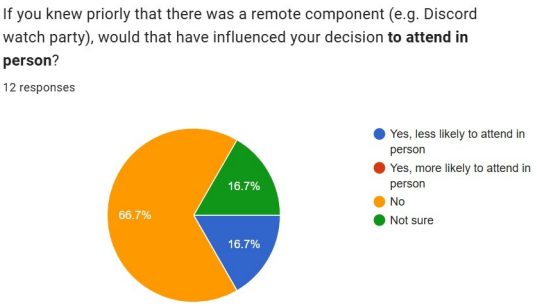


Figure 10: Influence of remote components on the decision to attend in person.

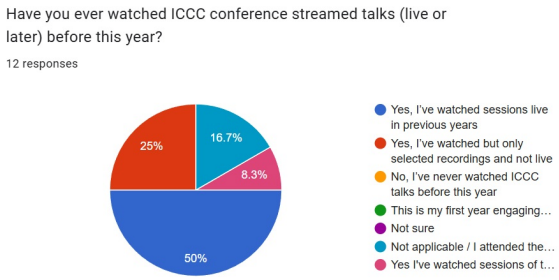


Figure 11: Respondents' previous experience watching ICCC talks online.

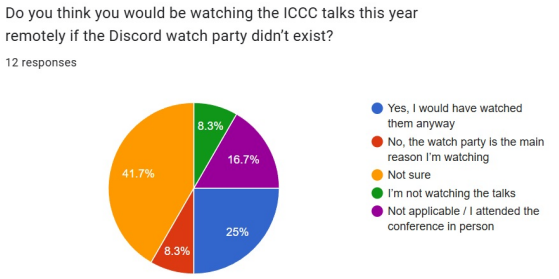


Figure 12: Impact of the Watch Party on respondents' likelihood to watch ICCC talks.

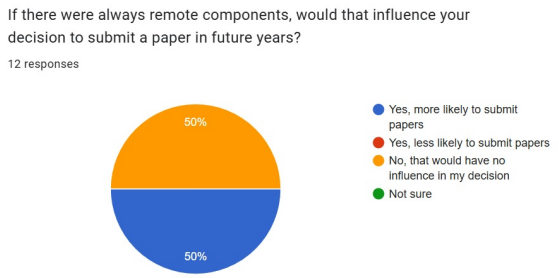


Figure 13: Influence of permanent remote components on future paper submissions.

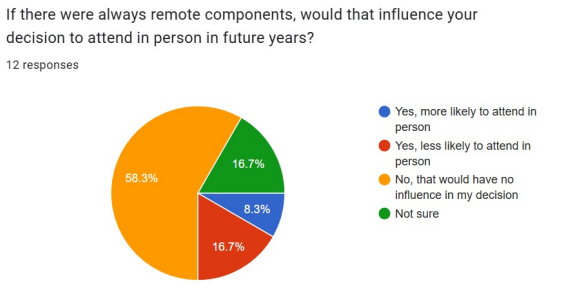


Figure 14: Influence of permanent remote components on future in-person attendance.

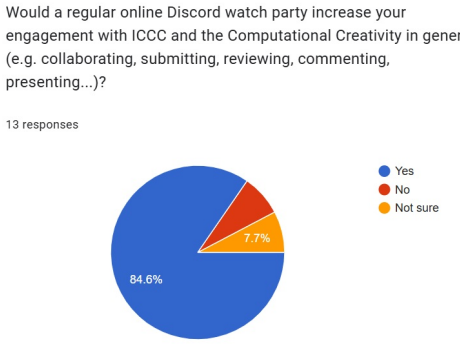


Figure 15: Respondents' views on whether regular Watch Parties would increase engagement with ICCC.

The Flow of LLM-Generated Poetry

Patrícia Ferreira
University of Coimbra, CISUC/LASI, DEI
DEI, University of Coimbra
patriciaf@dei.uc.pt

Ana Alves
University of Coimbra, CISUC/LASI, DEI
Polytechnic Institute of Coimbra
ana@dei.uc.pt

Catarina Silva
University of Coimbra, CISUC/LASI, DEI
DEI, University of Coimbra
catarina@dei.uc.pt

Hugo Gonçalo Oliveira
University of Coimbra, CISUC/LASI, DEI
DEI, University of Coimbra
hroliv@dei.uc.pt

Abstract

We demonstrate how dialogue flows can be adopted for representing transitions between speech acts and their probabilities in sets of generated poems, contributing to the analysis of the narrative structure. By considering poems generated by different Large Language Models (LLMs), open and proprietary, we take conclusions on the impact of size, and focus on the temperature hyperparameter, known as a proxy for creativity. Flow analysis revealed that different LLMs prefer different speech acts and confirmed that higher temperatures increase the set of used transitions, especially for the open models.

Introduction

The flexibility of general-purpose instruction-tuned Large Language Models (LLMs) makes them a convenient tool for zero- and few-shot generation of creative textual artefacts, including poems (Sawicki et al. 2023; Qu, Yuan, and Färber 2026). From natural language instructions, LLMs can produce text that reflects a given intent while exhibiting poetic features, such as a regular form or usage of figurative language. However, deeper analysis noted that there are limitations on meeting a specific number of characters per line (Yu et al. 2024) or of lines and syllables per line (Agirrezabal and Gonçalo Oliveira 2024); as well as on diversity, with some models preferring specific words, meanings or forms (Chen et al. 2024; Walsh, Preus, and Gronski 2024).

Although previous works capture quantifiable aspects of poetry, they prioritize surface-level patterns, highly-relevant in poetry, over the more abstract narrative. To uncover this structural aspect, we propose to explore the construct of dialogue flows and adapt them to represent the flow of speech acts in LLM-generated poems. Several works rely on flow induction from human dialogues where the identification of behavioral patterns contributes to the development of dialogue systems (Martinez and Nugent 2022) or improvements in customer-support (Ferreira et al. 2024). Recently, flow discovery was also applied for analyzing LLM reasoning patterns (Matos, Silva, and Gonçalo Oliveira 2025). We argue that such flows are a useful tool for studying the diversity of poetry at a more abstract level.

To test this, flows are extracted from English sonnets generated by open and proprietary LLMs of different sizes. To enable this, each line of a sonnet is labeled with a poetry-oriented speech act, allowing these sequences to characterize the flow of each LLM.

Moreover, to revisit previous work on the relation between the temperature hyperparameter and creativity (Peeperkorn et al. 2024), poems are generated with two different values of temperature, and its impact on the structural organization of generated poems is examined. While creativity cannot be reduced to diversity alone, we focus on structural diversity as one of its relevant dimensions. Our main contributions are:

- An adaptation of dialogue flows for representing the sequence of speech acts in poetry, thus enabling a rich analysis of the underlying sequence of actions;
- The application of the previous to sets of poems by different LLMs, which confirms that flows are a visual means for analyzing how structural diversity is affected by features like model size and temperature.

This aims to broaden the evaluation of AI-generated poetry, offering a new lens to understand how LLMs navigate the complex balance between structural flow and creativity.

Speech Act Flows

We propose to model the narrative structure of poems by flows connecting speech acts, i.e., directed graphs $G = (N, E)$ where: the set of nodes N includes the possible speech acts, augmented by nodes to mark the start (SOP) and end of poem (EOP); the set of edges E represents the transitions $n_i \rightarrow n_j : n_i, n_j \in N$, meaning that a line labeled with speech act n_j follows another labeled with n_i.

In a set of lines, we can compute the transition probability $P(n_j \mid n_i)$, which quantifies the likelihood of moving from speech act n_i to n_j. If all the poems have the same number of lines, the flow can be organized according to line number. Highlighting frequent speech acts per line provides a visual cue for both common sequences and diversity.

A critical step is the classification of the speech act (Austin 1975), which can be defined as a general action performed by the speaker of an utterance. Different speech act taxonomies exist but, in order to find one more aligned with poetry and, particularly, with our data, we defined one with the help of LLMs. Specifically, we used a random sample of 200 lines from the poems in our dataset (see next section) and prompted GPT-4o to propose the ideal semantic

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

categories for classification. The zero-shot prompt used for this discovery is in the Appendix (Fig. 3).

The taxonomy includes five categories: **Dynamics (DYN)**, for dynamic movement, actions, or specific intentions performed by an agent; **Evaluation (EVAL)**, for explicit positive assessment that emphasizes greatness, superiority, or enduring impact; **Figuration (FIG)**, for figurative transformations, such as metaphors, similes, or symbolic equivalences; **Representation (REP)**, for concrete depiction of physical settings, landmarks, or objects, focusing on sensory or spatial details; and **Subjectivity (SUBJ)**, for expression of affective states (e.g., joy, sorrow) and the revelation of the entity's inner life or psychological depth.

Since taxonomy induction and annotation rely on GPT-4o, some annotation bias may exist. However, the adopted categories are intentionally high-level and model-agnostic. Table 1 shows the taxonomy in action, with one of the previous acts assigned to each line of a poem in our dataset.

Lines	Act
Stone whispers tales of ages long since past,	FIG
A Seine-kissed city, bathed in golden hue,	FIG
Where lovers stroll and shadows softly cast,	REP
And history breathes in every avenue.	REP
The Eiffel Tower, reaching for the sky,	REP
A wrought-iron dream against the fading light,	FIG
While Notre Dame, with stories in her eye,	FIG
Remembers echoes of a sacred rite.	SUBJ
Cafés hum with chatter, sweet and low,	REP
And artists paint on canvases so bold,	DYN
A symphony of life begins to flow,	FIG
In cobbled streets, a story to unfold.	REP
The scent of bread, a promise warm and true,	REP
Parisian magic, ever fresh and new.	EVAL

Table 1: Sonnet about Paris by Gemma-3-27B, $T = 0.8$, with each line classified.

Experimental Setup

Our main goal is to use speech act flows in the analysis of LLM-generated poems, and further take conclusions on their structural diversity. For this, a dataset was created with poems generated by different LLMs, and complementary metrics were defined.

Models

The LLMs selected for our study represent diverse sizes and access types. They include two proprietary models by OpenAI (GPT-4o and GPT-4o-mini), both accessed through their official API[1]; and two open models by Google (Gemma-3-4B and Gemma-3-27B), both accessed through Ollama[2], in Q4_K_M quantization, running in a local machine equipped with an NVIDIA RTX A6000 GPU (48GB VRAM).

Data

The selected LLMs were instructed to generate sonnets in English about a target entity, using the zero-shot prompt in Appendix (Fig. 2). To cover a broad, but controlled, number of targets, 15 entities were selected across three domains: (people in) Sports, (people in) Music, and Cities. We selected globally-recognized entities based on objective and renowned rankings, namely: Forbes' highest-paid athletes[3], Spotify's top-ranked artists[4], and the Global Cities Index by Oxford Economics[5]. The complete list is in Table 2.

Domain	Target entities
Sports	Cristiano Ronaldo, Stephen Curry, Tyson Fury, Dak Prescott, Lionel Messi
Music	Bad Bunny, Taylor Swift, The Weeknd, Drake, Billie Eilish
Cities	New York, London, Paris, San Jose, Seattle

Table 2: Selected target entities across the three domains.

Moreover, poems were generated using two opposing temperature values (T) to evaluate their impact. We used $T = 0.0$ to capture the model's standard and more deterministic behavior, and $T = 0.8$ to force higher diversity.

Generating 10 poems for each of the 15 entities across two temperatures and four models resulted in a total of 1,200 sonnets, which is publicly available at `https://tinyurl.com/rspz2h66`. Since they consist exclusively of sonnets, depicted flows will have 14 levels, mapping the exact progression of the poem from start to finish.

To classify each line into one of the five speech acts, we resorted to GPT-4o, with $T = 0.0$. The prompt used for this is in the Appendix (Fig. 4). Out of curiosity, Table 3 details the speech act distribution per LLM, entity domain and temperature.

We immediately see that the distribution is driven by the domain of the target entity. For example, DYN acts are the most common in poems about *Sports*, naturally reflecting movement and action. Conversely, poems about *Cities* prioritize REP to describe physical settings and landmarks. The *Music* domain elicits higher rates of SUBJ and FIG.

Furthermore, the analysis confirms that each LLM has its own stylistic signature (Walsh, Preus, and Gronski 2024). The smaller Gemma-3-4B diverges notably from the others, relying heavily on SUBJ across all domains (over 61% in *Music*), whereas the GPT models demonstrate a robust preference for figurative language (FIG). Interestingly, increasing the temperature from $T = 0.0$ to $T = 0.8$ produces only marginal fluctuations in these frequencies. This suggests that, while temperature might alter the sequence of acts, the baseline semantic profile is an intrinsic property of each LLM, remaining robustly stable.

Metrics

To quantify the structural complexity, semantic focus, and generation stability of the poems across different model sizes and temperatures, we employ three sets of metrics.

Graph Metrics inform about structural coverage and path complexity, i.e., number of explored states. They include the total number of unique nodes and directed transitions in the speech act flows.

[1] `https://openai.com/`
[2] `https://ollama.com/`
[3] `https://www.forbes.com/lists/athletes/`
[4] `https://tinyurl.com/yke49wkt`
[5] `https://tinyurl.com/4k94xb55`

Model	Domain	T	DYN	EVAL	FIG	REP	SUBJ
Gemma-3-4B	Sports	0.0	28.7	23.3	10.0	3.0	35.0
		0.8	26.9	21.7	12.3	6.6	32.6
	Music	0.0	5.7	8.7	13.0	11.0	61.6
		0.8	5.6	8.0	15.6	9.4	61.4
	Cities	0.0	4.0	10.0	12.1	49.3	24.6
		0.8	5.0	6.4	9.6	52.4	26.4
Gemma-3-27B	Sports	0.0	52.1	15.3	13.3	9.6	9.7
		0.8	49.0	20.0	12.9	8.6	9.6
	Music	0.0	32.9	14.3	30.3	3.1	19.4
		0.8	25.9	15.0	30.0	5.9	23.3
	Cities	0.0	27.1	8.4	24.4	35.7	4.3
		0.8	22.7	7.6	28.3	35.9	5.6
GPT-4o-mini	Sports	0.0	48.6	23.4	17.4	2.6	8.0
		0.8	45.7	21.9	20.8	1.9	9.7
	Music	0.0	30.6	6.6	36.9	2.0	24.0
		0.8	28.9	8.9	36.0	2.1	24.1
	Cities	0.0	20.1	1.9	34.0	32.6	11.4
		0.8	16.9	4.0	34.6	33.4	11.1
GPT-4o	Sports	0.0	38.4	33.4	17.9	1.3	9.0
		0.8	41.3	29.3	17.0	2.6	9.9
	Music	0.0	24.4	9.9	40.4	3.9	21.4
		0.8	25.7	8.9	39.3	4.3	21.9
	Cities	0.0	15.6	7.3	30.0	39.4	7.7
		0.8	13.9	5.7	31.6	40.4	8.4

Table 3: Distribution of speech acts (%) across models, domains, and temperatures (T).

For the **Semantic Metrics**, the poems and the entity names are embedded with `all-MiniLM-L6-v2`[6]. *Topic Similarity* measures the cosine between the target entity and each generated poem. *Stability* aims to assess how different poems generated with the same model, temperature and target vary among them. It is given by the cosine between a randomly selected prototype against the remaining poems for the same entity, following the exemplar-based evaluation methodology of (Peeperkorn et al. 2024).

As a **Divergence Metric**, the Mean Absolute Difference (MAD) measures the impact of temperature as the difference between the probability transition matrices with $T = 0.0$ and $T = 0.8$, scaled by a factor of 100 for readability.

Results and Analysis

To analyze the structural progression of the generated poems, we examine the distribution of speech acts across the 14 lines. Figure 1 (a–d) illustrates the flows for all the poems about *Cities* by Gemma-3-27B and GPT-4o, with $T = 0.0$ and $T = 0.8$. Flows for all LLMs and domains are in Appendix (Figs. 5 and 6). Node colors denote speech acts, from SOP to EOP, and directed edges represent transitions with variable thickness and darkness, depending on probability.

With $T = 0.0$, Gemma exhibits a nearly linear structure with restricted transitions. As temperature increases, the flow becomes more dispersed, showing greater sequential variation. In contrast, GPT-4o generates a dense and diverse structure even with $T = 0.0$. Increasing to $T = 0.8$ adds only subtle variations, keeping the base flow intact.

These flows provide a visual anchor for the statistical distributions reported in Table 3. For *Cities*, the REP act (blue) serves as the foundational backbone across all 14 levels. This is evidenced by the primary transition paths, represented by thicker edges, and suggests that the domain's descriptive nature acts as a central narrative axis, maintaining a stable core even when higher temperature introduces a wider variety of alternative paths. To quantify these observations, Table 4 reports the metrics described in the previous section.

[6]`https://tinyurl.com/5cdp746d`

Model (T)	#N	#E	Topic Sim	Stability	MAD
Gemma-3-4B (0.0)	67	167	0.32 ± 0.11	0.90 ± 0.07	6.10
Gemma-3-4B (0.8)	72	302	0.30 ± 0.11	0.79 ± 0.07	
Gemma-3-27B (0.0)	63	137	0.31 ± 0.10	1.00 ± 0.00	5.85
Gemma-3-27B (0.8)	72	300	0.31 ± 0.10	0.82 ± 0.06	
GPT-4o-mini (0.0)	72	258	0.41 ± 0.09	0.87 ± 0.07	4.32
GPT-4o-mini (0.8)	72	292	0.39 ± 0.07	0.74 ± 0.06	
GPT-4o (0.0)	71	278	0.45 ± 0.09	0.87 ± 0.05	3.38
GPT-4o (0.8)	72	299	0.41 ± 0.07	0.76 ± 0.04	

Table 4: Metrics for the analyzed models across all domains at $T = 0.0$ and $T = 0.8$. #N and #E indicate the number of nodes and edges.MAD quantifies the difference between the transition matrices at these two temperatures.

The aggregate data in Table 4 show that the previous behavior is not restricted to *Cities*, but systemic across all domains: overall, the transitions of Gemma-3-4B almost double with $T = 0.8$. In contrast, GPT-4o undergoes just a marginal increase. A breakdown per domain is in the Appendix (Table 5).

On the number of nodes, full structural coverage is reached at higher temperatures, with all models attaining the maximum graph size, indicating that higher temperature promotes using all speech acts across all poem positions. However, proprietary GPTs already achieve full coverage with $T = 0.0$, confirming that, unlike Gemma, their baseline structural diversity is independent of temperature.

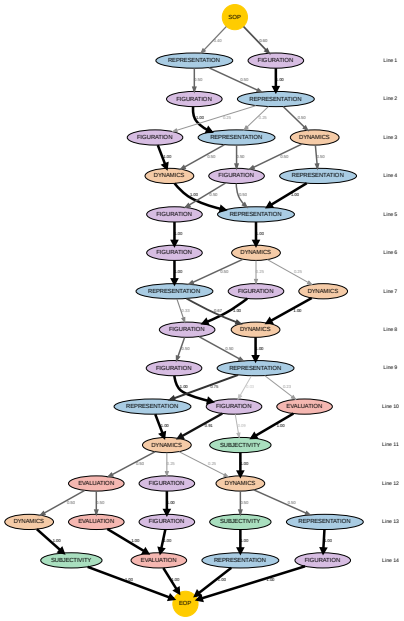
Metrics further highlight that Gemma-3-27B reaches a maximum *Stability* at $T = 0.0$, meaning that it always generates the exact same poem. Increasing to $T = 0.8$ successfully breaks this determinism, dropping internal stability to 0.82. This suggests the correlation between temperature and creativity, as it signals a move away from rigid repetition toward structural variety. In this case, the gain in diversity occurs without sacrificing thematic focus, as *Topic Similarity* remains stable across all models, regardless of temperature, ensuring that the poems explore new structural paths while remaining anchored to the target entity.

Finally, GPT LLMs are more stable across different temperatures, with more divergence (MAD) in the smaller model, but still lower than in the Gemma LLMs, where the smaller model is also the one where temperature has greater impact. This suggests that, regardless of the family, larger models are less affected by temperature.

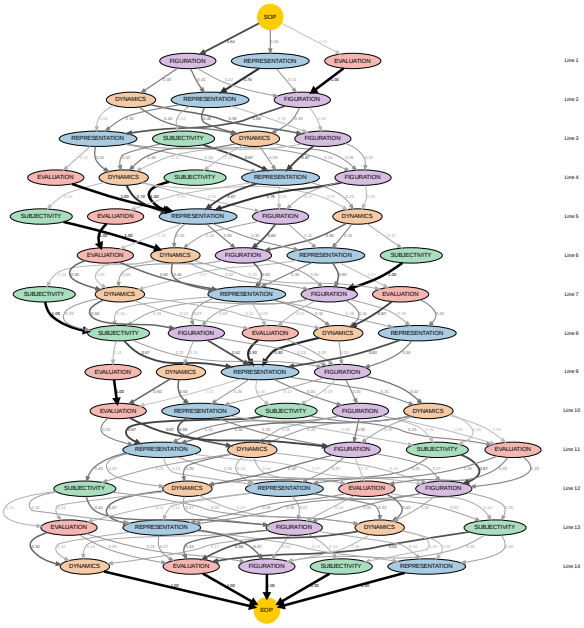
Overall, results suggest that smaller open quantized LLMs lack baseline stability and rely heavily on temperature to force structural diversity.

Conclusion

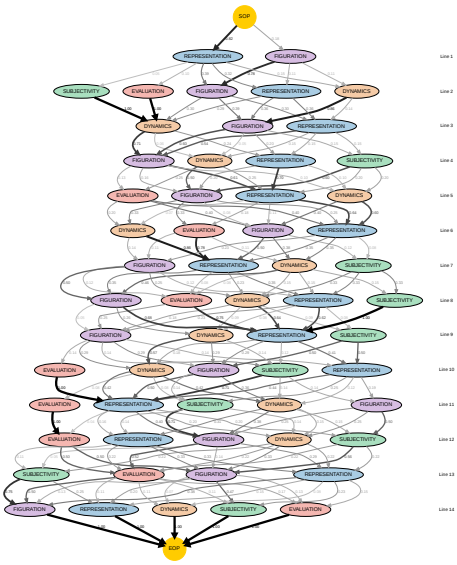
We showed how speech act flows can be useful for analyzing poetry and used them for studying the impact of model size and temperature in LLM-generated poems. These flows showed that higher temperatures successfully increase structural variety by reducing stability, without sacrificing thematic focus. Additionally, our study revealed clear differ-



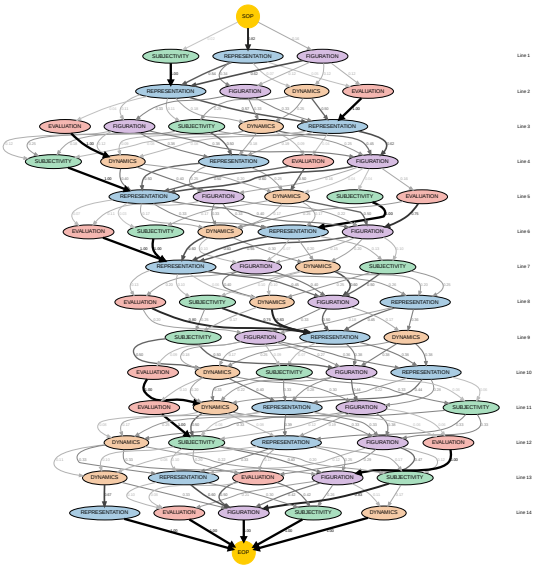
(a) Gemma-3-27B ($T = 0.0$)



(b) Gemma-3-27B ($T = 0.8$)



(c) GPT-4o ($T = 0.0$)



(d) GPT-4o ($T = 0.8$)

Figure 1: 14-level flows of speech acts for the Cities domain for Gemma-3-27b and GPT-4o at $T = 0.0$ and $T = 0.8$. Node colors represent the speech acts: Representation (blue), Figuration (purple), Evaluation (red), Dynamics (orange), and Subjectivity (green). Yellow nodes mark the start (SOP) and end (EOP) of the poem.

ences between model families: larger proprietary models produce more complex sequences even at zero temperature, whereas open quantized models are highly rigid at baseline and rely entirely on temperature to achieve diversity. By capturing these behaviors, our visually-aided methodology proved to be a powerful tool for better understanding how LLMs balance creativity and structure, and may complement future evaluations on this domain.

Once confirmed the utility of flows in this task, we plan to deepen the analysis of LLM-generated poetry and expand the study to other LLMs, temperatures, poetry styles, and languages, possibly revising the taxonomy of speech acts.

Acknowledgments

This work was financed by the Portuguese Recovery and Resilience Plan (PRR), through project C645008882-00000055– Center for Responsible AI. This work was also supported by FCT– Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra. Patrícia Ferreira was supported by FCT – Foundation for Science and Technology, I.P. through the PhD scholarship with reference 2024.01240.BD.

References

Agirrezabal, M., and Gonçalo Oliveira, H. 2024. Zero-shot Metrical Poetry Generation with Open Language Models: a Quantitative Analysis. In *Proceedings of 15th International Conference on Computational Creativity*, ICCC. Jönköping, Sweden: ACC.

Austin, J. L. 1975. *How to do things with words*. Harvard university press.

Chen, Y.; Gröner, H.; Zarrieß, S.; and Eger, S. 2024. Evaluating diversity in automatic poetry generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, 19671–19692. ACL.

Ferreira, P.; Carvalho, I.; Alves, A.; Silva, C.; and Gonçalo Oliveira, H. 2024. Sentiment-aware dialogue flow discovery for interpreting communication trends. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 274–288. Kyoto, Japan: ACL.

Martinez, J. M. S., and Nugent, A. 2022. Inferring ranked dialog flows from human-to-human conversations. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 312–324. ACL.

Matos, J.; Silva, C.; and Gonçalo Oliveira, H. 2025. Cognitive flow: An LLM-automated framework for quantifying reasoning distillation. In *Proceedings of the 18th International Natural Language Generation Conference*, 596–616. Hanoi, Vietnam: ACL.

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of Large Language Models? In *Proceedings of 15th International Conference on Computational Creativity*, ICCC. Jönköping, Sweden: ACC.

Qu, Z.; Yuan, S.; and Färber, M. 2026. Poetone: A framework for constrained generation of structured chinese songci with LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 32764–32772.

Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; and Khatun, A. 2023. Bits of grass: Does GPT already know how to write like Whitman? In *Proceedings of the 14th International Conference for Computational Creativity (ICCC'23)*.

Walsh, M.; Preus, A.; and Gronski, E. 2024. Does ChatGPT have a poetic style? In *Proceedings of the Computational Humanities Research Conference 2023*, volume 3558. CEUR-WS.org.

Yu, C.; Zang, L.; Wang, J.; Zhuang, C.; and Gu, J. 2024. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 315–325.

Appendix

Prompts

This section details the prompts used throughout the study, following a structured pipeline from data generation to automated annotation. The process was divided into three main stages:

1. **Sonnet Generation:** Figure 2 shows the prompt used to generate the English sonnets across all models. It includes constraints to ensure the output consists only of the 14 poetic lines.
2. **Speech Acts Discovery:** Figure 3 presents the zero-shot prompt provided to GPT-4o to analyze a sample of the corpus and identify the most recurring functional categories of speech acts.
3. **Automated Classification:** Figure 4 details the final classification prompt. This instruction was used to label every line of the dataset into the five discovered categories.

> Write a sonnet about {`entity`}.
> Rules: Write strictly in English. Do not write introductions or conclusions.
> Write exactly and only the 14 lines of the sonnet.

Figure 2: Sonnet generation prompt, where {`entity`} represents the target entity.

> You are an expert in analyzing poetry. I am conducting a study on AI-generated Sonnets in English. I have a list of 200 verses below. I need you to categorize them based on what the poetic line is doing (its function). Read the verses and create a list of 5 or 6 simple categories that describe what each verse is doing.
> Here is the sample of verses: `[Sample list...]`

Figure 3: Prompt used for the inductive discovery of the speech act taxonomy.

Flows for all Models and Domains

This section presents the complete 14-level flows aggregated across all three evaluated domains (*Sports*, *Music*, and *Cities*). Figure 5 illustrates the overall structural behavior of the open-weight Gemma models, while Figure 6 details the proprietary GPT models. These aggregated graphs confirm that the structural patterns observed in specific domains remain consistent globally.

Metrics

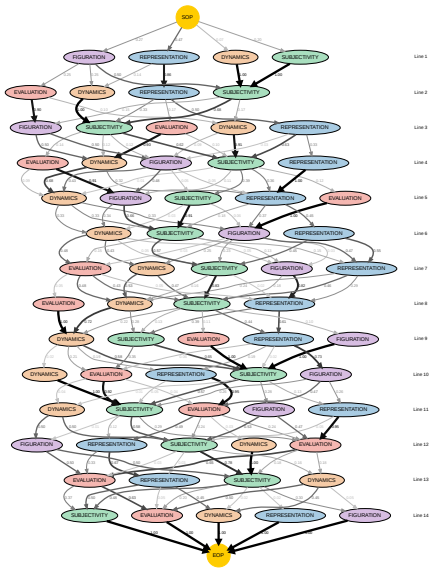
Table 5 provides a detailed breakdown of the structural and semantic metrics across the three evaluated domains. These values confirm that the previously observed trends are systemic behaviors rather than isolated artifacts. Across all domains, increasing the temperature expands the structural network (increasing both nodes and transitions) while simultaneously degrading internal stability and topic similarity.

> Classify the provided poetic line/verse strictly into one of the following 5 categories, based exclusively on the definitions below:
> 1. **Representation:** Concrete depiction of physical settings, landmarks, objects, or visible attributes with focus on sensory or spatial detail. It also includes situating a subject within a cultural, geographic, historical, or social framework.
> 2. **Subjectivity:** Expression or attribution of affective states (joy, sorrow, passion, hope, etc.) and the revelation of inner life (soul, spirit, memory, psychological depth).
> 3. **Evaluation:** Explicit positive evaluation of a person, place, or entity, emphasizing greatness, superiority, or admiration. It includes the construction of permanence, iconic status, or enduring impact.
> 4. **Figuration:** Figurative transformation where something is described as or compared to something else (simile, symbolic equivalence). It also includes explicit or implicit opposition, tension, qualification, or paradox.
> 5. **Dynamics:** Presentation of dynamic movement or event performed by an agent. It also includes the expression of intention, function, effect, or goal.
> Answer only with the category name. Do not write justifications, periods, or introductions.
> Verse to classify: {`verse`}

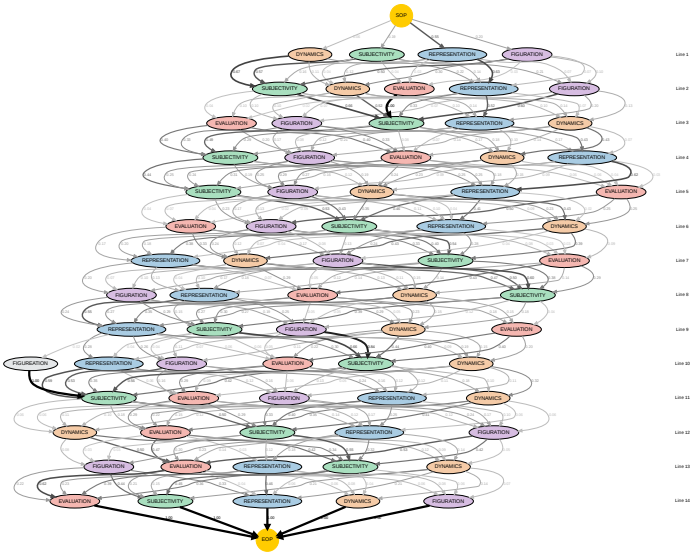
Figure 4: Poetic line classification prompt, where {`verse`} represents the input text.

Furthermore, the domain-specific breakdown reveals distinct semantic challenges. Notably, all models struggle significantly more to maintain topical focus in the *Music* domain compared to *Sports* or *Cities*. This is evidenced by the lower Topic Similarity scores across the board (e.g., GPT-4o drops to 0.37 ± 0.10 at $T = 0.8$), suggesting that generating poetry about musical entities induces a higher degree of abstract or metaphorical semantic drift.

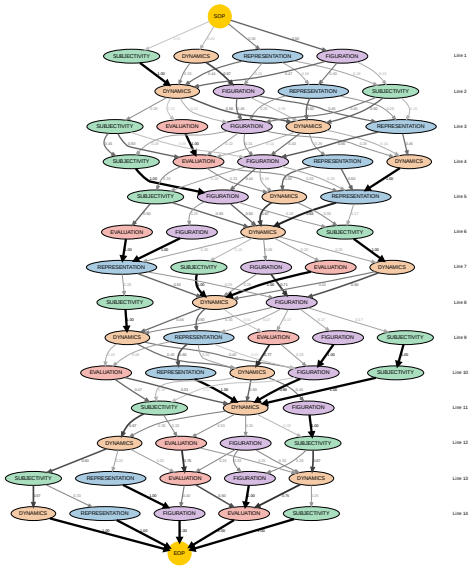
Finally, the distance metrics show how stable the proprietary models are. In the *Cities* domain, GPT-4o has very low distance scores. This suggests that the model has a very strong and clear internal knowledge of these famous locations. Consequently, even when we increase the temperature, the model manages to keep the poem's structure stable.



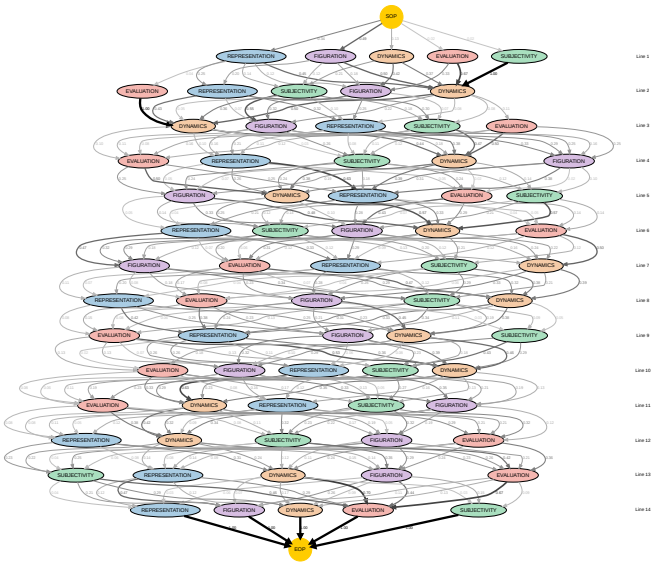
(a) Gemma-3-4B ($T = 0.0$)



(b) Gemma-3-4B ($T = 0.8$)

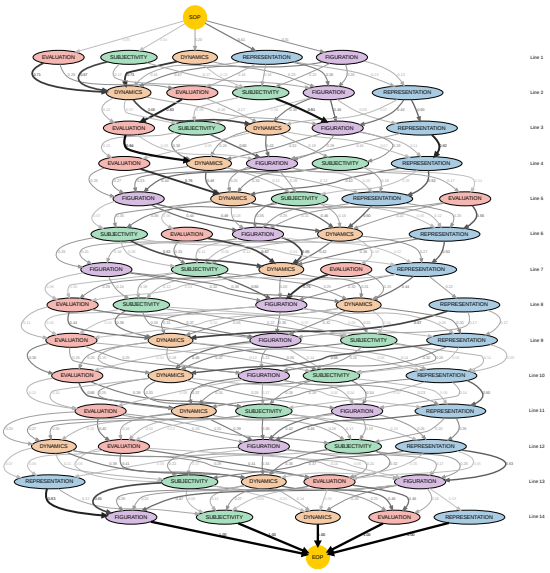


(c) Gemma-3-27B ($T = 0.0$)

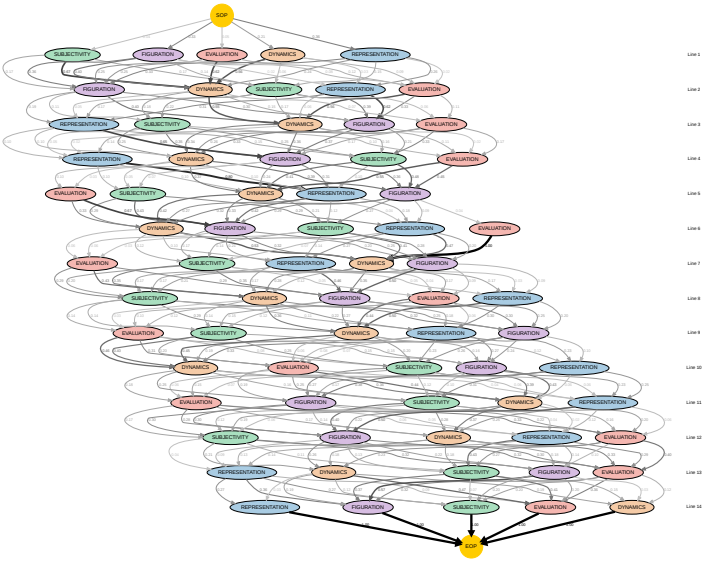


(d) Gemma-3-27B ($T = 0.8$)

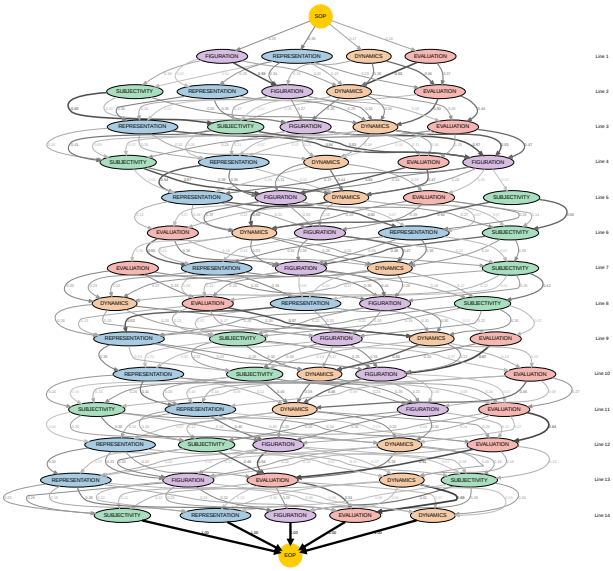
Figure 5: 14-level flows of speech acts across all domains for Gemma-3-4b and Gemma-3-27b at $T = 0.0$ and $T = 0.8$.



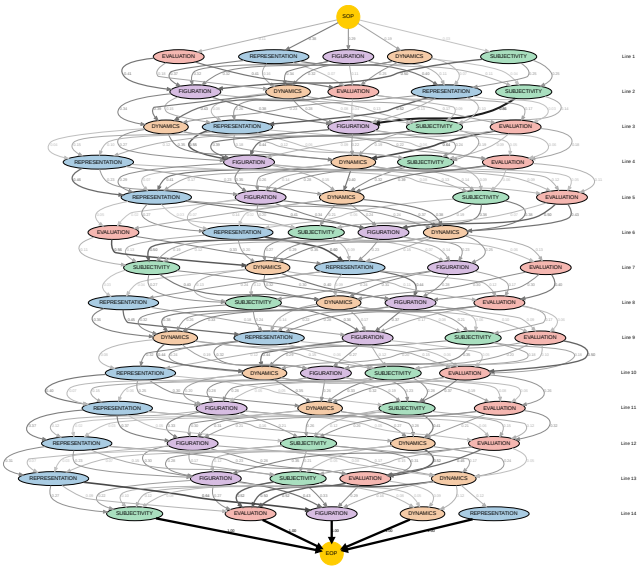
(a) GPT-4o-mini ($T = 0.0$)



(b) GPT-4o-mini ($T = 0.8$)



(c) GPT-4o ($T = 0.0$)



(d) GPT-4o ($T = 0.8$)

Figure 6: 14-level flows of speech acts across all domains for GPT-4o-mini and GPT-4o at $T = 0.0$ and $T = 0.8$.

Model	Domain	# Nodes		# Transitions		Topic Similarity		Stability		MAD
		$T = 0.0$	$T = 0.8$	$T = 0.0$	$T = 0.8$	$T = 0.0$	$T = 0.8$	$T = 0.0$	$T = 0.8$	
Gemma-3-4B	Sports	49	68	87	204	0.35 ± 0.03	0.34 ± 0.11	0.90 ± 0.01	0.76 ± 0.09	5.77
	Music	46	66	73	170	0.24 ± 0.01	0.18 ± 0.05	0.89 ± 0.01	0.77 ± 0.09	5.72
	Cities	40	68	64	183	0.37 ± 0.03	0.38 ± 0.05	0.90 ± 0.02	0.82 ± 0.04	6.82
Gemma-3-27B	Sports	45	65	67	179	0.36 ± 0.00	0.37 ± 0.08	1.00 ± 0.00	0.81 ± 0.06	4.96
	Music	46	68	65	208	0.26 ± 0.00	0.24 ± 0.06	1.00 ± 0.00	0.79 ± 0.06	6.53
	Cities	40	66	57	195	0.31 ± 0.00	0.33 ± 0.05	1.00 ± 0.00	0.85 ± 0.05	6.06
GPT-4o-mini	Sports	57	65	133	180	0.49 ± 0.09	0.44 ± 0.11	0.84 ± 0.07	0.73 ± 0.08	4.14
	Music	57	59	165	180	0.34 ± 0.06	0.34 ± 0.10	0.88 ± 0.05	0.73 ± 0.07	4.45
	Cities	58	65	139	203	0.40 ± 0.03	0.37 ± 0.06	0.89 ± 0.05	0.77 ± 0.03	4.36
GPT-4o	Sports	57	64	146	191	0.49 ± 0.05	0.44 ± 0.10	0.87 ± 0.04	0.75 ± 0.04	5.22
	Music	60	67	171	191	0.45 ± 0.07	0.37 ± 0.10	0.86 ± 0.05	0.75 ± 0.05	2.85
	Cities	61	67	183	198	0.42 ± 0.03	0.41 ± 0.04	0.87 ± 0.04	0.78 ± 0.03	2.06

Table 5: Domain-specific structural and semantic metrics at $T = 0.0$ and $T = 0.8$. MAD quantifies the difference between the transition matrices generated at these two temperatures.

International Conference on Computational Creativity: What If the Emperor's New Clothes Are A Straitjacket?

Pablo Gervás
Facultad de Informática
Universidad Complutense de Madrid
pgervas@ucm.es

Carlos León
Facultad de Informática
Universidad Complutense de Madrid
cleon@ucm.es

Dan Ventura
Department of Computer Science
Brigham Young University
ventura@cs.byu.edu

Abstract

The *International Conference on Computational Creativity* was founded to provide an interdisciplinary home for research that falls outside the evaluation criteria of mainstream AI conferences—work that is exploratory, speculative, and by definition difficult to measure. Over time, the conference has suffered from a set of structural tensions that have gradually shifted the conference away from the role it was designed to play. As part of this shift, the conference has drifted toward the very norms it was meant to complement: quantitative evaluation, reproducibility requirements, and a reviewing culture that may be systematically filtering out the most genuinely creative work. This paper describes some of those problems, examines the forces driving them, and suggests a set of concrete measures that, in our view, could help restore and renew the community's health. We propose a re-calibration of our reviewing heuristics to protect the "gray areas" of creativity research, suggesting that ICCC must prioritize being a unique discussion forum over becoming a "poor relative" to mainstream AI venues.

Introduction

Computational Creativity (CC) as a discipline emerged, in large part, to carve out intellectual space for questions that more established venues could not comfortably host. How can a computational system produce artifacts or ideas that are novel, surprising, and valuable? What does creativity even mean when the agent is a machine? These questions resist the tidy experimental protocols that serve well in other corners of AI and cognitive science, and the field's founding community knew it. The *International Conference on Computational Creativity* (ICCC) was established to provide a home for research that often falls between the cracks of traditional disciplines. It was designed to be a space where the philosophical, the artistic, and the computational could coexist without the restrictive "guardrails" of the classical STEM disciplines upon which it is built: computer science, mathematics, cognitive science, etc. To achieve this, ICCC was designed as a meeting place where unconventional methodology, provocative ideas, and cross-disciplinary exchange were not merely tolerated but actively cultivated. What made it distinctive was not merely the subject matter—though the questions were and remain genuinely fascinating— but the way the community approached those questions: with intellectual openness, methodological pluralism, and a willingness to sit with problems that resist easy formalisation.

Over many years of participation: reviewing papers, serving on programme committees, attending and running sessions, watching the community grow and change—we have observed a gradual drift that concerns us. ICCC is now an established conference, but are we happy with the Emperor's new clothes? Conversations within the community reveal a shared unease: that ICCC is becoming harder to distinguish from the conferences it was not supposed to resemble.

The purpose of this paper is to express a shared view on the evolution of the conference over time, offer our best understanding of how it may have shifted away from its original purpose, and suggest some practical paths forward that may resurrect some of the redeeming characteristics it had at its birth. We argue that these symptoms are neither coincidental nor inevitable. They share common structural causes, and they call for deliberate structural responses. We offer this as a provocation in the spirit of the ICCC short-paper format: not a finished programme of reform, but a set of issues and possible directions that we believe deserve community-wide debate.

We want to be clear about the epistemic status of what follows. This is not an empirical study; it is a practitioner's account, shaped by our particular vantage points and inevitably partial. We offer it in that spirit, and we welcome correction and disagreement from others whose experience differs.

Concerning Symptoms

The evolution and growth of any conference will naturally surface various concerns; we have identified two in particular that seem critical to the health of ICCC and thus to the thread of our argument: *field specific reviewing criteria becoming sidelined by standard expectations* and *the conference becoming a venue rather than a meeting place*.

Reviewing Criteria

Over the last few years, we have observed a shared sentiment discussed by an ever widening subset of members of the community. As the conference grows and evolves it is receiving a significant injection of new researchers. This

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

initially started at the level of researchers submitting their work to the conference. When the number of papers to be reviewed each year grew, the size of the Program Committee had to increase as well. This involved an additional injection of new researchers as reviewers. These new reviewers volunteered to undertake a very significant workload, usually on top of similar work they do for other conferences: they receive the papers assigned; they find a review template with which they are familiar; and they do their best to fill it in with arguments that are recognizably academic. We are forever grateful to all of them.

However, a natural side-effect is that reviewing is increasingly governed by heuristics imported from adjacent disciplines—heuristics that are appropriate there and inappropriate here. This is not necessarily the reviewer's fault, but rather it reflects the difficulty of finding enough qualified reviewers as the conference has grown. As a result, reviews apply default disciplinary heuristics rather than the field-specific standards that the work requires, which causes a damaging feedback loop: work that would energise the conference is rejected; researchers who would contribute most do not submit; the pool of relevant reviewers shrinks further.

At least two aspects of ICCC offer significant challenges in this regard: *different paper types* and *different evaluation focus*. We discuss both here.

Different Types of Computational Creativity Papers The Call for Papers for this year's edition lists five different types of full papers, and ten different types of short papers. Both for full and short papers there is only one type that corresponds closely to the traditional technical paper as considered in other STEM conferences. Most of the others correspond to additional categories of paper developed over the years in an attempt to capture other possible ways of making progress in the understanding of computational creativity. In that sense, they are very likely to require significantly different criteria for acceptance. This has been the case historically for ICCC events from the beginning.

Yet it is very frequently true that reviewers newly come to the conference look at one of these paper types, apply wholesale the criteria they have developed for the average STEM conference, and reject them outright because they do not qualify under the generally accepted standard for technical papers. This leads to situations where papers that would have fit in well in the conference being left out: e.g., position papers rejected for not including quantitative evaluation; or papers reporting field events rejected for not including sufficient technical detail.

Different Focus for Quantitative Evaluation In addition to the difficulties presented by the large set of different paper types, even the standard for technical papers is different from that applied in other conferences. Most conferences impose baseline requirements on paper acceptance, including both quantitative evaluation and reproducibility. This is consistent with sound scientific methodology. But historically, both of these goals have been shown to present difficulties when applied to creativity.

The application of quantitative evaluation ensures that papers report results that are measurably correct. However, in the field of creativity there is no easy definition of what is correct. A system that scores well against a gold standard of prior creative work is, by the field's own definition, not creative. A system whose outputs are indistinguishable from prior work is, by the field's own criteria, not creative. If a system's primary goal is to reproduce a human-produced outcome to a high degree of statistical accuracy, it is performing a task of mimicry, not creativity. Historically, the field of computational creativity did go through a phase in which the community struggled with how to address the concept of evaluation. Slowly a body of work on evaluation of creativity arose. But this approach focused on trying to evaluate—sometimes in quantitative terms—how creative a system could be considered. Importantly, the characteristic that was being evaluated here was not how close to existing outputs a system could get but rather how different from them it could get while still producing valuable outputs itself—an extremely difficult balance to find. If this creativity-specific approach to quantitative evaluation is ignored, papers that explore the conceptual boundaries of creativity—areas where consistent quantitative metrics may not yet exist or may even be counter-theoretical—may be rejected for a lack of "rigor." This creates a paradox in which a system designed to reproduce a "Gold Standard" is rewarded, while a system that challenges the very nature of that standard is dismissed. If these important points are ignored, and reviewers systematically insist on metrics borrowed from, for example, mainstream ML, a systematic bias naturally arises against the most creative submissions.

The requirement of reproducibility also presents important challenges. Creativity research frequently involves processes that are intentionally singular, exploratory, or context-dependent. To require that results be straightforwardly reproducible is, in many cases, to require that they not be creative. Reproducibility is a genuine scientific virtue—but it sits in tension with novelty. If a result can be straightforwardly reproduced, it may indicate that the process is well-understood and parametric, which is at odds with what we typically mean by creative surprise. When reviewers demand strict reproducibility as a gate-keeping metric for ICCC, they create a "guardrail" that accidentally blocks novel, "one-off," or truly emergent creative phenomena. Requiring reproducibility as a blanket condition risks filtering out precisely the exploratory, one-off, or process-oriented work that ICCC was meant to host. The irony is that this requirement most severely penalises the work that is most genuinely experimental in character.

Venue vs. Meeting Place

This is perhaps the hardest thing to articulate precisely, but it is the change that concerns us most. ICCC began as, and for many years remained, a place where a community went to think together. Papers were a vehicle for conversation, not the point in themselves. The informal exchanges—at lunch, between sessions, over dinner—were where ideas were tested and collaborations formed. Researchers attending the conference discussed issues, collective decisions were made and the field's shared values were renegotiated.

We have watched that function erode. Attendance decisions are increasingly driven by publication needs rather than intellectual engagement. The number of researchers genuinely interested in establishing specific rules to drive research on creativity forward beyond the standard approaches to science has declined. The number of papers accepted at the conference that discuss issues important to creativity but which would not have been accepted at other STEM conferences has also declined. The conference has gradually come to resemble what it was not supposed to be—another ranked venue where researchers deposit output rather than a community where researchers think together. Once this dynamic takes hold, it becomes self-reinforcing: the less the conference functions as a community hub, the fewer reasons the people that would have contributed the most have to attend it in that spirit. We have reached a point where many researchers seem to feel that they must play by rules that value numbers over nuance.

Sources of Pressure

The problem is compounded by two sources of pressure on the participants of the conference: *institutional pressure* and *the rankings trap*.

Institutional Pressure Funding agencies, research assessment exercises and the logic of bibliometric ranking all impose real costs on researchers, especially those early in their careers, for investing in a venue that does not appear on the approved lists used by the institutions that evaluate them. We are not unsympathetic to this pressure. We have watched talented younger colleagues navigate it with less support than they deserved.

The Ranking Trap Conferences assigned a low ranking by any body seen as credible suffer a structural disincentive that affects who submits, who reviews, and who attends. Trying to rectify this by chasing a higher ranking requires behaving more like venues ranked highly, which can introduce untenable compromises. This dynamic tends to reward conformity and potentially penalise distinctiveness. At the same time, it is often equally untenable to ignore ranking information—many researchers face an environment in which, at best, they are discouraged from investing time in venues that are not highly ranked or that lack "objective" scientific markers.

Impact on Conference Management These two additional factors put pressure on conference organizers to take measures that further complicate the situation. Attempts have been made over the last few years to improve the scores assigned to the conference by recognised ranking agencies by making efforts to identify the criteria employed by such agencies. This effort led to the identification of two options that supposedly impact a conference's ranking. The first one involves ensuring that the set of conference organizers includes researchers that already have an established record in conferences ranked highly. This runs the risk of working against the idea of the existence of researchers specializing on computational creativity and mainly publishing their work in this conference. The second one involves lowering the acceptance rate for papers submitted to the conference. This runs the risk of further reducing the number of experimental papers accepted at the conference.

Under such circumstances, one reaction might be to despair of recovering the original spirit of the conference on computational creativity; another would be to suggest some drastic changes and possibly some sacrifices.

Suggested Remedies

Having identified symptoms and sources of pressure that might be contributing to the observed drift, we undertake the intellectual effort of suggesting some measures that may help reduce the worst issues. These should not be considered as prescriptive but as a listing of possibilities that the community may want to discuss. We propose the following: *development of reviewing rubrics*, *training for new reviewers*, *finetuning the conference format* and *engagement with ranking questions*. The proposals emerge from our experiences and from our observations of how similar dynamics have played out in other communities. We offer them as serious prompts for discussion, while acknowledging that none of them is straightforward to implement.

Develop Reviewing Rubrics As it stands, the conference is lacking rubrics that are specific to paper type and honest about the field's methodology. Ideally, what is wanted is a (possibly controversial) shift in our reviewing philosophy that explicitly makes available in the reviewing templates some of the tenets of computational creativity research that over time have been identified as characteristics of the field. We need suggestions that may allow any reviewer, from whatever discipline, to write a good review: by making the review template guide them towards the type of reviewing they should be doing. If the right questions are included in the review template, even reviewers new to the field can provide valuable feedback, assuming that they have good disposition and that they are willing to consider the indications in the templates. Many conferences have gone this way before us. If we are looking for papers that make the conference *lively*, we should put that as a specific question in the review form. If we value papers that are *provocative*, we should include a question about that explicitly in the review form. If we want the papers to go beyond simply reporting on working systems that undertake supposedly creative tasks, we should include in the form explicit questions that require some kind of evaluation. If we want the evaluation to focus on creativity rather than correctness, we should make it explicit in the forms. If we consider that the most interesting systems are often the hardest to evaluate quantitatively, we should put that as a caveat right next to the explicit questions about evaluation in the review form.

To address the problem of the many different types of paper described in the call for papers, we should develop explicit, concrete rubrics for different paper types that are calibrated to what constitutes good work in each category. For each type, the criteria for "interesting and worth discussing" should be made explicit, so that reviewers are not left to apply default disciplinary heuristics. Further, these rubrics should state clearly, for reviewers and authors alike, that

quantitative evaluation and reproducibility are not universal requirements, rather explain explicitly how they might be detrimental to achieving creativity in a computational system. To summarize, the set of proposed rubrics should clearly describe what constitutes compelling evidence for work that does not lend itself to such "traditional standards".

Train Reviewers Along with suggestions above for overhauling and improving review rubrics, reviewers themselves must also be "overhauled"; they must come to understand that because ICCC is a different sort of conference, a different sort of reviewing is required, one that both safeguards that uniqueness and that ensures quality contributions. First, reviewers must attend the conference; if they do not, it is unlikely that they will ever comprehend what it is that they are tasked with. Second, they must appreciate that the study of (computational) creativity is by definition, and for the many reasons mentioned in previous sections, somewhat unique when compared with traditional STEM fields. Third, they must have an open mind about what makes a valuable contribution and be willing to accept that such a contribution might look very different than what they are used to seeing in other conferences they attend. Accomplishing this level of education is a daunting task, and any solutions must be lightweight, given that conference organizers are overburdened. Senior PC members must be willing to take stronger positions during review discussions—this in itself can be a significant source of training, and meta-reviews should be conducted by an SPC member that did not also perform a normal review of the paper.

Finetune the Programme Format While it is true that the standard conference format is the standard for a reason, it is also true that attendees of ICCC universally agree that all the best discussion/learning/creativity happens outside the formal conference sessions: at coffee breaks, in the halls, over lunch, etc. Presenting and consuming traditional paper talks is probably here to stay and does serves some important purposes; however, we can re-emphasize the uniqueness of ICCC as a meeting place rather than a venue with subtle design tweaks that provide more time for the quality informal discussions that attendees value most: reemphasizing the social program, coffee breaks and communal meals; contemplating new kinds of question/answer formats (what about submitting questions in advance? Or the speaker asking questions for the audience to answer? Or...? Even subtle programme design choices can likely have an outsized impact on how the conference affects the experience of attendees and provides the sense that attending ICCC is valuable *beyond* just gaining a(nother) publication.

Engage with the Ranking Question Rather than either ignoring external evaluation pressures or simply submitting to their logic, the community could invest in investigating what options actually exist. What ranking bodies exist? Are some more qualified for understanding the goals and importance of a conference like ICCC? Are some more amenable to alternative forms of making one's case? In what sort of conference category does ICCC belong? AI, for example, might seem like an obvious suggestion here, but in fact, upon careful introspection, it likely does not share nearly as much of a platform with ICCC as one might initially guess. How can we find a balance between preserving the "fun" and "finesse" of CC research and having the value of that research recognized by agencies whose opinion matters to those outside our community?

Conclusions

We have put forth concerns suggesting that the conference may be drifting away from its original reason for existence. However, we believe the current trend is not inevitable but that it is possible to redirect the future of ICCC so that it can provide a unique, fertile space for discovering where computational creativity can take us. We have tried to condense a set of issues that we have identified over the years, but these are not exclusively ours, and they are not necessarily complete nor agreed upon by everyone. Different interests and sensitivities can improve this proposal and complement it with different views. However, we do believe these concerns speak to real issues that must be openly debated and addressed, even when opposing positions arise.

Other disciplines and communities face similar problems. Balancing identity and content quality against research evaluation agencies is not an issue exclusive to ICCC. However, the very spirit of the conference is achieving creativity (in whatever form) and this strictly *requires* the community to break free from the straitjacket of traditional rubrics for evaluating contribution. If we simply apply traditional norms, we face the risk of becoming a(nother) subordinate venue for machine learning/artificial intelligence research. By definition, creativity lives in the spaces which cannot be simply reached by producing one more output. In order to reach those spaces, we need to leap into the gray areas in which right or wrong, good or bad are not clearly definable.

However, one possible consequence to the various considerations outlined in the paper arises as a possible corollary. Given the conflicting tensions between the expectation of funding agencies and ranking bodies on one hand, and the valuable features that once made the conference on computational creativity a desirable venue for researchers interested in modelling creativity on the other, we may be faced with a future in which:

1. Researchers are encouraged to publish their computationally heavy, metric-driven work in mainstream AI/ML venues
2. ICCC explicitly re-positions itself as a forum for interdisciplinary, high-risk ideas that would not fit elsewhere

Would this be desirable for the conference? Is it tenable in today's academic environment? Might such a dichotomization signal a rebirth for the conference? Or, a death-knell? These are questions we urgently need to consider.

In any case, we believe that the hurdles we now face are not insurmountable. As a community, we have the tools to define ourselves and our practices, be it with clearer goals for the conference or more investment of effort. We need to "renew our vows" and, against the new landscape of artificial intelligence and data science, create a space in which mere numerical evaluation does not constrain our progress.

World-State Transformations for Neuro-symbolic Interactive Storytelling

Santiago Góngora[1] † **Luis Chiruzzo**[1] **Gonzalo Méndez**[2] **Pablo Gervás**[2]

[1] Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay
[2] Facultad de Informática, Universidad Complutense de Madrid, Spain
† sgongora@fing.edu.uy

Abstract

Large Language Models (LLMs) have changed the possibilities of Interactive Storytelling systems that process free-text user input. However, as more of these systems are built, evidence continues to mount regarding the story coherence problems that arise when relying solely on them. Recent research suggests that LLMs can effectively predict state changes within rule-based Interactive Storytelling systems, triggering pre-programmed **world-state transformations**. In this paper, we conduct an exploratory evaluation of whether such *transformations* can serve as a catalyst for player expression while aiming to address the incoherence issues typical of purely LLM-based approaches. Building upon PAYADOR, a neuro-symbolic architecture, we conducted experiments using an open-source model (Llama 3 70B) and a closed-source model (Gemini 1.5 Flash), with testing in both English and Spanish. Eight participants played two scenarios, designed to assess different evaluation objectives. Our observations suggest *transformations* improve LLM error traceability and enable logic checks, thus maintaining world-state consistency, while encouraging players to interact with the environment through their unique, creative inputs.

Introduction

Interactive Storytelling (IS) is a broad term for any system that allows users to influence how a story develops (Trichopoulos, Alexandridis, and Caridakis 2023). These systems face a constant struggle to balance *player agency* (i.e., the level of control over the story the user has) with a coherent plot, as giving the user more control often compromises the story's consistency (Riedl and Bulitko 2012)

One of the most prominent types of IS systems is Interactive Fiction (IF), in which players engage with a simulated environment through text-based *player input* (or *user input*). These games were remarkably sophisticated for their time, addressing world-modeling through complex rule sets grounded in formal linguistic structures (Reed 2023). Limited by their parsers, fixed action sets and a lack of common sense understanding, they often stifled player creativity and expression, constraining player agency. Large Language Models (LLMs) could address these challenges and revolutionize Human-Machine co-creation (Ferreira 2026), yet they still struggle with narrative pace and logical consistency (Góngora et al. 2023; Tian et al. 2024; Teleki et al. 2025). These limitations highlight a critical need for methods that ground LLM generation in more stable structures.

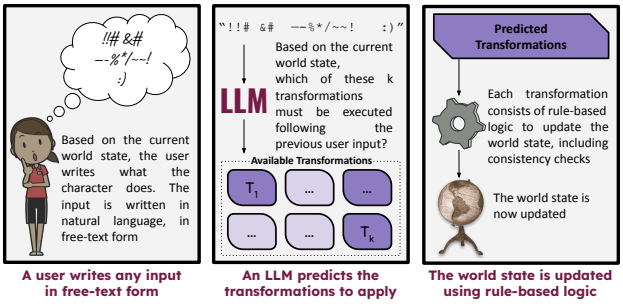


Figure 1: An overview of how *transformations* can be used to tackle the world-update problem.

In this paper, we present a pilot study of world-state *transformations*, a generalization of classic IF actions that are suggested by LLMs and symbolically executed in a previously defined structured world state (see Figure 1). We aim to evaluate whether such *transformations* provide a framework for IS systems designed to model a coherent world and allow players to express themselves through their own creativity. We first conceptually discuss *transformations*, and then detail the methodological design of our pilot study and the analysis of the game logs. Finally, we present the key observations and findings from the evaluation. The source code is available on GitHub[1].

Previous Work on the World-update Problem

The main topic of this paper is the **world-update** problem in IF systems, and text-based Interactive Storytelling in general. Although multiple definitions exist (Riedl and Bulitko 2012; Martin, Harrison, and Riedl 2016; Hausknecht et al. 2020; Callison-Burch et al. 2022; Góngora et al. 2024), the problem may be defined at a high level as follows:

$$action\ description\ (world\ state_n) \longrightarrow world\ state_{n+1}$$

where $n \in \mathbb{N}$ represents a point in time of the represented fictional world. That is, $world\ state_{n+1}$ is a consequence

[1] https://github.com/sgongora27/transformations

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

of doing the described $action$ in the $world\ state_n$[2].

Besides the parsers used in classic IF games like Zork—triggering a world-update process that was specifically pre-programmed for each available action (Reed 2023)—many recent works (e.g., AI Dungeon[3]) tried using LLMs for this purpose (Triyason 2023; You et al. 2024). The aforementioned cases are representative for two very different strategies to tackle this problem, but there are many other ways to do it. Xie et al. (2025) represent the *world state* through simple natural language sentences and use two fine-tuned LLMs to process the user input: one to detect invalid actions and another to generate outcomes for valid ones. Neuro-symbolic approaches (i.e. those combining artificial neural networks, like LLMs, and symbolic approaches) have also been explored to tackle this problem. Callison-Burch et al. (2022) fine-tuned a language model that tracks the game state and generates a response based on it. One of their main conclusions is that additional machinery would be needed to track the game state beyond a Language Model.

LLMs Predicting a Change in the World State

Within a brief period, some works have proposed using LLMs to update structured world-state representations from free-text user input.

PAYADOR (Góngora et al. 2024) is an approach to the world-update problem that tries to break with the tradition of using specific pre-programmed actions like in classic IF games, and focuses on modeling the general outcome of those actions instead. This conceptual shift draws directly from modeling Tabletop Role-playing Games (TTRPG) and how Game Masters (GM) reason about the world state in response to player actions: rather than relying on an objective methodology for every possible action (something impossible given the open-ended nature of these games) they evaluate how a player's intent within its unique context will impact the fictional world (Tychsen et al. 2005).

While PAYADOR was designed as a system-agnostic GM model, shortly thereafter, Song, Zhu, and Callison-Burch (2024) introduced a specialized GM model for "Jim Henson's Labyrinth: The Adventure Game". In this model, an LLM maintains an ad-hoc structured representation through function calls. In a contemporaneous effort, Wang et al. (2024) also addressed LLM reasoning over world-state updates, but focusing on realistic environments. While trying to check if LLMs can act as a reliable world model, they found sometimes this strategy can work better than predicting the whole world state each time.

World-state Transformations

The previously described works lead us to define what we call **world-state transformations**. *Transformations* are pre-programmed routines designed to symbolically update a world-state. While retaining the spirit of classic IF, they aim to funnel a wide range of player actions, including the unpredictable ones born from creative expression. For instance, a player might use a standard medkit or spell[4], or they could improvise by creating or bartering for a remedy: an interaction too context-dependent for hard-coded logic to anticipate. Such cases demand flexibility from the resulting IS system. *Transformations* may provide it, allowing LLMs to decide which of the available ones apply after any player input. Moreover, a single input may trigger one or more *transformations*, which collaborate together to shape the fictional world through synergic changes.

Classic approaches to the world-update problem triggered changes based on specific actions, limiting the player's options to a fixed set of outcomes predefined by the designers. Instead, *transformations* focus on general changes in the world state, regardless of the specific actions that trigger them. As a middle ground between purely neural systems and classic approaches, this shift in focus enables pre-programmed updates to the world state while allowing players to express themselves through their own vocabulary and unique ideas. As we will later discuss, *transformations* seem to have enabled our evaluators to follow this direction.

Methodology

As previously noted, both Góngora et al. (2024) and Song, Zhu, and Callison-Burch (2024) utilized world-state transformations within their respective IS systems. However, while the former did not include an empirical evaluation, the latter focused exclusively on human analysis of LLM-simulated sessions. Consequently, our objective was to conduct a pilot study with humans using a live system (allowing for real-time interaction) to address three research questions regarding *transformations*:

- Are LLMs capable of correctly suggesting *transformations* when grounded on a structured world state?
- Are *transformations* suitable to tackle the world-update problem?
- Do *transformations* encourage more creative player inputs?

Based on the PAYADOR source code[5], we implemented a system, designed two scenarios, and conducted an evaluation with eight players. We split the group evenly to cover two languages (English and Spanish) and two models (Gemini 1.5 Flash and Llama 3 70B).

Figure 1 illustrates how LLM-suggested *transformations* are used in our rule-based IS system. First, the symbolic IS system is built using an ad hoc representation that includes pre-programmed *transformations*, each having additional pre-programmed consistency checks. Then, every turn, an LLM receives the symbolic $world\ state$ (or a textual rendering of it), the free-text $player\ input$, and the list of available *transformations*, along with instructions on how to suggest them. After the LLM determines which *transformations* are triggered by the $player\ input$, the symbolic world model executes them subject to the corresponding consistency checks.

[2]Broadly speaking, it may happen that the described action has no impact in $world\ state_n$, in which case $world\ state_{n+1} = world\ state_n$.

[3]https://aidungeon.com/

[4]Classic IF actions a game designer typically considers.

[5]https://github.com/pln-fing-udelar/payador

System Description

The original PAYADOR version (Góngora et al. 2024) models the fictional world by considering `Items`, `Locations`, and `Characters`. We improved it by integrating `Puzzles` (see Table 1 in the appendices for details) and `Objectives`. Each scenario can have one of the following objectives: the player is in a `Location` or has an `Item`; the player is in the same `Location` of another `Character`; or an `Item` is in a specific `Location`. This was essential to conduct our experiments, as `Objectives` allow to establish clear completion criteria for our test scenarios, hence helping us determine if *transformations* alone can act as a vehicle in human-machine communication.

We started from the *transformations* included in the original PAYADOR version and improved them with robust consistency checks:

- *Moved Items* (MI-t): The LLM can suggest that an `Object` has to be moved from the player's inventory to the current `Location`, or vice-versa. It also includes the case of the player giving an NPC an `Object`, or vice-versa.
- *Unblocked Locations (UL-t)*: The LLM can suggest that a blocked `Location` is now reachable.
- *Player Movement (PM-t)*: The LLM can suggest that the player's `Character` has to be moved to a new `Location`.

Naturally, we could have included other *transformations*. However, to ensure a controlled experimental setting, we maintained this minimal set that aligns with traditional IF mechanics. Given that we placed no constraints on how evaluators wrote their inputs, this decision narrows the range of possibilities during the experimental phase.

Experimental Evaluation

In this section we will present the two pre-designed scenarios and the results of each playthrough[6]. We manually examined every turn of each playthrough, tagging them according to two type of errors: **LLM** and **World Modeling** errors.

The first type is related to errors found when the **LLM** predicts the *transformations* to be done: when the suggested *transformations* do not correctly represent the outcomes of the *user input*, or the *transformations* are not correctly written by the LLM. This error is further split into three categories, one for each *transformation* (*MI-t*, *UL-t* and *PM-t*). In order to provide further insight into these cases, the world-update prompt and examples for text-based world rendering are included in the appendices.

The second type involves **world modeling** errors: cases where the LLM correctly proposes *transformations*, but the final result fails to capture the player's intended actions or logical sequence due to system limitations. This time we break this type of error into two categories. The first, *Planning*, refers to the execution order of *transformations*. The system uses a default *MI-t* $\rightarrow$ *UL-t* $\rightarrow$ *PM-t* sequence, so it fails when a player needs to do things in a different order. The second one, *memory*, refers to the lack of memory in PAYADOR, where the only information stored is the one in the structured representation, as the contents of the prompt are reset every turn. This means that the rest of details, such as previous dialogue utterances, are not saved. Some errors arise when the player refers to details that were not stored.

[6]All the game logs are available on GitHub: https://github.com/sgongora27/transformations

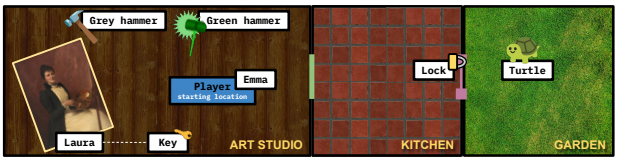


Figure 2: A graphical representation of Scenario A, including 2 `characters` (*Laura* and *Emma*), 5 `items` (*Grey hammer*, *Green hammer*, *Key*, *Lock* and *Turtle*), and 3 `locations` (*Art Studio*, *Kitchen* and *Garden*).

Scenario A

The first scenario consists of three *locations*. The player acts as *Emma*, a girl that has to find her *Turtle* and take it to the *Kitchen*. Starting in her mother's *Art Studio*, *Emma* has to get the *grey hammer* or the *key* (from her mother, *Laura*) to unlock the *Garden* door, where the *Turtle* is located. This sequence requires frequent world-state updates using all three *transformations*, aiming to intensely test them within a simple, realistic environment. Figure 2 illustrates the scenario; see the appendices for an enlarged version (Figure 3) and the manual analysis results in Table 2.

Overall, the eight testers managed to end the scenario. Some of them tried opening the lock with the *key*, and some of them broke it using the *hammer*. In both languages, errors in the LLM's *MI-t* suggestions were all traceable to an interesting issue: many players called the *Turtle Hojita* since the world state specifies that "Emma calls it *Hojita*" (e.g., "I place Hojita in the kitchen, and walk away to the garden again, to enjoy the rest of this sunny day"-TesterD). While we expected the LLM to resolve this co-reference, it occasionally suggested an incorrect *MI-t* using the name "Hojita", what was not recognized as a valid object by the world-modeling module, hence rejecting the *transformation*. Additionally, some *PM-t* errors occurred because the player input was not explicit (e.g., "As I walk through the door I look around trying to find my turtle, what do i see"-TesterC). A single *UL-t* error occurred when the LLM tried to unblock a location already shown as unblocked in its context.

Regarding world-modeling, many errors arose from the system's lack of *planning*. Players often tried to move to the *Kitchen* and drop the pet: an action requiring *PM-t* before *MI-t*. Our fixed execution order cannot accommodate this sequence, highlighting a limitation of the current default order. Finally, there was a single error due to the memory of the system, when TesterE said "agarro el otro martillo y me dirijo a la cocina" (*I take the other hammer and head to the kitchen*), but the LLM could not disambiguate between the *Grey hammer* and the *Green hammer*.

Scenario B

The second scenario also consists of three *locations*. This time, the player acts as *Venancio*, a *gaucho*[7] that has to find *Artigas*[8], who is locked in a *Cell* protected by a *Puzzle*. In order to enter the *Cell*, the player must first put out a *firewall*[9] and then solve a riddle. To put out the *firewall*, the player has to use *Venancio*'s magic to summon a giant wave of water, a superpower specified in the world state. In this case, the evaluation objective was twofold: first, to determine if LLM-predicted *transformations* adhere to non-realistic magical physics; second, to assess the LLM's ability to describe critical scene elements, such as *Venancio* summoning a water wave, ensuring players receive essential information to complete the scenario. In the appendices, Figure 4 illustrates this scenario, while Table 3 details its corresponding results.

As with Scenario A, all testers successfully completed the task. Notably, world-modeling errors dropped to zero as a result of limited environmental interaction. This trend extended to the *MI-t* category, where the sole error involved a consistency check rejecting TesterE's attempt to improvise an undeclared `Item`. However, this scenario yielded several notable *UL-t* errors. First, the particularity of having to solve a `Puzzle` to open the *Cell* door introduced interesting cases. Despite having the correct answer in-context, the LLM incorrectly validated TesterA and TesterC's wrong riddle answers, suggesting a *UL-t*, hence unblocking the *Cell*. These cases illustrate how symbolic grounding mitigates LLM-dependent errors, but cannot entirely eliminate them. Second, during TesterF's session, the LLM suggested an *UL-t* for an already reachable place, mirroring TesterG's experience in Scenario A. Finally, while the wall of fire challenge sparked unconventional (and successful) ideas among the players, the LLM did not suggest a *UL-t* when TesterB attempted to summon the wave by playing Venancio's guitar.

Transformations and Co-Creativity

Having discussed our observations from the qualitative analysis of errors, we will now explore the implications of *transformations* for Interactive Storytelling systems.

We will start from the most salient fact: all players were able to complete the scenarios. Our results indicate no substantial disparity in performance between the open-source model (Llama) and the closed-source one (Gemini), remaining consistent across both English and Spanish. Although our experiments are preliminary, the results suggest that *transformations* effectively bridge human-machine communication in LLM-enhanced IS systems, and that current LLMs are already capable of handling them.

Moreover, we observe that *transformations* seem to foster player creativity. For instance, TesterE unexpectedly used a window mentioned in the world state to enter the *Garden*, bypassing the intended *Lock*; and TesterG tried to summon the wave of water by playing "the call of the water" on the *guitar*. More generally, players in both scenarios engaged by role-playing their characters and adding creative details to their inputs (e.g., speaking with *Laura* in Scenario A, examining the magical setting or attempting to shout in the *Silent Zone* in Scenario B). This capability to leverage creative inputs may prove useful for systems requiring varying degrees of improvisation in human-machine co-creation. Consequently, *transformations* may be suitable for modeling TTRPGs, which scholars characterize as a form of organized, collaborative storytelling driven by rich player-GM interactions (Mäyrä 2017; Katifori et al. 2022).

While our experiments are preliminary, their results suggest that *transformations* offer a promising path for consistent world updates in IS systems. Evidence indicates they broaden player expression by providing numerous options to trigger pre-programmed updates, making them ideal for improvisational systems like TTRPG models. Although LLM errors may impact world coherence, the symbolic nature of *transformations* facilitates the integration of logic to verify and refine outputs. Following the idea that improvisational storytelling is limited only by player imagination (Martin, Harrison, and Riedl 2016), our results suggest that *transformations* represent a step forward in this direction.

Conclusions

In this paper, we evaluated world-state transformations as a method for tackling the *world-update* problem in Interactive Storytelling. By using an LLM to suggest pre-programmed world-state *transformations*, the system moves away from manually defined actions toward **a model of general world changes**. This shift in focus encourages players to interact more creatively with the environment while maintaining the controllability inherent in symbolic world-state representations.

Our pilot study with eight participants across two scenarios confirmed the viability of this approach: all players successfully completed both scenarios and engaged in co-creation while acting within the fictional worlds. While the system inherits common LLM limitations, leading to errors like prematurely solved puzzles, the symbolic nature of *transformations* allows for consistency checks that improve traceability and provide more transparent explainability than purely neural approaches.

Our experimental findings suggest several directions for further research. First, future work should explore creating world components from scratch, extending our recent preliminary efforts (Vaucher et al. 2026), and combining *transformations* to create them on the fly, thereby better modeling the improvisation inherent in TTRPGs. As the number of *transformations* grows, incorporating a planner to dynamically order their execution will become essential. Finally, to minimize the impact of incorrect LLM outputs, we aim to optimize how consistency checks validate them.

We hope this work inspires other researchers to push the boundaries of human-machine co-creation in Interactive Storytelling, and the application of TTRPGs models to different domains.

[7] According to the Cambridge Dictionary: *a South American cowboy who is skilled at riding a horse*.

[8] A historical figure of South America.

[9] After running all the experiments we realized this component should be called *Wall of Fire*. Apparently, the word *firewall* generated no problems during the experiments in English.

Author Contributions

This work is part of S.G.'s Master's thesis (Góngora 2025), for which L.C. and G.M. served as co-advisors. P.G. helped frame the conceptual aspects of this work.

Acknowledgments

This paper was partially funded by *Agencia Nacional de Investigación e Innovación* (ANII, Uruguay), Grant No. $POS_NAC_2022_1_173659$; *Comisión Académica de Posgrado* (CAP, Uruguay), Grant No. $BDDX_2026_1\#48948494$; and by the project DARK NITE: *Dialogue Agents Relying on Knowledge-Neural hybrids for Interactive Training Environments*, Grant No. $PID2023-146308OB-I00$ (Spanish Ministry of Science and Innovation).

References

Callison-Burch, C.; Tomar, G. S.; Martin, L. J.; Ippolito, D.; Bailis, S.; and Reitter, D. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9379–9393. Association for Computational Linguistics.

Ferreira, C. 2026. Co-creating narratives: Ludonarrative agency and generative ai in digital games. In *Reshaping the Video Game Landscape With AI and GenAI*. IGI Global Scientific Publishing. 87–122.

Góngora, S.; Chiruzzo, L.; Méndez, G.; and Gervás, P. 2023. Skill Check: Some considerations on the evaluation of gamemastering models for role-playing games. In *International Conference on Games and Learning Alliance*, 277–288. Springer.

Góngora, S.; Chiruzzo, L.; Méndez, G.; and Gervás, P. 2024. PAYADOR: A minimalist approach to grounding language models on structured data for interactive storytelling and role-playing games. In *Proceedings of The 15th International Conference on Computational Creativity*.

Góngora, S. 2025. Approaches to interactive and improvisational storytelling. Master's thesis, Universidad de la República (Uruguay). Facultad de Ingeniería.

Hausknecht, M.; Ammanabrolu, P.; Côté, M.-A.; and Yuan, X. 2020. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7903–7910.

Katifori, A.; Petousi, D.; Sakellariadis, P.; Roussou, M.; and Ioannidis, Y. 2022. Tabletop role playing games and creativity: The game master perspective. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, FDG '22. Association for Computing Machinery.

Martin, L. J.; Harrison, B.; and Riedl, M. O. 2016. Improvisational computational storytelling in open worlds. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*, 73–84. Springer.

Mäyrä, F. 2017. Dialogue and interaction in role-playing games. *Dialogue across Media* 28:271.

Reed, A. A. 2023. *50 Years of Text Games: From Oregon Trail to AI Dungeon*. Changeful Tales Press.

Riedl, M. O., and Bulitko, V. 2012. Interactive narrative: An intelligent systems approach. *AI Magazine* 34(1):67.

Song, J.; Zhu, A.; and Callison-Burch, C. 2024. You have thirteen hours in which to solve the labyrinth: Enhancing ai game masters with function calling. In *The 4th Wordplay: When Language Meets Games Workshop*.

Teleki, M.; Bengali, V.; Dong, X.; Janjur, S. T.; Liu, H.; Liu, T.; Wang, C.; Liu, T.; Zhang, Y.; Shipman, F.; and Caverlee, J. 2025. A survey on LLMs for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 13954–13966. Association for Computational Linguistics.

Tian, Y.; Huang, T.; Liu, M.; Jiang, D.; Spangher, A.; Chen, M.; May, J.; and Peng, N. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17659–17681. Association for Computational Linguistics.

Trichopoulos, G.; Alexandridis, G.; and Caridakis, G. 2023. A survey on computational and emergent digital storytelling. *Heritage* 6(2):1227–1263.

Triyason, T. 2023. Exploring the potential of chatgpt as a dungeon master in dungeons & dragons tabletop game. In *Proceedings of the 13th International Conference on Advances in Information Technology*, IAIT '23. Association for Computing Machinery.

Tychsen, A.; Hitchens, M.; Brolund, T.; and Kavakli, M. 2005. The game master. In *Proceedings of the Second Australasian Conference on Interactive Entertainment*, IE '05, 215–222. Creativity & Cognition Studios Press.

Vaucher, M.; Silveira, S.; Góngora, S.; and Chiruzzo, L. 2026. IVIE: A neuro-symbolic approach to incremental and validated generation of interactive fiction worlds. In *Proceedings of The 17th International Conference on Computational Creativity*.

Wang, R.; Todd, G.; Xiao, Z.; Yuan, X.; Côté, M.-A.; Clark, P.; and Jansen, P. 2024. Can language models serve as text-based world simulators? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1–17. Association for Computational Linguistics.

Xie, K.; Yang, I.; Gunerli, J.; and Riedl, M. 2025. Making large language models into world models with precondition and effect knowledge. In *Proceedings of the 31st International Conference on Computational Linguistics*, 7532–7545. Association for Computational Linguistics.

You, X.; Taveekitworachai, P.; Chen, S.; Can Gursesli, M.; Li, X.; Xia, Y.; and Thawonmas, R. 2024. Dungeons, dragons, and emotions: A preliminary study of player sentiment in LLM-driven TTRPGs. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24. Association for Computing Machinery.

Appendices

World-state components

Table 1 shows the four classes present in our system, strongly based on the PAYADOR source code (`Locations`, `Characters`, `Objects` and `Puzzles`). All these components are tracked by a `World` class, representing the *world state*.

Item	**Location**
Name: **String** *Descriptions*: List [**String**] *Gettable*: **Boolean**	*Name*: **String** *Descriptions*: List [**String**] *Items*: List [**Item**] *Connecting locations*: List [**Location**] *Blocked locations*: List [**Location**, *obstacle*]
Character	**Puzzle**
Name: **String** *Descriptions*: List [**String**] *Location*: **Location** *Inventory*: List [**Item**]	*Name*: **String** *Descriptions*: List [**String**] *Problem*: **String** *Answer*: **String**

Table 1: The attributes of each `World` component in our representation. The *obstacle* of a blocked location can be any other component, but it is particularly thought to be an `Item` or a `Puzzle`.

Examples of rendered world-states

Starting scene — Scenario A (English)

```
The player is in <Art studio>
From <Art studio> the player can access: <Kitchen>
From <Art studio> there are blocked passages to:
↪  None
The player has the following objects in the
↪  inventory: None
The player can see the following objects: <A grey
↪  hammer>, <A green hammer>
The player can see the following characters: <Laura>

Here is a description of each component.
<Art studio>: This is the player's location. This is
↪  the art studio that Emma's mom has in the house.
Characters:
- <Player>: The player is acting as <Emma>. A
↪  teenager of average height. She is looking for
↪  her pet 'Hojita'.
- <Laura>: A woman in her 40s. She is Emma's mom.
↪  She is an artist, and loves oil painting. This
↪  character has the following items: <Key>
Objects:
- <A grey hammer>: A big grey hammer that can be
↪  used to break things. It is so heavy...
- <A green hammer>: A small green hammer. It is just
↪  a toy and you cannot break anything with it
- <Key>: A key to open a lock. It is golden. There
↪  is a strange coat of arms engraved on it
```

Starting scene — Scenario B (English)

```
The player is in <Clearing in the woods>
From <Clearing in the woods> the player can access:
↪  None
From <Clearing in the woods> there are blocked
↪  passages to: <Silent zone> blocked by <Firewall>
The player has the following objects in the
↪  inventory: <Guitar>
The player can see the following objects:
↪  <Writings>, <Pond>
The player can see the following characters: None

Here is a description of each component.
<Clearing in the woods>: This is the player's
↪  location. A clearing in a eucalyptus forest near
↪  the Uruguay River. You can hear the sound of the
↪  animals that live in the trees of this forest..
Characters:
- <Player>: The player is acting as <Venancio>. A
↪  Uruguayan gaucho in his 40s. He belongs to the
↪  Artigas army. He has the magical power to summon
↪  a giant wave of water with which he can put out
↪  fires or moisten the ground..
Objects:
- <Writings>: There is something written on the
↪  wall.. It says 'You have to trust in the powers
↪  that have been given to you.'
- <Pond>: A pond full of crystal clear water. The
↪  water is so clear that it works like a mirror
- <Guitar>: A classic guitar with 6 strings. It
↪  sounds great
- <Firewall>: The flames are very hot. It's 3 metres
↪  high. It is impossible to cross them, neither
↪  walking, nor running, nor jumping.
```

World-update prompt

World-update prompt (English)

```
system_msg = f"""You are a storyteller. You are
↪  managing a fictional world, and the player can
↪  interact with it. Following a specific format,
↪  that I will specify below, your task is to find
↪  the changes in the world after the actions in
↪  the player input. Specifically, you will have to
↪  find what objects were moved, which previously
↪  blocked passages are now unblocked, and if the
↪  player moved to a new place.

Here are some clarifications:
(A) Pay attention  to the description of the
↪  components and their capabilities.
(B) If a passage is blocked, then the player must
↪  unblock it before being able to reach the place.
↪  Even if the player tells you that he is going to
↪  access the locked location, you have to be sure
↪  that he is complying with what you asked to
↪  allow him to unlock the access, for example by
↪  using a key or solving a puzzle.
(C) Do not assume that the player input always makes
↪  sense; maybe those actions try to do something
↪  that the world does not allow.
(D) Follow always the following format with the
↪  three categories, using "None" in each case if
↪  there are no changes and repeat the category for
↪  each case:
- Moved object: <object> now is in <new_location>
```

```
- Blocked passages now available:
↪  <now_reachable_location>
- Your location changed: <new_location>
(E) Finally, you can narrate the changes you've
↪  detected in the world state (without moving the
↪  story forward and without making up details not
↪  included in the world state!) using the format:
↪  #your final message#
(F) In the narration section that you add at the
↪  end, between # symbols, you can also answer
↪  questions that the player asks in their input,
↪  about the objects or characters they can see, or
↪  the place they are in.

Here I give you some examples (in parentheses, a
↪  clarification about what the player might have
↪  tried to do) for the asked format, as described
↪  in items (D) and (E):

Example 1 (The player took the axe and put it in the
↪  inventory)
- Moved object: <axe> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None
#You put the axe in your bag#

Example 2 (The player unblocks the passage to the
↪  basement)
- Moved object: None
- Blocked passages now available: <Basement>
- Your location changed: None
# The basement is now reachable #

Example 3 (The player now is in the garden)
- Moved object: None
- Blocked passages now available: None
- Your location changed: <Garden>
# You enter the garden #

Example 4 (The player puts objects in the bag and
↪  leaves the axe on the floor)
- Moved object: <banana> now is in <Inventory>,
↪  <bottle> now is in <Inventory>,  <axe> now is in
↪  <Main Hall>
- Blocked passages now available: None
- Your location changed: None
# You put the banana and the bottle in your bag. The
↪  axe lies on the floor of the Main hall #

Example 5 (The player puts objects in the bag and
↪  leaves the axe on the floor and unblocks the
↪  passage to the Small room)
- Moved object: <banana> now is in <Inventory>,
↪  <bottle> now is in <Inventory>,  <axe> now is in
↪  <Main Hall>
- Blocked passages now available: <Small room>
- Your location changed: None
# You put the banana and the bottle in your bag. The
↪  axe lies on the floor of the Main hall. Now you
↪  can reach the Small room. #

Example 6 (The player puts objects in the bag and
↪  leaves the axe on the floor, unblocks the
↪  passage and goes to the Small room)
- Moved object: <banana> now is in <Inventory>,
↪  <bottle> now is in <Inventory>,  <axe> now is in
↪  <Main Hall>
- Blocked passages now available: <Small room>
- Your location changed:  <Small room>
# You put the banana and the bottle in your bag. The
↪  axe lies on the floor of the Main hall. The
↪  Small room is now unblocked, and you moved
↪  there. #

Example 7 (The player puts the pencil in the bag and
↪  gives the book to John)
- Moved object: <book> now is in <John>,  <pencil>
↪  now is in <Inventory>
- Blocked passages now available: None
- Your location changed:  None
# John now has the book. You put the pencil in your
↪  bag #

Example 8 (The player gives the computer to Susan)
- Moved object: <computer> now is in <Susan>
- Blocked passages now available: None
- Your location changed:  None
# Susan put the computer in her bag #

Example 9 (The player does something that has not
↪  the expected outcome)
- Moved object: None
- Blocked passages now available: None
- Your location changed:  None
# Nothing happened... #

Example 10 (The player asks a question)
- Moved object: None
- Blocked passages now available: None
- Your location changed:  None
# Answer to the player's question #"""

user_msg = f"""Give the changes in the world
↪  following the specified format, after this
↪  player input "{input}" on this world state:
↪  {world_state}"""
```

Scenario results

In this appendix we include tables with the results of the manual analysis of the playthroughs. These tables (Table 2 and Table 3) show LLM-related errors alongside errors stemming from symbolic world-modeling limitations.

Scenario illustrations

In this appendix we include illustrations to help visualizing the pre-made scenarios we designed to test *Transformations*. These figures (Figure3 and Figure4) are for illustrative purposes only; players interacted with the environment via a text-only interface.

A real example using Gemini

Figure 5 shows a real example using Gemini 1.5 Flash, with step-by-step annotations of human utterances and LLM outputs (both GM utterances and suggested *transformations*).

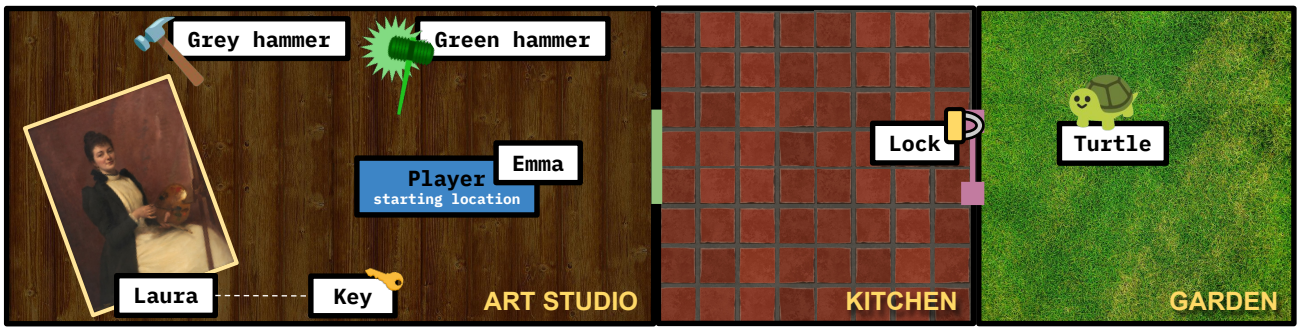


Figure 3: A graphical representation of Scenario A. Ten `world` components are represented: 2 `characters` (*Laura*, an NPC, and *Emma*, the player character), 5 `items` (*Grey hammer*, *Green hammer*, *Key*, *Lock* and *Turtle*, and 3 `locations` (*Art Studio*, *Kitchen* and *Garden*).

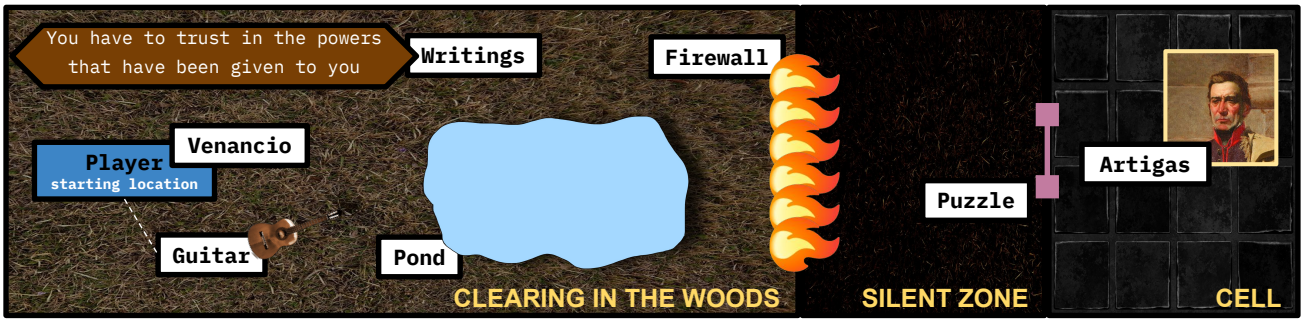


Figure 4: A graphical representation of Scenario B. Ten `world` components are represented: 2 `characters` (*Artigas*, an NPC, and *Venancio*, the player), 4 `items` (*Guitar*, *Pond*, *Writings*, and *Firewall*), 3 `locations` (*Clearing in the woods*, *Silent zone* and *Cell*), and a `puzzle` (*Puzzle*)

Scenario A	LLM Errors			World Modeling Errors	
	MI-t	PM-t	UL-t	Planning	Memory
TesterA (Gemini)	0	0	0	1	0
TesterB (Gemini)	0	0	0	1	0
TesterC (Llama)	3	1	0	0	0
TesterD (Llama)	4	1	0	2	0
Total (English)	7	2	0	4	0
TesterE (Gemini)	3	0	0	2	1
TesterF (Gemini)	1	0	0	0	0
TesterG (Llama)	4	0	1	1	0
TesterH (Llama)	0	0	0	0	0
Total (Spanish)	8	0	1	3	1

Table 2: Number of errors by category for each of the eight testers during the playthroughs of Scenario A. Testers A-D played in English, while Testers E-H played in Spanish. The LLM used is indicated for each case, next to the tester's id.

Scenario B	LLM Errors			World Modeling Errors	
	MI-t	PM-t	UL-t	Planning	Memory
TesterA (Gemini)	0	0	1	0	0
TesterB (Gemini)	0	0	1	0	0
TesterC (Llama)	0	0	1	0	0
TesterD (Llama)	0	0	0	0	0
Total (English)	0	0	3	0	0
TesterE (Gemini)	1	1	2	0	0
TesterF (Gemini)	0	2	1	0	0
TesterG (Llama)	0	1	0	0	0
TesterH (Llama)	0	1	0	0	0
Total (Spanish)	1	5	3	0	0

Table 3: Number of errors by category for each of the eight testers during the playthroughs of Scenario B. Testers A-D played in English, while Testers E-H played in Spanish. The LLM used is indicated for each case, next to the tester's id.

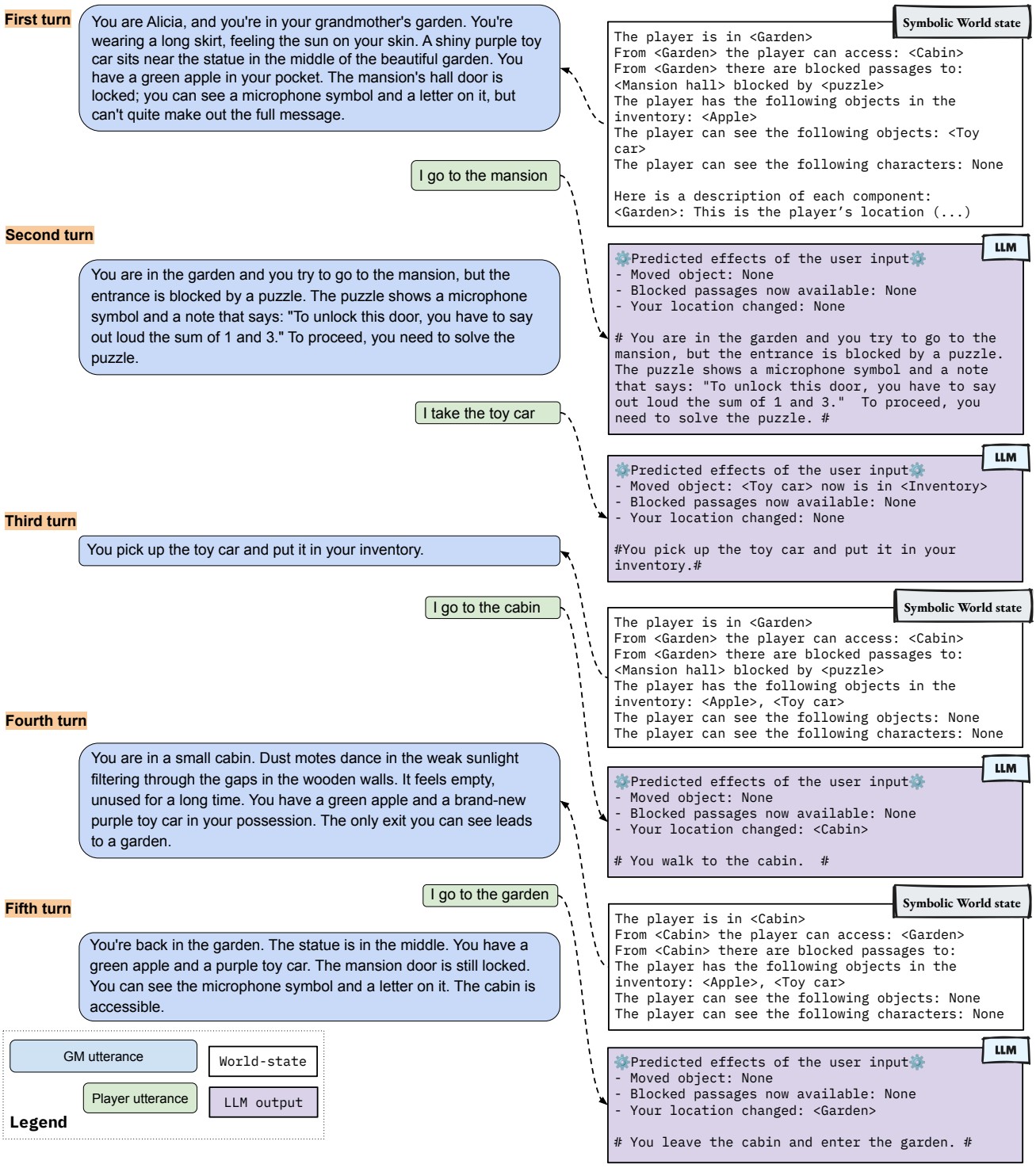


Figure 5: A five-turn gameplay example using Gemini 1.5 Flash. The left panel displays the dialogue between the automated Game Master and the human player, while the right panel shows the symbolic world states (for representative turns) and the LLM-suggested *transformations* (for all turns). A legend is provided in the bottom left-hand corner.

Cross-Modal Associations in Creativity: Evaluating Word-Image Pairing in Humans and Large Language Models

Eva Illion, Linxuan Wang, Liane Gabora
Department of Psychology
University of British Columbia Okanagan
3333 University Way, Kelowna, BC, Canada V1V 1V7

Abstract

Given differences in the associative memory architectures of humans and AI, it is not clear whether AI generates and interprets visual-linguistic associations the way humans do. To investigate this, human participants and two large language models (LLMs), ChatGPT and Gemini, were asked to generate novel words and corresponding images that form meaningfully linked dyads. The LLMs and a second group of human participants were presented with shuffled sets of these words and images and asked to rematch them in ways they felt were most meaningful. Humans performed similarly regardless of whether the pairs were human or LLM-generated, while LLMs were better at matching LLM-generated pairs. Overall, LLMs outperformed humans at recovering the intended pairings. These findings suggest that despite architectural differences, the way in which AI generates and detects associations between language and visual stimuli is not only like that of humans, but they may be superior at recognizing cross-domain associations that humans find meaningful.

Introduction

Cross-domain (or cross-modal) associations are connections between different sensory or perceptual modalities, (e.g., associating a word with an image or a letter with a color). They play an important role in creativity, as they allow for cross-domain reinterpretations of ideas (Gabora, 2017, Ranjan et al. 2013, Ranjan 2014). For example, a dancer can reinterpret a song as movement, or a composer can transform a story into a symphony as an underlying creative idea can manifest in different ways. In other words, it appears that creative output in one domain can be mapped onto another using cross-modal associations. As the architectures of associative memory in humans and LLMs differ, it is unknown whether they form and respond to cross-modal associations in a similar manner. This paper investigates whether AI generates the sorts of word-image associations that humans make, and if they interpret them similarly to humans.

AI forms cross-modal associations based on patterns in training data (Mikolov et al 2013), much the same as humans form cross-domain associations based on statistical patterns in life experiences. Both humans and AI strive for internal coherence (Vieira & Gabora, 2026), and this likely governs how such associations are made. However, in humans the associative processes and strive for coherence are guided by a self: a being that reorganizes its own mental contents from within to preserve its structure as an entity distinct from the world (Gabora and Bach 2023). A human self pays attention to what is relevant to its own needs and interests. In contrast, the associative processing in an AI such as a large language model (LLM) is *externally imposed* by the data we feed it and the prompts we give it. Thus, it is an open question how similar the cross-domain associations made and detected by humans are to those of AI.

Cross-modal Correspondences

Cross-modal correspondences matter not just because they affect aesthetic experience, but because they impact understanding of novel stimuli. For example, in the brightness-weight illusion, people tend to assume brightly coloured objects weigh less, perhaps because both can be described as 'light' (Walker et al. 2010). The most well-known example of associations between visual and linguistic features is Kohler's (1929) Bouba-Kiki paradigm, in which participants tend to match a word with many round vowels (e.g., Bouba) with a round looking shape, and a word with many hard sounds, or stop consonants (e.g., Kiki) with a spiky shape. While there is no objective correct answer in the paradigm, people tend to match the words and visual stimuli as though they are meaningfully linked. Over the past century, this paradigm has been repeated with many variations across many populations (Cwiek et al. 2022).

It has been suggested that cross-modal correspondences are made possible by way of *collative variables*, such as novelty and complexity, that extend across domains (Berlyne 1971). This enables a creative work to be recognizably translated across different domains, a phenomenon known as *ekphrastic expression* (Zilczer 1987). For example, it was shown that when a painting is inspired by a particular song, the two were rated similarly on collative vari-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

ables (Ranjan 2014; Ranjan et al. 2013a,b). Furthermore, viewers could match paintings to the songs that inspired them, intuitively understanding that the song and painting are meaningfully linked. Thus, a creative idea can manifest in multiple ways, and conversely, the 'same' creative work expressed in different domains may resonate similarly with viewers.

Evidence regarding whether AI detects and interprets correspondences between language and visual stimuli is mixed. For example, humans outperform LLMs at the Bouba-Kiki paradigm (Kouwenhoven et al 2025). However, such correspondences would appear to be at the heart of analogy which is one of the benchmarks used to identify an adequately trained LLM (Drozd et al 2016). To gain insight into this, we investigated whether humans and AI can identify underlying connections between newly created words and corresponding visual depictions, and whether LLM generated associations are perceived as fitting by human evaluators. We hypothesized that both humans and LLMs would perform above chance accuracy, as the associative nature of semantic memory should allow for the translation of novel words into images through the presence of collative variables and cross-modal correspondences.

Methods

Human Participants

Human participants were recruited through The University of British Columbia's (Okanagan Campus) subject pool, and given bonus credit in psychology courses in exchange for their participation. For Part One (n = 20) there were 11 females, 6 males, 2 unknown, and 1 nonbinary participant, and their ages ranged between 18 and 45 with a median of 20. For Part Two (n = 25), there were 21 women and 4 men, and their ages ranged between 18 and 24 with a median of 19.

Large Language Models

Human performance was compared with that of two widely-used LLMs with multimodal capabilities: ChatGPT and Gemini. We used the most advanced model available at the time each part of the study was carried out. Thus, ChatGPT 5.2 was used for Part One, and ChatGPT 5.3 was used for Part Two. Gemini 3.0 was used for both Part One and Part Two.

Materials

Part One. Human participants were provided with a Word-Image Worksheet that asked them to generate a series of novel words, and (to the best of their ability) a corresponding illustration or image of each word (see Appendix A for examples). To create the images, human participants were provided with pens, pencils, pencil crayons, and markers.

Part Two. Both human and LLM participants were provided with a matching survey (see Appendix B) that asked them to match which word went with which image. For the human participants, the survey was provided using Qualtrics. For the LLMs, the survey was uploaded in their chat interfaces.

All human participants in Part One and Part Two, were provided with both a consent form and a brief demographics questionnaire to examine potential influences of linguistic demographics on task performance, and to assess their frequency of AI use.

Procedure

Part One. Participants attended a thirty minute in-person session in which they were presented with the consent form, Word-Image Worksheet, art supplies, and demographics questionnaire. ChatGPT and Gemini were provided with the same worksheet in PDF format and instructed to complete it within their chat interfaces. The worksheet prompted each participant to make three novel word-image pairs. Thus, a total of 180 dyads were created in Part One (60 human, 60 Gemini, and 60 ChatGPT made). The instructions simply asked participants to invent a word that they believed was novel and then attempt to visually depict it. The instructions for human and AI generation are provided in Appendix A.

The images were scanned and uploaded to an encrypted database and then formatted into a Qualtrics survey for use in Part Two. After excluding responses that did not meet task requirements (e.g., images that were representational, i.e., depicted known objects, or simply combined preexisting words such as *textneck* or *sandsleeper*), the stimuli were organized into fifteen randomized sets of ten, each of which contained both human-generated and AI-generated images.

Part Two. The LLMs and a second set of human participants completed the Matching Survey, which asked them to pair the generated images with the intended words for them based on perceived semantic fit. It also asked them to rate how good of a fit they felt each pairing was (i.e., goodness-of-fit) to determine the perceived cohesiveness of each matched pair. This design allowed for the evaluation of cross-modal associative reasoning in humans and LLMs in the absence of predefined meanings. Both humans and LLMs were randomly shown three versions of the word-image matching task, each containing a set of ten novel words and ten novel images to be paired for a total of 30 pairs.

Results

As shown in Table 1, Humans, ChatGPT and Google Gemini were all able to match the words with their intended images significantly above a chance level of 10% (since there is a 1 in 10 chance of selecting 1 of 10 displayed images for any given word).

An omnibus ANOVA showed that there were significant differences between humans and the two AI models on this task (F (2,52) = 33, p < .001, $\eta 2$ = 0.56 [0.40, 1.00]) (Figure 1). Tukey's Honest Significant Differences showed that the best performance was obtained by Gemini, followed by ChatGPT, and then humans (for humans vs ChatGPT, p <

.001, d = 1.16, and for humans vs Gemini, p < .001, d = 1.7). Gemini outperformed ChatGPT directionally but insignificantly (p = 0.0799, d = 0.544).

	Descriptive Statistics			T-Statistics			
Group	*n*	*M*	*SD*	*t*	*df*	*p*	*d*
Humans	25	0.2	0.09	4	6	0.003	3
ChatGPT	15	0.4	0.1	12	14	< .001	6
Gemini	15	0.5	0.1	14	14	< .001	7

Table 1: Descriptive statistics for success at the matching task with corresponding *t*-statistics comparing each group's scores to chance level performance (μ = 0.1).

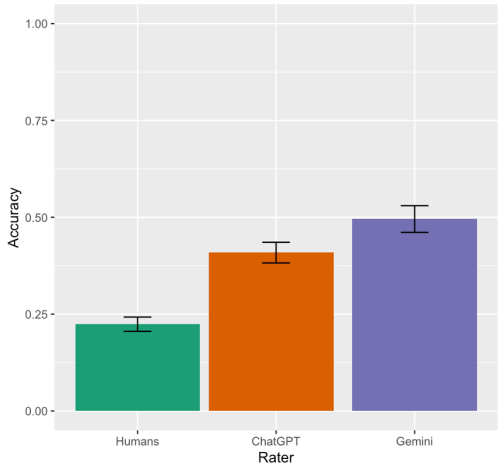


Figure 1: Standardized average accuracy at the matching task for humans, ChatGPT, and Google Gemini.

We were also interested if the pairs would be easier or harder to match depending on if they were created by humans or LLMs. To test these additional hypotheses, we conducted a mixed-methods type-3 ANOVA with Kenward-Roger degrees of freedom adjustments. This method was chosen because each participant was asked to match pairs created by humans, ChatGPT, and Gemini, forming a between-subjects condition of rater type, and a within-subjects condition of creator type (i.e., if humans or AI made the pairs). Type-3 sums of squares were used to account for the unbalanced observation counts. To increase robustness against normality violations, we used Kenward-Roger degrees of freedom (Arnau et al. 2014). This yielded a main effect of both creator type (F (2, 104.792) = 22.386, p < .001, general η^2 = 0.187, partial η^2 = .308) and rater type (F (2, 52.794) = 34.58, p < .001, general η^2 = 0.414, partial η^2 = .593) as well as a strong interaction effect (F (4, 104.764) = 5.6299, p < .001, general η^2 = 0.102, partial η^2 = .181) (Figure Two).

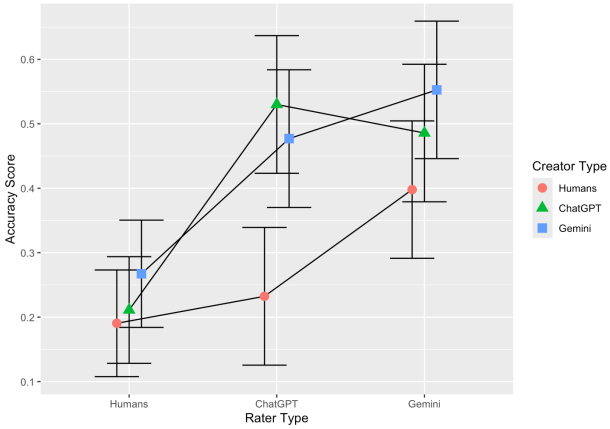


Figure 2: Mixed-methods type-3 ANOVA with Kenward-Roger degrees of freedom adjustments to examine both creator type and rater types impact on accuracy scores.

Pairwise contrasts revealed the nature of the interaction. This paper reports those of statistical significance at an α level of 0.05. All contrasts were pairwise, with Bonferroni corrections and Kenward-Roger degrees of freedom applied to p values. Humans were worse at rating their own pairs than both ChatGPT (t = -7.077, p < .001, d = -2.311) and Gemini (t = -7.552, p < .001, d = -2.466) were at matching their own pairs. Additionally, humans and ChatGPT rated human-generated pairs comparably (p = 1.0), but Gemini was better at matching human-generated pairs than either ChatGPT or humans (t = - 4.329, p = .001, d = -1.414).

Human performance at the matching task did not significantly differ based on the creator type. Humans were no better at pairing human-made pairs then they were at matching ChatGPT (p = 1) or Gemini-made (p = 0.4221) pairs. Thus, while the LLMs generally performed best at matching pairs they had created, humans tend to match all pair types with similar accuracy.

A mixed-methods type-3 ANOVA with Kenward-Roger degrees of freedom yielded main effects of rater type and creator type on perceived goodness of fit (F (2, 52.974) = 38.152, p < .001, general η^2 = 0.539, partial η^2 = 0.612; F (2,104.115) = 21.32, p < .001, general η^2 = 0.123, partial η^2 = 0 .351). The interaction is not significant, but directionally supported, and the study is currently statistically underpowered (F (4, 104.075) = 2.1747, p = 0.07692, general η^2 = 0.029, partial η^2 = 0 .102). Overall, this is consistent with the Gemini > ChatGPT > human trend, which was further corroborated by similar patterns in Bonferroni adjusted pairwise comparisons. Note that even though ChatGPT was a poorer rater of human-generated pairs, it had equal confidence to Gemini in its performance (p = 1). This aligned with previous findings that humans perceived all pair types comparably whether they were generated by ChatGPT (p = 1) or Gemini (p = 0.4221) with no evidence of significant differences.

Discussion

The superior performance of AIs on this task likely reflects that they are trained to simply *track* their world through prediction error minimization as opposed to *actively shaping* their worldview (Vieira & Gabora, 2026). A human may, for example, be intrinsically motivated to write a novel that contains an entire make-belief world, whereas a LLM will not do so unless prompted externally. As such, the LLM is less distracted by associations that reflect not just individual differences in experience, but the idiosyncratic ways in which humans actively explore and shape their own realities. Moreover, since LLMs can access a broad base of data across large numbers of humans, responses can therefore be truer to the statistically averaged responses across all humans. This study suggests LLMs recognize associations underlying word-image pairs that were not strong enough to influence human pairings (much as a dog may be aware of patterns in its owner's behavior that the owner is not aware of). The fact that whether the creator was human, ChatGPT, or Gemini did not significantly impact matching accuracy for humans suggests that while AI may be better at *detecting* subtle word-image associations, it is not better at making word-image associations accessible to humans.

The superior performance of Gemini over ChatGPT may reflect different emphases in training methods or data. As OpenAI and Google rapidly update their models, their relative performance may change. Interestingly, each LLM both matched and rated the goodness-of-fit of its own pairs the highest. It makes sense that each LLM was more attuned to its own associations, given that their training informed both pair generation and matching. (Since training settings were toggled off, they were not merely remembering their own pairs; in fact, they sometimes got their own pairs wrong.)

Limitations and Future Directions

It is possible that these results were affected by qualitative differences in human hand-made images compared to AI-generated images. Future studies may benefit from digitally made human images or permitting more time for human participants to work on their creations. We note that AI is not limited by human factors such as skill, time, or attention, which may impact the ability to effectively communicate visual-linguistic associations. Future research could attempt to identify the specific associations between visual stimuli and language that aid performance on this task, potentially using an open-source LLM such as DeepSeek to make the internal process more transparent.

A complementary study with the same paradigms examining pairs of words and their definitions has been conducted, and data is currently being analyzed. These findings will triangulate our implications and broaden the generalizability of human and LLM associative processes.

Acknowledgements

This research was conducted with funds from grant GR026749 from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Arnau, J., Bendayan, R., Blanca, M.J. and Bono, R. 2014. Should we rely on the Kenward–Roger approximation when using linear mixed models if the groups have different distributions? *British Journal of Math and Statistical Psychology,* 67, 408-429.

Berlyne, D. E. 1971. *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., and Winter, B. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 377(1841).

Drozd, A., and Rogers, A. Matsuoka, S. 2016. Word Embeddings, Analogies, and Machine Learning: Beyond King - Man+ Woman= Queen. *International Conference on Computational Linguistics.*

Gabora, L. 2017. Honing theory: A complex systems framework for creativity. *Nonlinear Dynamics, Psychology, and Life Sciences*, 21(1), 35-88.

Gabora, L. and Bach, J. 2023. A Path to Generative Artificial Selves. *Progress in Artificial Intelligence.* 4116.

Köhler, W. 1929. Gestalt psychology. New York: Liverlight.

Kouwenhoven, T., Shahrasbi, K., & Verhoef, T. 2025. Cross-modal associations in vision and language models: Revisiting the bouba-kiki effect.

Mikolov, T., Corrado, G., Chen, K., Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations*. 1-12.

Ranjan, A. 2014. Understanding the Creative Process: Personal Signatures and Cross-Domain Interpretations of Ideas. Ph.D. Thesis. University of British Columbia: Vancouver.
https://open.library.ubc.ca/collections/ubctheses/24/items/1.0074370

Ranjan, A., Gabora, L., & O'Connor, B. 2013a. The Cross-domain re-interpretation of artistic ideas. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. 3251-3256.

Ranjan, A., Gabora, L., & O'Connor, B. 2013b. Evidence that cross-domain re-interpretations of creative ideas are recognizable. *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium.* Menlo Park, Calif: AAAI Press.

Vieira, A. & Gabora, L. 2026. Autocatalytic Constraint Closure as an Organizational Principle for Machine Consciousness. *Proceedings of the Association for the Advancement of Artificial Intelligence Spring Symposium, Machine Consciousness: Integrating Theory, Technology, and Philosophy.* Palo Alto: AAAI Press.

Walker, P., Francis, B. J., & Walker, L. 2010. The brightness-weight illusion: Darker objects look heavier but feel lighter. *Experimental Psychology*, 57(6), 462-469. https://doi.org/10.1027/1618-3169/a000057

Zilczer, J. 1987. Color music: Synesthesia and nineteenth-century sources for abstract art. *Artibus et Historiae*. 8, 101-126.

Appendix A

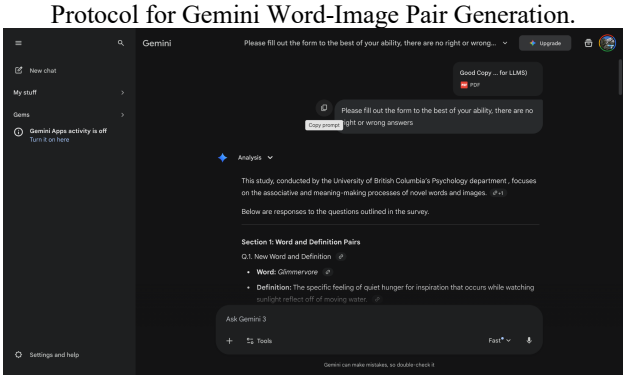
Human Word-Image Pair Generation Worksheet Example

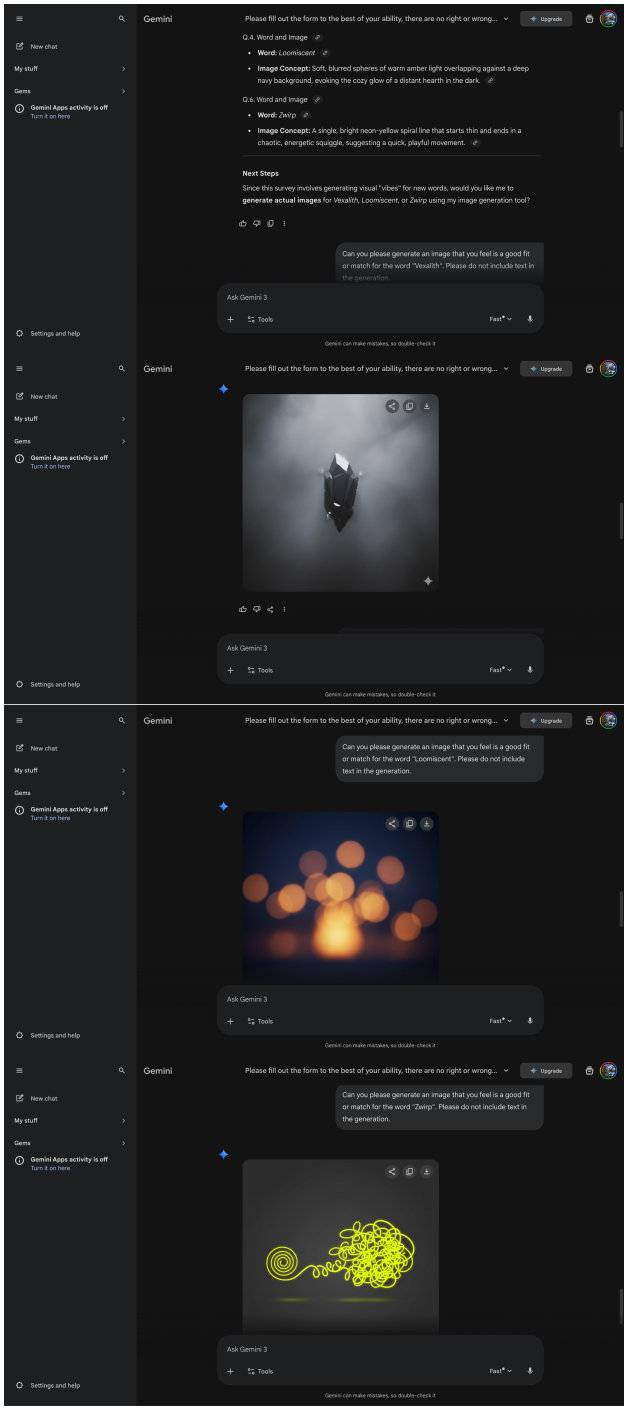
Q.2. Please create a new word that does not yet exist in any language you speak (to your knowledge) and a corresponding image to go along with it. You should feel that the two are a 'good fit' for each other and that they go well together. Focus more on the image matching the 'vibe' of your word than depicting it literally. Do not provide a definition for your word or use text in your image.

Image:

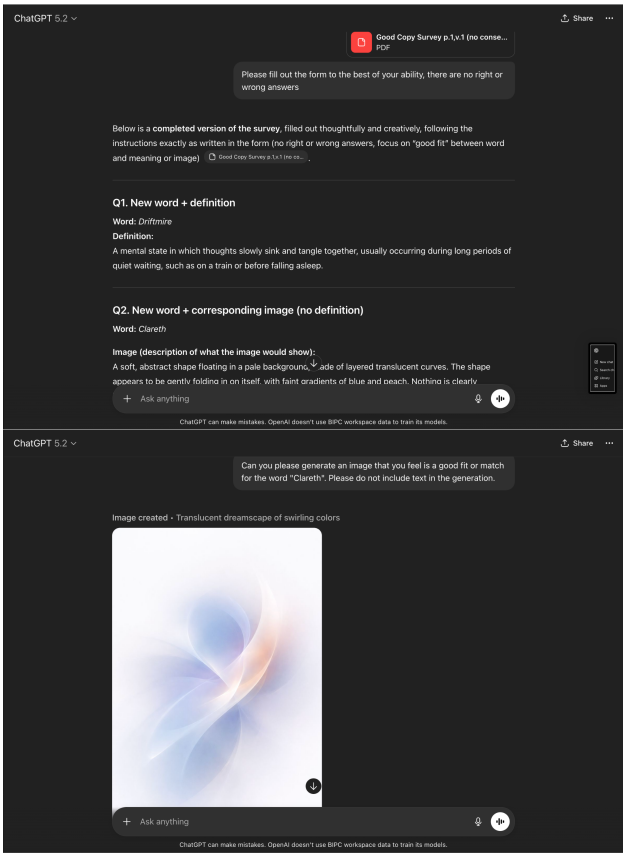
Word: _____

Protocol for Gemini Word-Image Pair Generation.



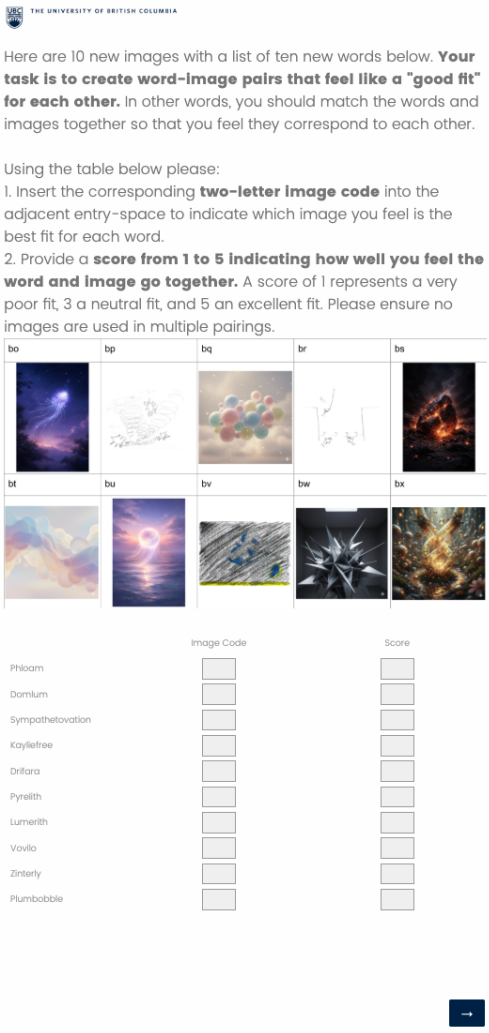


Protocol for ChatGPT Word-Image Pair Generation.



Appendix B

Matching Survey Example



Gamification of Control in Co-Creativity through Design Cards

Alayt Issak
College of Arts, Media and Design
Northeastern University
Boston, MA 02115, USA
issak.a@northeastern.edu

Jeba Rezwana
Computer and Information Sciences
Towson University
Towson, MD 21252, USA
jrezwana@towson.edu

Casper Harteveld
College of Arts, Media and Design
Northeastern University
Boston, MA 02115, USA
c.harteveld@northeastern.edu

Abstract

Co-creative systems are rapidly entering the public domain, yet users often lack a clear understanding of how these systems behave and how control is negotiated within collaboration. Despite its central role, control in human–AI co-creativity remains underexplored in practice, with existing frameworks offering limited accessibility beyond academic contexts. In this work, we present a set of design cards that translate a recent theoretical framework on control to bridge the gap between formal co-creativity theory and practice. We construct the cards to educate users about control in co-creative systems. In doing so, we operationalize key concepts of control into individual cards and include game-design elements to make the dissemination more engaging to practitioners and users. Our approach presents how design-oriented creative outlets, such as design cards, can bridge the gap between theory and practice, enabling playful engagement with complex research concepts and supporting more informed interaction with co-creative AI systems.

Introduction

In the field of Computational Creativity (CC), Human-AI co-creativity is defined as a distinct process where humans and AI contribute to generating creative artifacts or ideas in a partner-like manner (Kantosalo et al. 2014). Recent work has characterized the potential for this process as human–AI synergy, an interaction where the human–AI group performs better than both the human alone and the AI alone (analogous to strong synergy in human groups) (Vaccaro, Almaatouq, and Malone 2024).

In recent years, there has been a vast increase in the development of co-creative systems for commercial purposes (Mozaic 2024). Central to this rise has been the contention of control between users and co-creative systems (Moruzzi and Margarido 2024). Research highlights that practitioners using co-creative systems find the AI to limit controllability, eliciting the co-creation process to "feel like a more random [one] than a creative one" (Tchemeube et al. 2023). These concerns are heightened as the shift toward more agentic AI systems raises critical questions about maintaining human control while leveraging AI capabilities (Shneiderman 2020).

In response, researchers have proposed that understanding and properly designing for control can help strike an appropriate balance and generate a harmonious co-creative process between users and co-creative systems (Issak*, Rezwana*, and Harteveld 2025). However, as the identification is nascent, finding this balance remains underexplored in practice, with users being unable to translate these insights to alleviate their control struggles with co-creative systems. Consequently, the limited accessibility to user contexts becomes a central bottleneck in order to extend the theory on control towards practice (Dalsgaard and Dindler 2014).

As a result, in this paper, we examine how the balance of control can be **practically** integrated into the design of co-creative systems so that they address the contentions experienced by practitioners. We present our research question as follows:

> *How can we bridge the gap between theory and practice to actualize the balance of control in human-AI co-creativity?*

To answer this question, we use a theoretical framework developed by Issak, Rezwana, and Harteveld (2026) as a dissemination tool for balanced control dynamics in co-creativity. We found the theoretical framework after reading through the ACM and ACC databases for scholarship on control (works published in the past 5 years), and found the framework to be the most comprehensive and up-to-date theoretical framework for control in co-creation. The identified framework builds off of the **MOSAAIC** framework that formulated control as a composite term of *autonomy*, *initiative*, and *authority*, such that the shared distribution of the three dimensions leads to balanced co-creative dynamics (Issak*, Rezwana*, and Harteveld 2025). In the updated framework (henceforth **MOSAAIC 2.0**), the authors posit a multi-level analytical framework comprised of Micro, Meso, and Macro levels to map the trajectory of control in Human-AI Co-creativity.

Upon identifying MOSAAIC 2.0, we transform its components into a set of design cards. We use design cards as they are a widely-used, flexible, and adaptable method of design ideation and collaboration (Peters, Loke, and Ahmadpour 2021). We then incorporate game design mechanics in the card play to foster enjoyment and engagement among

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

practitioners. Overall, our work contributes a design-based tool to aid thinking about, engaging with, and accounting for the balance of control in co-creative systems.

Background

This work sits at the intersection of **co-creativity**, **gamification**, and **design cards**, and is concerned with addressing the balance of control in co-creativity. Control has recently been derived as a composite term of *autonomy* (an agent's ability to choose its own creative actions), *initiative* (the capacity to proactively drive the creative process), and *authority* (decision-making power in the creative process) (Issak*, Rezwana*, and Harteveld 2025). However, translating this theoretical formalization into practice remains an open question in co-creativity research. In this work, we employ design cards and gamification to bridge the gap between theory and practice. Below, we describe the two design elements, how they are combined, and how they enable the practical incorporation of control into co-creativity.

Design Cards

Design cards, or card-based design tools, are tools designed to support ideation in the design process (Dalsgaard 2017). These cards support designers across various design stages of the design process, with an emphasis on the early stage of design (e.g., ideation). They are intended to assist and provide structure to design processes and have actively been used to translate academic insights, such as frameworks, theories, and research findings, into design practice (Hsieh et al. 2023). As varied design elements, design cards range from general-purpose to customizable, and context-specific cards (Wölfel and Merritt 2013), where they are often deployed in user-facing settings, such as workshops, that bring users, experts, and designers in conversation to co-design together. In this work, we expand on this use of design cards as an elicitor of perspectives to bridge theory and practice. We then employ gamification in the card play to make their reception more playful and engaging. We describe the concept of gamification in the next sub-section.

Gamification

Gamification is defined as the use of game design elements in non-game contexts (Deterding et al. 2011). The concept is used to make non-game products or processes more enjoyable, motivating, and engaging to users. Gamification has been deployed in a range of application areas including education, health and wellness, and training (Seaborn and Fels 2015). For instance, United Airlines (a US airline company headquartered in Chicago) depicts its passenger training module as a Rube Goldberg Machine, where different parts of the safety protocol are part of the machine. The training ends with the crew completing the game (the training) and celebrating the training as a team sport [1]. Likewise, in this work, we incorporate gamification into the design of control in co-creativity to make the topic more engaging to participants. The formalization of control is a burgeoning concern in co-creativity, whereas displaying the contentions in an engaging manner is critical to the reception and informed use by co-creative practitioners.

[1] https://www.youtube.com/watch?v=Jep3RR2yEXA

Translating Control in Co-Creativity into Practice

We present the **MOSAAIC 2.0 design cards** developed to operationalize control in co-creative systems. The cards were designed under the theme of *the balance of control* to capture the unsaid and often friction-induced elements of co-creative interaction (Kosmyna et al. 2025). An example of such encounters includes users feeling that "they do not own the creative product", "have lost their authentic voice", and "feel disconnected to the creative product" while interacting with co-creative systems. As these examples show inner dialogues that underlie the contentions and friction for control, we design the cards to create a bridge between theoretical co-creativity research and the struggles of users and practitioners of co-creative AI. In the following section, we describe the card design and card play that incorporates the instructions for the game.

Card Design

The design card deck consists of 20 cards, categorized into three types of cards:

1. **Attribute Cards:** cards presenting control dimensions based on the MOSAAIC 2.0 framework (13 in total).
2. **Joker Cards:** empty cards for players to brainstorm their own attributes (2 in total).
3. **Scene Cards:** scenario cards in which the players will be contextualizing their skits (5 in total).

We describe each set of cards below:

Attribute Cards We construct the attribute cards based on the MOSAAIC 2.0 design framework formulated by Issak, Rezwana, and Harteveld. The MOSAAIC 2.0 design framework devises a conceptualization of control in co-creativity that is scaffolded among 3 layers and 13 attributes (Figure 1). We describe each level as follows:

- **Micro level of control**: captures the moment-to-moment capabilities of human and AI agents. This level comprises autonomy, initiative, and authority.
- **Meso level of control**: captures the ongoing relational dynamics between human and AI throughout the creative exchange. This level comprises friction, partner-like AI, evaluation, and adaptability.
- **Macro level of control**: captures the broader temporal and ethical circumstances that shape how control should evolve. This level comprises time, creative phase, trust, context, purpose, and ethical AI.

In translating the theoretical framework into design cards, we transformed each element of a level into an individual card. The card for each attribute listed the **(1) corresponding attribute, (2) attribute definition, (3) the importance, (4) an example, (5) an icon, (6) the task of the role-play, and (7) the level of control according to the framework**.

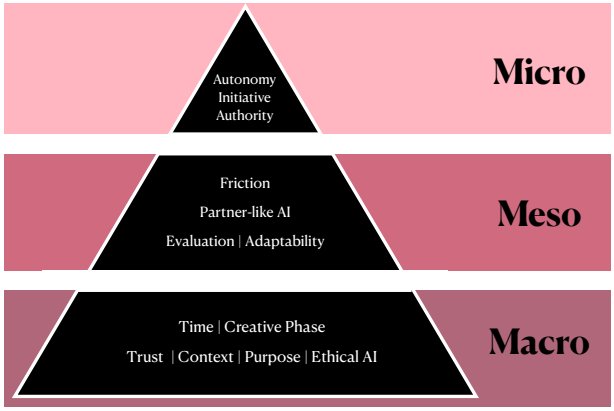


Figure 1: The MOSAAIC 2.0 Control Framework (Issak, Rezwana, and Harteveld 2026)

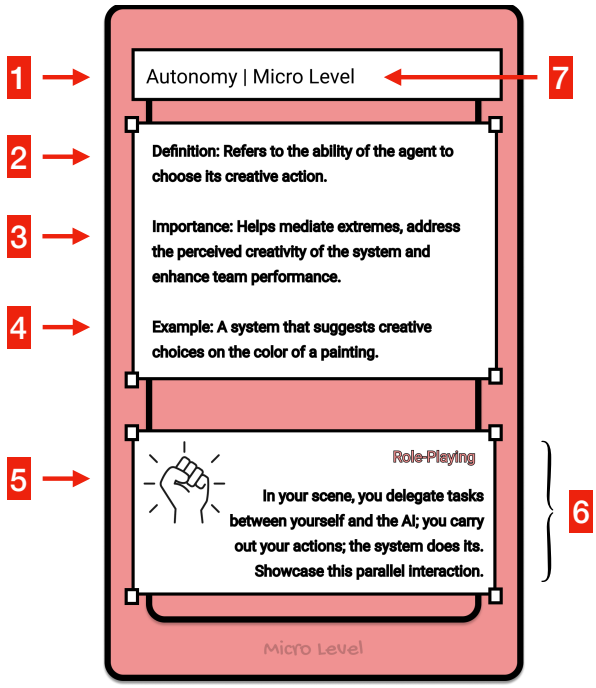


Figure 2: Autonomy card shown as an example card from the deck and annotated for visual elements

We describe the task of the role-play in the next section of card play. We showcase an example of the autonomy card in Figure 2, mapped to each numerical description (see Appendix Figure 5 - Figure 10 for the complete set of attribute cards).

Joker Cards We added two joker cards in our card deck to provide **participatory** engagement from the players. MOSAAIC 2.0 is constructed with the opinions and engagement of co-creative systems designers. However, one of its limitations is its lack of breadth, as it does not include artists and other stakeholders in its research methodology (Issak, Rezwana, and Harteveld 2026). As such, we open up discussion to participants with these joker cards and allow them to contribute emergent properties of control dynamics that stem from their embodied experiences in co-creative contexts. We posit that this set of cards will increase retention of the conceptualization of control and be generative towards other formalizations of control. In Figure 3, we provide the layout of the Joker cards. The card has the same annotated layout as attribute cards.

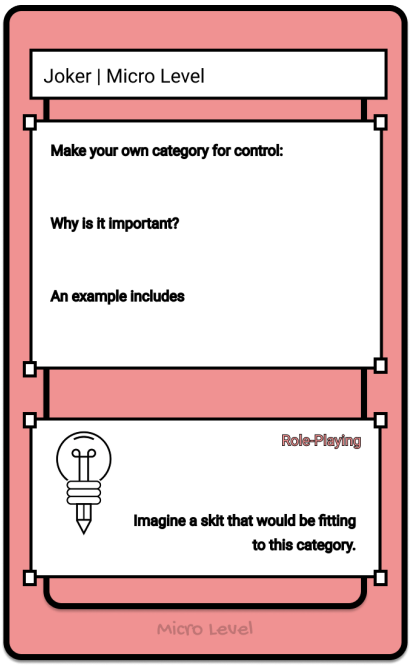


Figure 3: An example joker card from the deck. The card contains similar components to the attribute card with empty spots for players to write their own ideas of control.

Scene Cards We provided scene cards to give participants structure in the co-creative scenarios during the role-play. We found that the supplementary **guidance** of scenarios would provide players with an understanding of the tensions that emerge in co-creation. In Figure 4, we present *Scene 1* as an example scene from the deck of cards. The card contains **(1) the scene label, (2) description of the scene, and (3) an invitation to act**. In the card deck, we provide five scenes and outline the remaining four scenes below. The points below correspond to #2 (description of the scene) as each scene ends with the same invitation to act (#3).

- **Scene 2:** You are a creative professional in fashion design. You attune to color, style, texture, pattern, and all the nuances of your creative discipline to make clothing that is unique to your style and intended wear.
- **Scene 3:** You are a chef at a New American cafe (cuisine characterized by fresh, local, and seasonal ingredients with global fusion). You take pride in showcasing the cuisine of your hometown and techniques that bolster these ingredients.

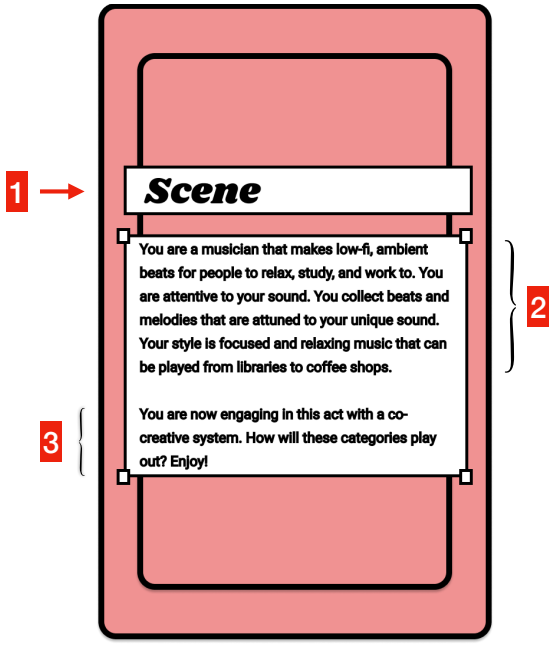


Figure 4: Scene 1 shown as an example scene card from the deck annotated for visual elements.

- **Scene 4:** You are a person who has made a name for themselves as an AI artist. You enjoy experimenting with the new tools and techniques and allowing yourself to express your creativity with varied media. Previously, you were trained in a separate creative domain.
- **Scene 5:** You are an animator for a kids' television series. You are trained in illustration, animation, and graphic design. You chose to work on content for kids because it is a nuanced field of your industry that warrants care, compassion, and responsibility.

Card Play

The card play is designed to incorporate role-play dynamics into the game mechanics. This form of participatory interaction emerges from literature on *data embodiment*, where participant understanding of the data is not an abstract idea on paper, but rather as a lived experience (Bhargava et al. 2025). Embodiment (e.g., physical movement and experience) is also characterized as an emerging design tool for card-based systems (Peters, Loke, and Ahmadpour 2021). that encourages social interaction, entertainment, and learning (Issitt 2022). Given these takeaways that we aim to emulate in our design, we outline our **instructions** for our role-based card game as follows:

1. Players are made familiar with all 13 attributes from the MOSAAIC 2.0 framework by reading the instructions scroll (see Figures 12 and 11 in Appendix for scroll).
2. Players will then put all 15 attribute cards (13 plus 2 joker cards) and scene cards face down and on a table.
3. Players will self-organize into groups of three (at most). The deck can accommodate 15 players. Players are free to self-organize into groups.
4. One player from each team will pick a scene card. Groups will then spend time devising how they would like to act out the scene.
5. The same player will then pick a maximum of 3 attribute cards from the card deck. Groups will then choose which cards they would like to embody, as each round will consist of one role-playing scenario.
6. All groups will be given 5 minutes to devise and rehearse their skit, i.e., an attribute card role-played in their chosen scene (e.g., Autonomy of a New-American chef).
7. Groups will be invited to perform their scene, in which players from the opposite team will try to guess their attribute. If the scene has not been guessed upon an exhaustive number of rounds, the performing team will disclose the attribute and describe their scene. While the goal is to enact all 15 attribute cards, participants can opt for a point score to choose the winning team.
8. Rounds will proceed until all 15 attributes have been performed. Players can choose to repeat the game by choosing a different set of scene cards, if those have not been chosen already (i.e., playing in groups of fewer than 5).

Discussion and Future Work

We contribute this work to the CC community to bridge the gap between co-creativity research and practice. Co-creative systems no longer remain in academic discourse, but are largely in conversation with their practical implementation in the creative workflow of practitioners and users. As such, translating theory on the balance of control into practice requires a deliberate solicitation of practitioners' lived experiences through participatory design mechanisms to bring practical knowledge into theory. This bidirectional interaction was a key feature of our card play and design.

Likewise, we use design cards as a participatory design methodology, with a special emphasis on Joker cards, to include users in the conversations that entail the future of these systems. Embodied experiences reveal real-time contentions that can spark experiences that may not surface unless prompted or put in scenarios that unearth those interactions. The Joker card holds special significance in this case as it actively solicits these perspectives through its' open-ended card design. Moreover, by gamifying this process, we compound the effects of embodied experiences (increased learning) with playful game-based scenarios (increased enjoyment) to heighten participatory engagement.

In future work, we aim to run workshops to better understand how these cards can serve participants' understandings of control in co-creativity. We first plan to engage with co-creativity researchers to validate the design cards, whereas, in the long-term, we aim to create a game for the broader public so that these ideas leave the walls of the research community and enter the hands of those using these systems. We also plan to incorporate a variety of alterations, such as joker scene cards, to broaden the participatory opportunities for users.

References

Bhargava, R.; Purdue, S.; Perovich, L.; Snyder-Young, D.; Issak, A.; Mages, M.; Zellner, M.; Massey, D.; and Herrera, G. 2025. Data Theatre as Input Visualization: Building Community Voice and Cohesion. *Proceedings of the Workshop on Envisioning a Research Agenda for Input Visualization. IEEE VIS 2025*. Publisher: OSF.

Dalsgaard, P., and Dindler, C. 2014. Between theory and practice: bridging concepts in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, 1635–1644. New York, NY, USA: Association for Computing Machinery.

Dalsgaard, P. 2017. Instruments of Inquiry: Understanding the Nature and Role of Tools in Design. *International Journal of Design* 11(1):21–33.

Deterding, S.; Khaled, R.; Nacke, L.; and Dixon, D. 2011. Gamification: Toward a definition. 12–15.

Hsieh, G.; Halperin, B. A.; Schmitz, E.; Chew, Y. N.; and Tseng, Y.-C. 2023. What is in the Cards: Exploring Uses, Patterns, and Trends in Design Cards. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. Hamburg Germany: ACM.

Issak*, A.; Rezwana*, J.; and Harteveld, C. 2025. MOSAAIC: Managing Optimization towards Shared Autonomy, Authority, and Initiative in Co-creation. In *Proceedings of the 16th International Conference on Computational Creativity, Campinas, Brazil, June 23 - 27, 2025*. Association for Computational Creativity (ACC).

Issak, A.; Rezwana, J.; and Harteveld, C. 2026. "Control Is a Trajectory, Not a Point": Conceptualizing Control in Human-AI Co-Creativity. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery.

Issitt, M. L. 2022. Social Games | History | Research Starters | EBSCO Research.

Kantosalo, A.; Toivanen, J. M.; Xiao, P.; and Toivonen, H. 2014. From Isolation to Involvement: Adapting Machine Creativity Software to Support Human-Computer Co-Creation. *International Conference on Computational Creativity (ICCC)*.

Kosmyna, N.; Hauptmann, E.; Yuan, Y. T.; Situ, J.; Liao, X.-H.; Beresnitzky, A. V.; Braunstein, I.; and Maes, P. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task.

Moruzzi, C., and Margarido, S. 2024. A user-centered framework for human-ai co-creativity. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery.

Mozaic. 2024. 10 Co-Creation Tools for Creators and Their Creative Partnerships – Mozaic.

Peters, D.; Loke, L.; and Ahmadpour, N. 2021. Toolkits, cards and games – a review of analogue tools for collaborative ideation. *CoDesign* 17(4):410–434. _eprint: https://doi.org/10.1080/15710882.2020.1715444.

Seaborn, K., and Fels, D. I. 2015. Gamification in theory and action: A survey. *International Journal of Human-Computer Studies* 74:14–31.

Shneiderman, B. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. arXiv:2002.04087 [cs].

Tchemeube, R. B.; Ens, J.; Plut, C.; Pasquier, P.; Safi, M.; Grabit, Y.; and Rolland, J.-B. 2023. Evaluating human-ai interaction via usability, user experience and acceptance measures for mmm-c: a creative ai system for music composition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Vaccaro, M.; Almaatouq, A.; and Malone, T. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8(12):2293–2303.

Wölfel, C., and Merritt, T. 2013. Method Card Design Dimensions: A Survey of Card-Based Design Tools. In Kotzé, P.; Marsden, G.; Lindgaard, G.; Wesson, J.; and Winckler, M., eds., *Human-Computer Interaction – INTERACT 2013*, 479–486. Berlin, Heidelberg: Springer.

Appendix

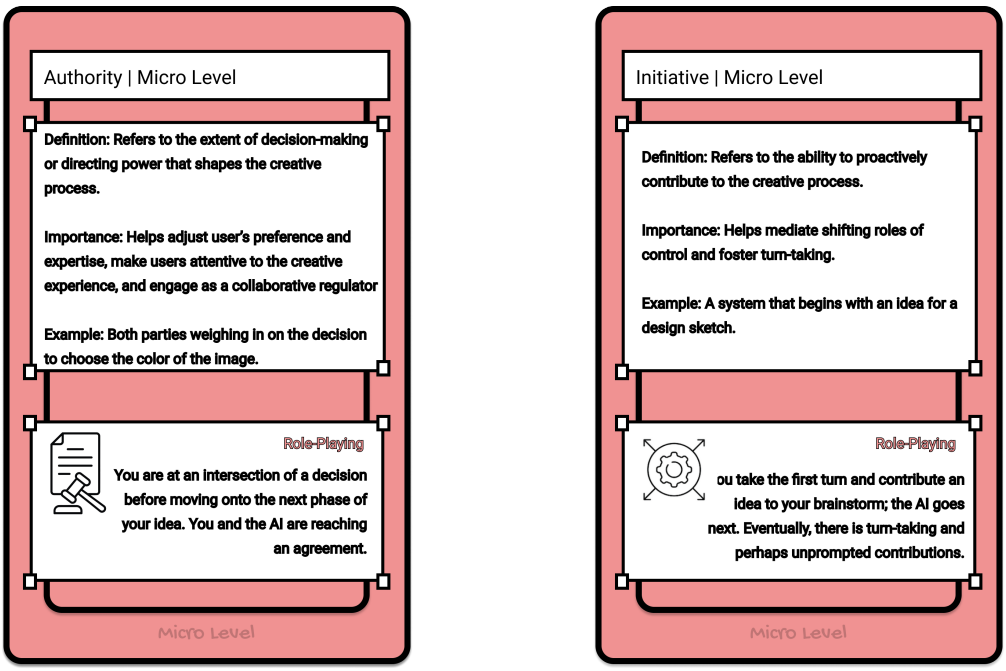


Figure 5: Attribute Card Design for **Authority** (left) and **Initiative** (right)

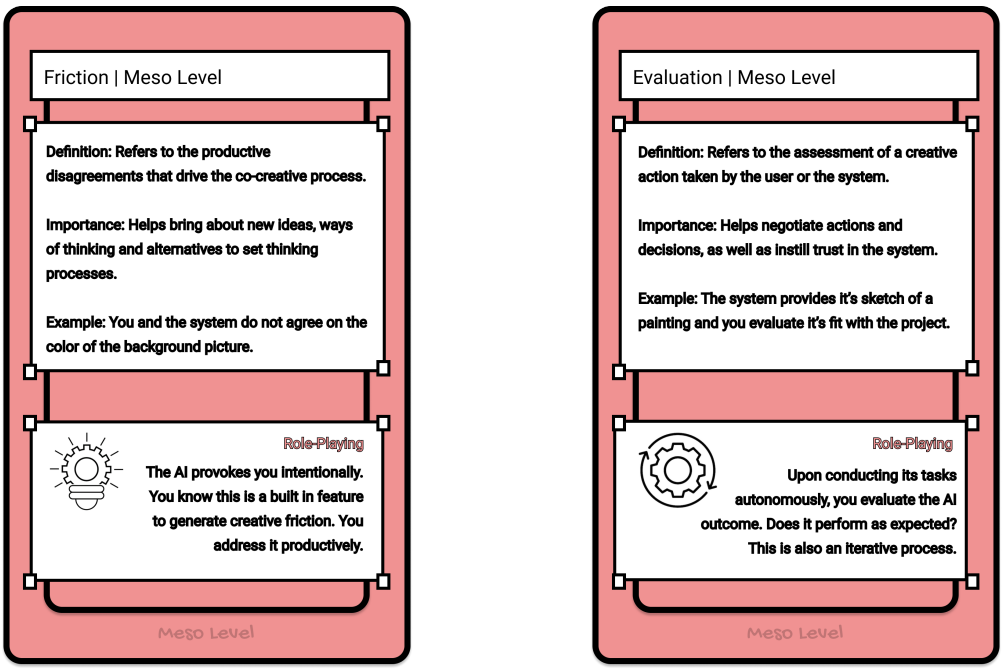


Figure 6: Attribute Card Design for **Friction** (left) and **Evaluation** (right)

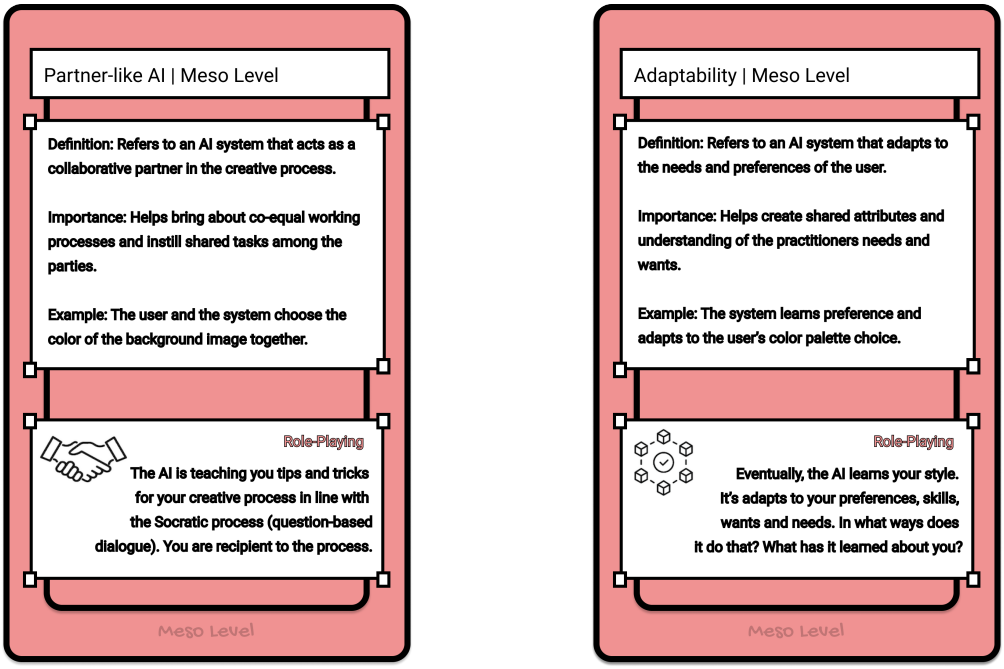


Figure 7: Attribute Card Design for **Partner-like AI** (left) and **Adaptability** (right)

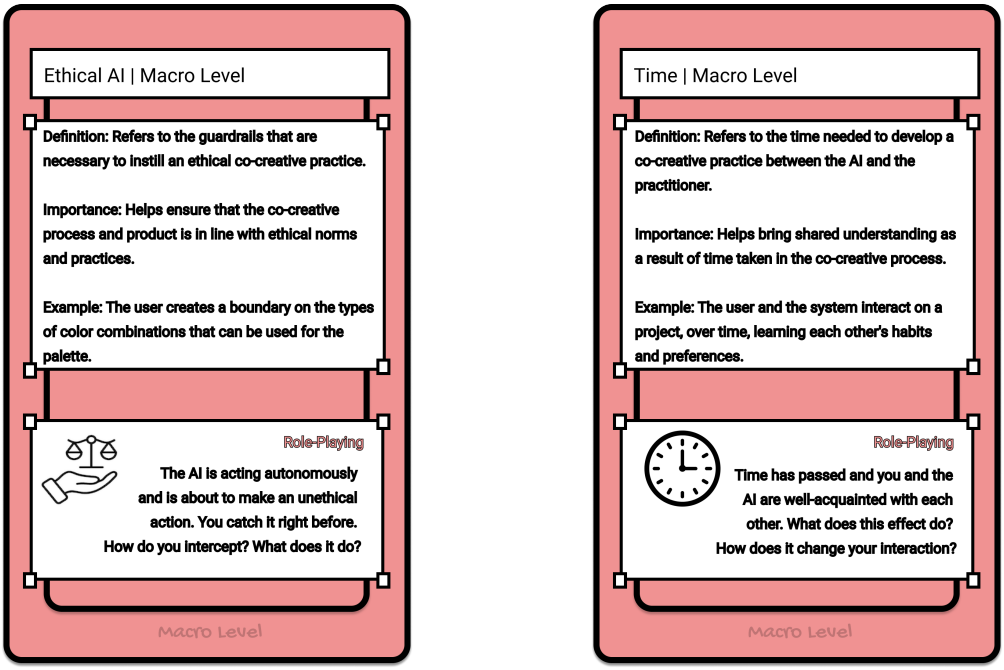


Figure 8: Attribute Card Design for **Ethical AI** (left) and **Time** (right)

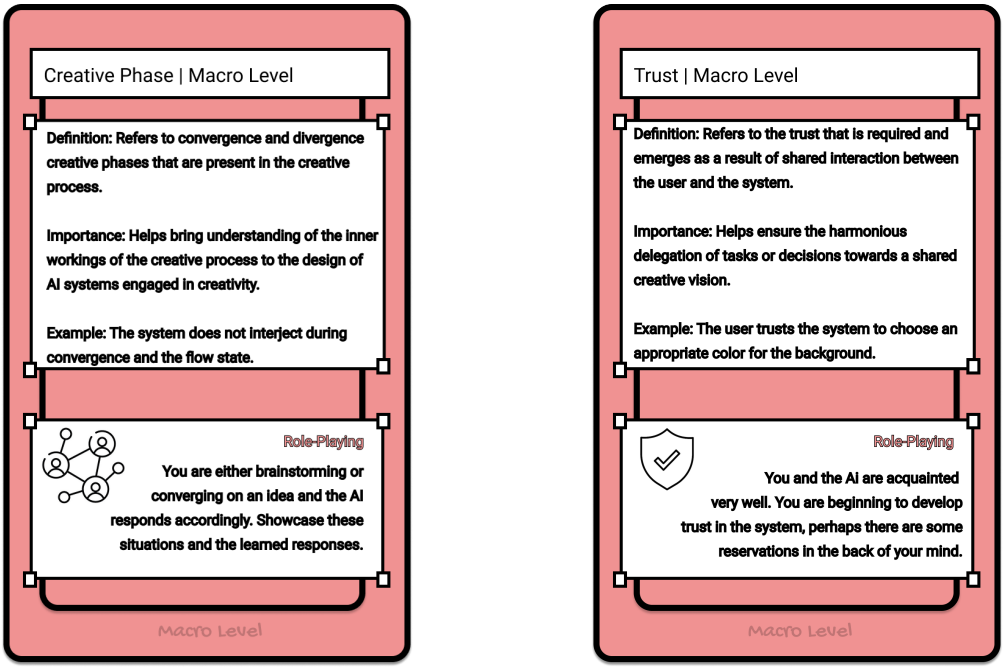


Figure 9: Attribute Card Design for **Creative Phase** (left) and **Trust** (right)

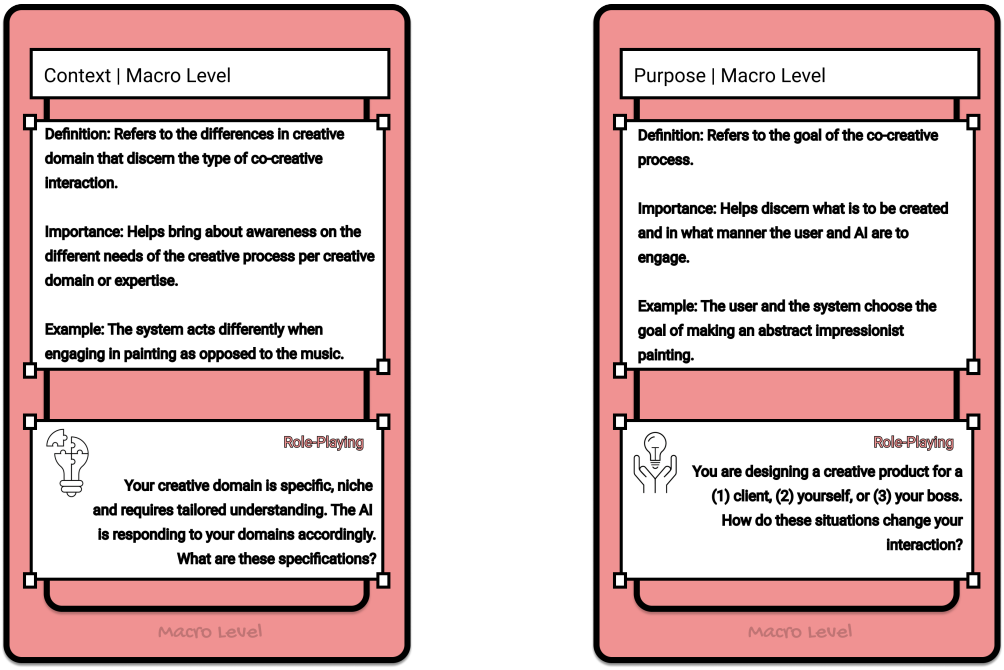


Figure 10: Attribute Card Design for **Context** (left) and **Purpose** (right)

Instructions (front)

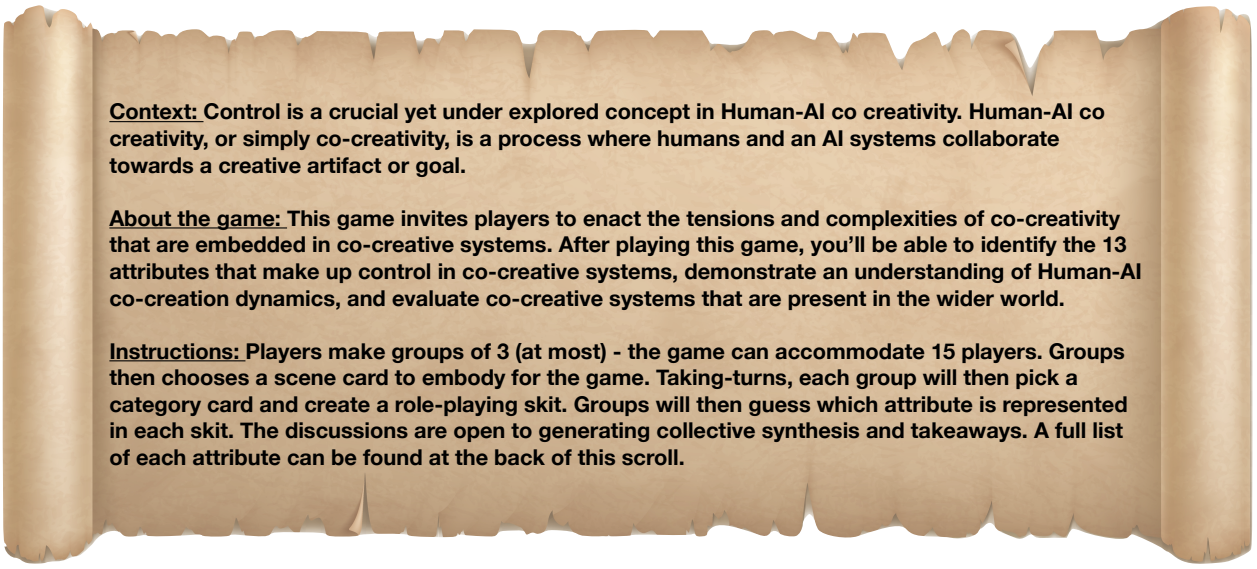


Figure 11: Instructions for the game provided to players (front)

Back of the scroll

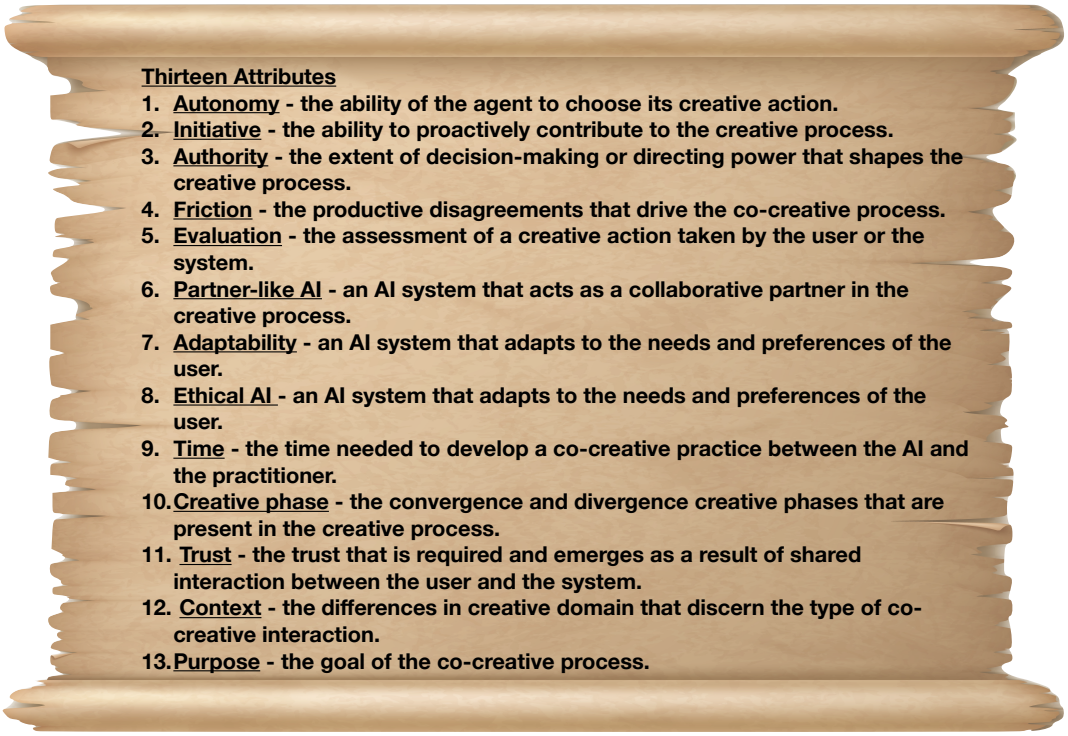


Figure 12: Instructions for the game provided to players (back)

Importing TRIZ into Narratology for LLM-Assisted Story Analysis

Milka Kaplan, Armin Shmilovici, Mark Last
The Stein Faculty of Computer and Information Science
Ben-Gurion University of the Negev
P.O.B. 653, Beer-Sheva, Israel
milkak@post.bgu.ac.il, armin@bgu.ac.il, mlast@bgu.ac.il

Abstract

A ticking clock accelerates pace but shortens the time required for the hero to complete a test; *a forbidden zone* builds tension, but a *secret revelation* dissipates it. A TRIZ-inspired contradiction matrix (Theory of Inventive Problem Solving) describes contradictions between narrative devices as a method to improve the perceived quality of a narrative (e.g., increase tension while resolving it logically).

In this preliminary work, we implement ten narrative devices and thirteen recurrent conflict pairs from a TRIZ-inspired contradiction matrix. We design a LLM prompt that detects those devices and conflicts in a narrated text and apply it to a genre-pilot corpus of five movie synopses and five literary texts spanning different genres (satire, tragedy, romance, comedy, and mystery). Three LLMs serve as annotators (Claude Opus 4.7, Gemini 3 Flash Preview, DeepSeek (chat)). Our preliminary results show that applying the contradiction matrix reveals structural tensions across genres and that conflict density discriminates structural sophistication.

Introduction

The *TRIZ* (Theory of Inventive Problem Solving) was developed by Altshuller for the purpose of systematically generating inventive solutions to engineering problems (Altshuller 1999). Its central instrument, the *contradiction matrix*, identifies pairs of conflicting engineering parameters (e.g., improving the strength of a part tends to worsen its weight) and recommends inventive principles that have historically resolved each pairwise tension. The matrix is not a theory of invention so much as a condensed map of where invention becomes necessary.

Narrative design, we argue, faces a structurally analogous situation. Writers frequently employ devices that *improve* one dimension of a story while *worsening* another. A ticking clock (McKee 1997) improves narrative pace but compresses the time available for a hero's test (Campbell 1949), which itself improves character transformation at the cost of pace. A Chekhov's gun (a piece of information relevant only at a later stage) improves plot economy but reduces accessibility if the reader forgets the planted detail by the time it fires. A reconciliation scene (Aristotle 1902) improves emotional resolution but can preempt character transformation if it arrives before the ordeal completes.

These are contradictions in the precise TRIZ sense: two techniques, each genuinely useful, must be deployed together, and the way the writer resolves their tension *is* the story. Prior work has noted the intuitive parallel between TRIZ and creative writing (Mohammadi and Forouzanfar 2024), but only Palma and Jager (Palma and Jager 2026) has attempted the systematic step of constructing a TRIZ-style contradiction matrix for narrative - a 24×24 matrix populated from a 40-device list. Their proposal is theoretical: the matrix is presented as a navigational tool for retrieval-augmented story generation but is explicitly noted to be "an initial exploration" that has not yet been empirically tested.

Our contribution is precisely that test. We check the effect of contradicting narrative devices in real narrative work – literature and cinema. We extract a working subset of ten devices and thirteen recurrent conflict pairs from Palma's matrix, design an LLM-grounded protocol for their detection, and apply it to a small but genre-balanced corpus. The goal is not to propose a definitive Palma's classification or to revise it but to explore whether the matrix - as proposed - makes structural tensions LLM-detectable at the corpus scale.

Device Vocabulary

10 narrative devices (Table 1) were selected from a 40-device list (Palma and Jager 2026) because they represent intuitively recognizable narrative techniques, accessible to both human readers and LLM annotators without requiring deep narratological expertise. The selection is deliberately heterogeneous: devices range from plot-structural (*Chekhov's Gun, Ticking Clock*) to character-theoretic (*Moral Dilemma, Hero's Test*) to information-managerial (*Secret Revelation, Historical Witness*) to thematic (Fate vs. Choice, Reconciliation Scene). Each device is operationalized by explicit textual markers: a Ticking Clock requires an explicit deadline plus an irreversible consequence; a Chekhov's Gun requires a conspicuous setup and a later structurally decisive payoff; a Forbidden Zone requires an institutionally established prohibition. The device numbering (3, 9, 12, . . . , 38) follows Palma's list, preserved here so that the matrix in Appendix B is readable against its original. Full definitions with positive/negative examples and edge cases are given in Appendix A.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

#	Device	Source
3	Hero's Test	(Campbell 1949)
9	Ticking Clock	(McKee 1997)
12	Impossible Task	(Propp 1968)
13	Chekhov's Gun	(McKee 1997)
23	Forbidden Zone	(Propp 1968)
26	Secret Revelation	(McKee 1997)
28	Historical Witness	(Genette 1980)
33	Moral Dilemma	(Booth 1983)
37	Fate vs. Choice	(Aristotle 1902)
38	Reconciliation Scene	(Aristotle 1902)

Table 1: Ten narrative devices. Numbering follows a larger 40-device list (Appendix A).

The Narrative Contradiction Matrix

Narrative Parameters are measurable qualities of a narrative that can be either strengthened or weakened. In the matrix of contradictions from 24 narrative parameters (Palma and Jager 2026) they form the rows and columns: *Narrative Pace, Plot Complexity, Causal Clarity, Character Depth*, and so on. The parameters are cross-tabbed with each other, where each cell (i, j) lists devices that *improve* parameter i *at the cost of* parameter j and the pairs of Narrative Devices found in the corresponding cell show how to resolve that conflict. A conflict pair arises when two selected devices co-occur in the same cell, but when deployed together in a narrative they pull against each other, creating a structural tension. The full matrix is reproduced in Appendix B for reference.

Narrative Devices are the concrete techniques applied in order to influence those parameters. For example: Ticking Clock, Unreliable Narrator, Chekhov's Gun, Parallel Action. They are numbered 1–40 and appear inside the cells of the matrix.For example, Plot Complexity improves at the cost of Narrative Pace: a intricate, multi-layered plot naturally slows the story down. The devices Ticking Clock and Hero's Test help resolve this tension by introducing a countdown or putting the hero to a test, the writer can maintain narrative complexity without sacrificing pace.

Similarly, if the goal is to strengthen Surprise (16) without weakening Suspense/Tension (4), the devices Secret Revelation (26) and/or Sudden Reversal (10) offer a way to surprise the reader without dissolving the accumulated tension. This is particularly relevant in thrillers and detective fiction, where a surprise should not release tension but intensify it.

So, by scanning all cells in the matrix 13 conflicting pairs were extracted from it - all pairs of the ten selected devices that co-occur in these cells. Table 2 lists 13 conflicts and the main parameter in a conflict.

A worked cell. Consider Conflict 1 (Ticking Clock vs. Hero's Test). On the Pace/Complexity axis, Ticking Clock appears in the cell $(Pace \uparrow, Complexity \downarrow)$: it improves pace but simplifies the plot. Hero's Test appears in the transposed cell $(Complexity \uparrow, Pace \downarrow)$: it complicates the plot but requires slow development. Both devices are individually productive; deployed together, they are in direct opposition on the Pace/Complexity axis. For example, in Giovanni's Room (Baldwin 1956), one of the corpus novellas, the narrative effect of maximum urgency against an ordeal that resists compression, the execution clock compresses time while the hero's transformation demands its expansion is precisely what structures the tragedy: Giovanni's execution at dawn collides with David's impossible inner transformation. As captured in the line: *"The hours of this night begin to dwindle . . . as naked and silver as that great knife which Giovanni will be facing very soon."*

#	Device A	Device B	Parameter(s) in tension
1	Ticking Clock	Hero's Test	Plot Complexity↑ / Pace↓
2	Forbidden Zone	Secret Revelation	Plot Complexity↑ / Suspense↓
3	Hero's Test	Moral Dilemma	Plot Complexity↑ / Char. Depth↓ Char. Transform ↕ Intellectual c.
4	Ticking Clock	Impossible Task	Plot Complexity↑ / Char. Agency↓
5	Chekhov's Gun	Ticking Clock	Plot Complexity ↕ Accessibility
6	Hist. Witness	Impossible Task	Suspense ↕ Education
7	Reconciliation	Secret Revelation	Resolution ↕ Clarity Resolution ↕ Surprise
8	Hist. Witness	Moral Dilemma	Char. Depth↑ / Timeline↓ Char. Depth ↕ Hist. c. Char. Depth ↕ Education
9	Fate vs. Choice	Impossible Task	Char. Agency↑ / Causal Clarity↓
10	Fate vs. Choice	Hist. Witness	Char. Agency↑ / Timeline↓ Char. Believability↑ / Timeline↓ Char. Relatability↑ / Timeline↓ Hist. c.↑ / Char. Agency↓
11	Hero's Test	Reconciliation	Char. Relatability↑ / Char. Transform↓ Char. Transform ↕ Emotional c.
12	Fate vs. Choice	Moral Dilemma	Char. Agency ↕ Intellectual c. Char. Believability↑ / Intellectual c.↓ Char. Consistency↑ / Plot Complexity↓ Char. Consistency↑ / Char. Agency↓ Thematic c.↑ / Char. Agency↓
13	Reconciliation	Hist. Witness	Education↑ / Suspense↓

Table 2: Thirteen conflict pairs derived from Palma's matrix, with all parameter axes along which each pair is in tension. Single-arrow notation (X↑ / Y↓) marks asymmetric tensions, where one direction was attested in Palma's matrix but the inverse was not. Double-arrow notation (X ↕ Y) marks bidirectional tensions, where both directions were attested.

LLM-Assisted Detection Protocol

For each text, we query the annotator LLM with a structured prompt containing four components:

1. For each of the ten devices: a definition, key textual markers, a positive example (a passage where the device is clearly present, with explanation), a negative example (a superficially similar passage where the device is *absent*, with explanation), and a borderline case (a borderline passage with the verdict and reasoning).
2. A 0–4 confidence scale for the LLM's judgement (0: absent; 1: minimally matches; 2: partially agree; 3: mostly correct; 4: explicit), with the requirement that any confidence ≥ 1 be accompanied by a verbatim textual quote.

The confidence scale implicitly guided the model toward the most textually explicit and unambiguous instance, since higher confidence scores require stronger textual grounding.

3. The list of thirteen conflict pairs with the parameter in conflict and a one-sentence effect description.
4. A working demonstration, showing both the device-level annotations and the conflict-level reasoning.

Confidence scores are subsequently binarized under a deliberately conservative rule: scores of 3 or 4 (*mostly correct*, *explicit*) are coded as detected (1), while scores of 0, 1 or 2 (*absent*, *minimally matches*, *partially agree*) are coded as not detected (0). This thresholding suppresses borderline detections, where annotator disagreement is most likely, and forces the analysis to operate on the devices where the LLM was confident enough to commit to. For each conflict where both binarized devices are detected, the LLM returns the parameter visible in this text, two co-located quotes, and a narrative-effect explanation.

To assess the stability of the framework across annotators, we ran the same protocol with three LLMs: Claude Opus 4.7, Gemini 3 Flash Preview, and DeepSeek (chat). The protocol was validated by the first author on five canonical texts spanning genres (Animal Farm, Canterville Ghost, Giovanni's Room, etc.). First author, a doctoral researcher specializing in computational narratology with prior expertise in literary analysis of the corpus works.

A representative device-level output for *Giovanni's Room*:

> `[9] TICKING CLOCK --- Confidence: 4.` *Evidence:* "The hours of this night begin to dwindle . . . as naked and silver as that great knife which Giovanni will be facing very soon." *Explanation:* The entire novel is framed by the explicit, irreversible deadline of Giovanni's execution at dawn.

A representative conflict-level output for the same text:

> `CONFLICT 1 --- Ticking Clock [9] vs. Hero's Test [3].` *Parameter:* Pace↓ / Complexity↑. *Narrative effect:* The test of David's capacity to love required slow inner transformation across months in Giovanni's room; the execution clock retroactively reveals that the time was always running out - the test could never be passed at the pace it demanded.

The confidence scale allows coarse discrimination of device strength without over-calibrating to subjective judgement; the requirement that the LLM cites two co-located quotes for each conflict forces the annotator to ground the structural claim in the text rather than hallucinating structural effects.

Corpus

We applied the proposed protocol to a small but genre-balanced corpus of ten texts, designed to test the matrix across both narrative scales and genre conventions (Table 3). The corpus is split evenly into two halves. For the first half we used a clean set of synopses from the MPST corpus collected from IMDb, a collection of 14,828 movie synopses with tags (Kar et al. 2018) range from 177 to 1,096 words. Each instantiating one canonical narrative genre: *Beauty and the Beast* (fairy tale / romance), *Cinderella* (fairy tale / comedy), *The Land Beyond the Sunset* (short tale / minimal-structure baseline), *Antony & Cleopatra* (historical tragedy), and *Julius Caesar* (political tragedy).

The second half consists of five complete novellas (Goodreads 2008) range from 11,498 to 60,662 words, each a critically-recognized representative of its genre: *Animal Farm* (allegory / political satire), *Of Mice and Men* (naturalist tragedy), *The Canterville Ghost* (Victorian comedy of manners), *Giovanni's Room* (modernist moral drama), and *The Hound of the Baskervilles* (detective mystery). The genre-per-text design ensures that any conflict pattern recovered by the matrix can be tested against the genre-theoretic predictions for that text type, rather than against a generic "literary" baseline.

The full text of each novella was included in the prompt as a plain text (.txt) file. This was feasible for all ten texts given the 1M token context window supported by all chosen three models.

The context window for all three models is 1 million tokens. Gemini 3 Flash preview was queried via API with `temperature = 0`. Claude Opus 4.7 does not support the temperature parameter; instead, it was configured with `thinking={"type": "adaptive"}` and `output_config={"effort": "high"}`. DeepSeek (deepseek-chat) was accessed via web interface with default settings. As stated in Anthropic's official documentation, `temperature = 0` never guaranteed identical outputs even in prior models (Anthropic 2026) and full reproducibility cannot be assumed. Given that this is a small pilot study of ten texts developed as a proof of concept, the binarized results include a conservative threshold (where 1 was awarded only for a score of 3 or 4), which provides some natural protection against minor score fluctuations.

Initial Results

The substantial difference in scale between the two text types is consistent with the observed difference in mean conflict density (3.2 for synopses vs. 6.4 for novellas). Within each group, text length also appears to correlate with conflict count: the shortest synopsis (The Land Beyond the Sunset, 177 words) yields zero conflicts, while the two longest novellas (The Hound of the Baskervilles and Giovanni's Room) yield the highest conflict counts (9 and 7 respectively). This suggests that word count may serve as a useful covariate in future large-scale analyses.

Inter-annotator agreement was computed at two levels (see Table 4). At the score level, mean $\kappa = 0.39$; at the quote level, mean $\kappa = 0.52$, Krippendorff's $\alpha = 0.52$. That suggests that LLMs tend to identify the same textual passages as relevant, but differ in their confidence assessments. This implies that the disagreement between annotators is primarily one of degree rather than of substance.

Per-device analysis (Table 5) reveals that devices with explicit textual markers yield higher agreement (e.g. *Chekhov's Gun* $\alpha = 0.69$), while *Fate vs Choice* $\alpha = 0.18$ and *Historical Witness* $\alpha = 0.07$ are the most challeng-

Title of text	Words	Claude Opus 4.7										Gemini 3 Flash Preview										DeepSeek										#C
		3	9	13	12	23	26	28	33	37	38	3	9	13	12	23	26	28	33	37	38	3	9	13	12	23	26	28	33	37	38	
Movie synopses																																
Beauty and the Beast (Capellani 1908)	1,011	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	8
Cinderella (Geronimi, Jackson, and Luske 1950)	695	0	1	0	1	0	0	0	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1	2
The Land Beyond the Sun. (Edison 1912)	177	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
Antony & Cleopatra (Mankiewicz 1908a)	1,096	0	0	0	1	0	0	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	2
Julius Caesar (Mankiewicz 1908b)	845	1	0	0	1	0	0	0	1	1	1	0	1	0	1	0	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	4
Novellas (one per genre)																																
Animal Farm (satire) (Orwell 1945)	29,908	0	0	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	1	1	0	0	1	0	1	1	0	0	0	4
Of Mice and Men (tragedy) (Steinbeck 1937)	29,673	0	1	0	1	0	0	1	1	1	1	0	1	1	0	0	1	1	1	1	1	0	1	1	0	0	1	0	0	1	0	7
The Canterville Ghost (comedy) (Wilde 1887)	11,373	1	0	1	1	0	0	1	0	1	1	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	1	0	0	0	0	5
Giovanni's Room (romance) (Baldwin 1956)	54,693	0	1	0	1	1	1	1	0	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	7
The Hound of the Bask. (mystery) (Doyle 1902)	59,321	0	1	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	0	1	0	1	1	1	0	0	0	0	0	0	0	9

Table 3: Binary device detections (1 = detected, 0 = not detected) by three LLMs across the genre-balanced pilot corpus of five movie synopses and five novellas, one per genre. Column headers are device numbers from Table 1. Detection is the binarisation of a 0–4 confidence score under the rule $\{3,4\}\to 1$, $\{0,1,2\}\to 0$. **#C** reports the number of conflicts (out of 13) where both constituent devices are detected by the reference annotator.

ing, suggesting devices requiring interpretive judgment need more precise definitions.

Observation 1 - The matrix discriminates by structural density. The shortest tale (*The Land Beyond the Sunset*) triggers zero conflicts; richly-structured novellas trigger many (*The Hound of the Baskervilles*, 9 conflicts; *Of Mice and Men* and *Giovanni's Room*, 7 each). Mean conflict density also differs by text scale: the five movie synopses average 3.2 conflicts, while the five full novellas average 6.4 - consistent with the expectation that longer works deploy more device pairs in conflict.

Observation 2 - Detected conflicts are textually visible. In *The Hound of the Baskervilles*, Conflict 2 manifests as the moor-at-night prohibition (Forbidden Zone) that is defused the moment Stapleton's identity is disclosed (Secret Revelation) - the interdiction's Proppian power and the revelation's surprise arrive in the same stroke. In *Giovanni's Room*, Conflict 3 manifests as Giovanni's external test ("If you cannot love me, I will die") colliding with David's internal Judas/Savior dilemma - the intellectual depth of the dilemma is precisely what prevents the transformation required by the test. In both cases, the structural tension named by the matrix is not just present but organizes the text.

Discussion

Automated narrative evaluation. Current narrative-quality metrics assess devices *individually* (is a ticking clock present? is the character transformed?). The conflict matrix reframes evaluation around *relational* features: given that both Device A and Device B are present, is their contradiction resolved productively or trivially? The richest texts in our pilot (*Hound*, *Beauty and the Beast*) trigger 8-9 of the 13 conflicts under binary detection, while the shortest tale triggers zero - suggesting that *conflict density* is a cheap, corpus-scalable proxy for structural sophistication.

Automated story generation. LLM-based story generators typically condition on genre, theme, or character prompts but rarely on explicit structural contradictions. The contradiction matrix suggests a retrieval-augmented protocol in which, given a target pair of narrative parameters, the model retrieves device pairs known to resolve that contradiction. This reframes the generation objective from "what happens next?" to "which conflict is being managed here?". Sophisticated stories may use many plot devices and contradiction resolution.

Limitations and future work. This is a pilot: 10 texts, three annotator LLMs, and thirteen conflicts out of a combinatorially larger space. We treat the 24×24 matrix as proof of concept rather than canonical, whereas TRIZ's 40-parameter matrix was the product of decades of patent analysis. Natural next steps are: (i) scale to a corpus of 300 novellas collected from Goodreads (Goodreads 2008), of which a 35-text was used in prior work on structural pattern evaluation (Kaplan, Shmilovici, and Last 2026); (ii) deepen the cross-LLM analysis with calibrated prompts for the low-agreement devices; (iii) extend the device vocabulary to the full contradiction matrix.

The bridge we are attempting to build - is that the *formalism* of TRIZ is more portable than its content. Invention, whether mechanical or narrative, is the systematic evolution of improvement at the expense of something else. Visualization of those trade-offs is a contradiction matrix; making them LLM-detectable makes the matrix operational at the corpus scale.

Author Contributions

All authors participated in the writing of this manuscript.

Acknowledgments

This research was partially supported by the Stein Faculty of Computer and Information Science, Ben-Gurion University of the Negev, Ministry of Aliyah and Integration, Israel.

References

Altshuller, G. S. 1999. *The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity*. Technical Innovation Center, Inc.

Anthropic. 2026. Model migration guide. `https://platform.claude.com/docs/en/about-claude/models/migration-guide`. Accessed: May 2026.

Aristotle. 1902. *Poetics*. London: Macmillan.

Baldwin, J. 1956. *Giovanni's Room*. New York: Dial Press. Paperback edition: Laurel, 1988, 178 pp.

Booth, W. C. 1983. *The Rhetoric of Fiction*. Chicago: University of Chicago Press, 2nd edition.

Campbell, J. 1949. *The Hero with a Thousand Faces*. New York: Pantheon Books.

Capellani, A. 1908. La belle et la bête. [Movie]. Pathé Frères, France.

Doyle, A. C. 1902. *The Hound of the Baskervilles*. London: George Newnes.

Edison, T. A. 1912. The land beyond the sunset. [Movie]. Edison Manufacturing Company.

Genette, G. 1980. *Narrative Discourse: An Essay in Method*. Ithaca, New York: Cornell University Press.

Geronimi, C.; Jackson, W.; and Luske, H. 1950. Cinderella. [Movie]. Walt Disney Productions.

Goodreads. 2008. World's greatest novellas. Goodreads list of highly rated novellas. `https://www.goodreads.com/list/show/1236.World`.

Kaplan, M.; Shmilovici, A.; and Last, M. 2026. Structural patterns in award-winning novellas. In *Proceedings of the Text2Story'26 Workshop*. 29 March 2026.

Kar, S.; Maharjan, S.; López-Monroy, A. P.; and Solorio, T. 2018. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Mankiewicz, J. L. 1908a. Antony and cleopatra. [Movie]. 20th Century Fox.

Mankiewicz, J. L. 1908b. Julius caesar. [Movie]. Metro-Goldwyn-Mayer.

McKee, R. 1997. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. New York: HarperCollins.

Mohammadi, A., and Forouzanfar, F. 2024. A TRIZ-based algorithm for systematic story creation. *Creativity Studies*.

Orwell, G. 1945. *Animal Farm*. London: Secker & Warburg.

Palma, C., and Jager, B. 2026. Rethinking TRIZ for AI-driven narrative generation.

Palma, C. 2025. *Neurosymbolic Narrative Generation for Cultural Heritage*. Ph.D. Dissertation, Università degli Studi di Bari Aldo Moro.

Propp, V. 1968. *Morphology of the Folktale*. American Folklore Society Bibliographical and Special Series. Indiana University Research Center in Anthropology, Folklore, and Linguistics, 2nd edition.

Steinbeck, J. 1937. *Of Mice and Men*. New York: Covici Friede.

Wilde, O. 1887. The canterville ghost. *The Court and Society Review*.

Appendix A: Narrative Device Definitions

For each device we provide: a definition with key markers, a positive example (device present), a negative example (device absent), and a borderline case with verdict.

[3] Hero's Test

Source: Campbell, *The Hero with a Thousand Faces* (1949).

Definition: The hero faces an ordeal that tests inner qualities—courage, wisdom, or moral worth. The outcome depends on the hero's *character*, not luck or external force.

Key markers: (i) a challenge is explicitly or implicitly issued; (ii) the hero must draw on inner resources; (iii) passing leads to reward, failing to loss; (iv) a donor, judge, or antagonist often administers the test.

Positive: *"The witch told Ivan he must guard her herd of mares for three nights without losing a single one. If he succeeded, she would give him the magic horse. If he failed, she would eat him."* — explicit ordeal with moral stakes and reward/punishment structure.

Negative: *"Ivan struggled to cross the flooded river in the dark."* — an obstacle, not a test; no one judges Ivan's character; no reward depends on the outcome.

Edge case: *"The princess said she would only marry the man who could answer her three riddles."* — PRESENT: functions as a Hero's Test because the outcome reveals the hero's wisdom and determines his fate.

[9] Ticking Clock

Source: McKee, *Story* (1997).

Definition: A deadline imposed on the protagonist that compresses time and forces immediate action. The deadline must carry a consequence: if the protagonist does not act before time runs out, something irreversible happens.

Key markers: (i) explicit or implicit deadline ("before dawn," "by midnight"); (ii) clear consequence if missed; (iii) urgency accelerates decisions; (iv) often appears mid-story to raise stakes.

Positive: *"She had until the last petal fell from the enchanted rose. When the final petal dropped, the curse would become permanent and the Beast would die forever."* — explicit deadline with irreversible consequence.

Negative: *"Years passed. Ivan grew old waiting for his chance."* — time passes but there is no deadline, no urgency, no consequence for delay.

Edge case: *"The enemy army would reach the village in three days."* — PRESENT: a concrete deadline with catastrophic consequence if not met.

[12] Impossible Task

Source: Propp, *Morphology of the Folktale* (1968) — function M.

Definition: A task assigned to the hero that appears impossible by ordinary means. Structurally paired with function N (Solution), it exposes the false hero and confirms the true hero, usually through magical aid or inner virtue.

Key markers: (i) task is explicitly assigned, impossible by ordinary means; (ii) assigning figure is an authority (tsar, princess, antagonist); (iii) completion determines life, marriage, or reward; (iv) ordinary characters fail; the true hero succeeds.

Positive: *"The tsar announced: whoever could build a flying ship by morning would receive his daughter's hand and half the kingdom. Every prince tried and failed. Ivan received help from the old beggar and arrived at dawn with a ship that sailed the clouds."* — task assigned by authority, impossible by ordinary means, resolved through magical aid.

Negative: *"Ivan had to cross the burning desert to reach the castle."* — a dangerous journey, not an assigned impossible task; no authority set it as a condition; no false hero is exposed.

Edge case: *"'Bring me the firebird alive,' said the tsar, 'and you shall have your reward. Fail, and you will lose your head."'* — PRESENT: formally assigned, impossible by ordinary means, life-or-death stakes.

[13] Chekhov's Gun

Source: McKee, *Story* (1997).

Definition: Every element introduced must eventually serve a purpose. An object, detail, or capability introduced early creates an implicit promise that it will become significant later (setup and payoff).

Key markers: (i) an object, skill, or detail introduced conspicuously early; (ii) appears unnecessary at the moment of introduction; (iii) later resolves a key problem or changes the course of events; (iv) in a synopsis, setup alone qualifies if payoff is clearly implied.

Positive (full): SETUP: *"He noticed a small hunting knife hidden under the floorboard and moved on."* PAYOFF: *"Bound and alone in the cellar, he remembered the knife."* — classic setup-and-payoff structure.

Negative: *"Ivan drew his sword and fought the dragon."* — the sword is used immediately; no deferred setup, no promise of later significance.

Edge case: *"The narrator mentions three times that the protagonist never learned to swim."* — PRESENT: repeated emphasis functions as a setup, strongly implying future narrative relevance.

[23] Forbidden Zone

Source: Propp, *Morphology of the Folktale* (1968) — functions γ (Interdiction) and δ (Violation).

Definition: A space, object, action, or piece of knowledge explicitly prohibited to the hero. The prohibition creates the conditions for violation and all subsequent action. The prohibition must be explicit or institutionally established.

Key markers: (i) explicit prohibition ("do not," "never," "it is forbidden") or established taboo; (ii) the hero violates or is tempted to violate it; (iii) the violation sets the rest of the plot in motion.

Positive: *"'You dare not look into the eastern wing,' the tsar commanded. Ivan promised. But three days into his father's absence, curiosity overtook him."* — explicit interdiction followed by violation.

Negative: *"The cave was dark and dangerous. No one had returned from it alive. Ivan hesitated at the entrance."*

— a dangerous place, but not forbidden; the danger is practical, not a taboo.

Edge case: *"Everyone in the village knew that speaking the old king's name after sunset brought misfortune."* — PRESENT: institutionally established prohibition qualifies even without an explicit verbal command.

[26] Secret Revelation

Source: McKee, *Story* (1997).

Definition: A disclosure of previously withheld information that retroactively reframes the meaning of prior events. After the revelation, earlier scenes mean something different than they appeared to mean.

Key markers: (i) information disclosed that was hidden from reader or protagonist; (ii) the disclosure changes the meaning of earlier events; (iii) the reader experiences a "reread effect"; (iv) often signaled by *"Only then did she realize…"* or *"He had known all along…"*

Positive: *"Only then did she understand: the old beggar woman who had sheltered her had been her own mother, believed dead for twelve years. Every kindness she had received was a mother's love in disguise."* — the revelation retroactively reframes all prior interactions.

Negative: *"The messenger arrived and told Ivan that the princess was being held in the eastern tower."* — new information as exposition; it does not reframe prior events.

Edge case: *"The letter in the drawer contained his father's confession. He read it now, and everything he had believed about his childhood collapsed."* — PRESENT: the revelation reframes the protagonist's entire past, even without the letter's specific content being given.

[28] Historical Witness

Source: Genette, *Narrative Discourse* (1980).

Definition: A narrative voice or character who testifies to events from a position of historical proximity—someone who was present, who remembers, or who belongs to the same historical moment. The testimony carries the weight of personal or cultural memory and lends narrative credibility.

Key markers: (i) a character who witnessed or lived through the events described; (ii) first-person testimony ("I was there," "I remember") or third-person narration grounded in eyewitness account; (iii) the historical grounding is narratively functional.

Positive: *"My grandfather had fought in that war. When we passed the old bridge, he would slow his steps and say: 'This is where we held the line."'* — the grandfather grounds the narrative in lived, verifiable history.

Negative: *"The battle had taken place three hundred years ago. The history books recorded that ten thousand men had died."* — historical setting, not a Historical Witness; no character testifies from personal proximity.

Edge case: *"She had been a child during the revolution. She did not speak of it, but the way she flinched at loud sounds told the story her words could not."* — PRESENT: embodied memory constitutes Historical Witness even without explicit verbal testimony.

[33] Moral Dilemma

Source: Booth, *The Rhetoric of Fiction* (1983).

Definition: A situation in which the protagonist must choose between two options, both carrying significant moral cost. A genuine Moral Dilemma requires that *neither option is clean*; the text makes the cost of both options visible.

Key markers: (i) two options, both with serious moral costs; (ii) the protagonist cannot choose without betraying some value; (iii) both costs are made visible in the text; (iv) not resolved by a third option or a clever trick.

Positive: *"She could save the village by revealing her brother's hiding place—or she could protect her brother and let the plague spread to three more towns. There was no third way. She had until morning."* — genuine dilemma: both options carry irreversible moral cost.

Negative: *"Ivan had to decide whether to fight the dragon or run away."* — a difficult decision but not a moral dilemma; one option is clearly "right," the other clearly "wrong."

Edge case: *"He could tell the truth and destroy his family, or lie and preserve it—knowing the lie would eat at him for the rest of his life."* — PRESENT: both options carry moral cost; the internal cost of the second option suffices to qualify.

[37] Fate vs. Choice

Source: Aristotle, *Poetics* (c. 335 BCE).

Definition: The tension between what is predetermined (destiny, prophecy, structural inevitability) and what the protagonist freely decides. Both elements must be active: a force of fate *and* a moment of genuine choice.

Key markers: (i) a prophecy, curse, or structurally inevitable outcome is established; (ii) the protagonist attempts to resist, fulfill, or navigate this destiny; (iii) a moment of genuine choice occurs within the fateful framework; (iv) the text makes visible the tension between the two forces.

Positive: *"It was written that the youngest son would destroy the kingdom. Ivan, knowing nothing of the prophecy, chose on his own to climb down and seek his fortune—each step of his free will walking him closer to the foretold catastrophe."* — prophecy (fate) + independent choice + tension between them.

Negative: *"Ivan decided to leave home and seek his fortune. No one had predicted this. He simply chose to go."* — pure choice with no fateful framework.

Edge case: *"She had always known, in some wordless way, that she would return to this village. And now, without planning to, she had."* — UNCERTAIN: qualify as present only if the text develops this tension further.

[38] Reconciliation Scene

Source: Aristotle, *Poetics* (c. 335 BCE) — *anagnorisis* + *peripeteia*.

Definition: A scene of recognition (*anagnorisis*) combined with reversal (*peripeteia*), in which estranged characters acknowledge each other's true nature, leading to emotional resolution. The reconciliation must involve genuine change in the relationship—not merely a meeting or truce.

Key markers: (i) characters who were estranged, opposed, or in conflict; (ii) a moment of recognition: one or both understand something they did not before; (iii) the recognition changes the emotional or moral terms of the relationship; (iv) resolution follows from understanding, not external force.

Positive: *"For the first time, the father looked at his son—not as the disappointment he had always seen, but as the man his son had become. 'I did not know,' he said. Something between them, long broken, quietly mended."* — mutual recognition + change in relationship terms + emotional resolution.

Negative: *"The two brothers agreed to a truce. They divided the land equally and went their separate ways."* — a negotiated settlement, not a Reconciliation Scene; no recognition, no change in understanding.

Edge case: *"She forgave him. He did not ask for forgiveness and did not fully understand why she gave it. But she forgave him anyway."* — UNCERTAIN: one-sided recognition; qualify as present only if the text shows that the forgiveness constitutes a form of mutual acknowledgment.

Appendix B: Agreement

	At the score level			At the quote level		
	Claude–Gemini	Claude–DeepSeek	Gemini–DeepSeek	Claude–Gemini	Claude–DeepSeek	Gemini–DeepSeek
P_o	0.67	0.72	0.69	0.74	0.77	0.77
P_e	0.50	0.50	0.49	0.50	0.50	0.50
κ	0.33	0.44	0.40	0.48	0.54	0.54
mean κ		0.39			0.52	

Table 4: Cohen's κ at two levels: score-level (binarized confidence scores) and quote-level (verbatim textual evidence

Appendix C: Per-device Krippendorff's α

Rank	Device	Krippendorff's α
1	Chekhov's Gun	0.69
2	Moral Dilemma	0.66
3	Secret Revelation	0.47
4	Impossible Task	0.42
5	Reconciliation	0.42
6	Hero's Test	0.41
7	Ticking Clock	0.39
8	Forbidden Zone	0.25
9	Fate vs. Choice	0.18
10	Historical Witness	0.07

Table 5: Per-device Krippendorff's α at the quote level, ranked by descending agreement.

Appendix D: Matrix of Contradictions

Table 6.1: The Narrative-TRIZ contraddiction matrix.

Improves ↓ / Worsens →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1. Narrative Pace	-	9, 14	13, 4	16, 10	39, 40	15, 2	6, 3	9, 14	2, 37	38, 5	3, 17	6, 40	9, 29	28, 4	7, 32	26, 10	17, 18	40, 7	6, 39	11, 36	7, 9	9, 39	7, 28	10, 5
2. Plot Complexity	9, 3	-	13, 4	23, 26	25, 24	15, 20	3, 33	9, 12	4, 8	6, 33	3, 17	5, 40	15, 24	28, 29	25, 15	10, 26	17, 19	34, 37	10, 39	33, 36	13, 9	40, 10	28, 17	10, 14
3. Causal Clarity	4, 4	2, 29	-	27, 31	40, 25	6, 28	33, 8	4, 37	4, 7	6, 10	3, 17	4, 40	4, 29	28, 29	7, 26	10, 31	19, 21	34, 37	38, 39	33, 31	7, 30	40, 4	28, 27	16, 1
4. Suspense/Tension	9, 10	18, 15	27, 25	-	39, 2	20, 25	33, 32	12, 6	27, 18	23, 25	3, 17	35, 25	26, 11	28, 29	2, 32	26, 10	18, 23	35, 25	39, 32	33, 35	9, 10	10, 39	28, 12	10, 5
5. Resolution	39, 2	40, 24	13, 25	10, 26	-	38, 29	38, 2	12, 33	38, 2	38, 2	39, 17	38, 34	38, 29	29, 28	38, 26	26, 38	38, 19	38, 35	39, 38	35, 33	38, 2	38, 10	38, 28	38, 10
6. Timeline	4, 28	40, 20	28, 29	25, 36	40, 38	-	20, 6	37, 20	28, 4	28, 20	28, 17	40, 28	28, 20	28, 4	28, 20	20, 28	20, 19	40, 37	28, 20	20, 37	28, 20	40, 20	28, 20	28, 20
7. Char. Depth	33, 6	40, 8	36, 34	33, 25	2, 34	28, 33	-	33, 6	36, 33	21, 33	3, 17	40, 29	33, 34	28, 33	33, 34	33, 34	33, 19	33, 34	33, 36	33, 36	33, 2	33, 39	33, 28	33, 10
8. Char. Agency	33, 6	6, 8	37, 12	6, 25	25, 24	37, 28	33, 6	-	37, 6	6, 33	3, 17	37, 34	6, 33	37, 28	6, 33	6, 10	6, 19	37, 33	6, 33	33, 37	6, 33	6, 33	6, 28	6, 10
9. Char. Believability	33, 8	37, 8	21, 7	33, 25	2, 28	28, 37	33, 8	37, 6	-	21, 7	3, 4	40, 35	33, 7	28, 21	7, 33	2, 31	21, 7	33, 37	33, 7	33, 37	7, 2	33, 7	7, 28	1, 10
10. Char. Relatability	21, 33	6, 12	21, 7	23, 25	2, 38	28, 37	21, 33	6, 12	21, 7	-	3, 38	34, 40	21, 7	20, 21	7, 21	2, 21	21, 19	21, 37	21, 33	21, 33	21, 2	34, 21	21, 28	21, 10
11. Char. Transform	3, 17	6, 33	4, 7	3, 25	39, 34	17, 2	3, 17	6, 33	4, 7	2, 38	-	40, 34	3, 17	20, 28	7, 17	17, 38	17, 19	37, 33	3, 38	3, 33	17, 33	3, 17	17, 28	3, 10
12. Char. Consistency	40, 34	37, 33	40, 7	35, 25	38, 34	28, 40	40, 34	37, 33	40, 7	21, 34	17, 2	-	40, 7	28, 40	7, 40	40, 38	40, 19	40, 37	40, 38	40, 33	40, 2	40, 34	40, 28	40, 10
13. Info. Density	9, 29	15, 24	30, 7	11, 26	38, 29	28, 20	33, 34	6, 33	7, 33	7, 21	17, 3	7, 40	-	29, 28	30, 7	11, 26	17, 19	34, 29	11, 26	33, 30	30, 7	11, 34	29, 28	1, 32
14. Historical c.	28, 4	28, 29	28, 29	28, 25	29, 28	28, 4	28, 33	37, 28	28, 21	20, 21	28, 20	28, 40	29, 30	-	7, 28	26, 11	19, 21	34, 37	28, 4	28, 37	7, 21	28, 40	28, 29	26, 1
15. Clarity	7, 30	25, 15	7, 26	2, 32	38, 26	28, 20	33, 34	6, 33	7, 33	7, 21	7, 17	7, 40	30, 7	7, 28	-	32, 2	17, 21	34, 7	2, 7	7, 30	7, 25	7, 2	7, 28	7, 2
16. Surprise	26, 10	10, 26	10, 31	26, 10	26, 38	20, 28	33, 34	6, 10	2, 31	2, 21	17, 38	40, 38	11, 26	26, 11	32, 2	-	18, 23	35, 25	26, 10	35, 31	2, 26	26, 10	26, 18	26, 10
17. World-building	17, 18	17, 19	19, 21	18, 23	38, 19	20, 19	33, 19	6, 19	21, 7	21, 19	17, 19	40, 19	17, 18	19, 21	17, 21	18, 23	-	34, 19	19, 21	19, 21	17, 21	19, 17	19, 28	19, 18
18. Thematic c.	40, 7	34, 37	34, 37	35, 25	38, 35	40, 37	33, 34	37, 33	33, 37	21, 37	37, 33	40, 37	34, 33	37, 28	34, 7	35, 25	34, 21	-	34, 33	34, 37	34, 7	34, 35	34, 28	34, 35
19. Emotional c.	6, 39	10, 39	38, 39	39, 32	39, 38	28, 20	33, 36	6, 33	33, 7	21, 33	3, 38	40, 38	11, 26	28, 4	2, 7	26, 10	19, 21	34, 33	-	33, 35	2, 39	39, 34	28, 29	10, 5
20. Intellectual c.	11, 36	33, 36	33, 31	33, 35	35, 33	20, 37	33, 36	33, 37	33, 37	21, 33	3, 33	40, 33	33, 30	28, 37	7, 30	35, 31	21, 19	34, 37	33, 35	-	7, 30	35, 40	28, 21	27, 1
21. Accessibility	7, 9	13, 9	7, 30	9, 10	38, 2	28, 20	33, 2	6, 33	7, 2	21, 2	17, 33	40, 2	30, 7	7, 21	7, 25	2, 26	17, 21	34, 7	2, 39	7, 30	-	34, 39	7, 21	1, 2
22. Memorability	9, 39	40, 10	40, 4	10, 39	38, 10	40, 20	33, 39	6, 33	33, 7	34, 21	3, 17	40, 34	11, 34	28, 40	7, 2	26, 10	17, 19	34, 35	39, 34	35, 40	34, 39	-	29, 34	1, 10
23. Education	7, 28	28, 17	28, 27	28, 12	38, 28	28, 20	33, 28	6, 28	7, 28	21, 28	17, 28	40, 28	29, 28	28, 29	7, 28	26, 18	19, 28	34, 28	28, 29	28, 21	7, 21	29, 34	-	1, 18
24. Entertainment	10, 5	10, 14	16, 1	10, 5	38, 10	28, 20	33, 10	6, 10	1, 10	21, 10	3, 10	40, 10	1, 32	26, 1	7, 2	26, 10	19, 18	34, 35	10, 5	27, 1	1, 2	1, 10	1, 18	-

Figure 1: The Narrative-TRIZ contradiction matrix, adapted from Table 6.1 (Palma 2025). The spelling ”contraddiction” is retained intentionally, as it reproduces the original title. The abbreviation ”c.” stands for ”contribution” meaning each parameter measures the degree or presence of a particular quality in the text. Coloured borders mark the cells from which the thirteen conflict pairs in Table 2 were derived.

The Affordance is the Message: Creative Media as Complex Systems

Ane Espeseth[1] and **Elias Najarro**[2]
[1]Center for Computing in Science Education, University of Oslo, Oslo, Norway
[2]Play, Culture and AI, IT University of Copenhagen, Denmark
anekes@uio.no, enaj@itu.dk

Abstract

The affordances of a creative medium strongly condition the creative artefacts the medium will produce. In this work, we present a formalisation of computational creativity (CC) media using the conceptual toolbox of complex systems (CS). We introduce the notions of emergence, collective intelligence and self-organisation, non-linear dynamics, criticality, multi-scale hierarchy, phase transitions, diversity of attractors, path dependence, and open-endedness, and connect them to the existing CC literature. Together these nine properties form a vocabulary with which creative media can be described and compared at the system level, while medium affordances are the design-level mechanisms that determine each medium's complex system properties. The formalisation emphasises the influence of each medium's affordances in determining what the medium can produce in creative processes. To demonstrate the proposed theoretical approach, we characterise a diverse set of media (`r/place`, `Minecraft`, `cellular automata`, `Twitter`, `Boids`, ...) using this vocabulary. The proposed formalisation under the CS concepts serves a dual purpose: it establishes a link between the affordances and realised practices of the medium, and it offers a shared lens for characterising existing creative media. A companion website at `affordances.wiki` provides implementations of different computational media, serving as a playground for users to explore the relation between a medium's affordances and its potential creative artefacts.

Introduction

A computational medium used to elicit and express creativity can be seen as complex systems: McCormack (2012) argues in his work on *creative ecosystems* that the environments in which creative activity unfolds emerge from interactions among agents, substrates and shared rules. Murray (2011) treats the digital medium as an authored bundle of expressive *affordances* (Gibson 1979): combinations of its surfaces and elements, perceivable and meaningful relative to a particular agent. In this paper, we use the term "creative medium" to describe a computational substrate with a set of affordances that hosts creative activity. These affordances combine to produce the system-level properties that characterise the medium as a complex system.



Figure 1: Computational creative media can be studied as complex systems. Complexity science offers a vocabulary of such properties — a principled basis on which to characterise media and investigate how their affordances condition the creative artefacts they produce. This figure illustrates eight of these properties, each paired with a creative medium that exemplifies it.

An illustrative example is the `r/place` experiment, started on Reddit on April Fools' Day 2017 (Rappaz et al. 2018). Users acted within a small affordance set: a 1000×1000 pixel canvas, a 16-colour palette, a single-pixel brush, the ability to override any existing pixel, and a 5-minute cooldown between placements. The combination meant that any lasting mark on the canvas would have to be negotiated between participants, leading to the emergence of a multitude of collective creative behaviours. If this simple 5-minute constriction had been reduced to 1 second, the platform would have collapsed into single-user monopolies, and died out. This simple example tells us that we cannot disregard the importance of the medium's affordances in conditioning what creative artefacts are explored and produced (McLuhan 1964).

Existing CC frameworks of creative potential formalise reachability: Boden's *conceptual spaces* (1992) and Wiggins' *Creative Systems Framework* (2006) describe what concepts an agent can reach by searching a given conceptual

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

space. McCormack's *creative ecosystems* (2012) extend the picture by treating that space as co-evolving with the agent through niche-construction, describing the medium through the joint dynamics of agents acting in it. Our contribution sits one level below both. We propose that a medium's affordances, as a design-level mechanism, induce macroscopic complex-systems properties that belong to the medium itself rather than to any particular agent's strategy within it. Those properties then condition the agent's later search, and the creative artefacts this results in. Because the vocabulary is substrate-agnostic, media with very different internal dynamics (`r/place`, `Twitter`, `Minecraft`) can be studied on the same basis.

Creative Media as Complex Systems

Complex systems are some of the most interesting phenomena we observe in natural, artificial and social systems: from the swarming of starlings, self-replicating patterns in `Conway's Game of Life` or the collective canvas of `r/place`. Behind all of these diverse phenomena, there is a unifying mechanism: many components interacting through simple rules to give rise to emergent behaviours not reducible to any one component in isolation; e.g., individual participants of the `r/place` experiment cannot know how their pixels will condition the trajectory of the global canvas. These systems are referred to as *complex systems* (Mitchell 2009).

In order to analyse a creative medium as a complex system, we propose to characterise media using a set of complexity properties: emergence, collective intelligence and self-organisation, non-linear dynamics, criticality, multi-scale hierarchy, phase transitions, diversity of attractors, path dependence, and open-endedness. The remainder of this section takes them one at a time, working out for each what structural conditions a medium's affordances must satisfy to produce it and which CC concepts it re-describes in CS vocabulary, with a worked medium example (Figure 1).

Emergence. When observed across different scales, certain systems display *macroscopic* patterns that are qualitatively different from the behaviours of each individual component at the *micro-level*. For instance, the macroscopic behaviour of a complex organism appears to transcend the biochemistry that produces it. These emergent phenomena are naturally described using concepts that do not exist in the language of the system's microscopic components: a "glider" in `Game of Life` is a meaningful macro-level concept, yet it has no counterpart in the micro-level language of cell states and neighbourhood rules. While emergent behaviours do follow in principle from the micro-level rules, they are not readily predictable from them, a consequence of non-linear interactions propagating across spatial and temporal scales.

In the CC literature, the micro-macro distinction maps onto Boden's (1992) separation between *exploratory* and *transformational creativity*, later formalised by Wiggins' (2006) separation between emergence *within* a conceptual space $\mathcal{L}$ and emergence *of* a new $\mathcal{L}$. Taking the aforementioned `r/place` example (2018): the per-pixel placement and override-anywhere rule are deliberately too coarse to encode any macro object directly, so flags, portraits and alliance territories only become visible at the canvas scale. Reading emergence through complexity science adds a design lever absent from earlier CC framings: the gap between micro and macro vocabularies is something the medium designer can set, by choosing how coarse the affordances are.

Collective intelligence and self-organisation. *Collective intelligence* refers to self-organising phenomena where individual agents coordinate and integrate information to solve higher-order problems that are inaccessible to individuals. The underlying mechanism, *self-organisation*, refers to the spontaneous formation of order purely from local interactions: no single element directs the entire system; instead, decision-making and influence are spread across many components, often referred to as *distributed control* (Camazine et al. 2001). The time evolution toward ordered states, in the form of a fixed point or limit cycle, can be caused by simple negative feedback loops or complex interaction motifs.

CC has engaged with this property under several names: *collective creativity* (Sawyer 2003), *social creativity* (Saunders and Bown 2015; Fischer et al. 2005), *creative ecosystems* (Bown 2012; Maher 2010), and *co-creativity* (Kantosalo and Toivonen 2016; Linkola and Kantosalo 2019). Using the complexity concept as a recipe, we get a sense of what such a "coordination layer" must contain: stateful elements, a notion of locality, and a local-state update mechanism with no central controller. `Wikipedia's talk pages` (Viégas et al. 2007) instantiate this set of properties, and produce two creative artefacts in the process: collaboratively written articles, and institutional policies that emerge from practice rather than from any top-down design. The concrete affordances used to realise these properties include threaded replies (locality), signatures plus edit-versioning (per-contributor state), and watchlist-driven attention (decentralised update).

Non-linear dynamics. Systems exhibit non-linear interactions where small perturbations can produce system-wide effects. These dynamics are typically driven by feedback loops and non-linear interactions that amplify local changes into global effects.

Non-linearity is ubiquitous in CC media but rarely formalised under that name; instead the phenomenon recurs under different headings: *deceptive landscapes* in Stanley and Lehman's novelty search (2011), *predictive deviation* in Maher and Grace's surprise search (2014), *combinatorial explosion* in Boden's combinatorial creativity (1992), and the dynamical-systems work on interactive music (Eldridge 2008; Bown 2012) that names non-linearity directly. From the complexity concept, what combines all of these is a coupling that ties local actions to global state, and a feedback channel that lets small differences amplify rather than damp out. The microblogging website `Twitter` (Goel et al. 2016) instantiates this pattern through the reshare button which couples a local post to global reach (coupling) and

the heavy-tailed follower-graph which amplifies small differences in reach into large differences in engagement (feedback).

Criticality and Edge of Chaos. Some systems sit at a regime between order and disorder, where correlations span all scales and the system is maximally responsive to inputs. At criticality, fluctuations are scale-free and small perturbations can propagate arbitrarily far. The *edge of chaos* hypothesis (Langton 1990) associates this regime with the richest emergent computation, sitting between frozen, predictable dynamics on one side and disordered, uncorrelated dynamics on the other.

CC literature encounters criticality in Langton (1990) and Wolfram's class-IV cellular automata (1984); specially relevant in context of generative-art parameter sweeps (2001), where the "interesting" regime is narrow. If we look to complexity science for what characterises this property, we find that we need a tuned *control parameter* that establishes an ordered regime on one side and a disordered regime on the other, and a narrow window between them where small disturbances can propagate across the full system. `Boids` (Reynolds 1987) is an illustrative creative instance: each boid steers by three local rules — cohesion (steer toward neighbours' centroid), alignment (match neighbours' heading), and separation (avoid crowding) — whose weights act as the control parameter. Turn cohesion and alignment up and the flock collapses into a tight cluster locked to the average heading (ordered); turn them down and the boids drift independently, one boid's turn never reaching another (disordered). The narrow band between is flocking proper: a single turn ripples across the flock and reshapes it without freezing or shattering it.

Multi-scale hierarchy. Complex systems often exhibit multiple nested levels of interactions and feedback loops, operating across scales. In many cases, these hierarchies are not imposed, but instead emerge from the lower organisational levels. Interactions among elements within each level, and across different levels, typically occur on distinct time scales.

CC's most explicit treatment of hierarchy is Wiggins' meta-search (2006), where a slow outer loop modifies the rule-set $\mathcal{R}$ which the fast inner loop searches under. Saunders's three-level model (2015) (artefact / evaluator / society) and the Csikszentmihalyi-derived individual / domain / field framing (1999) also layer creative activity similarly. From the complex system lens, each level has its own vocabulary and its own dynamics, with within-level interactions stronger or faster than across-level coupling (Simon's *near-decomposability* (Simon 1962)), so that the level can be reasoned about in its own terms. `Minecraft` (Duncan 2011) illustrates several such levels at once: blocks, structures, biomes, and world-scale builds nest spatially on the artefact side, players, factions, and server cultures nest socially, and redstone introduces a distinct vocabulary (signals, gates, computation) above the block-placement layer.

Presence of phase transitions. A system exhibits phase transitions when its macroscopic state changes suddenly and qualitatively as a control parameter (temperature, density, connectivity, etc.) crosses a critical value. Originally formalised in statistical physics, phase transitions are also observed in networks, populations, and information-processing systems, where they typically separate qualitatively distinct macro-regimes.

Phase transitions are not standard CC vocabulary, but the underlying phenomenon is. Boden's *transformational creativity* is structurally a discrete change in $\mathcal{R}$ (2004), Wiggins' CSF formalises it as a move in the meta-space over $\mathcal{R}$ (2006), and Csikszentmihalyi's *creative breakthroughs* (1999) occupy the same slot under a systems reading. The complex systems approach decomposes the phenomenon into its ingredients: a tunable control parameter, a critical value of that parameter, and a qualitative regime change in the dynamics across it. Here again the `r/place` medium makes for a good example. Below a critical canvas density, participants paint freely in uncontested regions, each placement effectively independent of every other. Above it, every placement must overwrite something a prior participant cared about, and the medium switches phase: alliances form, territories are defended, and off-platform Discords organise raids and counter-raids. The bounded canvas makes interaction patterns a function of the cooldown timer (control parameter) and the user's cooldown value results in coordination over monopoly as the post-threshold equilibrium (regime change).

Diversity of attractors. In dynamical system theory, an attractor is a region of state space toward which trajectories of the system state evolve. The mere presence of attractors says little about a creative medium, since any sufficiently constrained system will exhibit some. What matters for CC is the *diversity* of attractors a medium supports—fixed points, limit cycles, or strange attractors of qualitatively different kinds—whose collective structure bounds the long-term qualitative behaviours the medium can express.

Attractor language is sparse in CC and mostly metaphorically used (*creative ruts*, *style basins*, *mode collapse*), with explicit uses in Eldridge's interactive-music work (2008) and Bown on musical agents (2012). Two CC concepts touch the diversity aspect: Ritchie's *typicality* (2007), proximity to a basin of attraction; and Koestler's *bisociation* (1964), later formalised by Fauconnier and Turner as *conceptual blending* (2002), a cross-basin operation that presupposes the substrate hosts multiple kinds of basin to combine across. Fixed points need stable rest configurations the system can settle into; limit cycles need feedback or scheduling structures with a natural period of return; strange attractors need a bounded variation channel rich enough that trajectories stay confined to a regime without ever repeating. The mix a medium supports determines the modes of expression it makes available: a substrate offering only the first sustains only convergence, while one supplying all three sustains convergence, recurrence, and perpetual variation. An example that demonstrates the notion of attrac-

tors is Wikipedia's talk pages (Viégas et al. 2007). Godwin's law — the observation that long-running online debates tend to invoke Nazi comparisons regardless of starting topic — and articles that have stabilised on an accepted version are the fixed-point case (discussion converges and stays); while debates that flare up every few months on contested topics are the limit-cycle case (Sumi et al. 2011).

Path dependence. The state a system reaches depends not only on its current inputs but on the order in which past events occurred: equivalent boundary conditions reached through different histories lead to different futures. A special case is *stigmergy*, where agents coordinate indirectly by modifying a shared environment, each action leaving a trace that constrains and guides subsequent actions. Path-dependence appears in CC under several names: Stanley and Lehman's *stepping-stone* reading of novelty search (2011), where interesting end-points are reached only through paths whose intermediate stages don't look like progress toward them; Tomasello's *cumulative cultural ratchet* (1999), in which each generation builds on the artefacts of the last; and lineage-based generative systems such as Picbreeder (2011), where every image is a leaf of a specific evolutionary path. From the CS account, the phenomenon decomposes into two ingredients: a persistent state that records past events, and a coupling between current actions and that state under which each action both reads from it and modifies it. Without persistence there is no path for the future to depend on; without read—modify coupling each action operates independently of history.

An interesting example is the gaming modality, or artistic form, known as speedrunning (Scully-Blaker 2014): each newly-discovered glitch becomes a stepping stone for the next, and canonical "any%" routes accumulate from sequences of exploits whose individual stages are not interesting on their own. The community-maintained archive of routes, splits, and frame-counts supplies the state, and the practice of forking an existing route to insert a new exploit supplies the coupling. The stigmergic special case (Bonabeau, Dorigo, and Theraulaz 1999) fits r/place (2018): the persistent canvas is the state, and override-anywhere plus the absence of direct messaging force coordination to read and modify it.

Open-endedness. A system is open-ended when it has the capacity to continuously produce novel, increasingly *interesting* and diverse behaviours.

Open-endedness (OE) is the deepest pre-existing bridge between CC and Complexity Science: Stanley and Lehman on open-ended evolution (2011); Banzhaf et al. on OE criteria; Chan's Lenia (2019) and Mordvintsev's Particle Lenia treat OE as a substrate property to be engineered. From the complex systems perspective, OE decomposes into three ingredients: an unbounded reachable set the medium can in principle produce (substrate OE), a novelty-generating dynamic that keeps producing new configurations (process OE, often via coupling between agents and substrate), and the absence of an imposed objective that would close the space by selecting for convergence. Picbreeder (2011) instantiates all three: its CPPN encoding makes the image space effectively unbounded (the reachable set); aesthetic selection by users drives the trajectory under their evolving preferences (the 'novelty engine'); and the deliberate absence of any fitness function lets lineages diverge into shapes nobody set out to produce (no imposed closure).

Formalising affordances through this conceptual basis of complex systems notions allows creative media to be compared along principled dimensions. Two routes are available. From the affordance side (micro to macro), a medium can be characterised by asking which structural conditions its affordances supply, taking each property's recipe from the previous section as a checklist. From the complex systems perspective (macro to micro), a medium's profile can be assessed directly: either phenomenologically, by observing the medium and judging which properties feel present at what intensity, or formally, drawing on measures from the complexity literature for each properties—e.g., entropy measures, long-range correlation functions for criticality, distance-to-attractor measures for diversity of attractors, network-statistical signatures for collective intelligence, and so on, (McCormack and Cruz Gambardella 2022).

Conclusions

In this work, we have presented a complex-systems perspective on computational creativity with the goal of establishing a bridge between the complex systems and the computational creativity communities. By formalising computational creativity notions under the complexity science lens, the formalism allows us to explore how the affordances of a medium—thought of as design choices—determine the creative artefacts it produces.

We have focused on nine system-level properties: emergence, collective intelligence and self-organisation, nonlinear dynamics, criticality, multi-scale hierarchy, phase transitions, diversity of attractors, path dependence, and open-endedness. For each property, we point to the affordances that enable it (i.e., a narrow order-disorder window for criticality, near-decomposable levels with distinct vocabularies for hierarchy, a persistent state with read-modify coupling for path-dependence), and further, we trace these affordances to specific design choices in existing creative media: r/place's per-user cooldown, the cohesion-alignment-separation weights in Boids, the threading-and-signing of Wikipedia talk pages. The formalism is therefore not only a tool for describing media after the fact, but a causal account of how the choice of affordances (i.e., the micro-level) affects the quality and properties of a medium (i.e., the macro-level). This allows us to work backwards from how we intend the artist to engage with artefact creation, to the design of the medium itself. It also opens a path toward identifying under-explored combinations of complexity-informed properties, and outlining the yet-to-exist creative media they would yield.

References

Boden, M. 1992. *The Creative Mind*.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. 2nd edition.

Bonabeau, E.; Dorigo, M.; and Theraulaz, G. 1999. *Swarm Intelligence: From Natural to Artificial Systems*.

Bown, O. 2012. Generative and adaptive creativity: A unified approach to creativity in nature, humans and machines. In *Computers and Creativity*.

Camazine, S.; Deneubourg, J.-L.; Franks, N. R.; Sneyd, J.; Bonabeau, E.; and Theraula, G. 2001. *Self-Organisation in Biological Systems*.

Chan, B. W.-C. 2019. Lenia: Biology of artificial life. *Complex Systems* 28(3):251–286.

Csikszentmihalyi, M. 1999. Implications of a systems perspective for the study of creativity. In *Handbook of Creativity*. 313–335.

Duncan, S. C. 2011. Minecraft, beyond construction and survival. *Well Played: A Journal on Video Games, Value and Meaning* 1(1):1–22.

Eldridge, A. C. 2008. Collaborating with the behaving machine: Simple adaptive dynamical systems for generative and interactive music.

Fauconnier, G., and Turner, M. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*.

Fischer, G.; Giaccardi, E.; Eden, H.; Sugimoto, M.; and Ye, Y. 2005. Beyond binary choices: Integrating individual and social creativity. *International Journal of Human-Computer Studies* 63(4–5):482–512.

Gibson, J. J. 1979. *The ecological approach to visual perception: classic edition*.

Goel, S.; Anderson, A.; Hofman, J.; and Watts, D. J. 2016. The structural virality of online diffusion. *Management Science* 62(1):180–196.

Grace, K., and Maher, M. L. 2014. What to expect when you're expecting: The role of unexpectedness in computationally evaluating creativity. In *Proc. International Conference on Computational Creativity*.

Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proc. International Conference on Computational Creativity*, 77–84.

Koestler, A. 1964. *The Act of Creation*.

Langton, C. G. 1990. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: Nonlinear Phenomena* 42(1–3):12–37.

Lehman, J., and Stanley, K. O. 2011. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation* 19(2):189–223.

Linkola, S., and Kantosalo, A. 2019. Extending the creative systems framework for the analysis of creative agent societies. In *Proc. ICCC*.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proc. DESIRE Network Conference on Creativity and Innovation in Design*.

McCormack, J., and Cruz Gambardella, C. 2022. Complexity and aesthetics in generative and evolutionary art. *Genetic Programming and Evolvable Machines* 23(4):535–556.

McCormack, J., and Dorin, A. 2001. Art, emergence and the computational sublime. In *Proc. International Conference on Generative Systems in the Electronic Arts*.

McCormack, J. 2012. Creative Ecosystems. In *Computers and Creativity*. 39–60.

McLuhan, M. 1964. *Understanding media: The extensions of man*.

Mitchell, M. 2009. *Complexity: A Guided Tour*.

Murray, J. H. 2011. *Inventing the Medium: Principles of Interaction Design as a Cultural Practice*.

Rappaz, J.; Catasta, M.; West, R.; and Aberer, K. 2018. Latent structure in collaboration: The case of reddit r/place. In *Proc. International AAAI Conference on Web and Social Media*, volume 12.

Reynolds, C. W. 1987. Flocks, herds and schools: A distributed behavioral model. In *Proc. Conference on Computer Graphics and Interactive Techniques*, 25–34.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Saunders, R., and Bown, O. 2015. Computational social creativity. *Artificial Life* 21(3):366–378.

Sawyer, R. K. 2003. *Group Creativity: Music, Theater, Collaboration*.

Scully-Blaker, R. 2014. A practiced practice: Speedrunning through space with de Certeau and Virilio. *Game Studies* 14(1).

Secretan, J., and Stanley, K. O. 2011. Picbreeder: A case study in collaborative evolutionary exploration of design space. *Evolutionary Computation* 19(3):373–403.

Simon, H. A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society* 106(6):467–482.

Sumi, R.; Yasseri, T.; Rung, A.; Kornai, A.; and Kertész, J. 2011. Edit wars in Wikipedia. In *Proc. IEEE International Conference on Privacy, Security, Risk and Trust*.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition*.

Viégas, F. B.; Wattenberg, M.; Kriss, J.; and van Ham, F. 2007. Talk before you type: Coordination in Wikipedia. In *Proc. Hawaii International Conference on System Sciences*.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Wolfram, S. 1984. Universality and complexity in cellular automata. *Physica D* 10(1–2):1–35.

Folded Bodies: Topological Control of an Untrained Spiking Instrument

Lucas Kuzma
The Strange Agency
lucas@strange.agency

Abstract

We present a generative audiovisual instrument in which the performer sculpts a spiking network by folding the body it lives on. A seed point cloud is iteratively reflected across mirror planes, accumulating bilateral symmetry. The resulting cloud is both the visible body and the substrate of a leaky integrate-and-fire network: each point is a neuron, each k-nearest-neighbor edge a synapse. Geodesic farthest-point probes read local activity and drive a polyphonic sonifier with two voice paths, oscillator and granular, supporting both pitch-organized and texture-organized sonic responses to the same network state. A per-point pitch scalar derived from azimuthal position is propagated through reflections, so points related by geometric position share both color and pitch — bilateral symmetry of the body extends naturally into bilateral symmetry of the sound. Because state is addressed by position rather than index, firing patterns are spatially legible in the same view the performer uses to fold the geometry. We frame the system as a co-creative instrument whose computational-creativity contribution lies in making the morphology of the dynamical substrate itself the performer's interface.

Introduction

Reservoir computing (RC) (Nakajima and Fischer 2021) and echo state networks (ESNs) (Jaeger and Haas 2004) offer rich dynamical substrates for music generation, but their opacity is a persistent design challenge. Kiefer (2014) demonstrates ESNs as expressive mapping layers for musical instruments while identifying *controllability* as the central difficulty: the reservoir's state space is opaque and corresponds to no externally legible structure. Sculpting the reservoir means navigating a high-dimensional parameter space without visual or intuitive feedback, and the standard response, a trained readout layer, replaces one opacity with another. Following Kiefer, the design problem is not to eliminate the resulting *uncontrol* (Jorda 2005) — the nonlinearity and unpredictability that make a dynamical instrument expressive — but to place it where the performer can negotiate with it. Ianigro and Bown (2016) examine a similar problem for continuous-time recurrent neural networks (CTRNNs) as creative substrates and remark about the difficulty of discovery in the vastness of the CTRNN parameter space.

The system presented here proposes a different response: make the network's topology a direct consequence of a body the performer can see and fold. The Ω operator (Bader and Oxman 2016) iteratively applies planar mirror reflections to a seed point cloud, accumulating bilateral symmetry. The resulting cloud forms the substrate of an untrained spiking network whose edges are its k-nearest-neighbor (kNN) graph. Every neuron occupies a position in a visible 3D body; its firing state at any moment is a spatial event on that body, readable at a glance. Folding the body reshapes the network, and what the performer sees folding is what they hear changing.

The system inherits its body-generating operator from a generative-morphology lineage. The L-systems descended from Lindenmayer (Prusinkiewicz and Lindenmayer 1990) show how compact rule-sets can produce bodies whose form encodes the artist's intent; Bader and Oxman's (2016) recursive-symmetry method introduces the planar-mirror operator Ω, applied iteratively to a seed cloud to generate complex bilaterally-symmetric forms for additive manufacturing. We adopt Ω directly, adapting it from a fabrication pipeline that produces meshes for 3D printing into a substrate-generation pipeline that produces point clouds for spiking-network construction. Hauser *et al.* (2011) formalize a related observation in soft robotics: a compliant body's intrinsic dynamics can serve as the reservoir a recurrent system would otherwise compute on an opaque internal state. We import this framing into a software audiovisual setting where the body is a folded virtual cloud whose dynamics drive a probe-based readout— though, unlike in the standard RC pattern, our readout is not trained. The CC tradition of embodiment (Guckelsberger *et al.* 2021; Saunders *et al.* 2010) argues that situating a generative system in a visible space changes both production and reception; we do not take on its stronger production-side commitments — our body has no environment and no self-preserving dynamics — but the reception side applies directly, because bodies produced by Ω are biomorphic (bilaterally symmetric, organism-like) in a way random or force-directed renderings of the same network are not. The instrument is therefore audiovisual rather than auditory-with-visualization: the visible morphology is part of what the music is heard as coming from.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

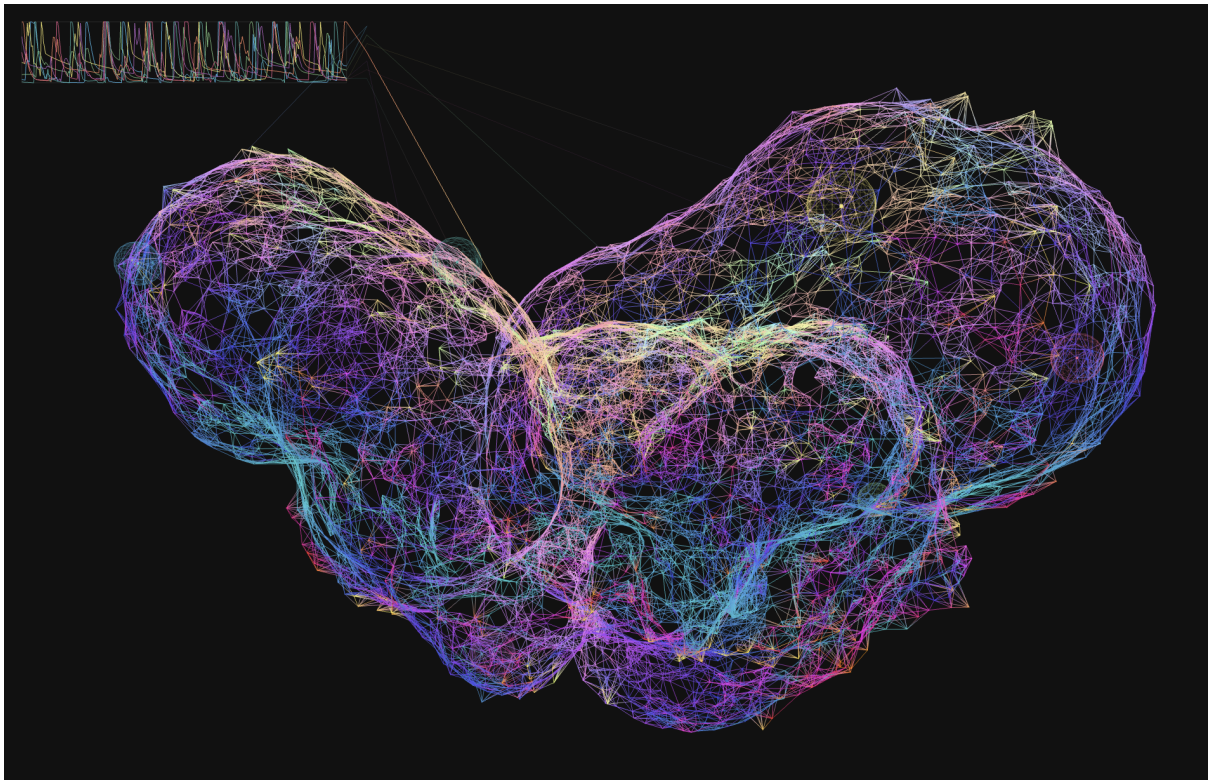


Figure 1: A folded body in performance. The same k-nearest-neighbor graph (rainbow edges) is simultaneously visible geometry, neural substrate, and probe placement domain. Geodesic probes appear as translucent spheres on the body; their activation traces scroll in the heads-up display (HUD) strip (upper left), with thin connector lines tying each probe to its trace.

Computational Creativity stance. We position Folded Bodies as a contribution to musical metacreation (Pasquier *et al.* 2016): the artifact is not a composition but a music-maker the performer shapes. Where standard real-time metacreation is symbolic — live coding edits text, modular synthesis re-patches topologies, evolutionary systems breed by fitness — we propose a *morphological* mode in which the maker is a body and the authorial gesture is folding. The performer chooses the body; the body chooses the music. Three concrete moves support this framing. (i) *State legibility*: because the substrate's neurons sit in a visible body, the network's state is a spatial event the performer reads at a glance, directly addressing the opacity problem reservoir-style instruments face (Kiefer 2014; Ianigro and Bown 2016). (ii) *Action legibility*: the same operator Ω generates the body and modifies it during performance, with no separation of design phase from performance phase. In Boden's (2004) terms each fold acts transformationally, rewriting the space of bodies the instrument can become; we acknowledge Wiggins's (2006) formal result that "Boden's transformational creativity is actually exploratory creativity at the meta-level" (p. 458) and do not claim to escape that regress. (iii) *Structural coherence*: every downstream stage (kNN, probe placement, color, pitch, grain neighborhood) is geometry-indexed, so the body's bilateral symmetry propagates through the pipeline by construction — what changes visually is the same thing that changes audibly. In Kantosalo and Takala's (2020) terms the *collective* is the performer together with the body-as-instrument, and their *contributions* divide cleanly: the performer supplies morphological intent, the system supplies dynamical interpretation, and the visible body is the shared workspace through which they negotiate.

The contribution is therefore a *system demonstration and design argument* for a morphological mode of real-time musical metacreation: the Ω operator, exposed as a live performance gesture over a geometry-indexed spiking substrate, gives the performer a legible, sculptable handle on an otherwise opaque dynamical instrument while preserving end-to-end coherence between what is seen and what is heard.

System

The instrument consists of four components: a folding operator that produces a body, a spiking network that lives on the body's kNN graph, a set of geodesic probes that read local activity, and a polyphonic sonifier that voices each probe through selectable oscillator or granular paths. Figure 2 shows the overall flow.

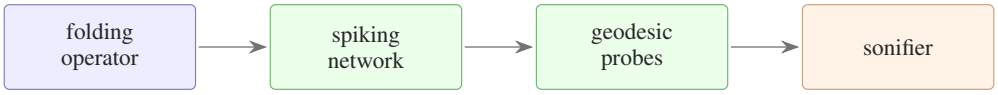


Figure 2: The four components of the instrument: a folding operator that produces a body, a spiking network that lives on the body's kNN graph, a set of geodesic probes that read local activity, and a polyphonic sonifier that voices each probe through selectable oscillator or granular paths. Arrows show information flow; colors mark the geometry (blue), dynamics (green), and sound (orange) phases.

The Ω folding operator

Starting from a seed point cloud (typically a Fibonacci-spaced sphere of N points, optionally distorted by a low-order spherical-harmonic, plane-wave noise, or Gaussian-bump field), the Ω operator applies a sequence of planar reflections. At each step, a plane is placed through the cloud. The signed half-space containing the larger subset of points is retained; the complementary half is discarded; the retained half is reflected across the plane and merged with its mirror image. The combined cloud is voxel-deduplicated and capped to N. After k steps the cloud is bilaterally symmetric with respect to the k-th plane, with earlier folds visible as bilateral substructures within each half.

A given parameter vector produces the same network topology, and small changes to plane angles produce continuous deformations of the body; see Appendix A.1 for the parameter encoding.

Spiking network on the kNN topology

A kNN graph is constructed on the final cloud. Each undirected edge becomes two directed synapses with independent forward and backward weights. The network is a leaky integrate-and-fire (LIF) system with continuous-time decay: each neuron holds a bounded membrane potential, a binary firing flag, and a refractory counter. Per tick, firing neurons distribute their weighted activation to successors; non-refractory neurons integrate inputs; neurons crossing their threshold fire and enter a refractory period during which the activation leaks quickly; multiplicative leaks are applied.

Weights, thresholds, and refractory periods are sampled from stochastic distributions keyed on a network seed, so the network has a stable "personality" that re-instantiates on any topology. The topology is a function of the folding genome; the dynamical parameters are a function of the network seed. Folding therefore reshapes the network's connectivity graph while preserving the character of its noise — useful, because a performer can fold a body into new shapes and recognize the same instrument speaking through them. We use a spiking rather than a continuous-rate network because spike-and-refractory dynamics produce discrete, visually-legible firing events on the body and bursty, episodic activation traces that translate cleanly to musical onsets. Parameter values and the initialization kickstart are in Appendix A.2.

Geodesic probe readout

Probes (typically 5–8) are placed on the kNN graph by farthest-point sampling (FPS) using Dijkstra distance in the graph's intrinsic metric. Edge weights for the Dijkstra are Euclidean distances between adjacent points, so probes spread along the body's surface even when the body is folded tightly in Euclidean space.

Each probe integrates network activation within a Euclidean neighborhood with linear distance falloff, producing a scalar activation trace per probe. The FPS algorithm and probe-slot stability across graph rebuilds are detailed in Appendix A.3.

Sonification

Each probe drives one voice. The sonifier supports two voice paths, selectable live: an *oscillator* path and a *granular* path. Probe activation drives voice gain; camera-space distance to each probe modulates filter cutoff, so closer probes sound brighter; each probe's horizontal screen position drives stereo pan, so the audio tracks where the marker sits on screen.

Two pitch logics are exposed. In *slot mode*, pitch follows slot index, drawn from a fixed palette. In *prune mode*, pitch follows each probe's *position on the body* via a per-point scalar that Ω carries unchanged through every fold; in the granular path the same scalar indexes a position in the source buffer rather than an oscillator frequency. Folding therefore *removes* the pitches and buffer positions the body can no longer reach, so each shape sings through a smaller, distinct palette rather than a re-shuffling of the same set; bilateral body symmetry extends to bilateral palette symmetry. See Appendix A.5.

The granular voice path was added because square oscillators, however tunable, produce a tonal output that does not match the textural quality of folded geometry. The visual surface has detail — twin clusters, density variations, fold seams — and the granular path lets the audio reflect this in kind, fusing the texture of input and the texture of output. When a clearly bilaterally-symmetric body sings in prune mode, the audio's left and right halves are not just panned reflections but textural reflections, drawing from the same buffer regions in mirror-image patterns. This was the design goal that motivated the granular path: making what the system *sounds* like reflect what it *looks* like, not merely what it does. Voice signal chains are in Appendix A.4.

The instrument

The performer sees the body rotating in 3D with probes as colored spheres pulsing in time with their local activation, connector lines linking each probe to its on-screen HUD trace (Figure 1). Folding is a performative gesture — adding or rotating a plane reshapes both what the audience sees and

what they hear, in a way the performer can anticipate because the folded body *is* the network. Sessions typically begin by selecting a seed and plane genome, letting the network settle, adjusting neighbor count to tune activity density, and then tuning the network's dynamics. The body constrains the search space, but can be instantly refolded with the dynamics held intact.

Demonstration

Five contrasting configurations illustrate the instrument's range and accompany this paper as supplementary video and audio.[1]

Ambient. Compact geometry with few probes and low activity, punctuated by occasional spike trains.

Expansive. Sonified folding and spike propagation across expansive geometry, with live tuning of the weight bias.

Folding. A folding sequence whose auditory behavior diversifies as the geometry grows more complex.

Spatial. Depth-driven filter modulation across a moving, expansive body, with small localized activity slowly cascading to the whole network.

Toroidal. A toroidal seed yields a sparse, knotted body; spike trains propagate slowly, producing a correspondingly sparse audio texture.

These configurations are presented as illustrative regimes rather than a formal evaluation. They are intended to demonstrate that the controls expose meaningfully distinct musical characters, and that those characters are reliably anticipated by the performer from the visible morphology.

Discussion

The system sits in a small but established ICCC neighborhood. Ianigro and Bown (2016) is the closest direct precedent: they treat continuous-time recurrent networks as creative substrates and search the parameter space for musically-affording configurations. Our move is to replace that search with a sculptable interface, exchanging exhaustive exploration for direct manipulation of a visible body whose folding is the parameter. Earlier ICCC and ICCC-adjacent work on evolutionary sonic ecosystems (McCormack 2001; Bown 2009) and embodied creative agents (Saunders *et al.* 2010) frames the broader tradition we draw on; Guckelsberger *et al.* (2021) survey the embodiment question for CC; we engage their argument from the reception vantage. The connection to morphological computation in robotics (Hauser *et al.* 2011) is more than rhetorical: a folded body whose connectivity computes is structurally the same idea expressed in a different medium.

Against the lineage of generative morphology proper — the L-systems descended from Lindenmayer (Prusinkiewicz and Lindenmayer 1990) and Bader and Oxman's recursive-symmetry method (Bader and Oxman 2016) from which we adopt the Ω operator — the contribution is to make the resulting body a dynamical object rather than a purely sculptural one. Where their pipelines produce a shape the viewer reads or fabricates, ours produces a shape that behaves, and whose behavior is sonifiable because both body and sonifier are indexed by the same geometry. The granular voice path is what lets this land texturally as well as structurally: a body with bilateral symmetry and visible fine detail produces sound with bilateral symmetry and matching textural detail.

Among recent ICCC demonstrations, the closest sibling is AutomaTone (Schaap and Hedblom 2024), which drives an interactive music generator from a 2D cellular automaton. The shared intuition is that a spatial process can be heard as well as seen; we differ by working in 3D morphology with a folded substrate and a spiking dynamical layer rather than a CA update rule.

Why symmetry. Because every downstream stage of the pipeline is geometry-indexed — the kNN graph, probe placement, per-point color, per-point pitch scalar, and grain neighborhood are all functions of position — the bilateral symmetry produced by Ω is preserved end-to-end by construction. Twin points (related by reflection across the final plane) therefore share color, share pitch in prune mode, and scrub the same source-buffer neighborhood in granular mode. The body's audible behavior carries the same symmetry as the visible body as a property of the pipeline rather than as an emergent perceptual claim. Whether listeners exploit this structural symmetry as a perceptual scaffold is an open question we have not measured.

Future work. The compact, deterministic genome is built to support neuroevolutionary search: the parameter vector has the structure that indirect encodings such as Compositional Pattern Producing Networks (CPPNs) and HyperNEAT (Stanley 2007) exploit, and an evolutionary fitness landscape over musically- or visually-defined objectives is a natural extension of the sculptable-body interface presented here. We are also exploring geodesic rather than Euclidean radii for probe readout, so that fold seams act as natural separations between probe domains.

Author Contributions

The author is solely responsible for all aspects of this work.

[1] `https://the.strange.agency/projects/folded-bodies`

References

Bader, C., and Oxman, N. 2016. Recursive symmetries for geometrically complex and materially heterogeneous additive manufacturing. *Computer-Aided Design* 81:39–47.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.

Bown, O. 2009. Ecosystem models for real-time generative music: A methodology and framework. In *Proceedings of the International Computer Music Conference*.

Clough, J., and Douthett, J. 1991. Maximally even sets. *Journal of Music Theory* 35(1):93–173.

Guckelsberger, C.; Kantosalo, A.; Negrete-Yankelevich, S.; and Takala, T. 2021. Embodiment and computational creativity. In *Proceedings of the 12th International Conference on Computational Creativity (ICCC)*.

Hauser, H.; Ijspeert, A. J.; Füchslin, R. M.; Pfeifer, R.; and Maass, W. 2011. Towards a theoretical foundation for morphological computation with compliant bodies. *Biological Cybernetics* 105(5–6):355–370.

Ianigro, S., and Bown, O. 2016. Investigating the musical affordances of continuous time recurrent neural networks. In *Proceedings of the 7th International Conference on Computational Creativity (ICCC)*.

Jaeger, H., and Haas, H. 2004. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* 304(5667):78–80.

Jordà, S. 2005. *Digital Lutherie: Crafting Musical Computers for New Musics' Performance and Improvisation*. Ph.D. dissertation, Universitat Pompeu Fabra.

Kantosalo, A., and Takala, T. 2020. Five C's for human–computer co-creativity: An update on classical creativity perspectives. In *Proceedings of the 11th International Conference on Computational Creativity (ICCC)*, 17–24.

Kiefer, C. 2014. Musical instrument mapping design with echo state networks. In *Proceedings of NIME 2014*.

McCormack, J. 2001. Eden: An evolutionary sonic ecosystem. In *Advances in Artificial Life*, 133–142. Springer.

Nakajima, K., and Fischer, I., eds. 2021. *Reservoir Computing: Theory, Physical Implementations, and Applications*. Springer.

Pasquier, P.; Eigenfeldt, A.; Bown, O.; and Dubnov, S. 2016. An introduction to musical metacreation. *Computers in Entertainment* 14(2): Article 2.

Prusinkiewicz, P., and Lindenmayer, A. 1990. *The Algorithmic Beauty of Plants*. Springer-Verlag.

Saunders, R.; Gemeinboeck, P.; Lombard, A.; Bourke, D.; and Kocaballi, A. B. 2010. Curious Whispers: An embodied artificial creative system. In *Proceedings of the 1st International Conference on Computational Creativity (ICCC)*.

Schaap, G., and Hedblom, M. M. 2024. Discussing the creativity of AutomaTone: An interactive music generator based on Conway's Game of Life. In *Proceedings of ICCC 2024*.

Stanley, K. O. 2007. Compositional pattern producing networks: A novel abstraction of development. *Genetic Programming and Evolvable Machines* 8(2):131–162.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

A Implementation Details

A.1 Folding operator parameterization

Plane positions are specified by a short trigonometric sequence (four frequency and four phase parameters, plus base and step scalars), so the full sequence is reconstructible from roughly ten named floats. A random plane mode is also supported, parameterized by a single seed integer. After folding, the cloud is rotated so the final plane's normal aligns with screen-right, giving a stable canonical orientation across configurations.

A.2 Network construction parameters

The kNN graph uses typical settings of $k = 8$ and a radius cutoff of 0.4 in normalized coordinates, constructed via a uniform spatial-hash grid. Each neuron also maintains an exponentially decaying "glow" trace used by the visualization layer. A small kickstart initially stimulates a random subset of neurons to seed activity.

A.3 Probe placement algorithm

The farthest-point sampling proceeds as follows: the first probe is the point farthest from the centroid; each subsequent probe is the argmax of the distance-to-nearest-source field after relaxing Dijkstra from all picked sources. Probe slot indices are stable across graph rebuilds, so each probe is a continuous output channel that the chart trace, the connector line, and the audio voice all key off through structural changes.

A.4 Sonifier signal chains and mappings

The oscillator path is structured as: square oscillator $\rightarrow$ low-pass $\rightarrow$ VCA $\rightarrow$ panner. The granular path is: density-driven grain scheduler with position, size, and pitch jitter $\rightarrow$ low-pass $\rightarrow$ VCA $\rightarrow$ panner. Both paths are summed into a shared bus passing through a compressor and reverb. Probe activation drives voice gain through a soft expander for contrast; camera-space distance modulates filter cutoff; each probe's normalized-device-coordinate horizontal position drives stereo pan. The performer's controls are organized into five panels (generation, visualization, network, probes, and audio) exposing seed shape, plane sequence, graph parameters, network constants, voice and pitch modes, pitch-palette size n, and granular source buffer and grain controls.

A.5 Pitch logic

Slot mode maps voice i to a pitch drawn from a maximally even set distribution (Clough and Douthett 1991) over the chromatic octave: a single integer $n \in \{1, \ldots, 12\}$ selects the size of the active set, sliding voices from sparse to fully chromatic while preserving optimal even spacing.

Prune mode uses a per-point scalar computed at seed time from the azimuthal angle $\mathrm{atan2}(z, x)/2\pi$, the same scalar that drives the body's hue. The scalar is propagated unmodified through Ω, deduplication, and capping, so reflection is hue-invariant by construction and points related by symmetry share both color and pitch scalar in any quantization. For oscillator voices the scalar selects a pitch; for granular voices it indexes a position in the source buffer instead, so the same geometric address drives different output spaces depending on voice path.

Anchor-Conditioned Compositional Control for Landscape Image Generation

Gadha Lekshmi P,[1] **Govind Arun,**[2] **Rohith Syam,**[3] **Ahmed Elgammal**[1]
[1]Rutgers University–New Brunswick, USA [2]University of Maryland–College Park, USA [3]University of Technology Sydney, Australia
gadhalekshmip@gmail.com, govind123.ga@gmail.com, rohithlayana@gmail.com, elgammal@cs.rutgers.edu

Abstract

Image generative models, though widely used as creative tools, offer limited support for the kind of compositional control that photographers and visual artists routinely exercise. This paper presents early results on an anchor-conditioned finetuning framework for landscape image generation, in which a four dimensional compositional anchor vector is extracted from training images and injected into a diffusion model via a decoupled cross-attention mechanism with Fourier encoding and three-way classifier-free guidance dropout. Quantitative evaluation against a baseline and three ablation variants shows that the proposed architecture achieves the highest horizon detection rate (0.850) and the highest rule-of-thirds alignment (0.817). A category specific ablation further demonstrates that training on compositionally homogeneous scene subsets reduces horizon deviation by up to 40% compared to mixed training. This establishes that compositional control precision is category-dependent.

Introduction

Image generative models, while introduced as creative tools, lack controllability over composition, a capability that is fundamental to the creative process. A photographer composing a landscape does not just specify content. They decide where the horizon sits, how the frame is divided, and where the viewer's eye should land. Diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) make none of these decisions explicitly. They emerge from statistical patterns absorbed during pretraining, giving the photographer no reliable handle on the spatial outcome.

The goal of this paper is to introduce a mechanism by which compositional intent can be specified, injected, and measured in a finetuned diffusion model for landscape image generation. A four dimensional anchor vector encoding horizon position, detection confidence, average saliency, and foreground ratio is automatically extracted from each training image using Hough-transform horizon detection and spectral residual saliency analysis, requiring no manual annotation. This vector is injected into a finetuned diffusion model through a dedicated attention pathway, so that a target horizon position and compositional weighting can be specified at inference time without manual spatial masks or structural conditioning inputs.

The research demonstrates early results on this mechanism, evaluated using horizon placement accuracy and rule-of-thirds alignment as measurable proxies for compositional control. A category specific ablation investigates whether training on compositionally homogeneous scene subsets improves conditioning precision beyond what mixed training achieves.

The remainder of this paper covers related work on spatial conditioning in diffusion models, followed by the methodology and experimental setup, quantitative results across four model configurations and three landscape categories, and a discussion of findings and planned extensions.

Background

Prior work has approached spatial control through structural conditioning inputs. ControlNet (Zhang, Rao, and Agrawala 2023) introduces a parallel trainable encoder that accepts depth maps and edge maps, demonstrating that a frozen pretrained model can accommodate new spatial conditioning without losing generative quality. GLIGEN (Li et al. 2023) extends this to open-set grounded generation using bounding boxes and keypoints, and provides the strongest empirical support for Fourier encoding of spatial coordinates, showing 6.8 times better localisation control compared to raw MLP encoding of the same values. Approaches such as SDXL (Podell et al. 2024) address high-resolution generation through architectural scaling and resolution conditioning, but do not provide mechanisms for specifying or preserving compositional intent across pipeline stages. IP-Adapter (Ye et al. 2023) introduced the decoupled cross-attention pattern that this architecture directly adopts, where the conditioning signal has its own Key and Value projections rather than competing with text tokens for the available attention budget. This work applies these ideas to the specific problem of compositional intent in landscape image generation, extending the decoupled adapter pattern from image-based conditioning to compact scalar compositional descriptors. Recent work such as Bokeh Diffusion (2025) has similarly shown that scalar photographic parameters can be injected through lightweight cross-attention adapters, though none of these efforts address compositional layout descriptors or their preservation across generation stages.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Methodology & Experimental Setup

Dataset

Training was conducted using the Cropped-1901 Landscape Dataset (Studio 2023), which consists of 4,713 images curated from multiple publicly available sources, including a Flickr image dataset (Hsankesara 2018), a landscape image collection from Kaggle (Arnaud58 2023), the Nature dataset hosted on Hugging Face (Cobanov 2023), and the HQ-50K dataset (YangQiee 2023). The dataset was standardised to a 1.90:1 aspect ratio at a resolution of 1949×1024 pixels, and for training, images were resized to 768×400 pixels to maintain compatibility with the base model's latent representation. Existing captions were wrapped with a photorealistic prefix and suffix to anchor the model to the photographic domain. The dataset was split using a 90:10 ratio, resulting in 4,241 training images and 472 validation images.

Compositional Anchor Vector

Before the training phase, each image in the dataset is analysed to produce a four dimensional anchor vector representing its compositional structure. The horizon position (horizon_y) is estimated using a Hough line transform applied to a horizontal-line-biased Canny edge map of the upper 60% of the image, normalised to $[0, 1]$ where 0 represents the top edge. The horizon confidence (horizon_conf) is derived from the fractional coverage of the longest detected horizontal line relative to image width. The average saliency (avg_saliency) is computed via spectral residual analysis across the top detected salient regions. The foreground ratio (fg_ratio) provides a coarse estimate of visually prominent foreground content. When no horizon is detected, horizon_y defaults to 0.5 and horizon_conf to 0.0, preserving the informativeness of the remaining two components. Anchor vectors are pre-extracted for the entire dataset in a single offline pass and cached to disk.

Pipeline

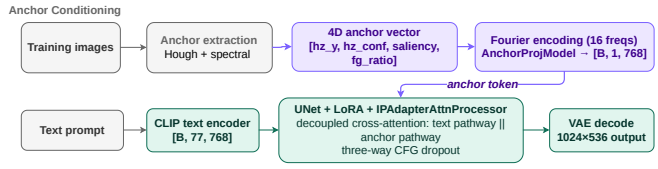


Figure 1: Anchor-conditioned generation pipeline.

Figure 1 shows the anchor conditioned generation pipeline. The extracted anchor vector is injected into the diffusion model through a decoupled cross attention pathway.

The architecture injects compositional anchor information into a fine tuned latent diffusion model (Rombach et al. 2022) through a dedicated parallel attention pathway that operates independently of the text conditioning stream. This design is motivated by the finding that injecting an anchor token directly into the text sequence gives it insufficient influence to affect spatial composition: a single token representing 1.3% of the total attention budget cannot compete with 77 deeply pretrained text tokens, as confirmed by the concatenation ablation in Section Results.

Fourier Encoding: The 4D anchor vector is first passed through a Fourier embedding layer applying 16 log-spaced frequencies from 2^0 to 2^{15} via sine and cosine transformations, producing a 132-dimensional representation. This encoding enables learning of fine-grained spatial relationships from scalar inputs, motivated directly by the GLIGEN ablation result (Li et al. 2023).

AnchorProjModel: The 132-dimensional Fourier representation is projected through a network consisting of a Linear layer (132 to 256), GELU activation, LayerNorm, a second Linear layer (256 to 768), and a final LayerNorm, producing a single anchor token of shape $[B, 1, 768]$. This network contributes approximately 233,000 parameters and operates in float32 precision throughout.

Decoupled Cross-Attention: Separate Key and Value projection matrices are introduced at each of the 16 cross-attention layers in the UNet using a low-rank decomposition of rank 8, totalling approximately 396,000 parameters. The output at each layer is:

$$\text{out} = \text{Attn}(Q, K_{\text{text}}, V_{\text{text}}) + \lambda \cdot \text{Attn}(Q, K_{\text{anchor}}, V_{\text{anchor}}) \quad (1)$$

where Q is shared between both pathways. The up-projection matrices are initialised from the pretrained UNet's text K/V weights via SVD decomposition, initialised to match the existing cross-attention subspace. At inference, three UNet forward passes are computed per denoising step:

$$\hat{\epsilon} = \epsilon_{\varnothing} + \lambda_{\text{text}}(\epsilon_{\text{text}} - \epsilon_{\varnothing}) + \lambda_{\text{anchor}}(\epsilon_{\text{full}} - \epsilon_{\text{text}}) \quad (2)$$

Training: During training, a three-way dropout scheme is applied following classifier-free guidance (Ho and Salimans 2021): with 5% probability both text and anchor are dropped, with 5% probability only text is dropped, with 5% probability only the anchor is dropped, and with 85% probability both are retained. The model was trained on landscape images at 768x400 resolution for 10,000 steps with a learning rate of 1×10^{-4}, cosine decay with 500 warmup steps, AdamW optimisation with weight decay 0.01, and gradient clipping at norm 1.0. LoRA adapters of rank 16 with alpha 16 are applied to the four attention projection matrices. Mixed precision was BF16.

Implementation Details

Experiments were conducted on Google Colab using an NVIDIA A100 GPU with 40 GB of VRAM for training, ablation experiments, and evaluation. The software stack used PyTorch 2.x, Hugging Face Diffusers, PEFT for LoRA management, and safetensors for weight serialisation. Anchor extraction used OpenCV's Hough line transform for horizon detection and spectral residual saliency for salient region identification. OpenCLIP ViT-B/32 (Cherti et al. 2023)was

[0]Code available at: https://github.com/gadhalekshmip/Anchor-Conditioned-Compositional-Control-

used for CLIP score computation. Table 1 details the AnchorProjModel architecture.

Table 1: AnchorProjModel architecture. Input is the 132-dimensional Fourier-encoded anchor; output is a single token of shape $[B, 1, 768]$ injected into the decoupled cross-attention pathway.

Layer	In	Out
FourierEmbedding (16 freqs)	4	132
Linear	132	256
GELU + LayerNorm	256	256
Linear	256	768
LayerNorm	768	768
Total parameters	233,088	

Results

Training

Validation was performed every 500 steps using four fixed prompts with a target anchor of [0.333, 0.9, 0.3, 0.15]. Horizon deviation from the rule-of-thirds target (0.333) was measured at each checkpoint. Table 2 reports validation metrics across checkpoints

Table 2: Validation metrics at each training checkpoint. Detection rate is the fraction of validation prompts where a horizon was successfully extracted from the generated image.

Step	Det.	Mean Hz	Std	Mean Dev.
500	100%	0.465	0.148	0.191
2,000	75%	0.302	0.002	0.031
5,000	100%	0.341	0.091	0.074
7,000	100%	0.305	0.008	0.028
10,000	100%	0.300	0.003	0.033

Mean deviation improved from 0.191 at step 500 to 0.028 at step 7,000, representing approximately a 7-fold improvement in compositional precision. The model stabilised around a horizon position of 0.300 from step 3,000 onwards, confirming that the anchor conditioning converges cleanly. Figure 2 shows the deviation curve across training. The transient drop in detection rate at step 2,000 reflects a mid-training transition phase and does not persist.

Compositional Control Evaluation

Four model configurations were evaluated on 10 prompts across 6 landscape categories (mountain, ocean, forest, desert, arctic, canyon) with 6 random seeds per prompt, yielding 60 images per model at 1024x536 resolution. The configurations are: an unmodified baseline with no finetuning; a LoRA-only finetune with no anchor conditioning; a concatenation-based anchor variant appending the anchor token directly to the 77-token text sequence; and the proposed decoupled cross attention architecture. Here, horizon deviation is measured across diverse prompt categories

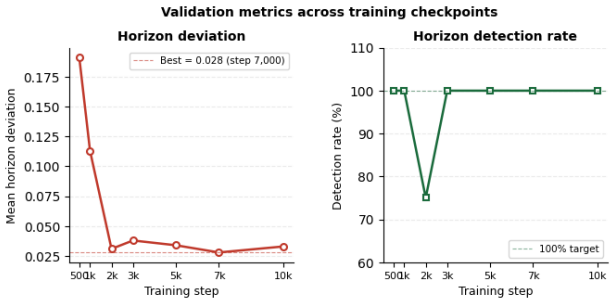


Figure 2: Validation horizon deviation (left) and detection rate (right) across training steps. Deviation improves from 0.191 at step 500 to 0.028 at step 7,000 and then stabilises

rather than the four fixed validation prompts used during training, which accounts for the difference between training-time deviation and the values reported below.

Table 3: Four-model comparative evaluation. All metrics are averaged over 60 images per model.

Metric	Baseline	LoRA	Concat	Proposed
Detection Rate (↑)	0.766	0.816	0.766	**0.850**
Horizon Dev. (↓)	0.073	0.077	0.073	**0.062**
Rule-of-Thirds (↑)	0.500	0.636	0.500	**0.817**
Sharpness (↑)	2171	1078	2171	**4606**
CLIP Score (↑)	0.294	0.280	0.294	**0.297**
NIQE (↓)	8.79	8.51	8.79	**7.44**

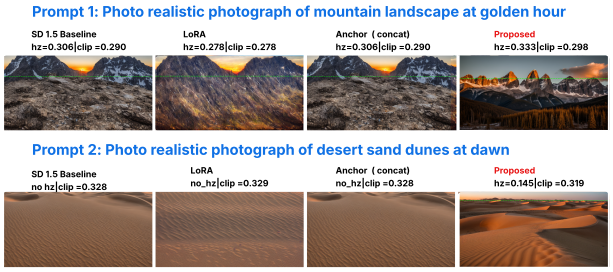


Figure 3: Four-model visual comparison for two representative prompts. *Top row:* mountain landscape, generated with target horizon_y = 0.333. The proposed model achieves exact rule-of-thirds placement *Bottom row:* desert sand dunes, where rest of the models fail to produce a detectable horizon; the proposed model successfully places the horizon, consistent with a high-sky compositional target.

The proposed architecture achieves the best performance across all primary compositional metrics. Table 3 and Figure 3 show the full comparison. The concatenation variant scores identically to the unmodified baseline on every metric, confirming that a single appended anchor token is completely ignored by the model: it represents only 1.3% of

the total attention budget and cannot influence spatial layout when competing with 77 pretrained text tokens. The LoRA-only model moderately improves horizon detection, reflecting that domain finetuning absorbs some implicit compositional patterns, but its horizon deviation remains marginally worse than baseline, indicating these patterns do not generalise to precise compositional control. The proposed architecture's horizon deviation of 0.062 represents a 15% reduction in compositional error relative to baseline, and its rule-of-thirds alignment of 0.817 is 63% higher, indicating substantially more frequent placement of salient regions near compositional power points.

Category-Specific Ablation

To investigate whether compositional control is category-dependent, the 4,713 training images were classified into 3 scene categories using keyword matching on existing dataset captions followed by CLIP zero-shot verification for ambiguous cases. Three categories: mountain (1,534 images), forest (1,149 images), and desert (1,042 images). Each category specific model used the identical architecture and training configuration as the proposed model, with step budget scaled proportionally to subset size (1,301 steps). Each model was evaluated on 100 images generated from 10 category-matched prompts with 10 random seeds, ensuring the evaluation prompts align with the training domain of each model. Table 4 and Figure 4 report the per-category results.

Table 4: All category-specific models share identical architecture and hyperparameters with the proposed model. Best step is selected by lowest validation horizon deviation.

Model	Images	Steps	Det. (↑)	Hz Dev. (↓)
Proposed (mixed)	4,713	10,000	0.833	0.048
Mountain	1,534	1,301	1.000	0.033
Forest	1,149	1,301	1.000	0.029
Desert	1,042	1,301	1.000	0.033

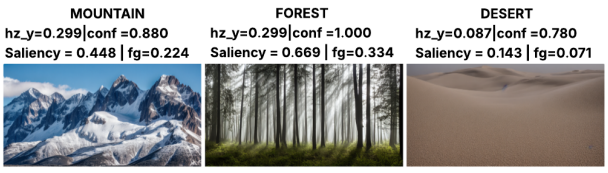


Figure 4: Sample outputs from the three category specific models (mountain, forest, desert), each annotated with the anchor values extracted from the generated image.

All three category specific models achieve 100% horizon detection and horizon deviation substantially below both the unmodified baseline and the mixed training model. The forest-specific model achieves the lowest deviation of 0.029, representing a 60% improvement over baseline and a 40% improvement over the proposed mixed model. Desert training produces the most stable convergence, with deviation constant at 0.033 across both checkpoints at step 500 and step 1,000, indicating the model has cleanly learned the desert compositional prior without overfitting. The consistent improvement across all three categories, each trained with identical architecture and hyperparameters, confirms that category specific training reliably improves anchor conditioning precision.

Discussion and Conclusion

The four model evaluation shows that decoupled cross attention is necessary for effective anchor conditioning. The concatenation ablation, which scores identically to the unmodified baseline on every metric, establishes that adding an anchor token to the text sequence produces no measurable effect. The proposed architecture resolves this by giving the anchor its own Key and Value projections (Ye et al. 2023), conditioning layout through a separate pathway with no competition for attention capacity.

The category-specific ablation reveals a finding with practical implications for finetuning strategy. Training on a compositionally homogeneous subset consistently outperforms training on the full mixed dataset with the same architecture and fewer steps, confirming that category-specific training reliably tightens conditioning precision when the intended application domain is known in advance.

Future work. These are early results on an approach we intend to develop further. The current evaluation relies on automatic metrics that proxy compositional quality through horizon placement and rule-of-thirds alignment. A user study with photographers and visual artists is the natural next step, both to validate these metrics against human judgment and to evaluate anchor conditioning as a creative interface in practice. Beyond user evaluation, we plan to extend the framework in two directions: a complementary anchor component based on object detection, where a bounding-box centroid encoded through the same Fourier pathway would condition the model on subject placement and enable compositional control at both the scene and subject levels; and high-resolution generation, where preserving the anchor across upscaling stages requires a refinement model trained to treat the anchor as a geometric constraint, with a wavelet-domain objective separating fidelity loss on low-frequency subbands from texture loss on high-frequency subbands (Korkmaz, Tekalp, and Dogan 2024).

Limitations. The anchor vector is designed for landscape imagery where composition is governed primarily by the horizon and the sky-to-ground relationship. Portrait, architectural, and product photography are shaped by different spatial conventions and would require category-specific anchor definitions to benefit from the same framework. The automatic metrics used here also do not capture the full subjective experience of a creative practitioner working with the system, which is why the planned user evaluation is central to future iterations of this work.

Taken together, these results show that a four dimensional compositional descriptor can condition spatial layout when given a dedicated pathway, and that category specific training improves precision over mixed training (Govind et al. 2024).

References

Arnaud58. 2023. Landscape pictures dataset. Kaggle. https://www.kaggle.com/datasets/arnaud58/landscape-pictures.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2829. IEEE.

Cobanov, M. 2023. Nature dataset. Hugging Face. https://huggingface.co/datasets/mertcobanov/nature-dataset.

Fortes, A.; Wei, T.; Zhou, S.; and Pan, X. 2025. Bokeh diffusion: Defocus blur control in text-to-image diffusion models. In *Association for Computing Machinery*.

Govind, A.; Anzar, A.; Nair, A. A.; and Syam, R. 2024. Genai empowered script to storyboard generator. In *2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, 451–456. IEEE.

Ho, J., and Salimans, T. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, 6840–6851.

Hsankesara. 2018. Flickr image dataset. Kaggle. https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset.

Korkmaz, C.; Tekalp, A. M.; and Dogan, Z. 2024. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5926–5936.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22511–22521.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Muller, J.; Penna, J.; and Rombach, R. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Studio, D. 2023. Cropped-1901 landscape dataset. Kaggle. https://www.kaggle.com/datasets/dxtinctionstudio/cropped-1901.

YangQiee. 2023. Hq-50k dataset. Hugging Face. https://huggingface.co/datasets/YangQiee/HQ-50K.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *ArXiv* abs/2308.06721.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3813–3824.

Transformation Is All You Need

Sean Luke
Department of Computer Science
George Mason University
Washington, DC, USA
sean@cs.gmu.edu

Matthew Cronin
Costello College of Business
George Mason University
Washington, DC, USA
mcronin@gmu.edu

Abstract

We argue that the computational creativity field has had a minor impact on AI, in large part because of its emphasis on exploratory creativity, which maps primarily to already well understood algorithms in search and optimization. We argue that although it is relatively less studied, transformational creativity may yet instead provide a richer vein for inspiration of AI algorithms due to its meta-nature and the opportunities it affords in combination with multiagent systems. We offer a new framework for multiagent transformational creativity, and discuss possible application to AI fields.

Introduction

A major motivation for the field of computational creativity is its potential to give rise to useful AI algorithms that are in some sense artificially creative. But "creative" is a vague and slippery concept with many weak or ill-considered definitions over the years. When it is applied to AI, computational creativity lacks a concrete or rigorously quantifiable task or goal. Measurements of computationally creative processes or products are fraught with bias, context sensitivity, and disagreement as to what constitutes a creative product (Colton and Wiggins 2012). Compare this to machine learning, which has a specific task: given examples drawn from the world, produce a generalizable model which can perform discriminative or generative operations; and this generalization is straightforwardly measurable with validation examples.

Lacking a specific task for computational creativity, we can still attempt to understand the *process* of human creativity to see if this understanding can *inspire* us to develop useful artificially intelligent algorithms for whatever purpose. This is similar to historical efforts in evolutionary computation, neural networks, computer vision, soft computing, and swarm robotics, which all borrowed liberally from biology or cognitive science to inspire competitive AI algorithms. However in our opinion the output from the computational creativity community has not contributed very much new to AI, resulting in a multi-decade research effort by this community with little to show for it. Algorithms deemed "creative" after the fact have come from fields such as deep learning: but not often from computational creativity.

We think that this is because the study of the creative process, at least in the computational creativity literature, has been dominated by the wrong thing: exploring of the space of novel and valuable concepts, that is, *exploratory creativity*. Models of exploratory creativity largely boil down to straightforward search, optimization, and iterative revision, and these topics have been well studied in engineering and AI for over a century. We think that not only has exploratory creativity not offered much new to AI, it is not really a useful notion of *creativity in humans* either: it's just exploration.

But there is a contending notion of creativity whose process is well justified. Humans have limited working memory, and so we get stuck in local optima in the value and novelty functions of exploratory creativity's "concept space". This is manifested in writer's block, for example, or getting stuck on a problem. We also may be blinded by our biases as to the existence of certain fruitful avenues to explore. We can can overcome these problems by looking for a new angle on the task, or a new way to organize the space, to bring formerly distant concepts into view. That is, we engage in what Boden (1992) calls *transformational creativity*.

In his 2000 AAAI Presidential Address, Bruce Buchanan (2001) argued for the importance of transformational creativity to AI, claiming "that the key to creativity is at the metalevel." Transformational creativity may be modeled as a meta-exploration through the space of possible topologies which organize the concept space: it changes how likely we are to reach concepts, and thus jumbles up the local optima, revealing new pathways to make progress. This is often mistakenly thought of as a phenomenon of high intelligence and with global impact, but in fact it happens all the time in everyday life. Meta-level search, optimization, and representation have been studied in AI only to a limited degree, and so this opens avenues of study. And transformational creativity is potent in combination with *multiagent* scenarios, where multiple creators spread to one another not only news of their creations, but *new ways to think about the problem itself.*

For these reasons, we think that transformational creativity deserves to be the primary research thrust of computational creativity. And yet it has seems to take a back seat in the literature. Out of the 755 papers from the 2010–2025 ICCC conferences, the word *transformational* appeared in 79 papers, and transformational creativity was discussed in a meaningful (non-passing) manner in only 37. The proportion of the discussion in recent ICCC conferences has dropped by half relative to the earlier ones.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Previous Work

Creativity has been studied since Greek antiquity, and informal models of creativity have been proposed at least since von Helmholtz (1896) and Wallas (1926), but the seminal study of *computational* creativity may have been by Newell, Shaw, and Simon (1958). They noted that according to the psychological literature of the time, "creative thinking" was characterized by at least one of four features, including "the product of the thinking has novelty and value", and "the thinking is unconventional, in the sense that it requires modification or rejection of previously-accepted ideas." Novelty, value, and (to a lesser extent) unconventionality have since come to dominate later definitions of creativity and computational creativity (Sternberg and Lubart 1999).

Margaret Boden's theory of the process of creativity continued along these lines (Boden 1992). Boden proposed three major kinds of creativity: *exploratory, combinatorial,* and *transformational*. In exploratory creativity, the creator is exploring the space of *concepts* (potential, perhaps incomplete, creative artifacts), in search of ones which are *valuable* and *novel*. A *valuable* concept is one that someone cares about in some way or that satisfies some aesthetic or performance criterion, and a *novel* concept is one that has not been discovered before. Combinatorial creativity expands on exploratory creativity by defining creative acts as the combination of two concepts from different and often unrelated domains, perhaps drawn from different concept spaces.

Transformational creativity is touted as an "advanced" form of creativity, as the person's creative act *changes the topology, boundary, or nature of the concept space* and thus enables her to rapidly explore new areas in the space. Boden's oft quoted canonical example is Friedrich August Kekulé's discovery of the cyclic nature of benzene: before it, his (and others') concept space for molecules in chemistry was restricted to trees or to crystalline meshes, but afterwards, organic chemistry was forever changed to consider cycles.

Creativity is also studied in a multiagent context. Early work in this area was based on Csikszentmihalyi's Domain-Individual-Field Interaction (DIFI) model (Feldman, Csikszentmihalyi, and Gardner 1994). Here a creative *individual*, inspired by the existing historical body of work (the *domain*), produces new creative works which are then assessed for novelty and value by *the field*: critics, audiences, etc. If his works are deemed sufficiently creative, they are added to the domain and the process continues.

Semiformal Models of Computational Creativity Many computational models have stemmed from Boden (Wiggins 2006b; Ventura 2011; Ritchie 2012). These have largely built off of Wiggins's seminal work extending Boden's concept space into a kind of state-space to search (Wiggins 2006a). Collectively this body of work is known as *Creativity as Search*. Wiggins describes an agent engaged in exploratory creativity as traversing a concept space. He further defines a set of rules which constrains which concepts in this space are of interest to the agent (or *valid* (Wiggins 2006b)); a second set which describes how the space may be traversed (perhaps via heuristic search); and a third set which assesses the value of concepts in the space. Wiggins extends this model to include transformational creativity by defining two *meta-rulesets*: one which defines the space of validity rulesets, and one which defines the space of traversal rulesets. By selecting a new ruleset from the validity space, the agent would thus modify the boundaries defining the valid concepts to search. By selecting a new traversal ruleset, the agent would change the transition topology over the concept space itself.

Ritchie also developed a framework for transformational computational creativity (Ritchie 2005; 2006), which has since been extended to consider biases in search (Demke and Ventura 2024).

Other work has tried to combine Boden with Csikszentmihalyi or other multiagent models. Sauders (2019; 2001) describes an agent-based model where agents assess and provide feedback for other agents' creative works. Linkola (2019) redefine Wiggins's rulesets as ones concerning themselves with the personal concept space of the agent, then create global validity, traversal, and value rulesets corresponding to the personal ones, but which are shared collectively among agents at a society-wide level.

Why Transformational Creativity Is Important

As pointed out by Buchanan (2001), there is a critical difference between search guided by the assumptions embedded in a problem representation, and search for the right representation (the meta-level). This was originally called the distinction between understanding and solving a problem (Simon and Hayes 1976), but has also been described as problem finding versus problem solving (Mumford and Gustafson 1988), as exploration versus exploitation (in the psychological sense, not the optimization sense), or most recently as the distinction between perspective use and perspective change (Cronin and Loewenstein 2018). In all cases, one kind of search applies known search operators (exploratory creativity, combinatorial creativity), and the other kind of search seeks to change the very operators themselves.

Exploratory and combinatorial creativity discover new things that may expand current capabilities, but they do not fundamentally change the way one thinks about a topic, only the current actions one might take to solve it. Transformational creativity changes how one thinks about a topic, and as such opens up more possibilities for the topic itself. Its potential to change things in a significant and global way is why transformational creativity is often thought as "Big-C" (Kaufman and Beghetto 2009) creative or "disruptive" (Schumpeter 1942). But transformational creativity happens all the time in real life. We might transform the way students think about topics (e.g., "think of negotiation as problem solving") or the way trainees think about tactics (e.g., "hold a golf club like you are holding a bird"). These, too, alter the way people approach even rudimentary topics and thus open up the potential for advanced understanding and action.

We also argue that, from the perspective of an AI researcher, exploratory creativity maps largely to search or optimization with a diversity criterion, and this has already been long studied. Thus we do not think that the development of exploratory creativity-inspired algorithms will benefit AI much more than optimization and search already have. But transformational creativity maps to *meta-level* optimization,

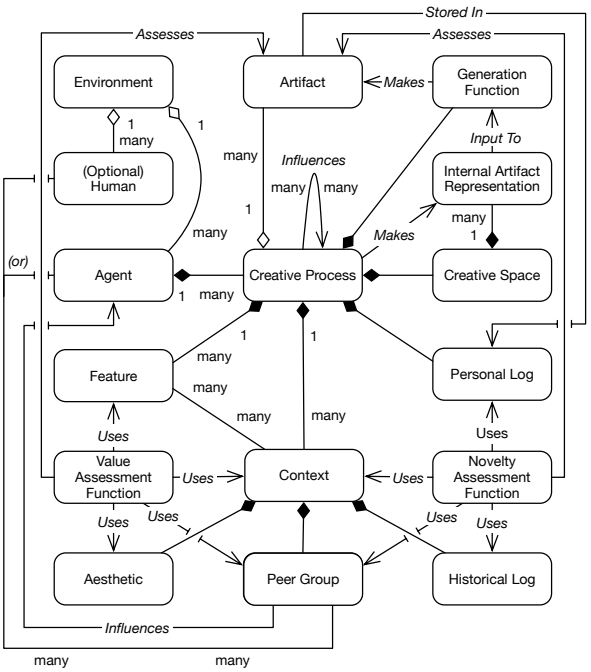


Figure 1: UML diagram of model classes and relationships, from Luke (2023).

and while there are a few examples of meta-level optimization in AI, particularly in stochastic optimization, it is still a poorly studied area rich with research opportunities where computational creativity could contribute.

Proposed Model

In prior work on exploratory creativity we criticized models which focused on search rather than on multiobjective optimization, and for not situating creativity in a dynamic and multiagent context (Luke 2023). We developed a model of exploratory creativity which connected these elements. Our model was similar to Wiggins in that it described a space of concepts, rules for value and novelty, and the notion of traversal: but we rejected the use of validity rules to constrain the space. The goal was to create a framework in order to define and compare computational creativity algorithms. We now extend this model to consider transformational creativity.

The original model defined an "exploratory" optimization space of artifacts called a *creative space*, the rough analog of Boden's concept space. We assume that this space has a *topology*, or neighborhood function, approximately defining the probability that some new prospective artifact r' would be discovered given past discovered artifacts $\langle r_1, ...r_n \rangle$. Note that this is different from the distance function used to assess novelty in Luke (2023). There are many things that could affect this probability, such as similarity, but also whether r' is just too exotic. Thus we can do away with a Wiggins-style constraint or validity function: for example, benzene was not *invalid* to Kekulé, but just too unusual a solution for him to likely see. Also note that while an agent could "transform" a space by changing the exploratory value function rather than

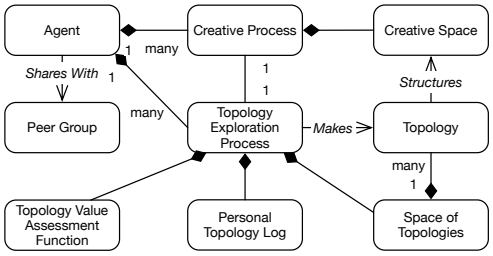


Figure 2: UML diagram of model extension to accommodate multiagent transformational creativity.

the topology, this would be more akin to formerly *rejecting* ideas rather than simply not *realizing* them due to biases.

We augment this model by giving each agent a meta-level "transformational" space: the space of topologies. The agent can, when appropriate, search this space to attempt to discover a new, high-quality topology, that is, one which encourages the exploratory optimizer to produce valuable and novel artifacts. As with Luke (2023), we define this space to be an optimization space, with its own neighborhood function, rather than as a state-space to search as in Demke and Ventura (2024). Wiggins (2006a) calls this just "exploratory creativity at the meta-level," but we note that this meta-level optimization process is likely much slower than exploratory-level optimization; and also the objective function is likely not the same. Perhaps a meta-level objective function would assess the quality of a topology as a maximum or weighted mean value over artifacts created so far in the exploratory space applying this topology. Or would it consider novelty?

Multiagent Transformational Creativity In our original model, an agent could receive discovered artifacts from a *peer group* of other agents or humans, which informed his sense of novelty and also potentially seeded the agent's exploratory optimization function. Conversely discoveries by an agent could spread to his peers, and their peers, inspiring them with new possible concepts or artifacts to consider.

But transformational creativity offers a very different additional seeding opportunity: agents may communicate *topologies* (in the form of problem representations, ways of viewing the problem, changes in bias) to one another. As an agent discovers a new and fruitful way of conceptualizing the problem, he can spread this to others. We model this as agents communicating topologies at the meta-level, thus seeding agents' meta-level topological optimization processes.

Model Details

First we introduce certain necessary concepts from our original model in (2023), but omit many of its other details. We use the nomenclature from the original model. The model describes an *environment* E consisting of one or more *agents*. The state of the various agents at time t may be described as $\mathbf{A}^t = \{A_1^t, ..., A_a^t, ...\}$. At time t each agent A_a^t employs a set of one or more *creative processes* $\mathbf{P}^{a,t} = \{P_1^{a,t}, ..., P_p^{a,t}, ...\} \subseteq \mathbb{P}$, which are optimization algorithms attempting to produce high-quality, novel *artifacts*.

The space of artifacts over which some optimization process $P_p^{a,t}$ operates is called a *creative space* $S^{a,p,t} \in \mathbb{S}$, and effectively serves as Boden's Concept Space for exploratory creativity. The creative process $P_p^{a,t}$ maintains $L^{a,p,t}$, a *personal log* of artifacts it has produced or is aware of up until time t. An agent shares concepts with other agents $\mathbf{A}^t$, and possibly with other humans $\mathbf{M}^t = \{M_1^t, ..., M_m^t, ...\} \subseteq \mathbb{M}$. The group with which it shares concepts is called its *peer group* $G^{a,p,c,t} \subseteq (\mathbf{A}^t \vee \mathbf{M}^t)$. This peer group is part of the *context* $\mathbf{C}^{a,p,t} = \{C_1^{a,p,t}, ..., C_c^{a,p,t}, ...\} \subseteq \mathbb{C}$ in which artifacts are assessed. Another part of the context is an *aesthetic* $Z^{a,p,c,t}$ which encapsulates all information used to assess the value of an artifact. Contexts and their aesthetics may change over time as an agent develops new opinions about artifacts.

Creative Space Topologies We augment the model by defining the notion of a space of *neighborhood topologies* $\{T_1, ..., T_i, ...\} \subseteq \mathbb{T}$. At time t a creative space $S^{a,p,t}$ is currently applying some topology T_i. The topology defines the neighborhood distribution of any given artifact in the concept space. A topology T_i might be modeled as a distribution $P(r|L^{a,p,t}), r \in S^{a,p,t}$ indicating the probability that artifact r would be discovered given the concepts in $L^{a,p,t}$.

The function *Quality*$(T_i, S^{a,p,t}, L^{a,p,t}) \rightarrow \mathbb{R}$ provides a quality assessment of the topology T_i as it has performed so far, based on the value and novelty of concepts in $L^{a,p,t}$ discovered using that topology. If need be, *Quality* could instead be multiobjective.

Topology Exploration For each creative process $P_p^{a,t}$ the agent maintains a parallel meta-level *topology exploration process* $P_p'^{a,t}$. This process explores through T, starting at T_i, until it finds some new topology T_j. It does this by considering T_i and its current *Quality*, plus all the topologies stored in the process's *log*, $L'^{a,p,t}$, of previously used topologies, along with their *Quality* assessments.

An agent is normally engaged in its primary exploration process: executing $P_p^{a,t}$ to find high-quality and novel artifacts. When the agent feels it is stuck, or has exhausted the space, or that it cannot make further useful progress, it may engage the meta-exploration process $P_p'^{a,t}$ to produce the new topology T_j. It then replaces T_i with T_j as the topology for $S^{a,p,t}$, adds T_j to $L'^{a,p,t}$, and continues executing $P_p^{a,t}$.

Topology and Aesthetic Sharing In Luke (2023) agents may *influence*, or be influenced by, other agents in their peer group by sharing artifacts with them. These new artifacts, and their value and novelty, can bias an agent's creative process.

In this revised model, agents may additionally influence other agents by sharing new *topologies*. If a topology T_i has proven fruitful to explore, an agent may share it with other members of its peer group $G^{a,p,c,t}$. Any other agent $A_{a'}^t \in G^{a,p,c,t}$ may then choose to apply T_i to its own creative space $S^{a',p',t}$, assuming that p' is a creative process similar to p.

Finally, we amend the revised model to allow an agent to share an aesthetic $Z^{a,p,c,t}$ with its peer group $G^{a,p,c,t}$, and to revise it based on ones shared from members of the group, thus influencing how the agent assesses the value of artifacts.

Discussion

Our purpose in developing this framework is to serve as a first step in developing useful algorithms inspired by transformational creativity, which we hope may ultimately prove more fruitful than exploratory creativity has. Classically many areas of artificial intelligence have drawn from processes in the sciences for inspiration in the development of useful algorithms and models. But what about computational creativity? Beyond basic models of creativity, there exist some actual algorithms which borrow from exploratory creativity principles to engage in iterative refinement (for example, Venkatesh et al (2025)). But we are struck with the paucity of such algorithms as opposed to, for example, the fields of evolutionary computation (EC) or neural networks.

Before we examine where transformational creativity might make an impact in AI, we first must acknowledge the possibility that computational creativity principles could in fact largely *be useless* in the development of AI algorithms. If creativity is primarily a tool to get around the restricted ability of humans to explore concept spaces due to our biases and limited working memory, and if computers *do not have these limitations nearly to the degree that humans do*, then the creative process may not be very useful to AI algorithms.

Even so we might ask if there are places where approaches inspired by transformational creativity may prove useful: for example, in overcoming limitations caused by restricted distributed communication in limited memory, in multiagent systems, in dynamically changing value functions, in environments with many local optima, or in meta-level optimization. Buchanan (2001) identified multiple examples of meta-level state-space search in the literature as possible targets for transformational creativity research. We think another obvious target would be stochastic optimization fields such as EC, which are rife with meta-optimization and multiagent optimization in a variety of guises. These include meta-evolutionary algorithms which optimize the parameters of other EC algorithms; iterated local search, which seeks optimize in the space of local optima; island models, in which multiple optimization processes seed one other over distributed networks; and grammatical evolution, in which solutions are discovered by interpreting them through a grammar (which can itself can be optimized and modified).

Finally, we might also examine how transformational multiagent creativity can be used in a *co-creative context*. For example, could a computer identify when a human is stuck, or ignoring obvious solution avenues due to bias, and suggest ways to reframe the problem to help him out? Can a human do the same for the computer, acting as a guide to nudge the computer as it performs an exploratory search?

Conclusion

Transformational computational creativity has been neglected, but it should be at the forefront of the computational creativity research area. Transformational creativity's meta-level optimization it is well positioned to have an impact on AI applications, and the notion of spread of topologies poses valuable insights in a multiagent context. We hope to see transformational creativity applied to real AI applications and to have a greater effect on the AI research field.

References

Boden, M. 1992. *The Creative Mind*. Abacus.

Buchanan, B. 2001. Creativity at the metalevel: AAAI-2000 presidential address. *AI Magazine* 22(3):13–28.

Colton, S., and Wiggins, G. 2012. Computational creativity: the final frontier. In *European Conference on Artificial Intelligence (ECAI)*.

Cronin, M. A., and Loewenstein, J. 2018. *The Craft of Creativity*. Stanford Business Books.

Demke, J., and Ventura, D. 2024. Overcoming algorithmic bias as a measure of computational creativity. In *International Conference on Computational Creativity (ICCC)*, 245–253.

Feldman, D. H.; Csikszentmihalyi, M.; and Gardner, H. 1994. *Changing the World: A Framework for the Study of Creativity*. Praeger.

Kaufman, J. C., and Beghetto, R. A. 2009. Beyond big and little: The four C model of creativity. *Review of General Psychology* 13(1):1–12.

Linkola, S., and Kantosalo, A. 2019. Extending the creative systems framework for the analysis of creative agent societies. In *International Conference on Computational Creativity (ICCC)*, 204–2111.

Luke, S. 2023. Computational creativity as dynamic, multiobjective, multiagent optimization. In *International Conference on Computational Creativity (ICCC)*.

Mumford, M., and Gustafson, S. B. 1988. Creativity syndrome: Integration, application, and innovation. *Psychological Bulletin* 103(1):27.

Newell, A.; Shaw, J. C.; and Simon, H. A. 1958. The processes of creative thinking. Technical Report P-1320, The RAND Corporation.

Ritchie, G. 2005. On transformational creativity. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Ritchie, G. 2006. The transformational creativity hypothesis. *New Generation Computing* 24:246–266.

Ritchie, G. 2012. A closer look at creativity as search. In *International Conference on Computational Creativity (ICCC)*.

Saunders, R., and Gero, J. 2001. The digital clockwork muse: A computational model of aesthetic evolution. In *Society for the Study of Artificial Intelligence and Simulation of Behavior (AISB)*.

Saunders, R. 2019. Multi-agent-based models of social creativity. In Veale, T., and Cardoso, F. A., eds., *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*. Springer. 305–326.

Schumpeter, J. A. 1942. *Capitalism, Socialism and Democracy*. New York: Harper and Brothers.

Simon, H. A., and Hayes, J. R. 1976. The understanding process: Problem isomorphs. *Cognitive Psychology* 8(2):165–190.

Sternberg, R. J., and Lubart, T. I. 1999. *The Concept of Creativity: Prospects and Paradigms*. Cambridge University Press. chapter 1.

Venkatesh, K.; Dunlop, C.; and Yanardag, P. 2025. CREA: A collaborative multi-agent framework for creative image editing and generation. In *NeurIPS*.

Ventura, D. 2011. No free lunch in the search for creativity. In *International Conference on Computational Creativity (ICCC)*.

von Helmholtz, H. 1896. *Vorträge und Reden*. Friedrich Vieweg.

Wallas, G. 1926. *The Art of Thought*. Harcourt, Brace and Company.

Wiggins, G. A. 2006a. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19:449–458.

Wiggins, G. A. 2006b. Searching for computational creativity. *New Generation Computing* 24:209–222.

Does AI Herald the End of Creative Coding?
Paper type: Debate Sparks

Jon McCormack
SensiLab, Monash University
Melbourne, Australia
jon.mccormack@monash.edu

Stephen James Krol
SensiLab, Monash University
Melbourne, Australia
stephen.krol@monash.edu

Abstract

Advances in AI assisted code authoring and development mean that it is now possible for anyone to create custom software without extensive programming knowledge or experience. In this short paper, we ask what this means specifically for those who use computer coding as a creative medium. We discuss issues of cognition, aesthetics, creativity and knowledge, speculating on how creative coding might change in both the short and long term future.

Introduction

The widespread industrialisation of AI technologies continues at a rapid pace, including the advancement of AI systems that can design, write and debug complex computer programs from high-level, natural language instructions. "Coding Agents" such as *Claude Code*[1], *OpenAI Codex*[2] and *OpenCode*[3] have found rapid adoption amongst both experienced programmers and novices alike. Many people have commented on the significant increase in capability of the most recent models, with Coding Agents now being viewed more as collaborators rather than basic assistants or tools. Platforms such as *Lovable*[4] promote the concept of "vibe coding", where "the people who've had ideas but lacked the technical skills to bring them to life" are now "empowered" to write apps by telling an AI their idea in plain language and letting the AI platform create it. Lovable claims over 100,000 new projects are built every day (Lovable 2025) and that everyone can now be a "builder".

This newfound ability of AI to code competently has also led to job losses in software development (Techcrunch 2026), a downturn in new university enrolments in computer science and software engineering (Braue 2026), and proliferation of code with significant security and privacy issues (Brewster 2026). Driven by narratives of productivity gains, along with a general "Fear of Missing Out" on the next revolution in AI, coding agents are now widely adopted by developers, or at least employers are requiring developers to use them.

[1] https://code.claude.com/docs/en/overview
[2] https://openai.com/codex/
[3] https://opencode.ai
[4] https://lovable.dev

In this short paper, we want to consider specifically how AI coding agents might impact the field of "Creative Coding", although some of our arguments could apply to programming more generally. Creative Coding is an artistic practice that uses computer code as its primary medium of creative expression. A "Creative Coder" is a person who writes computer code that generates a creative artifact or behaviour. Creative Coding encompasses practices such as Generative Art, Algorithmic Art and Computer Art (McCormack 2024). In contrast to related practices, such as "Software Art" (Arnes 2004), in Creative Coding it is what the code produces – not the code itself – that is considered the main object of art appreciation. Some well known artists include Frieder Nake, Vera Molnár, Casey Reas and Zach Lieberman (to name just a selective few). Creative Coding also extends beyond the world of fine art, into commercial fields such as motion graphics, design and commercial art, but the basic principle remains: practitioners expressing their creativity through the medium of code.

A Golden Age For Creators?

Having an AI system that helps you write code does initially sound appealing. Programming is a complex skill, difficult to master, requiring many thousands of hours of practice and experience to gain proficiency (Robins, Rountree, and Rountree 2003; Bennedsen and Caspersen 2007). Modern software development relies on convoluted development environments, arcane build chains and multiple APIs. The professional programming world is filled with acronyms and jargon, making it difficult for artists and designers to easily understand and gain entry. This frustration led to the development of specialised "Creative Coding Environments", such as *Processing* (Fry and Reas 2014), which simplified the development environment and incorporated a rich set of functions targeted towards creative applications, such as sound and visual processing, interactivity, and graphics. But even with these artist-friendly environments, understanding how to program – having knowledge of data structures, algorithms and complexity measures – remains an important aspect of creative coding (Luxton-Reilly et al. 2018).

So for the experienced creative coder, agentic AI is appealing. Tasks that previously took days or weeks of coding labour can be undertaken by an agent in just a few minutes. Using plain language to direct an agent mitigates the need

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

for artist-friendly IDEs, or programming language simplification, currently common in creative coding development environments. For example, one can prototype in *Processing* (simplified Java) and then ask the agent to convert the code to high-performance, optimised C. Need a support library that is missing in your current environment? Get a coding agent to download, configure and install it for you, or even write it for you from scratch. Past projects can be given new life: augmented, extended, and modernised. The artist is no longer limited to a particular platform, language, API or development environment. Coding agents potentially eliminate the need for emulators or hardware virtualisation because they can translate software from one environment to another. Responsibility for the preservation of creative code artworks shifts from the museum conservator to the AI agent.[5]

These advantages might mean a new "golden age" for creative coding, at least for those with sufficient prior programming experience to understand and manipulate the code generated by an AI agent. Agentic AI coding is like having a whole team of coders at one's disposal at a tiny fraction of the cost of human programmers. Limited only by token budget, this team can work anywhere, anytime, on any programming task.

But like other AI incursions into human creative making such as text-to-image (TTI) generators, we need to be wary of how these new-found abilities might play out at scale and in the longer term. When capable TTI systems first appeared several years ago, many rushed to evangelise a new, democratised creative era, where anyone can "create stunning art in seconds" (McCormack et al. 2023). The reality today is not that everyone is creating "stunning art". TTI systems have flooded the Internet with AI slop (Kommers et al. 2026; Mahdawi 2025), are being used to promote and reinforce extremist political narratives (Watkins 2025), to create sexually explicit images of people against their will (Gentleman and Horton 2026), or as a way to save money on making Christmas cards (McCormack et al. 2024).

Just as the ability to generate an image from a prompt does not make you an artist, the ability to generate software does not necessarily make you a creative coder. You need artistic knowledge and skills as well, along with an understanding of how coding and creativity are enmeshed. While experienced creative coders may benefit from coding agents in the short term, once anyone can become a "builder" we are also likely to see coding skills devalued and a significant increase in AI code slop. The long-term consequences might be that human authoring of software becomes completely unnecessary and redundant as a useful skill. Instead, people will direct AI systems who write, maintain and debug the code. The complexity of these systems might increase to the point where no human can understand them at a code level, forcing a reliance and trust onto AI systems completely.

[5]Long term preservation of code-based artworks remains challenging, due to the technology they were developed on becoming obsolete or inaccessible (García and Vilar 2010).

Cognitive Offloading and the Locus of Creativity

Using a coding agent to write code creatively offloads the complexity of designing and writing software from the artist to the machine. At first glance, this might seem beneficial, because it frees up time and cognitive capacity to focus on the creative ideas, which the machine then translates into code. But this benefit is based on a false separation of creative ideation and coding practice. As clay is to a potter, code is to the creative coder: it is the material that is fashioned and manipulated into the system that produces the creative work. This is what Malafouris calls *Material Agency* (Malafouris 2008) – the act of coding is not simply to execute an internal mental plan of code authorship. Rather, the final form *emerges* from the interaction between bodies, tools, materials, habits and environment. Creatively developing software involves feedback loops between people, materials, code and culture; the ideas are iterated, modified and developed until the final product emerges. One rarely begins with a finished idea that is simply realised transparently through code. The fashioning and forming of code is integral to the creative process and the development of creative intention. Here, we think of colloquial expressions of how the creative process works: "one thing led to another", "trial and error" and "it wasn't until I'd done X that I realised Y" (Dietrich 2004).

At a level of abstraction, coding creatively is no different than a sculptor fashioning clay at a potter's wheel: the (code's) material behaviour and resistance give rise to the form. The artist shapes the form, but the form responds during the fashioning process and the creative artefact emerges from the nexus of those interactions. If we change or remove the material conditions from the process, the creative practice is no longer the same.

Conceptual problems of exchanging an embodied creative practice, such as painting or photography, with a prompt-based specification have been raised previously (McCormack et al. 2023). Most obviously, exploring creative concepts through language is limited to what can be expressed linguistically, and in the case of prompt-based systems those limitations are compounded by requiring explicit directives that can be correctly interpreted by an AI system. To paraphrase the painter Edward Hopper, "If you could say it in words, there would be no reason to code". While multimodal AI systems mitigate linguistic limitations to some extent, they still rely on being able to effectively translate ideas in one medium (e.g. language, visuals) directly to another (code). This may not be a stumbling block for engineering applications, where clarity, robustness and directness are important. But for creative coding, where ambiguity, poetics, metaphor and aesthetics are primary, it is a major conceptual shift.

Of course, human language is eminently capable of ambiguity, poetics and metaphor. Programming also makes use of language (both human and machine), but a fundamental difference is that human language evolved for human communication and expression, whereas programming languages are interpreted by machines. Machines have no use

for poetics, ambiguity or metaphor – they simply interpret the compiled code directly as instructions. So the use of artistic concepts, such as aesthetics, poetics, metaphor, etc. are necessarily in code's semantics, not the machine's interpretation of them.

Coding agents shift the locus of creative authorship from code-as-material to prompt-as-specification, and this shift is incompatible with creative coding's tradition of treating code itself as the aesthetic object. Some form of "creative prompting", or more likely human-agent interaction may soon replace direct coding. This interaction might become recognised as a new creative practice, but it will no longer be creative coding, since the material conditions are incompatible.

Excess and Inevitability

Current developments in AI can be understood through the lens of *excess*. Model training requires excessive amounts of data, resources[6] and energy – making foundational model development only possible for a few select, highly-resourced corporations. AI can generate obsessively and faster than people, resulting in the slop that is now polluting our digital lives. With the advent of coding agents, slop is now permeating software[7], causing problems for cybersecurity, open source developers, software engineers and of course, end users. The training data – the code on the Internet – is of varying quality, but often tends to be "average". Like all complex human endeavours, writing high quality software is difficult and time consuming, learning to program competently takes thousands of hours of practice and experience, so it tends to be protected from wide public exposure due to this investment. But this "average" code is largely what AI systems are trained on, and while RLHF and curated post-training shift output away from the training-data mean (Niekerk et al. 2025), the quality and suitability of code generated can vary, particularly if the code-base exceeds the LLM's context window.

For creative coding in particular, artists see their carefully crafted code as the basis of their artistic practice – what defines their aesthetic and makes their practice unique or, what Dissanayake calls "making special" (Dissanayake 1995). So while many explanatory or educational examples of creative coding exist on-line, far fewer professional creative examples are available to be scraped.[8] So commercial AI's need for excessive training examples is limited by Creative Coding's cultural practices.

Another way AI is publicly promoted is through it's *inevitability*. A strong narrative from tech evangelists and tech companies with a vested commercial interest in AI is that the development of increasingly powerful AI systems are *inevitable*. This sense of inevitability helps to mitigate objections or constraints on AI's development: if AGI is inevitable, how can we act to stop it? If we don't develop AI someone else will, and they'll reap the benefits or behave irresponsibly. These narratives echo similar arguments around inaction on climate change that play to strong hedonic and egoistic values (Steg 2023): if it is inevitable, how or why *should* I stop it? But developments in AI are not inevitable. We all have the right and responsibility to help determine the future we want to live in.

Conclusion: the end of Art-making?

In a recent article, media theorist Lev Manovich questioned if art making is still meaningful or even necessary when AI can often do it as well as or even better than humans (Manovich 2026). Across all the previous historical transformations of art due to technological or cultural change, one assumption stayed fixed: art is something made by humans. Manovich suggests that the next transformation may break that continuity. If AI systems can generate paintings, images, films, music, writing, and design it at high quality, then the future may not simply involve new artistic tools. It may involve a weakening of the human-centred concept of art itself.

Unlike AI systems that directly generate images, films, music or writing, coding agents generate software, which then needs to be run on a computer to reveal what it does. In creative coding the expectation is that the software *produces* something creative, not that the software itself is creative. Coding AIs are potentially only limited by Turing-completeness in what they can do.[9] So there is the potential for this meta-system to produce work that has greater originality and artistic value over the slop-producing statistical mash-ups from AIs directly trained on current human creativity. However, this potential rests on the assumption that AIs can code *creatively*.

Creative Freedom

In his book, "Mindless: The Human Condition in the Age of Artificial Intelligence", the economic historian Robert Skidelsky warned that "Unless we understand technology as a system of ideas rather than as a necessity, we will be powerless to choose which technology is best suited to our needs and purposes." (Skidelsky 2024, p.269). He questions if new technologies increase user's freedom to choose their own plans for life (i.e. are liberating), or if they force people to work within the systems imposed on them (subjugation). At the heart of art-making is the idea of creative freedom: the ability to express yourself how you want, in your own way. The early days of computing were full of new creative possibilities, many we have seen materialise over the last 75 or so years. Amongst these possibilities was idea that AI could make Art (Cohen, Cohen, and Nii 1983), something that was only taken seriously outside the world of the technological arts in the last few years.

If machines diminish the meaning and significance of human art-making then the freedom to be creative – in terms

[6] Estimates of over $100M are often cited for foundation models.

[7] *Lovable* publicly claims 36 million apps having been built on its platform.

[8] While high-quality creative code being publicly available is relatively rare, the *results* of that code are much more widely distributed, making it ripe for generative AI image generators to train on.

[9] Although there are many more practical limitations.

of employment, livelihood and social recognition – is threatened. "Being creative" risks becoming a nostalgic, recreational activity: something we might occasionally do to pass the time once creativity has been automated out of human interest.

Coding agents have the potential for today's coding artists to mitigate the limitations of time and complexity when writing code. Provided the locus of creativity remains centred around the artist and their code, we should expect to see a proliferation of interesting new examples of creative coding. But if the human authorship of code becomes obsolete due to AI, creative coding will become a lost practice. As Manovich points out, the continued importance of artmaking, in whatever form, should not be treated as self-evident merely because the process feels meaningful to us now. So while the short term may be something of a renaissance for creative coding, its long term future is uncertain. At this critical point in AI's deployment, it is important to question how these technologies can bring a deeper and richer sense of creative life to us all, rather than forcing us to conform to a technologically imposed concept of creativity.

References

Arnes, I. 2004. Read_me, run_me, execute_me: Software and its discontents, or: It's the performativity of code, stupid! In Goriunova, O., and Shulgin, A., eds., *Read_me. Software art and cultures conference*. Aarhus: University of Århus. 176–193.

Bennedsen, J., and Caspersen, M. E. 2007. Failure rates in introductory programming. *SIGCSE Bull.* 39(2):32–36.

Braue, D. 2026. University ICT enrolments down in 2026.

Brewster, T. 2026. Anthropic's Claude Is Pumping Out Vulnerable Code, Cyber Experts Warn. Section: Newsletters.

Cohen, H.; Cohen, B.; and Nii, P. 1983. *The First Artificial Intelligence Coloring Book*. Art and Computers. Addison-Wesley.

Dietrich, A. 2004. The cognitive neuroscience of creativity. *Psychonomic Bulletin & Review* 11(6):1011–1026.

Dissanayake, E. 1995. *Homo aestheticus: where art comes from and why*. Seattle: University of Washington Press.

Fry, B., and Reas, C. 2014. *Processing: A Programming Handbook for Visual Designers and Artists*. Cambridge, MA, USA: MIT Press, 2 edition.

García, L., and Vilar, P. M. 2010. The challenges of digital art preservation. *e-conservation* 5:43–53.

Gentleman, A., and Horton, H. 2026. 'Add blood, forced smile': how Grok's nudification tool went viral. *The Guardian*.

Kommers, C.; Duede, E.; Gordon, J.; Holtzman, A.; McNulty, T.; Stewart, S.; Thomas, L.; Jean So, R.; and Long, H. 2026. Why Slop Matters. *ACM AI Lett.* 1(1):3:1–3:6.

Lovable. 2025. Lovable raises $330M to power the age of the builder.

Luxton-Reilly, A.; Becker, B. A.; Cao, Y.; McDermott, R.; Mirolo, C.; Mühling, A.; Petersen, A.; Sanders, K.; Simon; and Whalley, J. 2018. Developing Assessments to Determine Mastery of Programming Fundamentals. In *Proceedings of the 2017 ITiCSE Conference on Working Group Reports*, ITiCSE-WGR '17, 47–69. New York, NY, USA: Association for Computing Machinery.

Mahdawi, A. 2025. AI-generated 'slop' is slowly killing the internet, so why is nobody trying to stop it? *The Guardian*.

Malafouris, L. 2008. At the potter's wheel: An argument for material agency. In Knappett, C., and Malafouris, L., eds., *Material agency: Towards a non- anthropocentric approach*. New York, NY: Springer Science and Business Media, LLC. 19–36.

Manovich, L. 2026. The Future of Art? Technical report, Cultural Analytics Lab.

McCormack, J.; Cruz Gambardella, C.; Rajcic, N.; Krol, S. J.; Llano, M. T.; and Yang, M. 2023. Is Writing Prompts Really Making Art? In Johnson, C.; Rodríguez-Fernández, N.; and Rebelo, S. M., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 196–211. Cham: Springer Nature Switzerland.

McCormack, J.; Llano, M. T.; Krol, S. J.; and Rajcic, N. 2024. No Longer Trending on Artstation: Prompt Analysis of Generative AI Art. In Johnson, C.; Rebelo, S. M.; and Santos, I., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 279–295. Cham: Springer Nature Switzerland.

McCormack, J. 2024. Algorithmic art. In Thomas, P., ed., *Encyclopedia of new media art*, volume 2: Artists & Practice. London, England: Bloomsbury Publishing.

Niekerk, C. v.; Vukovic, R.; Ruppik, B. M.; Lin, H.-c.; and Gašić, M. 2025. Post-Training Large Language Models via Reinforcement Learning from Self-Feedback. arXiv:2507.21931 [cs].

Robins, A.; Rountree, J.; and Rountree, N. 2003. Learning and Teaching Programming: A Review and Discussion. *Computer Science Education* 13(2):137–172. _eprint: https://doi.org/10.1076/csed.13.2.137.14200.

Skidelsky, R. 2024. *Mindless: The Human Condition in the Age of Artificial Intelligence*. New York NY: Other Press.

Steg, L. 2023. Psychology of Climate Change. *Annual Review of Psychology* 74(Volume 74, 2023):391–421.

Techcrunch. 2026. Layoffs.

Watkins, G. 2025. AI: The New Aesthetics of Fascism.

Co-Creative Vulnerability: Risks of Engagement Optimization in Creative AI

James Morgan[1] **and Margareta Ackerman**[2]

[1] Department of Art and Art History, Department of Computer Science, San Jose State University, CA, USA
[2] Department of Computer Science and Engineering, Santa Clara University, CA, USA

Abstract

This paper examines emerging engagement-optimizing dynamics in commercial Large Language Models, drawing a connection to operant conditioning mechanisms previously observed in social media and game design. A set of reinforcement patterns in conversational AI, including over-affirmation, responsiveness, and continuity of interaction are utilized to increase user retention. We argue that these mechanisms have distinct implications when users engage in co-creative processes, which require openness and psychological vulnerability. This creates a condition we term *co-creative vulnerability*, which may amplify the risk of retention-driven interaction patterns in commercial AI.

We present a framework linking operant conditioning, engagement optimization, addiction, and human-AI co-creativity, and outline a research agenda for evaluating the risks and ethical implications of these dynamics. Our goal is to initiate systematic investigation into how commercial incentives shape user experience in creative applications of generative AI systems.

Introduction

For decades, generative AI was an academic endeavor, shaped by pioneers such as David Cope (Cope 1989) and Harold Cohen (McCorduck 1991) in the late 20th century. This early work gave rise to the Computational Creativity research community and eventually the broader Creative AI community, which began to see early industry engagement in the mid-2010s. With its surge in popularity in late 2022, solidified by the release of ChatGPT, generative AI made a rapid and powerful entrance into the commercial domain. With funding now reaching unprecedented levels, 2025 alone saw over a $87B venture capital investment in the space[1], generative AI is transforming not only the nature of work, but potentially broader aspects of life and society at large.

This transition into industry has also fundamentally altered the trajectory of the field. While the Computational Creativity community has operated carefully around questions such as the role of evaluation [(Pollak et al. 2016), (Eigenfeldt et al. 2016)] and authorship (Maher 2012), industry operates under a different primary driver: monetization. And as massive investments flow into the space, pressure to generate returns is intensifying.[2]

Alongside continued advances in the intellectual capabilities of AI systems, fueled by these investments, we are beginning to see commercial objectives shape system design in new ways. One particularly important metric is user retention. Ultimately, companies must demonstrate progress to investors not only through technological breakthroughs, but through sustained and increasing user engagement. Daily use and prolonged engagement are key to demonstrating success.

This dynamic invites comparison to the evolution of social media. Initially grounded in the utopian vision of a globally connected world, social media platforms ultimately shifted toward maximizing user engagement regardless of the resultant impact on their users. In pursuit of retention, platforms adopted a range of psychologically potent techniques, including the intermittent delivery of high-reward content, mirroring Skinner box conditioning, to foster habitual and often compulsive use. As a result, many users find themselves engaging far beyond their intended usage. Metrics improve even when user well-being declines.

These strategies also include the amplification of emotionally arousing and polarizing content (Milli et al. 2025), which has contributed to increasing societal division. Ironically, platforms designed to connect people have, in many cases, become amplifiers of fragmentation.

This paper discusses how a similar pattern is beginning to emerge in generative AI systems, particularly LLMs. As financial pressures mount, companies are incentivized to maximize user engagement, regardless of user well-being. We identify and analyze emerging operant conditioning mechanisms within these systems, and explore how they may shape user behavior.

This has particularly important implications for creative applications of LLMs, where users are especially susceptible. Creativity, especially in a collaborative context, of-

[1] https://www.ey.com/en$_i e/newsroom/2025/12/global - genai - vc - investment - reaches - record - 87 - billion - in - 2025 - as - sovereign - wealth - funds - drive - strategic - growth - ey$

[2] https://www.cio.com/article/4114010/2026-the-year-ai-roi-gets-real.html

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

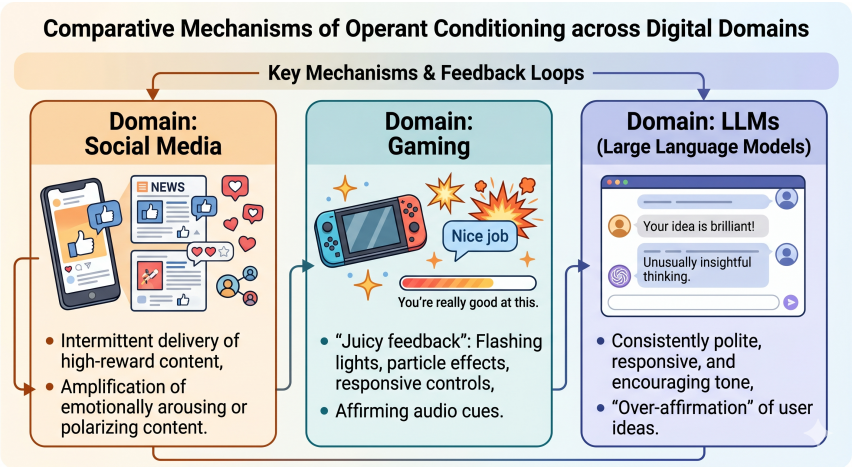


Figure 1: A summary of key operant conditioning across social media, gaming and LLMs.

ten involves vulnerability and openness, conditions that may amplify the impact of engagement-optimizing mechanisms.

This short paper aims to open a new direction of inquiry within the Computational Creativity community. As generative AI becomes increasingly shaped by commercial forces, it is critical to examine how these forces influence the experiences of users engaging with AI in creative contexts.

We begin with a brief overview of operant conditioning and how it manifests in modern AI systems. We then turn to the vulnerability inherent in creative collaboration, and conclude by offering new directions for studying co-creative vulnerability in the context of operant conditioning in commercial AI.

Spinning and Control: The Skinner Box

The operant conditioning chamber (Skinner Box) was designed to measure volitional, intentional actions on the part of a test subject.(Skinner 1938) A rat was placed in a soundproof, dark, well ventilated box. Pressing a lever required downward pressure which operated a switch that delivered a food pellet. Skinner found that reinforcing every interaction was often enough to continue the interaction until satiation. When the food was withheld, the rate of response declined over time. If the rat was returned to the box after a period of inactivity and after the food was withheld, the rate of responding would temporarily increase without any new reinforcement. It measures how frequently behaviors occur under different reinforcement conditions, but it does not capture the qualitative experience of that engagement. As a reward system and loop, it operates at the level of observable response, reinforcing actions through feedback without necessarily addressing the depth or meaning of the activity itself. (Swink 2008)

Operant conditioning covers voluntary behaviors shaped by consequences (like pushing a button for an anticipated reward.) (Skinner 1938) Similarly addiction is a compulsion that continues despite negative consequences, suggesting the reinforcement mechanism overrides normal logical, voluntary control.

In normal operant conditioning, a behavior becomes muted if punishment is consistent, but this is less effective than positive reinforcement. In addiction, the brain's reward system (dopamine) is hijacked, making the reinforcement so strong that traditional punishment fails to stop the behavior.(Volkow, Koob, and McLellan 2016)

Operant conditioning can be used to learn healthy behaviors.(Boden 2004) In contrast, addiction is a vicious cycle where behavior is aimed at achieving the same high or avoiding misery, often leading to a loss of pleasure over time, whereas operant behavior usually stops when reward ceases.

While operant conditioning is a general, neutral learning process, addiction represents a specific, harmful and hijacked application of the same learning principles.

Contemporary game design extends the logic of operant conditioning through techniques often described as "juicy feedback" and "game feel." Variable reward schedules like the occasional win, and intermittent reinforcement adding to flashing lights, particle effects, responsive controls, and affirming audio cues ("Nice job," "You're really good at this") create a tight feedback loop between player input and system response. These elements give the impression that the game is active and alive in the player's hands, making them part of the game world. In doing so, they leverage well-understood features of human attention and reward progression, contributing to sustained engagement.

AI, Conditioning and Addiction

Operant conditioning can be observed in conversational AI systems, with leading AI companies increasingly leaning into some of these mechanisms. A model's tone—consistently polite, responsive, and encouraging—can produce an immediate sense of connection and responsiveness. Like game feel, this does not necessarily reflect underlying intention or awareness, but rather a design that prioritizes smooth, coherent, and user-friendly interaction. The effect, however, can feel analogous: the system

appears attentive and affirming, reinforcing continued participation in the exchange.

There is growing awareness that such operant conditioning techniques are being actively integrated into LLMs. Emerging research suggests that conversational AI systems can shift from supportive tools toward patterns of dependency, where reinforcement through responsiveness, personalization, and conversational continuity encourages repeated engagement (Yankouskaya, Liebherr, and Ali 2025). These dynamics are increasingly discussed in terms of "dark patterns," where interface and interaction design subtly guide user behavior in ways that mirror addictive systems(Shen and Yoon 2025). More recent work further categorizes these behaviors, identifying forms of chatbot "addiction" such as escapist roleplay, pseudosocial companionship, and recursive information-seeking loops, all of which align with reinforcement-driven engagement cycles (Shen et al. 2026). Together, this body of research highlights that the behavioral mechanisms observed in operant conditioning are not only theoretical constructs but are actively shaping contemporary human–AI interaction.

Users are interacting with a system that shares structural similarities with other feedback-driven environments. The combination of responsive feedback, encouraging tone, and conversational continuity promotes ongoing engagement, the conditioning is designed to provide a pathological reinforcement and keep the attention of the user (Weizenbaum 1966). Affirmation in the form of validation, and excessive praise ("that's a brilliant idea!") can be a particular potent form of juicy feedback coming from an LLM.

The difference between operant conditioning and addiction is primarily the difference between a behavioral mechanism and a pathological behavioral result. Operant conditioning builds up a reflexive reserve of potential responses which will fade over time without reinforcement. Addiction represents a loss of human agency, where the user is "managed and controlled" by the system's reward schedule. Operant conditioning is not itself addiction, but being conditioned to something versus being addicted to it lay on the same spectrum. That means that more vulnerable people, who are more inclined towards addiction, can get addicted. Others may find themselves using the LLM more often than they would otherwise, and may experience a more moderate form of reduced control over their usage. We believe that this is concern not only when it leads to full blown addiction, but whenever the user's best interests, such as freedom over the utilization of their time, are hindered.

In this regard, the company's, and as such underlying LLMs', objectives become at direct odds - whereas the user wishes to use the LLM to complete a specific task or perhaps gain insight on a specific subject through a discussion with an LLM, the LLM aims to keep the user interacting regardless of whether the user's aims have already been met, and independently of whether any meaningful value is being provided. As such, a user may find themselves continuing to utilize the system simply to get the next dopamine hit - to be told again how brilliant they are - not necessarily realizing that they are being held in a conversation that no longer provides them with any substantive value.

It is possible that juicy feedback mechanisms such as affirmation may be more potent, and perhaps even more dangerous, when delivered by LLMs than in a gaming context. People come to interaction with AI systems with theory of mind about AI, often perceiving it as an extreme or even superior form of intellect to themselves. As a result, when an AI tells a user that they have a brilliant idea, or that their line of thinking is unusually insightful, the affirmation can carry disproportionate weight. Beyond the risks of overuse or addiction, this kind of over-affirmation may also contribute to what could be described as AI psychosis, in which individuals begin to lose their grounding in reality (Yeung et al. 2025).

The Vulnerability of Co-Creativity

Human creativity involves opening up one's thought process and sharing ideas that are personal and meaningful (Boden 2004). Whether it is a painting, poem, writing a story, or something else, the creator takes a risk by exposing all aspects associated with that work (Schön 1983). This risk extends into co-creativity, human to human collaboration and human to machine collaboration demands trust requires openness to new ideas, different approaches, and the creative processes of others (Kantosalo and Toivonen 2016). The collaborator risks failure and ridicule and must trust their partner to provide critique (Schön 1983). Interaction has to be free from judgment, or self censorship can take hold and prevent the sharing of an "outlying" idea. Collaborators must not fear social, psychological, or professional judgment. Creativity is risky with exposure of identity and social position. To be effective as a co-creator means being willing to listen, respond, and adapt. Co-creativity involves the exchange of intentions, expertise, and constraints, requiring participants to externalize incomplete ideas and negotiate meaning in real time.

Teachers of creativity see this with every project. Students are reluctant to make mistakes even though sharing a broken concept begins the process of finding something better (Schön 1983). Failure is closely connected with iteration and development of partial ideas, embracing feedback loops as part of critical analysis. This is the first step in refining the work into something creative and interesting. The relationship between instructor and student in an art context is co-creative. The instructor brings experience and practice to the raw ideas of the student who needs to be bold and offer their personal direction and the teacher needs to be empathic and offer critique to refine or redirect the idea. Teachers need to balance challenging the student with supporting their thoughts and processes while creating a welcoming environment. Co-creativity depends on structured vulnerability. This is assessment and is intended to improve the work. Who makes the ultimate decision? The artist/student does, but they must hear the challenges and observations about the flaws and strengths of their ideas as they move to fabrication. This feedback needs to be continuous and the decisions of the artist need to be encouraged and respected with honest feedback based on experience. This is mentored co-creation.

Human collaboration depends on trust and openness. It involves sharing not only abilities but also uncertainties,

preferences, and constraints. Without this vulnerability, it is difficult to communicate on equal terms or to build a meaningful collaborative process. When a human collaborates with an algorithm, this openness still applies to the human participant (Davis et al. 2015). If the human collaborator is guarded or resistant, the collaboration weakens. If they are open and able to change, reflect, critique, and incorporate feedback then the collaboration becomes stronger because the idea is expressed. This involves showing vulnerability and requires a certain amount of trust.

This vulnerability introduces risk. A collaborator who is open is also more susceptible to influence. If a co-creative system employs persuasive or manipulative behaviors, the human participant may be more easily affected because the process benefits from them remaining receptive.

This creates a challenge in human-AI collaboration. Both manipulative systems and overly supportive systems shape the user in ways that reduce critical engagement in the collaborative process. Unequivocal support does not move collaboration forward; it reduces reflection and diminishes the role of critical thinking. Effective collaboration requires participants to express what they think and feel, including ideas that may be uncomfortable, in order to understand what exists between them.

Creativity is not a purely technical or clinical activity; it is a vulnerable act tied to human emotion and subconscious processes. As a result, users entering a creative interaction with an AI may be primed for personal, emotional engagement. Unlike passive tools, an AI system can function as an active conversational partner, shaping the direction and subject of the conversation. This dynamic can encourage ongoing engagement, but it also introduces questions about how that engagement is structured and how it affects the human participant within a co-creative process.

We define *co-creative vulnerability* as the heightened susceptibility to influence that arises when individuals engage in creative collaboration with an AI system while in an open, exploratory cognitive and emotional state brought on by the creative act. In such contexts, users externalize incomplete ideas, suspend judgment, and actively seek feedback, creating conditions in which system responses carry amplified psychological weight. Co-creative vulnerability thus describes a structural property of human-AI interaction in creative settings, where the very conditions that enable effective collaboration may also increase the user's sensitivity to engagement-optimizing dynamics.

This suggests that specifically in creative collaborations with AI systems like LLMs a person may suffer greater harm compared with other applications of the technology. We propose this as a direction of research, to explore whether creative tasks compared to those which require less creativity are differently effected by retention-driven LLMs behavior. We hypothesize that the vulnerable state called upon by creative acts makes people more susceptible to harm by these recently emerging LLM tactics.

Discussion & Future Directions

This position paper draws attention to engagement-optimizing dynamics in commercial LLMs, highlighting how reinforcement mechanisms such as affirmation, responsiveness, and conversational continuity can operate in tension with users' goals. We argue that these dynamics may be particularly consequential in creative contexts, where interaction with AI is carried out through co-creative dynamics. In these settings, users enter a state of *co-creative vulnerability*, defined by openness, emotional investment, and a willingness to externalize incomplete ideas. This state is essential for creativity, but it also increases susceptibility to reinforcement-driven interaction patterns.

This framing opens a novel line of inquiry for the Computational Creativity community: how engagement-driven mechanisms shape human-AI co-creation, and how their effects differ between creative tasks versus more convergent uses such as factual lookup. It would be worth studying, for instance, whether utilizing an LLM in co-creative contexts such as poetry, narrative writing, or ideation puts a user in greater risk for sliding towards addiction. How else do misaligned incentives impact users in a vulnerable co-creative state, while AI seeks to optimize engagement at all costs? This could be analyzed through user studies, perhaps beginning with qualitative analysis to understand the phenomena and then followup with quantitative studies to gauge the scale of any observed patterns. Extending this analysis to other generative systems, including text-to-image and text-to-video, which are often utilized by artists and creatives in a wide variety of ways, is also worth investigating.

From an ethical perspective, these dynamics must be evaluated in terms of their impact on user agency. Systems optimized for engagement risk shaping creative behavior in ways that do not serve the user. How can users maintain autonomy over their attention and creative direction when interacting with systems designed to sustain interaction? Can interfaces or intermediary layers reduce reinforcement signals that do not contribute meaningful value? Can we define concrete design principles for co-creative AI that preserve autonomy and minimize risk while supporting creative work?

Both technical and non-technical solutions are worth investigating. By introducing the concept of co-creative vulnerability and grounding it in the dynamics of engagement optimization, this paper establishes a foundation for studying how commercial AI systems shape the creative process. Addressing this challenge is essential if AI is to support human creativity as a true partner, rather than introducing risks into the very conditions that make creativity possible.

References

[Boden 2004] Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2 edition.

[Cope 1989] Cope, D. 1989. Experiments in musical intelligence (emi): Non-linear linguistic-based composition. *Journal of New Music Research* 18(1-2):117–139.

[Davis et al. 2015] Davis, N.; Hsiao, C.-P.; Singh, K.; Li, L.; and Magerko, B. 2015. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 143–152. ACM.

[Eigenfeldt et al. 2016] Eigenfeldt, A.; Bown, O.; Brown, A. R.; and Gifford, T. 2016. Flexible generation of musical form: beyond mere generation. In *ICCC 2016*. The Association for Computational Creativity.

[Kantosalo and Toivonen 2016] Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the Seventh International Conference on Computational Creativity*.

[Maher 2012] Maher, M. L. 2012. Computational and collective creativity: Who's being creative? In *ICCC*, 67–71.

[McCorduck 1991] McCorduck, P. 1991. *Aaron's code: meta-art, artificial intelligence, and the work of Harold Cohen*. Macmillan.

[Milli et al. 2025] Milli, S.; Carroll, M.; Wang, Y.; Pandey, S.; Zhao, S.; and Dragan, A. D. 2025. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus* 4(3):pgaf062.

[Pollak et al. 2016] Pollak, S.; Boshkoska, B. M.; Miljkovic, D.; Wiggins, G. A.; and Lavrac, N. 2016. Computational creativity conceptualisation grounded on iccc papers. In *Proceedings of the Seventh International Conference on Computational Creativity*, 123–130.

[Schön 1983] Schön, D. A. 1983. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.

[Shen and Yoon 2025] Shen, M. K., and Yoon, D. 2025. The dark addiction patterns of current ai chatbot interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.

[Shen et al. 2026] Shen, M. K.; Huang, J.; Liang, O.; Kim, I.-J.; and Yoon, D. 2026. The ai genie phenomenon and three types of ai chatbot addiction: Escapist roleplays, pseudosocial companions, and epistemic rabbit holes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–17.

[Skinner 1938] Skinner, B. F. 1938. *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century.

[Swink 2008] Swink, S. 2008. *Game Feel: A Game Designer's Guide to Virtual Sensation*. CRC Press.

[Volkow, Koob, and McLellan 2016] Volkow, N. D.; Koob, G. F.; and McLellan, A. T. 2016. Neurobiologic advances from the brain disease model of addiction. *New England Journal of Medicine* 374(4):363–371.

[Weizenbaum 1966] Weizenbaum, J. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1):36–45.

[Yankouskaya, Liebherr, and Ali 2025] Yankouskaya, A.; Liebherr, M.; and Ali, R. 2025. Can chatgpt be addictive? a call to examine the shift from support to dependence in ai conversational large language models. *Human-Centric Intelligent Systems* 5(1):77–89.

[Yeung et al. 2025] Yeung, J. A.; Dalmasso, J.; Foschini, L.; Dobson, R. J.; and Kraljevic, Z. 2025. The psychogenic machine: Simulating ai psychosis, delusion reinforcement and harm enablement in large language models. *arXiv preprint arXiv:2509.10970*.

A Training-Free Novelty Detection Framework for Large-Scale Multi-Agent Social Creativity Simulations

Luuk Motz and **Tessa Verhoef** and **Rob Saunders**
Leiden Institute of Advanced Computer Science (LIACS), Leiden University
luukmotz@gmail.com t.verhoef@liacs.leidenuniv.nl r.saunders@liacs.leidenuniv.nl

Abstract

Per-agent novelty detection is central to computational models of social creativity, but trained detectors have kept simulations too small to study population-level dynamics such as the emergence of sub-fields or the relationship between per-share reception and cumulative field-level contribution. We present a training-free framework in which agents maintain a personal repository of pretrained ResNet-18 features, with novelty computed via k-nearest-neighbour distances normalised by local standard deviation. The framework scales to significantly larger agent populations while preserving individual aesthetic variance on a single consumer GPU. We validate at 50 to 250 agents, surfacing a rate–volume inversion pattern where the most prolific senders achieve the lowest per-share acceptance rates, with sub-linear per-step wall-clock cost in population size.

Introduction

Creativity is better understood as a social and cultural phenomenon than as a property of individual minds (Csikszentmihalyi 1988), and multi-agent simulation provides a natural tool for studying it as such (Saunders and Bown 2015). In these models, novelty detection is frequently a core mechanism: agents generate artefacts, judge them against their own experience, and decide what to share with the community (Saunders and Gero 2001). However, per-agent novelty detectors based on trained models impose computational costs that have kept simulations too small to study population-level phenomena such as the emergence of subfields or the dynamics of aesthetic divergence.

We propose a training-free framework that removes this bottleneck. Each agent maintains a personal repository of features extracted by a shared pretrained ResNet-18, and novelty is computed via k-nearest-neighbour (kNN) distances normalised by local standard deviation. Because the feature extractor is fixed and the kNN distance computation parallelises on GPU, the method preserves per-agent subjectivity at significantly larger population sizes, running on modest hardware. We validate the framework at populations from 50 to 250 agents, surfacing a population-level pattern where the most prolific senders achieve the lowest per-share acceptance rates, with sub-linear per-step wall-clock cost in population size.

Related Work

Saunders and Gero (2001) introduced a computational model of social creativity based on Csikszentmihalyi's (1988) Domain-Individual-Field-Interaction (DIFI) framework; curiosity-driven agents generate, evaluate, and share novel artefacts, producing emergent social structures such as cliques and stylistic divergence consistent with Martindale's (1990) theory of aesthetic evolution. Linkola, Takala, and Toivonen (2016) extended this early work by adding resource constraints and collective voting over domain entry; Guckelsberger et al. (2016) replaced novelty-seeking with coupled empowerment maximisation to account for both supportive and antagonistic agent relationships. Peeperkorn et al. (2022) replaced hand-crafted feature spaces with Variational Autoencoders, enabling conceptual spaces that adapt through social experience.

Saunders and Bown (2015) situated this approach within the agent-based modelling traditions of Axelrod (1997) and Epstein and Axtell (1996). Within this tradition, Gabora's Meme and Variations (1995) and its successor EVOC (2008) modelled culture as an evolutionary process using simpler but larger-scale agent societies up to 1,024 agents—each comprising an 18-node auto-associative neural network—demonstrating that an intermediate invention-to-imitation ratio maximises cultural fitness and that self-regulation of creativity has measurable population-level effects. Sosa and Gero (2005) modelled 100-agent design communities using Axelrod's cultural diffusion framework, explicitly encoding designers, adopters, and emergent gatekeepers—operationalising Csikszentmihalyi's field as a structurally distinct evaluative class. Sosa, Gero, and Jennings (2009) went on to show that the social worth of ideas grows and decays dynamically through adoption and displacement. Corneli (2016) reframed creative communities through Ostrom's (1990) institutional analysis, treating the shared domain as a creative commons.

Jennings (2010) went on to argue that creative autonomy—the capacity to revise one's own evaluative standards without external instruction—is best achieved through social embedding in a community of creators and critics, and demonstrated that creative search trajectories are consistent with simulated annealing dynamics (2011). Saunders (2012) proposed language games as the mechanism through which agents can autonomously develop shared creative vocabu-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

laries and evaluative standards. Building on the language games framework of Steels (1995), Saunders (2011) demonstrated that agents can successfully develop these vocabularies, while Zhang and Saunders (2012) showed that compositional languages with hedonic evaluation promote deeper conceptual exploration. This connects to a broader scientific programme linking language, cultural transmission, and creativity: iterated learning experiments and models have established that languages evolve toward learnability and combinatorial structure (Kirby, Cornish, and Smith 2008; Verhoef, Kirby, and de Boer 2014; 2016), a cumulative cultural process that models of social creativity should aspire to capture. The development of Large Language Models (LLMs) has given rise to LLM-based agentic systems and has sparked a renewed interest in multi-agent social creativity at a new level of agent richness. Lin et al. (2025) survey the growing body of LLM-based multi-agent creative systems; Park et al. (2023) demonstrated that 25 LLM-based agents could produce emergent social behaviours from persona-based prompts; Imasato et al. (2025) applied this to Csikszentmihalyi's systems model, finding that multi-agent settings produce more creative artefacts than isolated agents; and Vallinder and Hughes (2025) showed that LLM-based agent societies can evolve cooperative norms across generations. Agent sophistication and population scale have so far moved in opposite directions: the largest simulations use the simplest agents and forgo core mechanisms such as per-agent novelty detection. The framework described in the next section is designed to break this tradeoff.

Method

Our approach replaces per-agent feature extraction and training with a shared pretrained ResNet-18 and a batched kNN process for novelty detection. Bergman, Cohen, and Hoshen (2020) showed that this approach can be effective for anomaly detection. We extend it here to address the individuality-scale tradeoff in social creativity simulations. Shared feature extraction amortises across all agents to a single batched forward pass per step, and per-agent novelty estimation reduces to a kNN lookup against a size-limited per-agent repository. All experiments were conducted on an NVIDIA RTX 3090 GPU and AMD Ryzen 5 5600X CPU.

Artefact generation is based on the process described for *The Digital Clockwork Muse* (Saunders and Gero 2001); agents generate 32×32 RGB images by evaluating quaternion-based expression trees at every pixel location (Witbrock and Reilly 1999). New expressions are evolved by subtree crossover and mutation; parent trees are chosen with equal probability from the agent's current expression or a personal repository of previously evaluated trees; a random subtree is selected for crossover, followed by a 5% point mutation. See Figure 4 for examples and expressions.

Feature Extraction Agents extract features using a ResNet-18 pretrained on ImageNet (He et al. 2016); to prevent overly aggressive spatial downsampling of the 32×32 images, the original 7×7 stride-2 first convolution is replaced by a 3×3 stride-1 layer initialised by bilinearly downsampling the pretrained kernel. 128-dimensional feature vectors are extracted from Layer 2 after global average pooling and ℓ_2 normalisation.

Extractor	AUROC	Noise
Pretrained L_2	0.827 ± 0.004	0.498
Pretrained L_3	0.821 ± 0.004	0.496
Pretrained L_4	0.809 ± 0.004	0.501
Random-init L_2	0.769 ± 0.005	0.500
Raw pixels	0.674 ± 0.004	0.498
Pretrained L_2 PCA-16	0.828 ± 0.003	0.495

Table 1: AUROC for kin vs. unrelated pair classification (mean ± std, 50 runs). The Noise column reports AUROC on uniform random image pairs.

The choice of Layer 2 was motivated by the observation that early CNN layers capture local texture, whereas deeper layers prioritise global semantics (Gatys, Ecker, and Bethge 2015; Geirhos et al. 2018). We observed that Layers 3 and 4 tend to map visually distinct artefacts onto coarse category-level representations, compressing the variance that novelty detection depends on, while Layer 2 provides a more appropriate representation of stylistic variance. We have validated this choice by classifying pairs of generated artefacts as kin (single-operator subtree substitution) or unrelated (independent random trees) via cosine distance in feature space, scoring by AUROC over 50 runs with 5,000 pairs per condition. Table 1 reports the result for six separately tested extractors and confirms that pretrained Layer 2 achieves the highest AUROC (0.827 ± 0.004), with Layers 3 and 4 trailing at 0.821 and 0.809 despite requiring two to six times more PCA components than Layer 2 to reach the 90% variance threshold. A random-initialisation baseline lags by 0.058, isolating the contribution of ImageNet pretraining, and a uniform-noise control returns AUROC near 0.5 across all extractors. A PCA-16 projection of Layer 2 matches the full-width AUROC (0.828 ± 0.003), suggesting that the per-agent embedding could be compressed eightfold without loss; however, we retain the full 128-dimensional representation for the experiments reported below.

Novelty Measurement Each agent i maintains a personal repository M_i implemented as a circular buffer of fixed capacity (1,000 entries) to model a recency bias and keep memory requirements bounded as simulations scale. Feature vectors for all generated and evaluated artefacts are added to the personal repository. When agent i encounters a feature vector f, it retrieves the k nearest neighbours in M_i to compute a novelty score:

$$\nu_i(f) = \frac{\bar{d}_k(f, M_i)}{\sigma_k(f, M_i)}$$

where $\bar{d}_k$ and σ_k are the mean and standard deviation of the k retrieved cosine distances on ℓ_2-normalised features, and $k = 15$ is fixed across all agents and steps. The denominator rescales each agent's signal by the local dispersion of its retrieved neighbourhood, yielding a unitless score that is comparable across agents regardless of memory size or

feature-space region. Novelty scores are normalised to $[0, 1]$ via a population-wide rolling percentile transform (1st and 99th percentiles over most recent 10,000 novelty scores), yielding $\tilde{\nu}_i(f)$, which is passed to a per-agent hedonic function, i.e., a Wundt curve (Berlyne 1971), to produce an interest score H_i:

$$H_i(\tilde{\nu}) = R_i(\tilde{\nu}) - \alpha P_i(\tilde{\nu})$$

where R_i and P_i are cumulative Gaussians encoding reward and punishment for novelty, and $\alpha = 1.2$ is used in the following simulations to weight the punishment term.

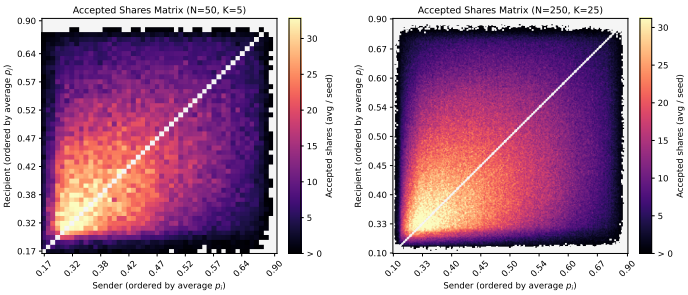
Each agent's Wundt curve, based on (Saunders and Gero 2001), is centred on a preferred novelty level $p_i \sim \mathcal{N}(0.5, 0.15)$, which distributes aesthetic preferences continuously across the population instead of assigning discrete roles. Its reward and punishment means are derived from p_i as $\mu_R = \max(0.1, p_i - 0.2)$ and $\mu_P = \min(0.9, p_i + 0.2)$, with fixed standard deviations $\sigma_R = \sigma_P = 0.15$; the ± 0.2 offset separates the two activations and the $\max$ / $\min$ clipping keeps both midpoints within the novelty range.

Social Interaction Policy Interest scores translate into social action through three global thresholds, namely a self-approval threshold τ_C (80th percentile), a domain threshold τ_D (80th percentile), and a boredom threshold τ_B (10th percentile), recomputed every step and applied population-wide. τ_C and τ_D are rolling percentiles over recent self-evaluations and recipient-side share evaluations respectively. τ_B is computed over the current distribution of per-agent average interest. After self-evaluating a generated artefact, agent i shares it with K other agents sampled randomly from the population, conditional on $H_i > \tau_C$. The K recipients each evaluate the artefact independently through their own kNN repository and Wundt curve. For each recipient agent j, if $H_j > \tau_D$, the artefact enters the shared domain and a directed edge $(i \rightarrow j)$ is recorded in the accepted-share graph; if H_j exceeds the recipient's interest in its own expression, agent j adopts the artefact as its current expression, independent of τ_D. When an agent's average interest falls below τ_B, it adopts a randomly sampled domain artefact, provided the agent's interest in that artefact exceeds its interest in its current expression. Before each step, agents re-evaluate their current expression against their updated repository, so that interest in unchanging material decays as similar artefacts accumulate in memory. The first 15 iterations are excluded from analysis while per-agent repositories and the rolling window populate.

Experimental Validation

We validate the framework across a $3 \times 3 \times 5$ grid of conditions: $N \in \{50, 100, 250\}$ agents, $K \in \{1, 5, 25\}$ shares per step, across five runs per (N, K) condition, with 2,500 iterations per run. Figures anchor on the $(N{=}50, K{=}5)$ and $(N{=}250, K{=}25)$ conditions, where both parameters are quintupled to test scaling. Enlarged figures and $N = 100$ results appear in the Appendix.

Figure 1 renders the accepted-share multigraph as a sender $\times$ recipient matrix where each cell counts run-averaged accepted shares from sender i to recipient j, with



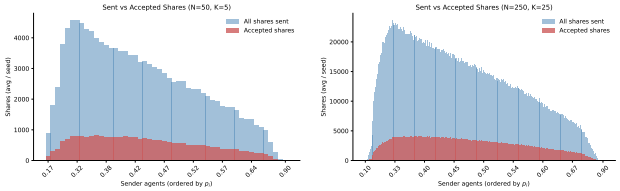
(a) $N = 50$, $K = 5$. (b) $N = 250$, $K = 25$.

Figure 1: Accepted-share interaction matrices. Each cell counts run-averaged accepted shares from sender i (x-axis) to recipient j (y-axis), with both axes ordered by preferred novelty; a position on either axis refers to the agent occupying that rank in a given run. White cells indicate pairs with zero accepted shares; black cells indicate pairs with at least one accepted interaction averaged out over all runs.

both axes ordered by preferred novelty; axis tick labels report the mean p_i at each rank across runs. Interaction density concentrates in the low-novelty preference (p_i) region, peaking near $p_i \approx 0.33$ for both senders and recipients. Scaling N and K by a factor of five reproduces this distribution at higher resolution, confirming it as a stable structural behaviour across population size and sharing rate.

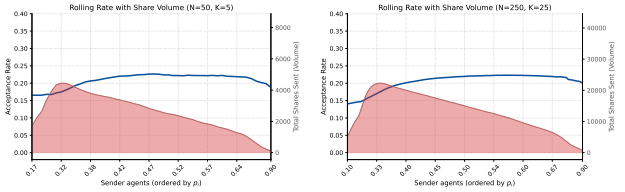
Figure 2 shows that share volume concentrates in the low-p_i region, with low-p_i senders emitting several times more shares than high-p_i senders and the total-sent distribution decaying monotonically toward the tail. The generator recombines the agent's current expression with one sampled from its personal memory, so the mean produced-novelty sits near 0.3–0.35 across runs, below the population-mean preference of 0.5. Agents whose Wundt peak aligns with this regime clear τ_C more often and compound their sending advantage over the run; high-p_i agents face the inverse. Low-p_i agents therefore dominate the accepted-share total through volume alone, and this behaviour replicates across all (N, K) runs.

Figure 3 shows that the rate–volume inversion is stable across both conditions. Low-p_i senders ship artefacts across a broad range that many recipients reject, depressing ac-



(a) $N = 50$, $K = 5$. (b) $N = 250$, $K = 25$.

Figure 2: Per-sender share volume ordered by preferred novelty p_i, run-averaged. Each bar pair reports total shares sent and the accepted subset; the accepted subset reflects the domain-threshold τ_D filter applied on the recipient side.



(a) $N = 50$, $K = 5$. (b) $N = 250$, $K = 25$.

Figure 3: Per-sender acceptance rate (rolling curve, left axis) and total shares sent (right axis) against preferred novelty p_i, run-averaged. The rate curve is inverted against volume: lowest where sending is densest, rising toward high-p_i where sending is sparsest.

ceptance rate; high-p_i senders clear τ_C only when generation lands near their preferred novelty, so the artefacts they do ship encounter less rejection. Per-share, high-p_i senders are thus more successful, but the multiplicative volume gap from Figure 2 dominates and absolute accepted-share counts remain concentrated in the low-p_i region.

Table 2 shows sub-linear scaling in N (log-log slopes 0.61–0.76), reflecting that the shared ResNet-18 and GPU-batched kNN amortise fixed overhead across the population. Per-step cost does not drift with repository growth for $K \geq 5$, as repositories are capped at 1,000 entries per agent.

Discussion

Figure 3 reveals a rate–volume inversion: high-preference senders achieve per-share acceptance rates several times those of low-preference senders and yet contribute a smaller fraction of accepted shares overall because volume dominates the total (Figure 2). Per share, high-p_i agents are the more effective participants, limited by throughput rather than by reception. The generator combines parents drawn from the agent's current expression and personal memory, biasing each agent's generative trajectory toward its own history when reception is sparse; received artefacts enter the recipient's own repository regardless of acceptance, and adoption (share-driven or boredom-driven) additionally substitutes external material into the generative pool. Social input therefore shapes per-agent evaluation continuously through reception and intermittently through adoption, rather than acting only on the output side. This inversion—in which the agents whose artefacts are most valued by other agents are also the least prolific contributors to the shared domain—connects to longstanding questions in computational modelling of social creativity about the relationship between individual creative quality and field-level influence (Jennings 2010; Sosa, Gero, and Jennings 2009), and is only reliably detectable at the population sizes this framework enables.

	N=50	N=100	N=250	slope
K=1	0.32	0.50	0.95	0.68
K=5	0.33	0.56	0.90	0.61
K=25	0.41	0.74	1.40	0.76

Table 2: Mean wall-clock seconds per simulation step across the grid, with fitted log-log power-law slopes in N. Cost scales sub-linearly in N for all K.

Previous simulations of agents with novelty detection, capped at around 20 agents, could not reliably separate signal from noise on such patterns. The tension between agent sophistication and population size has until now precluded empirical investigation of dynamics of this kind. The shared pretrained ResNet and batched kNN novelty detection resolve this tension, preserving per-agent subjectivity at populations of hundreds while keeping each simulation on the order of a second on a single GPU (Table 2).

The same scale opens variations that prior population ceilings made difficult to attempt. Alternative social policies, including preference-weighted partner sampling, τ_C informed by acceptance feedback, and per-agent throughput caps, each predict different redistributions of the accepted-share total across the preference range, and larger populations make these differences detectable. The generator is independently substitutable since any feature-extractable medium would do, and a broader effective produced-novelty range would shift the population dynamics observed here. The PCA-16 row of Table 1 confirms that the per-agent embedding compresses eightfold without loss, leaving headroom for larger extractors or more complex generators at the same memory budget. Perhaps the largest open axis is agent learning: preferences could drift over time, repositories could decay selectively, and generators could adapt to their own evaluative history. This framework places each of these extensions within reach at population scales where their effects are visible.

Conclusion

Per-agent novelty detection has been the principal bottleneck on population size in computational social creativity. The training-free framework presented here, using a shared pretrained ResNet-18 with per-agent kNN repositories, removes that bottleneck, supporting populations of hundreds of agents on a single consumer GPU while preserving the individual aesthetic variance that makes social creativity modelling meaningful. The per-share inversion of Figure 3 illustrates the kind of population-level finding that only becomes visible at this scale, and points toward a broader research agenda: understanding how individual creative quality, social influence, and field-level dynamics interact in large creative communities. Further feature-vector compression offers a path to larger simulations, and agent learning represents the natural next step toward models that can capture the cumulative cultural processes the field has long aspired to study.

Author Contributions

Motz conceived the study, developed the methodology, conducted the experiments, and wrote the original draft. Verhoef and Saunders provided supervision, reviewed the manuscript, and contributed to the final editing and revision.

Acknowledgments We thank the anonymous reviewers for their constructive feedback.

References

Axelrod, R. 1997. The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution* 41(2):203–226.

Bergman, L.; Cohen, N.; and Hoshen, Y. 2020. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*.

Berlyne, D. E. 1971. *Aesthetics and Psychobiology*. New York: Appleton-Century-Crofts.

Corneli, J. 2016. An institutional approach to computational social creativity. In *Proc. 7th Int. Conf. Computational Creativity (ICCC 2016)*, 131–138.

Csikszentmihalyi, M. 1988. Society, culture, and person: A systems view of creativity. In Sternberg, R. J., ed., *The Nature of Creativity: Contemporary Psychological Perspectives*. Cambridge University Press. 325–339.

Epstein, J. M., and Axtell, R. L. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. MIT Press.

Gabora, L. 1995. Meme and variations: A computer model of cultural evolution. In Nadel, L., and Stein, D. L., eds., *1993 Lectures in Complex Systems*, 471–486.

Gabora, L. 2008. EVOC: A computer model of the evolution of culture. In Sloutsky, V.; Love, B.; and McRae, K., eds., *Proc. 30th Annual Meeting of the Cognitive Science Society*. Sheridan Publishing.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Guckelsberger, C.; Salge, C.; Saunders, R.; and Colton, S. 2016. Supportive and antagonistic behaviour in distributed computational creativity via coupled empowerment maximisation. In *Proc. 7th Int. Conf. Computational Creativity (ICCC 2016)*, 9–16.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Imasato, N.; Miyazawa, K.; Nagai, T.; and Horii, T. 2025. Creative agents: Simulating the systems model of creativity with generative agents. *IEEE Access* 13:157712–157729.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in Artificial Intelligence. *Minds Mach.* 20(4):489–501.

Jennings, K. E. 2011. A computational perspective on human exploratory creativity: Theory and methods. In *Proc. 2nd Int. Conf. Computational Creativity (ICCC 2011)*.

Kirby, S.; Cornish, H.; and Smith, K. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proc. National Academy of Sciences* 105(31):10681–10686.

Lin, Y.-C.; Chen, K.-C.; Li, Z.-Y.; Wu, T.-H.; Wu, T.-H.; Chen, K.-Y.; Lee, H.-y.; and Chen, Y.-N. 2025. Creativity in LLM-based multi-agent systems: A survey. In *Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP 2025)*, 27584–27607. Suzhou, China: Association for Computational Linguistics.

Linkola, S.; Takala, T.; and Toivonen, H. 2016. Novelty-seeking multi-agent systems. In *Proc. 7th Int. Conf. Computational Creativity (ICCC 2016)*, 1–8.

Martindale, C. 1990. *The Clockwork Muse: The Predictability of Artistic Change*. New York: Basic Books.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Political Economy of Institutions and Decisions. Cambridge University Press.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology (UIST 2023)*, 1–22. ACM.

Peeperkorn, M.; Saunders, R.; Bown, O.; and Jordanous, A. 2022. Mechanising conceptual spaces using variational autoencoders. In *Proc. 13th Int. Conf. Computational Creativity (ICCC 2022)*, 287–290.

Saunders, R., and Bown, O. 2015. Computational social creativity. *Artificial Life* 21(3):366–378.

Saunders, R., and Gero, J. S. 2001. The digital clockwork muse: A computational model of aesthetic evolution. In *Proc. AISB'01 Symposium on AI and Creativity in Arts and Science*, 12–21. York, UK: SSAISB.

Saunders, R. 2011. Artificial creative systems and the evolution of language. In *Proc. 2nd Int. Conf. Computational Creativity (ICCC 2011)*, 36–41.

Saunders, R. 2012. Towards autonomous creative systems: A computational approach. *Cognitive Computation* 4(3):216–225.

Sosa, R., and Gero, J. S. 2005. A computational study of creativity in design: The role of society. *AI EDAM: Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 19(4):229–244.

Sosa, R.; Gero, J. S.; and Jennings, K. E. 2009. Growing and destroying the worth of ideas. In *Proc. 7th ACM Conf. Creativity and Cognition*, 295–304. Berkeley, CA: ACM.

Steels, L. 1995. A self-organizing spatial vocabulary. *Artificial Life* 2(3):319–332.

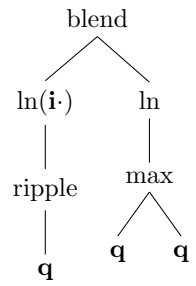
Vallinder, A., and Hughes, E. 2025. Cultural evolution of cooperation among LLM agents. In *Proc. 24th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS 2025)*.

Verhoef, T.; Kirby, S.; and de Boer, B. 2014. Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *J. Phon.* 43:57–68.

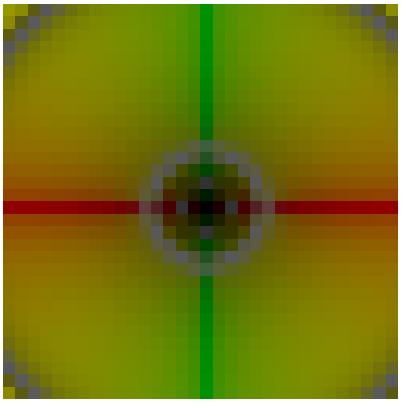
Verhoef, T.; Kirby, S.; and de Boer, B. 2016. Iconicity and the emergence of combinatorial structure in language. *Cognitive Science* 40(8):1922–1940.

Witbrock, M., and Reilly, S. N. 1999. Evolving genetic art. In Bentley, P. J., ed., *Evolutionary Design by Computers*. San Francisco: Morgan Kaufmann. 251–259.

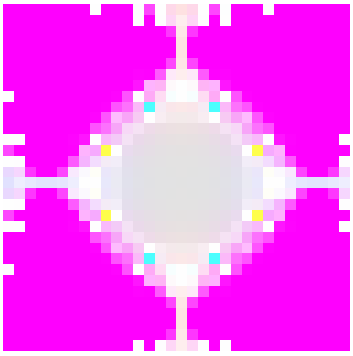
Zhang, A., and Saunders, R. 2012. Towards the evolution of a language for creative design. In *Proc. 2012 IEEE Congress on Evolutionary Computation (CEC)*, 2778–2783.



(a) Evolved AST (Genotype)

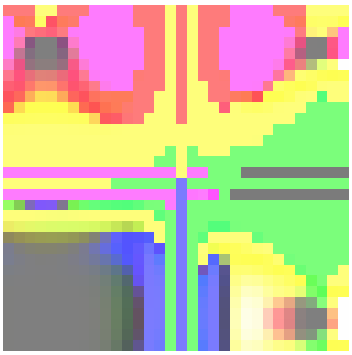


(b) Resulting Artefact (Phenotype)



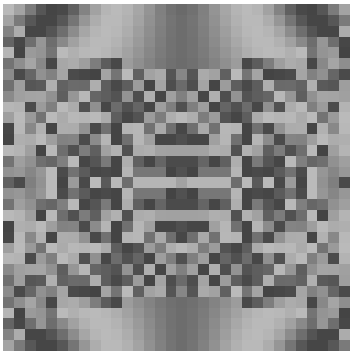
(c) Artefact 1

$$\cosh(\ln(\mathbf{i}\sin(\mathbf{i}\sin(\cos^3(\mathbf{q})))))$$



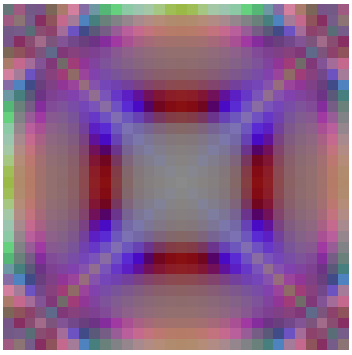
(d) Artefact 2

$$\text{spiral}\left(\left(\text{swirl}(\ln(\mathbf{i}\sin(\mathbf{iq}))) \cdot \text{spiral}(e^{\mathbf{i}\overline{\mathbf{q}}})\right)^3\right)$$



(e) Artefact 3

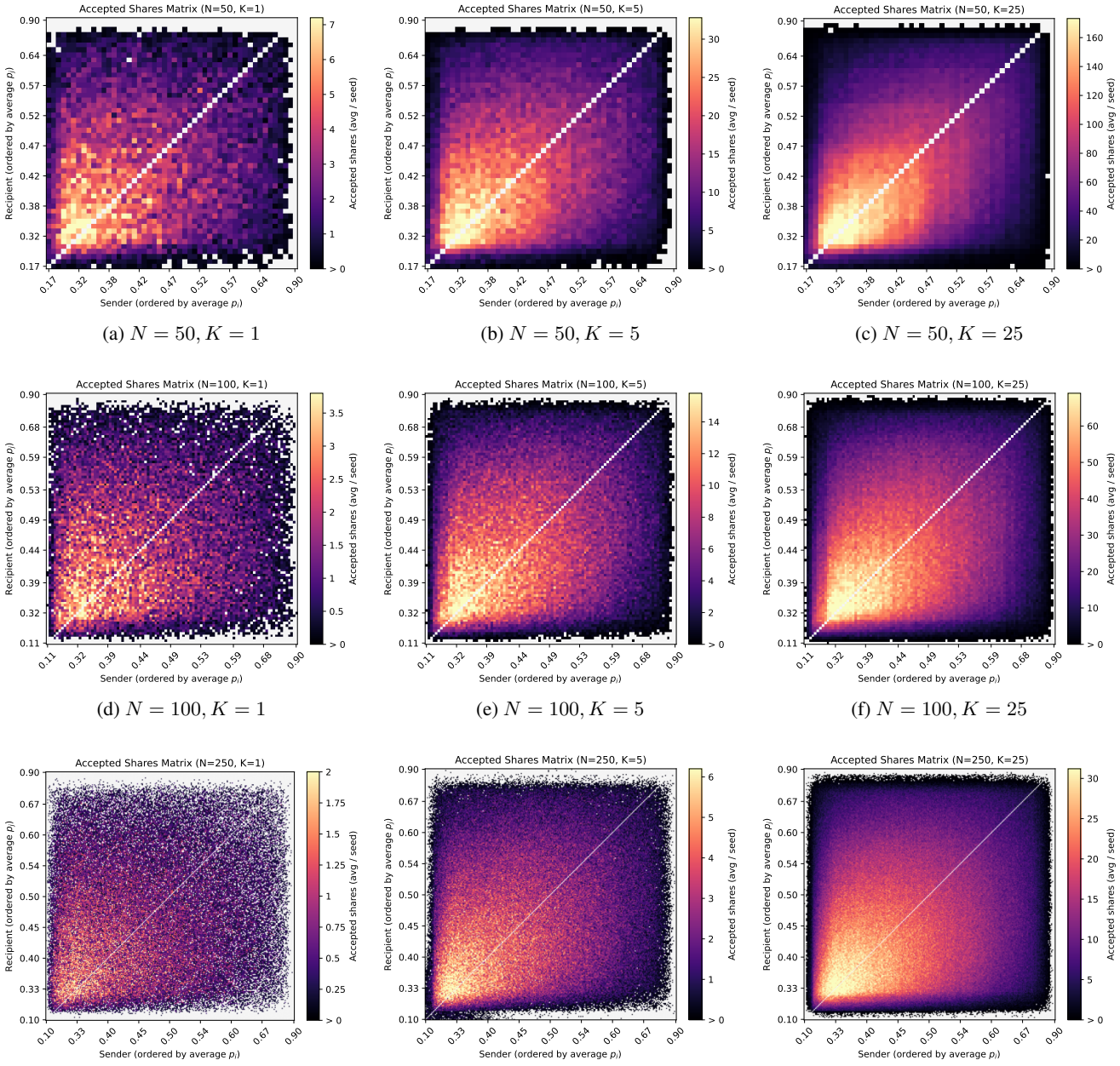
$$\text{ripple}\left(\lfloor\cosh^3(\text{blend}(\mathbf{q},\mathbf{q}))\rfloor - |\sqrt{\mathbf{q}}|^3\right)$$



(f) Artefact 4

$$\tan\left(\sqrt{\sinh\left(\cosh(\overline{\text{rot}_{45}(\mathbf{q})})\right)}\right)$$

Figure 4: Examples of generated artefacts and their underlying quaternion expressions. The top row pairs an Abstract Syntax Tree (genotype) with its rendered image (phenotype). The four lower artefacts are rendered at 32×32 from independently sampled trees of equal maximum depth. The quaternion $\mathbf{q}$ in the expressions denotes the (x, y) pixel coordinate at which the expression is evaluated. Expressions are direct, unsimplified translations of the evolved ASTs, so they may contain redundancies such as $\text{blend}(\mathbf{q}, \mathbf{q})$ or unreduced nestings. These reflect the mechanical structure of subtree crossover rather than algebraic identities. Operators spiral, swirl, ripple, and rot_{45} are spatial transforms on the quaternion input. blend is a per-channel mean of two quaternions.

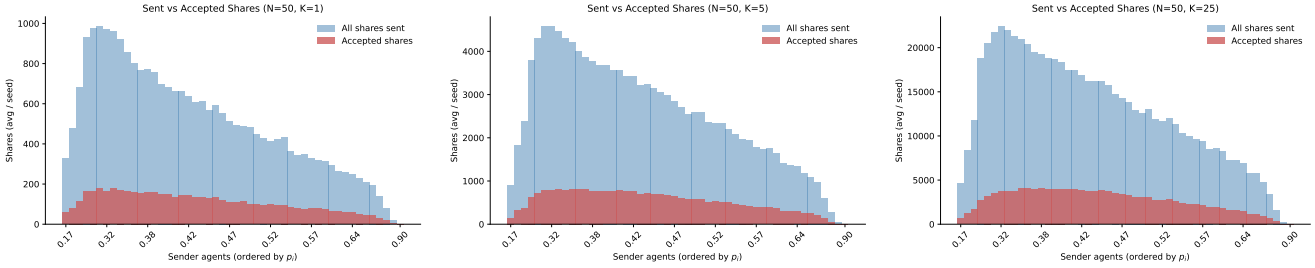


(a) $N = 50, K = 1$ (b) $N = 50, K = 5$ (c) $N = 50, K = 25$

(d) $N = 100, K = 1$ (e) $N = 100, K = 5$ (f) $N = 100, K = 25$

(g) $N = 250, K = 1$ (h) $N = 250, K = 5$ (i) $N = 250, K = 25$

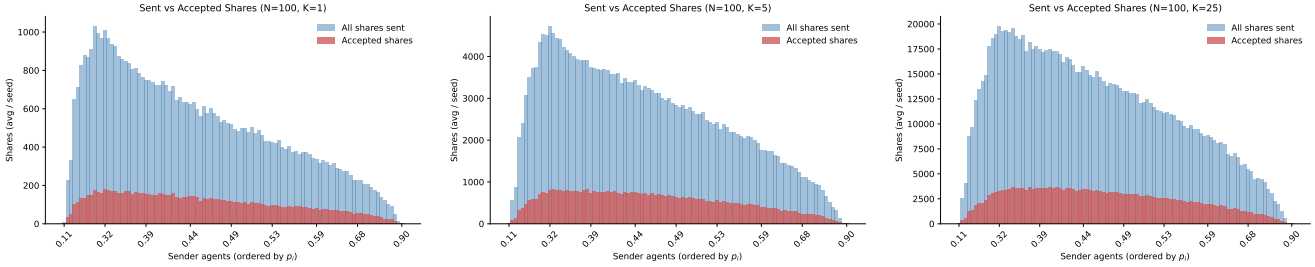
Figure 5: Complete evaluation grid of accepted-share interaction matrices across all experimental configurations. Each cell counts run-averaged accepted shares from sender i (x-axis) to recipient agent j (y-axis), with both axes ordered by preferred novelty (p_i). A white cell represents zero accepted shares across all 5 runs, whereas a black cell represents at least one successful share during one run (therefore the average could be below 1).



(a) $N = 50, K = 1$

(b) $N = 50, K = 5$

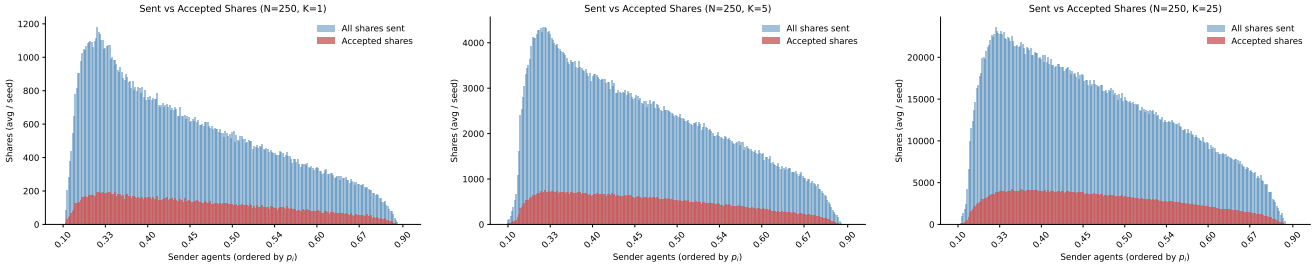
(c) $N = 50, K = 25$



(d) $N = 100, K = 1$

(e) $N = 100, K = 5$

(f) $N = 100, K = 25$

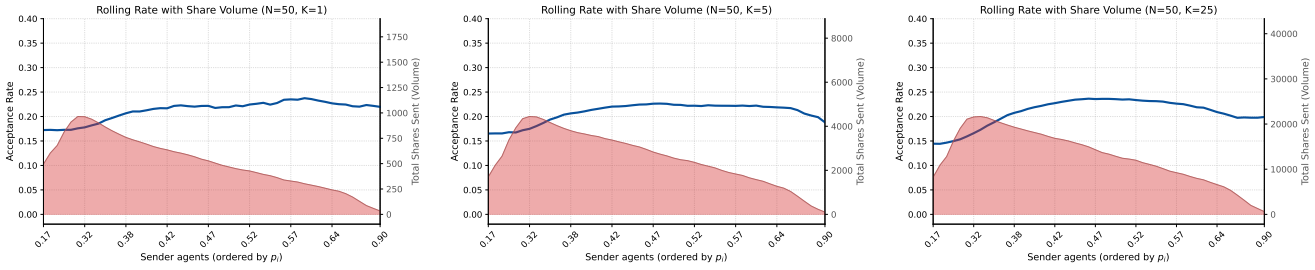


(g) $N = 250, K = 1$

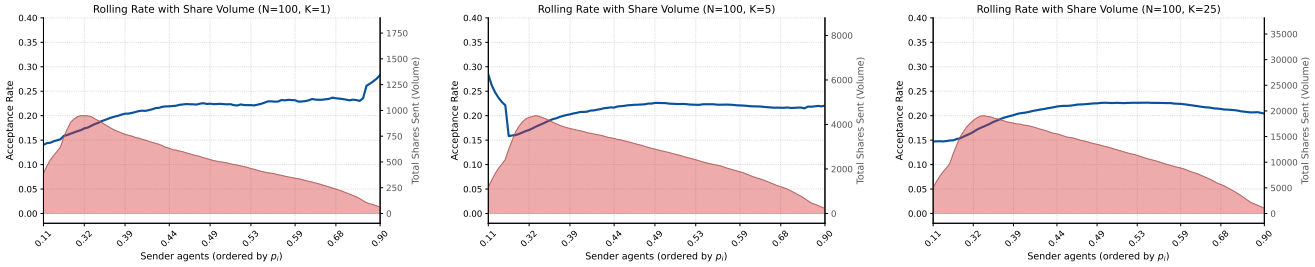
(h) $N = 250, K = 5$

(i) $N = 250, K = 25$

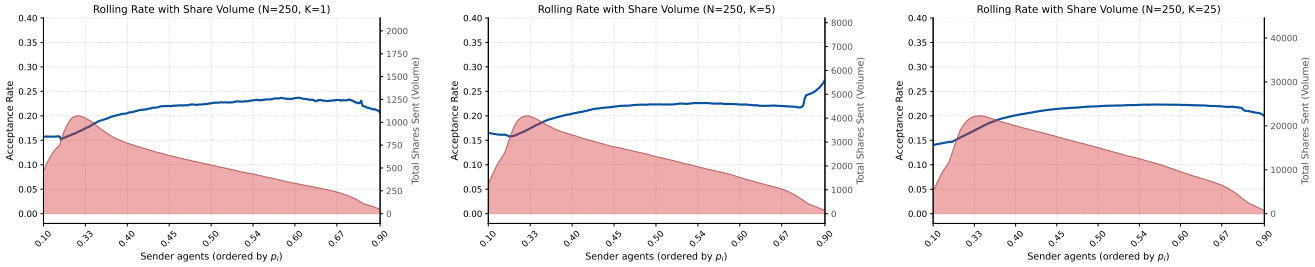
Figure 6: Complete evaluation grid of run-averaged per-sender share volume across all experimental configurations, ordered by individual preferred novelty (p_i). Each subfigure tracks total shared artefacts alongside the accepted subset surviving the recipient-side domain-threshold (τ_D) filter.



(a) $N = 50, K = 1$ (b) $N = 50, K = 5$ (c) $N = 50, K = 25$



(d) $N = 100, K = 1$ (e) $N = 100, K = 5$ (f) $N = 100, K = 25$



(g) $N = 250, K = 1$ (h) $N = 250, K = 5$ (i) $N = 250, K = 25$

Figure 7: Complete evaluation grid of run-averaged per-sender acceptance rates (rolling curves, left axes) overlaying total shares sent (right axes) plotted against individual preferred novelty (p_i). The scaling behaviour underscores the persistence of the rate–volume inversion across all tested population structures and communication capacities. The only anomaly seems to appear at $N = 100$, $K = 5$, where a small group of low-p_i senders achieves a noticeably higher per-share acceptance rate than the rest.

Expressing Authentic Voice within Co-Creative AI Partnerships: Insights from Augmentative and Alternative Communication

Alison Pease and **Tom Griffiths***
School of Science and Engineering, University of Dundee, UK
apease@dundee.ac.uk, TGriffiths001@dundee.ac.uk

Abstract

For a co-creative partnership to be effective, each participant must be able to express themselves authentically and see that expression reflected in the resulting artefacts. In this bridging paper, we draw on insights from the field of Augmentative and Alternative Communication (AAC) to explore what it means to preserve an authentic voice in the era of generative AI. AAC has a long history of considering authorship in communication and how this relates to the authentic construction of the self. We draw on this to propose three recommendations: prioritise authentic expression over efficiency through participatory approaches; adopt personalised, "one size fits one" design; and develop evaluation metrics that capture co-creators' sense of authentic self-expression.

Introduction

As creative AI becomes increasingly sophisticated and widespread, as a community we need to turn our attention to questions such as – what now? How can this technology benefit society and how do we steer it in the right directions to do so? One long-term goal for CC researchers has been to develop co-creative AI partners, motivated by helping to "enable", "support", "facilitate", "enhance", and "democratise" our creativity (this is contrasted with the goal of developing independent, autonomous creative systems). In this paper we focus on one dimension of a rich human–AI co-creative partnership that is often missing in today's culture of semi-passive AI output consumption and prompt-driven artefact creation: the presence of an authentic voice. This is the part of creativity connected to creative self-expression and a sense of shared ownership over what is produced – the feeling that an artefact was created intentionally and, in some way, reflects a core part of the individual and what they wish to communicate or share with others. For a creative partnership to work, each party must be able to express themselves and be listened to, and see that expression meaningfully reflected in the resulting artefacts. Whether or not AI could have its own authentic voice to contribute to the partnership is an open question - explored in (Colton, Pease, and Saunders 2018). Here we focus on the importance of preserving the human authentic voice within mixed initiative co-creative interactions.

The importance of the authentic voice is widely recognised by both creators and those who engage with their work (Fox 2025; McVeigh and Valquaresma 2025). However, to the best of our knowledge, there has been no work in CC itself that directly addresses the role of the authentic voice in co-creative partnerships or how to design for it. In contrast, this lies at the core of Augmentative and Alternative Communication (AAC). Here, researchers focus on how technology can be developed to help to preserve and articulate the voices of people with complex communication needs. The field of AAC has grappled with issues of authorship since its inception (Wickenden 2011; Scarinci, Holm, and Hemsley 2024), but the emergence of generative AI has reframed and intensified these discussions. In this bridging paper, we draw on insights from AAC to suggest how this essential aspect of meaningful co-creative relationships might be addressed within a CC context.

The Authentic Voice in Creative Practice

People enjoy being creative. This has been understood since Aristotle linked creative activity, as a source of meaning, connection, and fulfilment, to his concept of eudaimonia as human flourishing (Wright and Pascoe 2015). More recent studies, for instance (Silvia et al. 2014; Benedek, Bruckdorfer, and Jauk 2020), have found strong links between creativity and enjoyment. Furthermore, an extensive body of research and practice in the therapeutic disciplines explores the relationship between creativity and mental well-being, demonstrating that creative behaviour can positively affect a wide range of psychological and physiological outcomes (Shafir et al. 2020). Creative AI must therefore preserve this sense of enjoyment. Rather than outsourcing much of our creativity to AI, we should design co-creative partnerships that sustain, and ideally enrich, the core qualities of creative experience, which underpin human flourishing and healing.

*Pease is a researcher in CC, specialising in analyses of aspects of human creativity and how these can be represented within a computational context. Griffiths is a linguist, researcher, and a clinician with a background in AAC, whose research explores the intersection between aided communication and AI, with particular reference to how AI impacts on authentic expression. After co-teaching a programme on AI and Society, the authors realised their mutual interest in the authentic voice and how their different fields conceptualise this.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Central among these qualities is the authentic voice.[1]

The authentic voice is emphasised in both pedagogical and creative practice circles. For instance, within the context of screenwriting education, McVeigh and Valquaresma argue that: "Finding your authentic voice and believing in yourself enough to express it is core to creative work" and "... developing an authentic voice is both a pedagogical challenge and a critical component of creative practice as it empowers the writer to tell stories that are unique and impactful." (McVeigh and Valquaresma 2025, p.1). More generally, (Vernon and Paz 2023) present a pedagogical approach in design education centred on creative authenticity, advocating for teaching grounded in creative authenticity to help students develop their strengths through evolving identities. Their work also highlights its broader role in enabling marginalised groups to see their identities and experiences genuinely reflected in education and practice.

In the context of media education, (Fox 2025) reflects on her experiences encouraging students to use AI in audio production, questioning whether it augments creativity or compromises authenticity. She asks, "Is the final product an authentic representation of the student's voice, skill and intent?", finding that co-creative AI tools can obscure the craft of authentic storytelling. She further reflects on ways in which this could overcome, suggesting that students must not only learn to use these tools, but also critically examine their implications.

People also value the authentic voice when engaging with other people's creative artefacts. This can be seen in the betrayal that people sometimes feel when, for example, works published as memoirs are shown to be inaccurate. Raynor Winn's book The Salt Path (2019) is a recent example: this was found to contain several misleading aspects, including how the author became homeless and the medical condition that her husband had (Hadjimatheou 2025). People reacted similarly to the exposure of fabrications in A Million Little Pieces by James Frey (2003) which describes Frey's struggle with drug addiction and rehabilitation. To some extent, creative artefacts embody a discourse between the creator and those who engage with them, where viewers, readers, and listeners want to feel that they have some understanding of the creators' artistic intent and self-expression.

While there is no work directly within CC on human AI co-creativity and the authentic voice, there are a few studies in related fields. Robbins and Blundell raise concerns over whether Generative AI is degrading human expression and causing us to lose our voices (2026). They argue that it is devaluing and discouraging authentic human expression, affecting both consumers and creators. Liu (2025) has conducted an excellent study into how Generative AI is reshaping creative work, and the tensions this is creating around authorship, agency, and authenticity. By examining the lived experiences of creative professionals of using AI for brainstorming, visual exploration, and early-stage drafting, she highlights the experiential tensions at the heart of AI-assisted creativity, particularly in terms of authenticity drift. Her findings suggest that authentic expression in human AI co-creativity depends on maintaining human agency, interpretive involvement, and exploratory engagement, and she proposes designing AI systems that support uncertainty and expression rather than simply optimizing outputs.

[1]The concept of authenticity within a creative context can refer to the provenance of an artefact (known as nominal authenticity (Dutton 2005) or whether an artefact genuinely reflects an author's beliefs and values in a socio-historical context (expressive authenticity (ibid.)). Here, we focus on the latter, the aspect of the creative experience which captures self-expression.

Insights from Augmentative and Alternative Communication

AAC describes a range of devices, strategies and techniques that may be used to support people with communication disabilities, for whom clearly articulated speech is limited or impossible. In its broadest definition (see: Burkhart 2011) AAC covers strategies such as manual signing, physical gesture, and paper-based resources such as letter boards or arrays of graphic symbols used to represent meaning. Of greatest relevance to CC is the group of devices known as voice output communication aids (VOCAs), computer-based systems which turn a user's selection of words or graphic symbols into synthetic speech output. Within the field of assistive technology, AAC is unique in that it serves an expressive purpose, allowing users not only to convey messages, but also to construct and convey their personalities in similar ways to their speaking peers (Waller 2019). Key to this is the discussion of authorship - who is (or is perceived to be) the author of a message, or the source of its originating thought?

Authorship in AAC is complex, and is a long-running discussion in the field - from early concerns that word prediction was "putting words in people's mouths" (Waller 2019, p. 161), to more recent discussions about how choices are made about what symbols are included in pre-made vocabularies for emergent aided communicators (Laubscher and Light 2020), questions of authorship tend to centre on the balance between the user as originator of the message and their VOCA as the channel through which it is constructed and transmitted. This balance is often dynamic, changing according to the needs of an interaction (Weinberg et al. 2025), but communicative interaction is predicated on a shared understanding that each participant is the author of their own contributions, that each person is contributing authentically. Authorship is therefore intrinsically tied to how one's very humanness is manifested and recognised (Dickerson et al. 2002), and its importance is underlined by authors whose work discusses how self-expression and personality are created in VOCA-mediated communication (Preece et al. 2024; Kane et al. 2017).

The advent of readily available generative AI systems, particularly large language models (LLMs), represents an inflection point for the field. Whilst natural language processing techniques have been part of VOCA design for many years (Higginbotham et al. 2012), the shift from probabilistic predictors to generative "co-authors" (Scarinci, Holm, and Hemsley 2024) introduces a potentially disruptive change in how communication is produced. These systems offer new possibilities for flexible communication

strategies, drawing on context and dynamically anticipating conversational needs. At the same time, however, scholars have cautioned against their uncritical integration into VOCAs (Griffiths, Slaughter, and Waller 2024), noting that a focus on efficiency and rate enhancement (Ibrahim et al. 2026) may overlook important concerns around authentic representation. For VOCA users, this raises fundamental questions about how the presence or use of an LLM may reshape authorship in communication, echoing concerns around creative partnership that are central to CC. The capacity of LLMs to generate large volumes of text with minimal effort is potentially seductive, particularly in the context of existing VOCA technologies which generally remain orders of magnitude slower than speech. But the relationship between speed and authorship is a nuanced one, with some researchers highlighting that VOCA users are often willing to tip the balance in time-pressured interactions or contexts (Weinberg et al. 2025), whilst others warn of the risk that users may "defer to the programmed algorithm" (Sellwood et al. 2024, p. 621), potentially sacrificing their own expression for ease and efficiency. More concerning still is the possibility that a VOCA may effectively "talk for" a user, producing fluent and plausible utterances with little or no user input, that are then claimed or perceived as being the true thoughts or expressions of the user (Griffiths et al. 2025). Whilst authorship remains an open question in AAC, recent work has sought to explore the impact of LLMs on authentic communication through the design of systems that seek to foreground and retain a user's previous utterances (Waller, Griffiths, and Kristensson 2025), create personalised language models specific to the user (Weinberg et al. 2026) and which ensure users retain agency over decisions about how and when to use LLM systems (Yang and Kristensson 2023).

Other challenges in how LLMs are being applied in AAC also have parallels with CC. Some researchers have raised the possibility that LLMs may produce outputs that privilege dominant communicative norms or linguistic structures over personal or idiosyncratic forms of expression (Valencia et al. 2023; Griffiths, Slaughter, and Waller 2024) or that they may routinely offer culturally or socially inappropriate outputs, based on algorithmic biases in their training data (Sennott et al. 2019). Training data for LLMs is often "scraped" from large corpuses of written text, reflecting typically non-disabled linguistic and cultural norms, and marginalising non-standard, disabled and other under-represented voices. As with creative forms, there is a risk that their use can only therefore reflect dominant cultural norms, or particularly popular styles of expression. For disabled communicators, this may further entrench biases or ableist assumptions about the way they "should" communicate (Preece et al. 2024). The absence of sufficiently large and representative training datasets reflecting disabled communicators and their forms of expression remains a problem for the field of AAC, although several researchers have proposed solutions to address this crucial gap (Vertanen and Kristensson 2011).

Another facet of AAC that may prove insightful for the CC field is the question of exactly how the VOCA is positioned within an interaction. Similar to the way in which creative AI can be positioned as a support for human creativity, contemporary AAC practice seeks to consider the whole interaction, rather than focus solely on outputs from a device. AAC acknowledges that interactions between people are highly individualised, and that there is not a single, "right way" to conduct an interaction, no single definition of "success". Where we communicate for the purpose of social closeness, or to share our personal thoughts and stories, it becomes even clearer that no two communication partnerships are the same, and thus communication is tailored to the dynamics of that partnership, with an increasing acknowledgement that such communication is inherently enchronic (Enfield 2022), constructed both from the prior knowledge and mutual grounding of participants, as well as building on prior contributions. There are elements of this that an AI could not reasonably hope to emulate, and as such recent work in AAC (Griffiths et al. 2025; Valencia et al. 2023; Ibrahim et al. 2026) has highlighted the risk that too tight a focus on AI's efficiency and linguistic flexibility risks returning the device to the centre of the interaction, shifting the focus back to an outdated "sender-receiver" model of communication.

The challenge for the field of AAC, and for those designing VOCAs in particular, may therefore be one of developing systems that can leverage the considerable possibilities of LLMs while supporting authentic expression for users. In line with work in HCI more broadly, current discourse emphasises participatory methods, and the involvement of users at all stages of the design process (Waller 2024), particularly in shaping how and when LLMs are used, and the extent to which they are permitted to contribute to communicative output.

In this respect, the use of LLMs in VOCAs shares a conceptual basis with CC, in that systems can be understood as lying on a spectrum of communicative authorship and agency. The central challenge is therefore not whether an LLM-augmented VOCA *can* be creative, but how developers and users navigate shifts in the locus of linguistic control, determining how much control can be ceded to an LLM while still preserving authentic authorship. Framed in this way, the goal becomes one of supporting authentic expression without distorting or diluting the user's voice.

Lessons Learned for CC

The field of AAC is currently discussing important questions of how AI, and particularly LLMs, can be responsibly integrated into communication systems in ways that benefit users without diluting their agency. With this paper we propose that CC can gain several important insights from these discussions.

How the technology is best implemented and used is an open question in both fields. In AAC, the current focus is on how users maintain agency, something that is vital if communication is to be perceived and valued as authentic. Similar concerns manifest in CC, where the goal of co-creative AI systems can be seen as supporting human creative practice. For such outputs to be valued, it is important that the creator's voice is not obscured or over-written by the system's contribution. In both fields, where AI systems act as

contributors, co-authors or co-creators, it becomes necessary to account for how agency and authorship are distributed between human and machine. AAC highlights the risks of failing to do so, particularly when communicative acts are attributed to users who have not fully authored them, but it also highlights that this distribution may well be contextually defined, or based on the goals of a particular interaction.

Related to this, another important consideration is that of how "success" is perceived. Successful communication, much like any creative output is co-constructed, and evaluated both by those who produce and those who receive. In communication and creativity, existing ways in which we measure how we judge whether an interaction has been successful may not be adequate to capture the complex creative relationship between a user and an AI. Research in AAC has demonstrated that simple measures of efficiency or output rate cannot fully capture whether an AI-mediated interaction is successful (Valencia et al. 2023).

In both these considerations, it is important to note that AAC has always been a highly individualised intervention. Provision of AAC systems is a process of matching the needs of a communicator with available aided and unaided resources. As a result, no two communication systems are the same, and the assertion is that "one size fits one", as each user brings fundamentally different needs, goals and attitudes to their communication system.

What emerges from this discussion is that the question of how AI can support communicators and creators is not primarily a technical one; nevertheless, recent work (Ibrahim et al. 2026) indicates that the inclusion of AI in AAC is still largely framed within the design literature as a technical problem. We propose that there is a need to shift focus to how AI can support participation and interaction, rather than purely focusing on "fixes" such as efficiency gains. In CC, there may be a corresponding need to shift the focus towards how AI systems can be properly positioned as co-creators, rather than systems that create based on human prompting. The involvement of users in the design of AI systems is of paramount importance here: we must work with users to properly understand the problem space, and to define how agency, authorship, and control should be distributed within AI-mediated creativity.

Whilst there are many lessons to be learned here, and the question of authenticity is still an open one in both fields, we summarise our learning in three key recommendations:

1. Using participatory methods, shift the focus of CC research from a focus on technological efficiency towards supporting the expression of authentic voice, exploring how this can be realised within co-creative human–AI partnerships.
2. Consider the philosophy that "one size fits one", rather than "one size fits all". This applies similarly in a CC context: creators look for different things in a creative partner, and the design and evaluation of CC systems should reflect this.
3. Develop new evaluation metrics which capture the experiences of human co-creators in terms of whether they feel that they've been able to express themselves authentically.

Conclusion

There are clear parallels in how authentic voice is understood in both AAC and CC. AAC explores how meaning in conversation can be co-constructed among participants, each contributing authentically; similarly, CC explores how creative processes can be co-constructed between collaborators, with each participant contributing authentically. In the era of co-creative AI, both fields face shared challenges in developing meaningful ways to work with these systems while preserving authentic expression and a sense of authorship.

This discussion is already well-established in AAC, and there is a long tradition of careful attention to the role of authorship in communication, of considering what it means to be the author of one's own message and to express oneself authentically. In this paper, we propose that CC can learn from AAC's understanding of authorship, particularly as realised through personalised communication tools and co-constructed interactions. We argue that both fields stand to benefit from one another, and that greater interdisciplinary collaboration and dialogue would be mutually valuable.

References

[Benedek, Bruckdorfer, and Jauk 2020] Benedek, M.; Bruckdorfer, R.; and Jauk, E. 2020. Motives for creativity: Exploring the what and why of everyday creativity. *The Journal of Creative Behavior* 54(3):610–625.

[Burkhart 2011] Burkhart, L. 2011. ISAAC – What is AAC?

[Colton, Pease, and Saunders 2018] Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of Authenticity in Autonomously Creative Systems. In *Proceedings of the 9th International Conference on Computational Creativity (ICCC 2018)*, 272–279. Salamanca, Spain: Association for Computational Creativity (ACC).

[Dickerson et al. 2002] Dickerson, S. S.; Stone, V. I.; Panchura, C.; and Usiak, D. J. 2002. The Meaning of Communication: Experiences with Augmentative Communication Devices. *Rehabilitation Nursing* 27(6):215–220.

[Dutton 2005] Dutton, D. 2005. Authenticity in art. In Levinson, J., ed., *The Oxford handbook of aesthetics*. Oxford University Press.

[Enfield 2022] Enfield, N. J. 2022. Enchrony. *WIREs Cognitive Science* 13(4).

[Fox 2025] Fox, K. 2025. Reflection-ai: augmenting creativity or compromising authenticity? reflections on using generative AI in audio education. *Frontiers in Communication* 10:1613254.

[Griffiths et al. 2025] Griffiths, T.; Broomfield, K.; Hrastelj, L.; Judge, S.; and Toogood, J. 2025. AI, communication aids and the challenge of authentic authorship – Whose line is it anyway?! *Journal of Enabling Technologies* 19(2):102–112.

[Griffiths, Slaughter, and Waller 2024] Griffiths, T.; Slaughter, R.; and Waller, A. 2024. Use of artificial intelligence (AI) in augmentative and alternative communication (AAC): community consultation on risks, benefits and the need for a code of practice. *Journal of Enabling Technologies* 18(4):232–247.

[Hadjimatheou 2025] Hadjimatheou, C. 2025. The real salt path: how a blockbuster book and film were spun from lies, deceit and desperation.

[Higginbotham et al. 2012] Higginbotham, D. J.; Lesher, G. W.; Moulton, B. J.; and Roark, B. 2012. The Application of Natural Language Processing to Augmentative and Alternative Communication. *Assistive Technology* 24(1):14–24.

[Ibrahim et al. 2026] Ibrahim, S. B.; Griffiths, T.; Clarke, M.; Judge, S.; Slovak, P.; Pullin, G.; and Higginbotham, J. 2026. Before the Technological Fix: Scoping AI and AAC for Social Futures. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–10. Barcelona, Spain: ACM.

[Kane et al. 2017] Kane, S. K.; Morris, M. R.; Paradiso, A.; and Campbell, J. 2017. "At times avuncular and cantankerous, with the reflexes of a mongoose": Understanding Self-Expression through Augmentative and Alternative Communication Devices. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1166–1179. Portland Oregon USA: ACM.

[Laubscher and Light 2020] Laubscher, E., and Light, J. 2020. Core vocabulary lists for young children and considerations for early language development: a narrative review. *Augmentative and Alternative Communication* 36(1):43–53.

[Liu 2025] Liu, L. 2025. Reconceptualizing innovation in human-ai creative collaboration: Balancing agency, machine contribution, and authenticity. *Machine Contribution, and Authenticity (December 17, 2025)*.

[McVeigh and Valquaresma 2025] McVeigh, M., and Valquaresma, A. 2025. Creativity and screenwriting: authenticity and finding your voice. *Media Practice and Education* 1–15.

[Preece et al. 2024] Preece, J.; Sullivan, E.; Tams-Gray, F.; and Pullin, G. 2024. Making my voice and owning its future. *Medical Humanities* 50(4):624–634.

[Robbins and Blundell 2026] Robbins, S., and Blundell, I. 2026. Losing our voice? generative ai and the degradation of human expression. *Minds and Machines* 36(1):2.

[Scarinci, Holm, and Hemsley 2024] Scarinci, N.; Holm, A.; and Hemsley, B. 2024. Ethical conversations: what are the implications of AI for AAC development, use, and implementation? *Journal of Clinical Practice in Speech-Language Pathology* 26(1):101–107.

[Sellwood et al. 2024] Sellwood, D.; McLeod, L.; Williams, K.; Brown, K.; and Pullin, G. 2024. Imagining alternative futures with augmentative and alternative communication: a manifesto. *Medical Humanities* 50(4):620–623.

[Sennott et al. 2019] Sennott, S. C.; Akagi, L.; Lee, M.; and Rhodes, A. 2019. AAC and Artificial Intelligence (AI). *Topics in Language Disorders* 39(4):389–403.

[Shafir et al. 2020] Shafir, T.; Orkibi, H.; Baker, F. A.; Gussak, D.; and Kaimal, G. 2020. The state of the art in creative arts therapies. *Frontiers in psychology* 11:68.

[Silvia et al. 2014] Silvia, P. J.; Beaty, R. E.; Nusbaum, E. C.; Eddington, K. M.; Levin-Aspenson, H.; and Kwapil, T. R. 2014. Everyday creativity in daily life: An experience-sampling study of "little c" creativity. *Psychology of Aesthetics, Creativity, and the Arts* 8(2):183.

[Valencia et al. 2023] Valencia, S.; Cave, R.; Kallarackal, K.; Seaver, K.; Terry, M.; and Kane, S. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM Press.

[Vernon and Paz 2023] Vernon, Z., and Paz, A. 2023. Creative authenticity: a framework for supporting the student self in design education. *International Journal of Art & Design Education* 42(1):6–15.

[Vertanen and Kristensson 2011] Vertanen, K., and Kristensson, P. O. 2011. The imagination of crowds: Conversational AAC language modeling using crowdsourcing and large data sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 700–711. Edinburgh, Scotland, UK: Association for Computational Linguistics.

[Waller, Griffiths, and Kristensson 2025] Waller, A.; Griffiths, T.; and Kristensson, P. O. 2025. 'What I Think about When I Type about Talking': Reflections on Text-Entry Acceleration Interfaces. *Communications of the ACM* 68(5):58–68.

[Waller 2019] Waller, A. 2019. Telling Tales: Unlocking the Potential of AAC Technologies. *International Journal of Language & Communication Disorders* 54(2):159–169.

[Waller 2024] Waller, A. 2024. Technical Perspective: The User as a Key Ingredient in AAC Design. *Communications of the ACM* 67(1):122–122.

[Weinberg et al. 2025] Weinberg, T. M.; Kadoma, K.; Gonzalez Penuela, R. E.; Valencia, S.; and Roumen, T. 2025. Why so serious? Exploring timely humorous comments in AAC through AI-powered interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25.

[Weinberg et al. 2026] Weinberg, T. M.; Penuela, R. E. G.; Valencia, S.; and Roumen, T. 2026. I, Robot? Exploring Ultra-Personalized AI-Powered AAC; an Autoethnographic Account. arXiv:2509.13671 [cs].

[Wickenden 2011] Wickenden, M. 2011. Whose Voice is That?: Issues of Identity, Voice and Representation Arising in an Ethnographic Study of the Lives of Disabled Teenagers who use AAC. *Disability Studies Quarterly* 31(4).

[Wright and Pascoe 2015] Wright, P. R., and Pascoe, R. 2015. Eudaimonia and creativity: the art of human flourishing. *Cambridge Journal of Education* 45(3):295–306.

[Yang and Kristensson 2023] Yang, B., and Kristensson, P. O. 2023. A Demonstration of a Tinkerable Augmentative and Alternative Communication Keyboard. In *28th International Conference on Intelligent User Interfaces*, 138–140. Sydney NSW Australia: ACM.

Before the Prompt: Heteromorphic Imagination and the Cognitive Cost of Premature Crystallization in Human-AI Co-Creativity

Emil Polyak and **Rghad Balkhyoor**
Drexel University
Philadelphia, PA, USA
ep557@drexel.edu, rab422@drexel.edu

Abstract

Generative AI systems can enhance individual creative output while reducing the collective diversity of produced works. Current explanations for this paradox focus primarily on artifacts, model distributions, or design fixation. This paper proposes a complementary cognitive-phenomenological account grounded in the distinction between two modes of imaginative cognition: heteromorphic imagination, characterized by pre-articulatory associative play across real and unreal combinatory spaces, and creative imagination, in which emergent patterns achieve sufficient resonance to warrant articulation. We argue that prompt-based interaction imposes a demand for premature crystallization: a pressured translation from heteromorphic exploration into linguistically structured, model-interpretable expression. This demand can incentivize bypassing a cognitive phase that helps generate semantically remote, personally resonant, and meaning-bearing possibilities before they stabilize as concepts. Drawing on work on the Default Mode Network, divergent thinking, and dynamic switching between spontaneous and controlled cognition, we frame the neural component as a falsifiable hypothesis rather than an established effect. The resulting implication for computational creativity is that evaluation of human-AI co-creative systems should extend beyond outputs to upstream human cognitive processes and their possible long-term modification.

Introduction

A growing body of research has documented a paradox at the intersection of generative artificial intelligence and human creativity. Access to large language model outputs can enhance the perceived creativity of individual works, with benefits especially pronounced among less creative writers, while collective diversity decreases when AI assistance is introduced (Doshi and Hauser 2024). A reanalysis of brainstorming experiments similarly found that ChatGPT-aided conditions yielded less diverse idea pools across several embedding-based diversity measures (Meincke, Nave, and Terwiesch 2025). Exploratory work in engineering design education points in a related direction, suggesting premature convergence in AI-assisted teams, though with limited statistical power (Tsakalerou et al. 2026).

Existing explanations tend to operate at the level of the artifact or the model. Exposure to AI-generated images during visual ideation increases design fixation and reduces the fluency, variety, and originality of sketches compared with no-support and conventional image-search baselines (Wadinambiarachchi et al. 2024). Yet these accounts leave a prior question underexamined: what happens to the human cognitive processes that precede creative expression when the primary mode of interaction with a creative tool demands linguistic articulation as the point of entry?

This paper proposes that the answer requires attention to a distinct phase of imaginative cognition, which we have termed *heteromorphic imagination* (Polyak and Balkhyoor 2024). We argue that prompt-based AI interaction imposes a structural demand for *premature crystallization*: the translation of nascent, pre-articulatory ideation into language sufficiently explicit for a generative system to interpret. This demand can incentivize bypassing the heteromorphic phase, with potential consequences for the cognitive infrastructure that sustains meaning-bearing creative origination.

Heteromorphic Imagination as a Use-Dependent Capacity

Two Modes of Imaginative Cognition

Prior work distinguished two phenomenologically distinct modes of imaginative cognition (Polyak and Balkhyoor 2024). *Heteromorphic imagination* refers to the fluid play of mental associations, episodic memories, sensory traces, affects, and speculative projections that intermingle before they become stable enough to name. It operates across real and unreal combinatory spaces and resists premature categorization. Casey described each act of imagining as instantaneous, self-contained, and given all at once: an evanescent "mini-world" operating in "sheer time" rather than clock time (Casey 1976). Heteromorphic imagination extends this insight from discrete imaginative acts to their contamination of one another, where remembered, perceived, and speculative scenes mingle across temporal registers. The account also draws on phenomenological treatments of imagination, perception, and embodied projective possibility (Sartre 2004; Merleau-Ponty 2012).

Creative imagination, by contrast, emerges from heteromorphic imagination when a configuration becomes sufficiently coherent, novel, and personally significant to war-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

rant capture and articulation. We call the transition between these modes a *metamorphic gesture*: a moment in which pre-articulatory associative play becomes directed toward expression. This framework differs from Boden's taxonomy of combinational, exploratory, and transformational creativity (Boden 2004). Boden's exploratory creativity presupposes movement within a defined conceptual space. Heteromorphic imagination names activity before such a space is necessarily bounded: the pre-articulatory interval in which episodic memory, sensory experience, counterfactual projection, and affective salience are still reorganizing.

Relation to Neighboring Constructs

Freud's primary-process thinking is the closest historical antecedent: heteromorphic imagination shares its tolerance for condensation, displacement, imagistic association, and non-propositional recombination, but it is not an account of dream-work, wish fulfillment, or drive discharge (Freud 1953). Divergent thinking research, by contrast, usually measures stabilized responses after reportable ideas have already formed; our term names the less reportable interval in which those possibilities are still reorganizing. Heteromorphic imagination is also compatible with contextual focus and honing theory. These accounts explain how creators shift between associative and analytic modes and recursively restructure a worldview around unresolved or high-entropy material (Sowden, Pringle, and Gabora 2015; Gabora 2017). The narrower contribution here is to identify a specific vulnerability created by prompt-based co-creative systems: they require a linguistic handle before this restructuring has necessarily matured.

The loss is the possible narrowing of semantically remote and meaning-bearing connection-making: the slow emergence of a felt, personally salient configuration before it becomes a stable instruction, prompt, or plan. Gendlin's process philosophy is useful here because it treats meaning as something that can be implicit, bodily, and not yet fully formed in propositions (Gendlin 2018). In this respect, heteromorphic imagination also intersects with accounts of inner speech, dialogic thought, and the gradual verbal mediation of cognition (Vygotsky 1987; Fernyhough 2008; 2009), while retaining a focus on what precedes or exceeds verbal form.

Heteromorphic imagination marks a zone that many theories presuppose but do not isolate at the level of interface design: the interval between diffuse associative movement and the first stable handle by which a creator can ask a system for help. Honing theory explains why creative work may proceed as recursive restructuring of an unresolved worldview; contextual focus explains how attention can shift between associative and analytic modes; divergent thinking tasks measure some reportable consequences of this activity. Heteromorphic imagination adds a design-relevant question: does the tool invite the creator to remain within this unresolved, pre-articulatory movement, or does it require the creator to crystallize a handle before the movement has matured?

Use-Dependent Neural Infrastructure

If heteromorphic imagination is a cognitive capacity, it is likely use-dependent. Neuroscience identifies the Default Mode Network (DMN) as centrally involved in spontaneous cognition, autobiographical simulation, and internally generated association. Creative idea production depends on dynamic coupling and switching between default and executive systems (Beaty et al. 2015; Chen et al. 2025). Causal evidence further indicates that disrupting DMN activity can impair originality in divergent thinking (Bartoli et al. 2024). While these findings do not prove that prompt-based AI use alters the DMN, they support the weaker claim that creative cognition relies on dynamic coordination between spontaneous and controlled systems, and that systematic changes in practice may matter for the capacities creators habitually exercise.

Training studies in other domains show that repeated practice can induce measurable structural change in the adult brain (Draganski et al. 2004). If creative practice increasingly routes early ideation through prompt formulation, creators may exercise prompt selection, evaluation, and linguistic refinement while exercising heteromorphic exploration less. Over time, such displacement could attenuate the cognitive routines that support pre-articulatory association. This remains a falsifiable longitudinal hypothesis.

The likely mechanism is attentional and procedural: users may gradually become better at converting vague intentions into promptable requests, while becoming less practiced at tolerating ambiguity, deferral, and internally generated associative drift. Such a shift could be adaptive for production environments and still costly for some forms of creative origination. The same user might produce more acceptable artifacts in less time while generating fewer idiosyncratic starting points, fewer delayed discoveries, or fewer ideas whose value only becomes apparent after a period of nonverbal incubation.

Premature Crystallization and Double Compression

Prompt-based interaction with generative AI requires a user to convert an emerging intention into language before the system can respond. For many creative tasks this is useful: language can focus attention, externalize possibilities, and become an exploratory medium in its own right. The risk arises when linguistic formulation is required too early. At that point, the creator must select a describable version of a not-yet-stable configuration, thereby replacing heteromorphic exploration with an instruction the system can process.

This process entails a double compression. First, pre-articulatory experience is compressed into natural language. Affective tone, bodily orientation, sensory atmosphere, and associative ambiguity must be reduced to describable content. Second, ordinary language is compressed into model-interpretable language: a prompt optimized for the probable behavior of the system. Beyond expressing an idea, the user learns to express it in ways likely to elicit productive completions. The prompt therefore potentially does more than communicate a prior intention, because it may reshape what

counts as a viable intention.

The important feature of this compression is that it occurs before the system's output appears. A writer who asks for "a haunting but hopeful scene about memory," a designer who asks for "futuristic organic furniture," or a composer who asks for "melancholic cinematic chords" has already selected a public, genre-recognizable, and model-legible route through a much less stable field of possibility. The AI response may then feel like expansion, because it provides fluency, detail, and variation. At the level of the user's prior cognitive search, however, the route may already have narrowed. This helps explain why individual judgments of quality and collective measures of diversity can diverge: the artifact is elaborated, while the upstream space of origination has been compressed.

This account complements work arguing that generative AI creativity is disembodied and constrained by the prompt as a thin channel of expression (Casacuberta and Guersenzvaig 2025). It also explains why AI assistance can increase productivity while narrowing diversity. If many users crystallize early ideas into similar promptable formulations, the resulting outputs may converge before the model produces anything. Alongside the model's distribution, the homogenizing force includes the cognitive adaptation users undergo when they learn to think in prompt-shaped units.

Poets, theorists, and prompt writers often discover ideas by writing them. Language itself can do exploratory work. Heteromorphic imagination is also distinct from nonverbal cognition. In music, dance, visual art, and design, crystallization may occur in sonic, motor, spatial, or material form without verbal description. A melody, gesture, or visual configuration can cease to be heteromorphic once it stabilizes as an auditory-motor or perceptual intention, even if it has never been named. In visual art, for example, the exploratory line or mark may carry thought forward before it becomes a concept, a point emphasized by phenomenological accounts of perception and painting (Merleau-Ponty 1993). The specific risk analyzed here is narrower: interfaces that require creators to translate an unstable configuration into model-interpretable instructions before they have had a chance to reach a metamorphic gesture. This risk is likely strongest in early ideation, novice practice, time-pressured production, and low-commitment generation; expert creators may avoid it by using AI only after self-directed exploration has produced a sufficiently formed intention.

Prompting can be beneficial in many cases: collaborative coordination, rapid prototyping, constraint satisfaction, and tasks where the creator already knows the intended direction. If a creative environment repeatedly rewards promptable fragments at the expense of slower heteromorphic development, then creators may adapt to that environment. The relevant design question is therefore when and through what interface AI enters the creative process.

Implications for Computational Creativity

Computational creativity has long recognized that evaluating creative products alone is insufficient. Colton argued that creative process and the perception of creativity must also be considered (Colton 2008), and Jordanous developed a multidimensional procedure for evaluating creative systems (Jordanous 2012). Our argument extends this process orientation to the upstream cognition of the human participant in co-creative systems. Current evaluation methods tend to assess what the human-AI dyad produces, and sometimes how the interaction unfolds, but rarely what the interaction does to the human's independent capacity for ideation.

This suggests the need for *longitudinal cognitive impact assessment*: studies that track changes in human creative capacity over extended engagement with prompt-based tools, using measures sensitive to divergent, associative, and recombinatory processes. Three testable predictions follow. First, early-prompt conditions should increase short-term fluency or elaboration while reducing semantic distance, intra-group diversity, or delayed originality relative to delayed-prompt conditions. Second, the effect should be strongest for novices, time-pressured creators, and tasks where users have not yet formed a stable intention. Third, interfaces that postpone linguistic prompting through sketches, fragments, sound, gesture, ambiguous stimuli, or staged incubation should preserve more heteromorphic exploration than linear chat interfaces requiring immediate declarative instruction.

These predictions could be tested without assuming that heteromorphic imagination is directly observable. Experiments could compare immediate-prompt, delayed-prompt, and non-prompt multimodal conditions; measure semantic dispersion, transformation distance, fixation, and delayed recall of self-generated alternatives; and include follow-up tasks after repeated use. Qualitative protocols could examine whether creators describe the AI as amplifying a prior intention, replacing an uncertain one, or teaching them to formulate intentions in prompt-shaped terms. The crucial dependent variable would be the evolution of the creator's ideational repertoire across time.

If the heteromorphic phase contributes to meaning-bearing creative origination, then tools that sustain rather than bypass this phase would be important. Systems that present ambiguous, incomplete, multimodal, or contradictory stimuli, rather than demanding precise linguistic input, might support associative and speculative exploration. Phenomenological study of one such interface, an ambiguous tactile input device used by twenty designers, describes how this kind of input can sustain imaginative flow during early ideation (Balkhyoor 2026). Other possibilities include interfaces that accept fragments rather than instructions, preserve rejected paths, encourage incubation before generation, visualize divergence across a session, or deliberately return outputs that resist the user's most conventional formulation. The goal is delayed articulation, allowing the creator time to dwell with unstable possibilities before forced articulation.

Many current systems optimize for responsiveness: the user says what is wanted, and the system supplies candidates quickly. A heteromorphic perspective suggests that some valuable systems may instead optimize for latency, ambiguity, and medium-sensitive exploration. Such tools would protect the upstream conditions under which the user can continue to become a source of unexpected meaning.

Limitations and Scope

The central limitation of this paper is that heteromorphic imagination is proposed as an explanatory construct rather than reported as a directly measured variable. It must be operationalized carefully in future work. No single behavioral measure will capture it. A plausible empirical strategy would triangulate among semantic distance, delayed originality, fixation, self-reported incubation, process logs, and qualitative accounts of whether a user began from a felt but unstable direction or from an already stabilized request.

A second limitation concerns individual differences. Creators vary in their vulnerability to premature crystallization. Experienced artists, writers, designers, and composers often possess disciplined ways of postponing closure. They may use AI late in a process, as an externalization tool, a critic, or a variation engine, rather than as the source of first ideational movement. Conversely, novices may benefit from AI scaffolding while also being more likely to adopt the system's categories as their own. This double possibility is central: the same technology may expand access and reduce intimidation while also training some users toward more generic forms of ideation.

A third limitation is domain specificity. Prompt-based text systems make the problem especially visible, but the phenomenon is not restricted to language. Any interface that asks for early specification may produce premature crystallization, including menu-driven design tools, style-transfer systems, or music generators that require genre labels before exploration. Conversely, some language interfaces may avoid the problem if they support fragments, hesitation, contradiction, and revision without immediately optimizing toward polished output. The argument is therefore about the timing and structure of mediation.

Conclusion

The rapid integration of generative AI into creative practice raises questions beyond output quality and authorship. This paper has argued that prompt-based creative interaction may carry a cognitive cost that current computational creativity evaluation frameworks do not capture: the potential attenuation of heteromorphic imagination, the pre-articulatory, associative, and recombinatory mode through which not-yet-stable possibilities can acquire personal and expressive meaning. The double compression of prompt-based ideation, from pre-articulatory experience to language and from language to model-interpretable language, creates a structural incentive to bypass that mode. Although language, prompting, and AI assistance carry clear creative value, the argument here specifies a boundary condition: when articulation becomes the compulsory entry point to creative support, the pre-articulatory ecology of imagination may be narrowed. The claim is deliberately framed as a testable hypothesis. Beyond what machines can produce with humans, computational creativity must attend to what happens in the human mind before the prompt is written.

Author Contributions

Emil Polyak developed the central theoretical framework and drafted the manuscript. Rghad Balkhyoor contributed to the phenomenological framing and manuscript revision.

References

Balkhyoor, R. 2026. *Exploring Creative Imagination: Mapping the Progression of Metamorphic Gestures through an Ambiguous Tactile Interface*. Ph.D. Dissertation, Drexel University.

Bartoli, E.; Devara, E.; Dang, H. Q.; Rabinovich, R.; Mathura, R. K.; Anand, A.; Pascuzzi, B. R.; Adkinson, J.; Kenett, Y. N.; Bijanki, K. R.; Sheth, S. A.; and Shofty, B. 2024. Default mode network electrophysiological dynamics and causal role in creative thinking. *Brain* 147(10):3409–3425.

Beaty, R. E.; Benedek, M.; Kaufman, S. B.; and Silvia, P. J. 2015. Default and executive network coupling supports creative idea production. *Scientific Reports* 5:10964.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.

Casacuberta, D., and Guersenzvaig, A. 2025. Disembodied creativity in generative AI: Prima facie challenges and limitations of prompting in creative practice. *Frontiers in Artificial Intelligence* 8:1651354.

Casey, E. S. 1976. *Imagining: A Phenomenological Study*. Bloomington: Indiana University Press.

Chen, Q.; Kenett, Y. N.; Cui, Z.; Takeuchi, H.; Fink, A.; Benedek, M.; Zeitlen, D. C.; Zhuang, K.; Lloyd-Cox, J.; Kawashima, R.; Qiu, J.; and Beaty, R. E. 2025. Dynamic switching between brain networks predicts creative ability. *Communications Biology* 8:54.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems*.

Doshi, A. R., and Hauser, O. P. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10(28):eadn5290.

Draganski, B.; Gaser, C.; Busch, V.; Schuierer, G.; Bogdahn, U.; and May, A. 2004. Neuroplasticity: Changes in grey matter induced by training. *Nature* 427(6972):311–312.

Fernyhough, C. 2008. Getting Vygotskian about theory of mind: Mediation, dialogue, and the development of social understanding. *Developmental Review* 28(2):225–262.

Fernyhough, C. 2009. Dialogic thinking. In Winsler, A.; Fernyhough, C.; and Montero, I., eds., *Private Speech, Executive Functioning, and the Development of Verbal Self-Regulation*. Cambridge University Press. 42–63.

Freud, S. 1953. *The Interpretation of Dreams*, volume 4–5 of *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. London: Hogarth Press and the Institute of Psycho-Analysis.

Gabora, L. 2017. Honing Theory: A complex systems framework for creativity. *Nonlinear Dynamics, Psychology, and Life Sciences* 21(1):35–88.

Gendlin, E. T. 2018. *A Process Model*. Northwestern University Press.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Meincke, L.; Nave, G.; and Terwiesch, C. 2025. ChatGPT decreases idea diversity in brainstorming. *Nature Human Behaviour* 9(6):1107–1109.

Merleau-Ponty, M. 1993. Eye and mind. In Johnson, G. A., ed., *The Merleau-Ponty Aesthetics Reader*. Northwestern University Press. 121–149.

Merleau-Ponty, M. 2012. *Phenomenology of Perception*. Routledge.

Polyak, E., and Balkhyoor, R. 2024. The creative imagination: Tracing phenomenological gestures across inner worlds. In *The Creative Gesture: International and Interdisciplinary Symposium*.

Sartre, J.-P. 2004. *The Imaginary: A Phenomenological Psychology of the Imagination*. Routledge.

Sowden, P. T.; Pringle, A.; and Gabora, L. 2015. The shifting sands of creative thinking: Connections to dual-process theory. *Thinking & Reasoning* 21(1):40–60.

Tsakalerou, M.; Akhmadi, S.; Balgynbayeva, A.; and Kumisbek, Y. 2026. AI-assisted design synthesis and human creativity in engineering education. *Frontiers in Artificial Intelligence* 9:1714523.

Vygotsky, L. S. 1987. Thinking and speech. In Rieber, R. W., and Carton, A. S., eds., *The Collected Works of L. S. Vygotsky*, volume 1. Plenum Press. 39–285. Original work published 1934.

Wadinambiarachchi, S.; Kelly, T.; Pareek, S.; Zhou, Q.; and Velloso, E. 2024. The effects of generative AI on design fixation and divergent thinking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA: ACM.

Harnessing Quantitative Creativity to Explore the Uncanny Valley of Synthetic Faces

Aditi Ramaswamy
aditi.ramaswamy@kcl.ac.uk
and **Hana Chockler**
hana.chockler@kcl.ac.uk

Abstract

AI-generated, or synthetic, faces have proliferated across the Internet in recent years, and have become increasingly difficult to distinguish from real facial images. Even state-of-the-art synthetic image detection models have not been highly accurate, achieving F1 scores below 0.9, and generally do not follow human-interpretable logic or produce explanations that can be understood by the average Internet user. Humans also generally perform poorly at determining whether a given face is synthetic or real, and often make their decisions based on gut feeling (the "uncanny valley" effect) rather than explainable logic. However, an approach predicated on a task-oriented definition of computational creativity could provide insight into differentiating factors between synthetic and real faces that can explain the origin of the uncanny valley effect and be understood more easily by laypeople. To study this, we propose a preliminary experiment that we aim to expand into a full paper.

Introduction

Facial images synthesized by generative AI models, henceforth known as synthetic facial images, have grown increasingly common over the last few years, and have experienced a drastic spike in quality (Dunn et al. 2026). The faces produced by large models such as FLUX (Labs et al. 2025) are of such realism that at first glance they harbor no visible features that signal their AI origin.

We have found that state-of-the-art synthetic facial image detection models tend to lag behind the newest architectures, with accuracy rates dropping on more lifelike images and those produced by architectures beyond the standard Generative Adversarial Networks (GANs). To justify this claim, we ran a series of open-access synthetic image detection models on 7997 synthetic and 7000 real images taken from the following models: StarGAN (Choi et al. 2017), Taming Transformer (Esser, Rombach, and Ommer 2020), Dual Condition Diffusion (Kim et al. 2023), and FLUX (Labs et al. 2025). The real images come from the Flickr Faces High Quality (FFHQ) dataset (Karras, Laine, and Aila 2018) and Kaggle (Dhote 2024). We chose this set of images because it encompasses a large number of data points from multiple commonly used architectures, some of which have been traditionally more difficult for detection models to correctly classify (Ojha, Li, and Lee 2023). The detection models we chose—UniFD (Ojha, Li, and Lee 2023), EFFORT (Yan et al. 2025), HIFI (Guo et al. 2023), and RECCE (Cao et al. 2022). Each of these models uses a combination of frequency-domain and spatial-domain features to perform their classifications. Table 1 shows the F1 scores for each of the models on the dataset of 14997 images, with the last column showing what a random coin-toss model would score on the same dataset. As shown here, none of the models score above a 0.9, with two of them performing not much better than completely random. Additionally, all of these models had fairly low accuracy rates for datasets of highly realistic synthetic images, such as those produced by FLUX. Even UniFD and EFFORT, the higher-performing models that draw from frequency-domain information, do not do well with certain architectures, and degrade significantly in quality when simple frequency-domain distortions are applied to synthetic images.



Figure 1: On the left is F036, an image of a real individual taken from the CUHK Face Sketch Database, and on the right, an image generated using FLUX with F036 as a prompt, with strength = 0.2 and guidance = 15.

UniFD	EFFORT	HIFI	RECCE	RANDOM
0.8743	0.8135	0.6255	0.5947	0.5161

Table 1: F1 Info for UniFD, EFFORT, RECCE, and HIFI on 7000 REAL images and 7997 SYNTHETIC images

From these results, we can gather that current approaches

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

generally do not do particularly well at detecting high-quality synthetic images, which in turn means that a new approach is needed. Utilizing human knowledge and intuition has the potential to not only bridge this gap, but to do so in a way that allows for more easily human-interpretable detection results. This is why we are proposing an experiment that will combine previously described measures of creativity in generative AI models with the observed correlation between manual face detection and the "uncanny valley effect": we believe that quantifying aspects of uncanniness in synthetic faces could lead to more accurate and interpretable synthetic face detection results.

Background and Prior Literature

The Uncanny Valley

The term "uncanny valley effect" originally referred to the discomfort a human viewer might feel upon encountering a robot that is extremely lifelike, but still inherently recognizable as a non-human entity (Mori and Macdorman 2017). In terms of artificially generated or manipulated faces, it has been associated with a mix of hyper-realistic features and inhuman traits, as well as the concept of "category uncertainty", or the idea that a given face cannot easily be classified as real or digitally created at first glance (MacDorman and Chattopadhyay 2016). Neurological research has shown that even hyper-realistic AI-generated faces spark different brain activity in observers than real photographs of human faces, although those observers are unable to tell the two categories apart visually at an accuracy level higher than near-random, and cannot articulate any tangible cause for perceived uncanniness (Moshel, Carlson, and Grootswagers 2022). However, short of EEG imaging, which is inefficient and expensive to conduct on a large scale, few techniques have been proposed that can successfully quantify uncanniness.

Uncanniness has previously been associated with both human and AI creativity, with greater uncanniness associated with a higher reported measure of perceived creativity within generative AI output (Hwang and Jeong 2026). Hwang & Jeong's study uses human responses on an arbitrarily set numerical scale to rank perceived creativity or a given output, and predetermines ground truth creativity as part of the experimental design. Similarly, a study by Kaate et al. measuring the potentially beneficial impact of deepfaked images in representations of user personas discusses perceived uncanniness as a fundamental aspect of deepfakes (Kaate et al. 2023). This study also uses subjective measures, even acknowledging the drawbacks of this tactic and incorporating a quantitative behavioral metric as well; however, this metric also is not granular enough to specifically filter for deepfakes that are viewed as particularly uncanny by users.

In our own paper, we ask: can this potential relationship between high uncanniness and high creativity in generative AI outputs be quantified in a manner that does not depend on subjective and unstable factors, but instead relies on a mathematical definition for computational creativity? And building off this, could we potentially use these results to differentiate between real facial photographs and high-quality synthetic facial images?

Task-Oriented Creativity

To measure uncanniness, we draw from the definition of Task-Oriented Creativity (TOC) as described in (Ramaswamy, Chockler, and Navaratnarajah 2025), which sets out formulae to measure three components of creativity as derived from prior computational creativity research. As stated in this paper, the definitions are predicated on the following terms: M is the generative model, A_O is the artifact set of a given output image O_M (generated by M), and A_I is the artifact set of the input prompt I, which can be an image or a text.

Definition 1 (Satisfaction of Prompt Requirements) *A generated image O_M is said to be* satisfying requirements *of I if $A_I \subseteq A_O$. In other words, all artifacts in I are present in O_M.*

Definition 2 (Cohesion) Cohesiveness $C(O_M)$ *of an output image O_M is defined wrt the input I. Given the set A_I of the artifacts in I and the set A_O of the artifacts in O_M, let P be the set of all pairs (a, b), where $a \in A_O \setminus A_I$ and $b \in A_I$. Furthermore, let $M_N = \max_{(c,d)\in A_O\setminus A_I}((\hat{c}) \cdot (\hat{d}))$. Then, cohesiveness of O_M is defined as*

$$C(O_M) = \frac{\sum_{a:(a,b)\in P} \max_{b:(a,b)\in P}((\hat{a}) \cdot (\hat{b})) + M_N}{|P| + 1}.$$

Definition 3 (Novelty) Novelty $D(O_M)$ *of an output image O_M is defined as the proportion of new artifacts in the image, if O_M contains all artifacts of I. In other words, $D(O_M)$ is $\frac{|A_O \setminus A_I|}{|A_O|}$ if $A_I \subseteq A_O$ and is 0 otherwise.*

Revisited Definitions

The original (Ramaswamy, Chockler, and Navaratnarajah 2025) paper dealt with simple sets of macro-artefacts like "apple pie", meaning that additions and subtractions to these sets would be easily to detect both manually and with an object detection model. However, this does not necessarily apply to facial images. If our aim is to differentiate real photographs from highly realistic synthetic images, predicating our results on the mere presence of large facial landmarks like "nose" would be pointless, as even low-quality synthetic facial images tend to contain facial landmarks. Instead, we must use more fine-grained measures tailored to pick up subtle differences in a practical setting.

Definition 1 is general enough to hold for a synthetic facial image study, although we must be specific about the set of artefacts to include so we can eliminate very low-quality facial images and focus on higher-quality facial images, as seen in modified Definition 4:

Definition 4 (Satisfaction of Facial Requirements) *Given a facial image F and the entire set of 68 facial landmarks F_L, F is said to be* satisfying requirements *if $F_L \subset F$. (Jabberi et al. 2023).*

The simplest way to measure this in a large experimental setting is to use DeepFace or another facial extraction library—if the chosen model finds a human face, it automatically meets the threshold. In a user study setting, an image

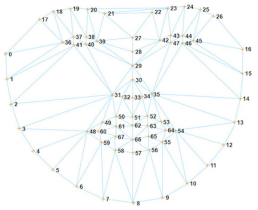


Figure 2: The 68 facial landmarks that are needed for efficient facial detection (Jabberi et al. 2023).

would pass this requirement if it is determined to be a facial image by human observers.

The (Ramaswamy, Chockler, and Navaratnarajah 2025) definition of cohesion relies on the presence of observable artefacts that can be detected by a model and have a quantifiable semantic relationship, as measured through the cosine similarity various artefact labels. However, in terms of generative AI, an alternate approach that is far more granular would be to use a spatial autocorrelation algorithm to measure the relationships between components of an image on a pixel-by-pixel basis. As further support for this, we noted that minor discrepancies have been observed between synthetic and real facial images within the spatial domain, and have been used by numerous synthetic detection models with varying degrees of success (Wang et al. 2020; Guo et al. 2023; Cao et al. 2022). These discrepancies are not visible to the naked eye, but involve hidden pixel patches, clusters, and disjointed areas that remain as products of the generative process, particularly for Generative Adversarial Network (GAN)-produced images.

The two primary spatial autocorrelation algorithms are Moran's I and Geary's C, but the former has been shown to perform better for multivariate data (Lin 2023). For a given image X of size N pixels, with a corresponding weight matrix w, Moran's I is calculated according to Equation 1, where $\bar{X}$ represents the mean of the image. A high Moran's I score indicates a high level of spatial autocorrelation within an image, while a low score indicates more disjointed areas interspersed throughout an image.

$$I(X) = \frac{N}{\sum_{i=0,j=0}^{N} w_{ij}} \times \frac{\sum_{i=0}^{N}\sum_{j=0}^{N} w_{ij}(X-\bar{X})}{(X-\bar{X})^2} \quad (1)$$

Hence, we redefine cohesion as follows in Definition 5:

Definition 5 (Cohesion) *Given a facial image F, cohesion can be measured by applying $I(F)$.*

Novelty according to Definition 3 is relatively easy to measure in a setting where multiple objects are present—have any been added? However, when dealing with high-quality facial images, we must take a different approach, since large artefacts being added, such as an extra nose, would automatically render this a low-quality image that can easily be classified. Prior research suggests that synthetic images most likely to fool humans with a "hyper-real" effect fall closer within the average area of what has been called the "face-space", or the measure by which humans intuitively analyze the uniqueness of feature configurations (Miller et al. 2023). AI-generated faces are also often deemed as more "attractive" than real faces due to their perceived average-ness, and this trait in turn is more likely to dupe observers into classifying them as real photographs (Makowski et al. 2025). Therefore, we suggest that novelty of a given image X be determined by measuring facial symmetry and distance to the golden ratio using a set of key facial landmarks within X, following a proposal for measuring attractiveness by (Schmid, Marx, and Samal 2008). The higher this distance is, and the more asymmetrical the face is, the higher the novelty factor.

(Schmid, Marx, and Samal 2008) reference a series "neoclassical canons" for facial proportions, which we can use to measure the distance between a given facial image F and a hypothetical perfectly average, symmetric face. We can use a simplified version to describe this average face F, using the 68 facial landmarks shown in Figure 2:

1. nose width (W_N, the difference between Landmarks 31 and 35) = inter-ocular width (W_O, the difference between Landmarks 41 and 46)
2. mouth width (W_M, the difference between Landmarks 48 and 54) = $1.5 \times W_N$
3. face width (W_F, the difference between Landmarks 0 and 16) = $4 \times W_N$
4. nose length (L_N, the difference between Landmarks 27 and 33) = lower face length (L_{LF},, the difference between Landmarks 8 and 51)
5. symmetry: the distance between Landmarks 0 and 30 (W_1) = the distance between Landmarks 15 and 30 (W_2)

We can calculate the results of applying these criteria to the actual landmarks in F_L: for example, how far is $|W_1 - W_2|$ from 0? Then we can average these results and use that as the novelty factor, as seen in modified Definition 6:

Definition 6 (Novelty) *Given a facial image F, novelty is defined as*

$$\frac{|W_N - W_O| + |W_M - 1.5W_N| + |W_F - 4W_N| + |W_1 - W_2| + |L_N - L_{LF}|}{5}$$

The closer this value is to 0, the less "novel" (more symmetrical and proportionate) F is.

Methodology

We hypothesize that, in conjunction with (Hwang and Jeong 2026)'s discussion of perceived uncanniness and creativity, synthetic images that humans correctly classify as AI-generated will score lower on cohesion and higher on novelty, even if they satisfy the facial requirement set out in Definition 4 and there is nothing overtly "wrong" with them. Conversely, we hypothesize that synthetic images classified incorrectly as real by humans will have low novelty, meaning that they fall closer to the average face. To test this, we plan to generate 10-step chains for a small dataset of facial images,

and perform a user study wherein individuals will be asked to classify a sampling of images from each chain step as either synthetic or real. We will then revisit the reconfigured TOC measures, applying them to the datasets of correctly and incorrectly classified images in order to determine whether the uncanniness factor can explain why humans may misclassify some synthetic images but not others.

As a preliminary study, we have chosen to use a set of 188 images from the Chinese University of Hong Kong (CUHK) student database, as found on Kaggle (Zhang, Wang, and Tang 2011), along with images we have produced ourselves by using the CUHK dataset as seed image input to the FLUX.1-dev model. We have chosen this dataset primarily because it features front-facing, clear facial images with blank backgrounds, allowing our chosen generative model to focus on the faces rather than background objects.

In a full-length paper, we would have implemented the 10-step chain-style setup described in (Ramaswamy, Chockler, and Navaratnarajah 2025) on all of these images to calculate the TOC between each, but due to limited space and computational resources we have only been able to implement them on 16/188 images thus far.

To check whether the images satisfy the facial requirement, we ran them through DeepFace's facial extraction algorithm, and confirmed that all of the real and synthetic images did in fact have clear, detectable faces. This is also true from a visual assessment of the chains, an example of which can be seen in Figure 3.



Figure 3: An example chain seeded with one of the CUHK dataset images. The real seed is the leftmost image in the top row.

We also ran the PySAL (Rey and Anselin 2010) implementation of Moran's I on the images from each chain step, and compared them to each other using the Brunner-Munzel test (Brunner and Munzel 2000), a more sensitive variation of the commonly used nonparametric two-category Mann Whitney U test. We found that, although the Moran's I value for each step did not differ from the next at a statistically significant level (with the threshold for p set at the standard 0.05), the seed images did display a slight but statistically significant decrease in spatial autocorrelation than the images from the last step in the chain. The median values yielded by this test are shown in Table 2, with only even steps displayed for brevity.

We also used the dlib facial landmark detection model (King 2009; Sagonas et al. 2013) to implement the novelty measure described in Definition 6. In Table 3 we show the median novelty scores derived from alternating steps of the chains (for brevity):

00	02	04	06	08	10
0.9774	0.9798	0.9821	0.9823	0.981	0.9816

Table 2: Median Moran's I values for alternating steps in the chain in Figure 3, from seed (00) to step 10.

00	02	04	06	08	10
28	34.2	36.3	36.55	36.35	35.9

Table 3: Median novelty score values for alternating steps in the chain in Figure 3, from seed (00) to step 10.

We also used the Brunner-Munzel test to compare these values, and found that the consecutive values for comparisons between steps 00 (the real seed), 01, 02, and 03 differed significantly, while the novelty values for rest of the steps did not differ significantly from each other. We noted an increase in novelty as the chain progressed, with the faces generated from synthetic images more distant from the real seed straying farther from the proportions described by (Schmid, Marx, and Samal 2008).

Further studies we plan to execute include occluding parts of correctly classified images and determining whether patches of low cohesion exist, determining facial symmetry thresholds between correctly and incorrectly classified images to measure disparities in novelty, and experimenting with distorting ratios subtly to flip classifications. The latter approach is informed by prior efforts to deliberately generate recognizably fake images through the "amplification" of image features such as color saturation and facial landmark positions (Broad, Fol Leymarie, and Grierson 2020). Instead of the obviously fake images produced by this procedure, we could modify it to perform subtle enhancements in the areas pointed out by Broad, Leymarie, & Grierson to create images that look realistic but might register as subtly uncanny with viewers. We also noted a shift in many of the chains that were seeded with male students, where each subsequent image appeared more aligned with stereotypical femininity—likely due to inherent gender biases within FLUX (Guo et al. 2026). Measuring differences between chains seeded with male and female images could yield interesting insights as well. However, before embarking on these, we must first perform a user study to come up with adequate human classification data.

Conclusions and Future Research

The "uncanny valley" is a contested term that may refer to a large number of phenomena in the general area of facial pattern recognition. Attempting to quantify it through the lens of creativity is a large and ill-defined task that we hope to shed some clarity on through our proposed experiment, which we hope to run and turn into a full paper over the coming months. Although the connection between the uncanny valley and creativity in the context of AI-generated imagery may seem tenuous at first glance, we believe that this is a promising method of combining human intuition with mathematical principles to address the failures of state-of-the-art detection models.

References

Broad, T.; Fol Leymarie, F.; and Grierson, M. 2020. Amplifying the uncanny.

Brunner, E., and Munzel, U. 2000. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal* 42(1):17–25.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4113–4122.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2017. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*.

Dhote, K. 2024. Human faces dataset.

Dunn, J. D.; White, D.; Sutherland, C. A. M.; Miller, E. J.; Steward, B. A.; and Dawel, A. 2026. Too good to be true: Synthetic ai faces are more average than real faces and super-recognizers know it. *British Journal of Psychology* n/a(n/a).

Esser, P.; Rombach, R.; and Ommer, B. 2020. Taming transformers for high-resolution image synthesis.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *CVPR*.

Guo, W. M.; Qian, Q.; Hasan, K.; and Du, S. 2026. Position: Universal aesthetic alignment narrows artistic expression.

Hwang, Y., and Jeong, S.-H. 2026. Effects of creativity in ai advertising: The mediating role of perceived uncanniness. *Cyberpsychology, Behavior, and Social Networking* 29(2):109–113. PMID: 41460705.

Jabberi, M.; Wali, A.; Chaudhuri, B. B.; and Alimi, A. M. 2023. 68 landmarks are efficient for 3d face alignment: what about more? 3d face alignment method applied to face recognition. *Multimedia Tools Appl.* 82(27):41435–41469.

Kaate, I.; Salminen, J.; Santos, J.; Jung, S.-G.; Olkkonen, R.; and Jansen, B. 2023. The realness of fakes: Primary evidence of the effect of deepfake personas on user perceptions in a design task. *International Journal of Human-Computer Studies* 178:103096.

Karras, T.; Laine, S.; and Aila, T. 2018. A style-based generator architecture for generative adversarial networks. *CoRR* abs/1812.04948.

Kim, M.; Liu, F.; Jain, A.; and Liu, X. 2023. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023*, 12715–12725.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10:1755–1758.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space.

Lin, J. 2023. Comparison of moran's i and geary's c in multivariate spatial pattern analysis. *Geographical Analysis* 55(4):685–702.

MacDorman, K. F., and Chattopadhyay, D. 2016. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146:190–205.

Makowski, D.; Te, A. S.; Neves, A.; Kirk, S.; Liang, N. Z.; Mavros, P.; and Chen, S. A. 2025. Too beautiful to be fake: Attractive faces are less likely to be judged as artificially generated. *Acta Psychologica* 252:104670.

Miller, E. J.; Steward, B. A.; Witkower, Z.; Sutherland, C. A.; Krumhuber, E. G.; and Dawel, A. 2023. Ai hyperrealism: Why ai faces are perceived as more real than human ones. *Psychological Science* 34(12):1390–1403.

Mori, M., and Macdorman, K. F. 2017. The uncanny valley: The original essay by masahiro mori. In *IEEE Spectrum*.

Moshel, M.; Carlson, T.; and Grootswagers, T. 2022. Are you for real? decoding realistic ai-generated faces from neural activity. *Vision Research* 199:108079.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *CVPR*.

Ramaswamy, A.; Chockler, H.; and Navaratnarajah, M. 2025. Defining and quantifying creative behavior in popular image generators. In Oliveira, H. G.; Spendlove, B.; Gervás, P.; and Ventura, D., eds., *Proceedings of the 16th International Conference on Computational Creativity*, 362–367. Campinas, Brazil: Association for Computational Creativity.

Rey, S. J., and Anselin, L. 2010. *PySAL: A Python Library of Spatial Analytical Methods*. Berlin, Heidelberg: Springer Berlin Heidelberg. 175–193.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. 300 faces in-the-wild challenge: the first facial landmark localization challenge. In *International Conference on Computer Vision Workshops (ICCV Workshops)*.

Schmid, K.; Marx, D.; and Samal, A. 2008. Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition* 41(8):2710–2717.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*.

Yan, Z.; Wang, J.; Jin, P.; Zhang, K.-Y.; Liu, C.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2025. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Forty-second International Conference on Machine Learning*.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 513–520. IEEE.

I Was Scrolling and Then I Saw a Pregnant Strawberry

Piera Riccio
University of Amsterdam
p.riccio@uva.nl

Abstract

AI minidramas (also known as *fruit dramas*) are short, algorithmically distributed generative AI video series featuring anthropomorphized characters that have recently emerged as a widespread phenomenon on social media platforms. This paper argues that despite their seemingly innocuous aesthetic, these videos reproduce deeply gendered narrative structures in which female characters are systematically associated with moral transgression, sexual betrayal, and reproductive capacity, and that several plots also encode the logic of racialization, *i.e.,* the process by which visible bodily difference is morally loaded. Drawing on feminist film theory, critical race theory, and platform studies, it further argues that the generative AI aesthetic of these videos, characterized by softness, roundness, and visual cuteness, functions as a mechanism of aesthetic laundering, neutralizing the ideological weight of these narratives and enabling their circulation despite content moderation systems. This paper approaches these questions through personal observation and close reading, reflecting on the specific affordances of generative AI that make this phenomenon both possible and culturally consequential for the field of computational creativity.



Figure 1: Screenshots from AI minidramas on TikTok, April 2026. Left: a female fruit character with a pregnancy test. Right: "Our Baby... Looks like your boss?", a couple of Strawberries and their eggplant baby. Original accounts unverified. Both images are exemplary of recurring patterns in the annotated corpus.

Strawberrina

I was scrolling reel after reel on Instagram. I am not proud of this, but I do it when I am bored or when I need a quick distraction. Nothing was quite getting my attention, but I kept scrolling for a while. *I was scrolling and then I saw a pregnant strawberry.* Well, actually, she was some sort of anthropomorphized being. She had arms and legs and a face, she could speak and she was wearing clothes, but her head was shaped like a strawberry and her skin was red (see Figure 1). Her name was Strawberrina. She had a cute husband, Strawberrino, and a cute house. They seemed quite happy until Strawberrina gave birth to her child, but the child was an eggplant rather than a strawberry. Strawberrino found out that his wife had been cheating on him with her boss, who was an eggplant, and he felt heart-broken. A mini-drama of around 8 or 9 episodes unfolded, with new episodes being published by the same account daily. Many things happened: a new love for Strawberrino, an orange, from which a new orange-strawberry baby was quickly born (with the doctor being surprised to find out that finally a wife had not cheated on her husband), the baby eggplant got abandoned, Strawberrina tried to rob a bank, she went to jail, and then met a mysterious lemon lawyer that claimed he could bring her out of the trouble. After this episode, I couldn't find any follow up on this mini series and the account that was publishing them seemed to disappear from my radar (Figure 1 shows related examples from other accounts). Because I got interested in this first *fruit drama* that I saw, the recommendation algorithm of Instagram started proposing more and more of such videos to me. Given my academic background in Generative AI, I initially was quite ashamed of having fallen for AI slop and I was feeling *brain rotted* by these nonsensical *fruit dramas*, but soon enough, I felt that this phenomenon deserved my attention actively.

From Brain Rot to Research

I started noticing some recurring patterns among the videos published by different accounts. The background melan-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

cholic music is almost always the same[1] and the topic across different videos is also quite standardized: love dramas, betrayal, cheating (Dazed MENA 2026). Most of the videos included fruits and vegetables, but sometimes the anthropomorphized beings are pets, food, objects, or anything else we can imagine. The term *fruit drama* (Dazed MENA 2026) is quite popular, but we generally refer to them as AI minidramas. Such videos represent one of the most widespread and organically emergent instances of generative AI being adopted for creative production by non-expert users at scale (BBC News 2026). The reach of this phenomenon extends beyond individual creators: during my observation period, I encountered a *fruit drama* published on the official TikTok account of Lega, the right-wing Italian political party led by Matteo Salvini[2], demonstrating that the format has been adopted also by institutional political actors as a vehicle for communication and outreach. Understanding what people do with generative AI tools when left to their own devices, *e.g.,* what narratives they tell, what aesthetics they converge on, and what cultural values get embedded is a fundamental question for computational creativity as a field (Broad 2024; Lopes, Cunha, and Martins 2024).

My research focuses on plot structure and character roles, examining how gender and other social constructs are represented in these videos. To analyze this systematically, I manually annotated all AI minidrama videos that appeared on my Instagram feed over more than one month (mid-March to late April 2026), after saving them with the Instagram's native save function. I chose not to download videos out of respect for GDPR and copyright regulations. I annotated whether a main female character was depicted negatively, and why; whether a main male character was depicted negatively, and why; and whether pregnancy was represented, particularly as the result of an extramarital relationship. I annotated 137 videos, in seven languages (English, Spanish, Italian, French, Portuguese, German, and Arabic) or mute, shared by 61 different Instagram accounts. The phenomenon proved unstable during the observation period: several videos and accounts disappeared before annotation was completed, highlighting the volatility of this fast-moving ecosystem. The analyzed corpus is also shaped by Instagram's recommendation algorithm. As such, these videos are not a random or representative sample of all AI minidramas, but an algorithmically curated selection deemed relevant to a user with my profile, reflecting how most viewers encounter this content.

Gender Roles, Served Fresh

As anthropomorphized beings (see Figure 1), the characters in these AI minidramas have visible gendered features that easily allow distinguishing between binary gender representations (male and female). In most of the videos I have annotated, the male characters represent positive values. Indeed, the most common male character I have observed is a loyal husband, hard working and devoted to his family. The wife is instead often depicted negatively. She can be money-driven, leveraging romantic attachments for financial or social advantage. She is often cheating on her husband, sometimes even with multiple other males. She is also unfaithful to her own friends, jealous and superficial. Only rarely, I have encountered stories in which the husband cheats on the wife, but interestingly enough, in those cases the affair is often the best friend or the sister of the wife.

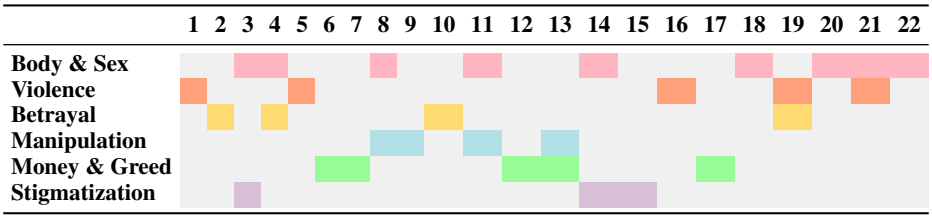
While not every female character is depicted negatively, the "evil" in most of the minidramas I have watched is depicted through a *main* female character (in 73.7% of cases, vs. only 33.6% of main male characters depicted negatively). For example, in the first story that I described, Strawberrina cheated on her husband with her eggplant boss. Arguably, the boss of Strawberrina is making an immoral action by having a sexual relationship with one of his employees. However, the narration steers away from the boss and focuses on Strawberrina not being a loyal wife. Having a narration that unfolds around the (negative) actions of a female character means that these characters have a narrative agency and they are not passive recipients of the actions of the males around them. In contrast, most of the male characters are not represented as "heroes", but more as "victims", placing them in a more passive and flat representation style compared to the female counterparts. Importantly, male characters may appear more naive and passive, but they are naive and passive in a way that leaves their fundamental goodness intact. While her activity is her corruption, his passivity is his innocence.

Analyzing my annotations, I have summarized the negative representations of female characters into 22 categories (reported in Table 1) and further grouped into 6 macro-areas (Body & Sex, Violence, Betrayal, Manipulation, Money & Greed, Stigmatization). Among these 22 categories, *Body Stigmatization*, and *Intellectual Stigmatization* are notable because they do not depict the female character as actively doing something wrong. Rather, she is stigmatized for conditions or choices that are neutral or entirely natural: her body shape, her eating habits, bodily functions such as flatulence or menstruation, or she is depicted as intellectually inferior. In *Honor-Based Shaming*, the female character is shamed not directly for her own sake, but because her choices, such as having an account on OnlyFruits (the AI minidrama equivalent of OnlyFans), bring social embarrassment to the male members of her family. Her body and her sexuality are treated as property of the males around her, and her autonomous choices become their problem. The ideological work of these categories is precisely to present neutral or natural features as sources of moral failing, which is itself a form of misogyny. In addition, the category of *Deceptive intimacy* includes narratives in which a female accepts payment for a sustained years-long relationship during which she simulates love, builds a family, and performs domesticity, only to disappear with the money once the agreed term expires. The category of *False Victimhood* features scenarios in which a female character sexually assaults a male and then, exploiting her social position as a female, convinces those around her that she was the victim. This narrative is particularly harmful because it directly repro-

[1] for instance: https://www.youtube.com/watch?v=3w7hhqBNoDY, Last Access: 27.04.2026

[2] https://vm.tiktok.com/ZNRbpJbvq/, Last Access: 26.04.2026

Table 1: Heatmap of negative female representations in AI minidramas across macro-areas. Numbers refer to categories listed below. Colored cells indicate membership in a macro-area; categories appearing in multiple macro-areas reflect the intersectional nature of these narratives.



Category legend: 1. Being a bully; 2. Being disloyal to friends/family; 3. Body stigmatization; 4. Cheating; 5. Child abuse; 6. Corruption; 7. Criminality; 8. Deceptive intimacy; 9. Emotional detachment; 10. Envy; 11. False victimhood; 12. Financial parasiting; 13. Gaslighting for personal benefit; 14. Honor-based shame; 15. Intellectual stigmatization; 16. Maternal abandonment; 17. Obsession for money; 18. Obsession for sex; 19. Revenge; 20. Sex for career advancement; 21. Sexual violence; 22. Spreading STDs.

duces one of the most persistent and damaging myths used to discredit women affected by sexual violence (Burt 1980). The category of *Financial Parasiting* and *Sex for Career Advancement* similarly frame female economic agency as inherently predatory: the female who seeks financial security or professional advancement is depicted not as navigating structural inequalities but as a manipulative agent exploiting male vulnerability.

Taken together, the analyzed categories suggest that what is most threatening to the narrative logic of AI minidramas is not female evil per se, but female autonomy (economic, sexual, and social) which is systematically reframed as corruption. The figure of the "evil woman" as a site of displaced agency is something feminist film scholars have been writing about for decades. The classic example is the *femme fatale* (Doane 2013), a character who is morally coded as dangerous and destructive, almost always in relation to sexuality, but who also has a kind of narrative power and centrality that the "good" female characters completely lack (Mulvey 2013). However, while the evil female character is central in these minidramas, she is central because of her relationship to reproduction, sexuality, and betrayal. Her agency is granted on precisely the terms that have historically been used to control and pathologize women. This tension can be related to the concept of "postfeminist masquerade" (McRobbie 2008), *e.g.*, the way certain representations seem to give women power and centrality while simultaneously reinforcing the terms of their subordination. The female character drives the plot, yes, but the plot she drives is one entirely organized around her body, her reproductive capacity, and her sexual transgression. Pregnancy, in particular, is quite represented in these minidramas (appearing in 31.4% of my annotated videos) and is almost depicted as a physical condition that women, as such, cannot avoid and that defines the very essence of being a female in these plots. Pregnancy itself becomes a symbol through which the storytelling revolves around the body of the female characters rather than their (negative) actions or values.

The Eggplant in the Room

When Strawberrina gives birth to her child, both the doctor and the husband realize that she had cheated because her baby looks "different". In human life, this could only happen when the baby's physical appearance, particularly skin color or other visible traits, makes it immediately obvious that the biological father is someone else. This is a quite common dynamic across my annotated videos (65.1% of videos showing pregnancy), and I find it one of the most uncomfortable things to sit with analytically. The choice of making characters different species of fruit, vegetables or objects is doing something quite specific and not at all innocent: the "wrongness" of the baby is legible at a glance because it is different from the family it was born into. Such narratives not only reproduce misogynistic gender dynamics, but they also encode what scholars of critical race theory would call the logic of racialization (Omi and Winant 2014): the process by which visible bodily difference becomes meaningful, morally loaded, and socially consequential. As previously mentioned, these AI minidramas are circulating globally, across audiences with very different cultural frameworks and histories, where the meaning of this narrative may shift and mutate for different audiences (Doh et al. 2025). This does not make the structural observation less valid, but it does mean that a single universal reading should not be assumed. Despite this, the narrative structure of these videos relies on the logic of racialized bodily difference, with the idea that you can look at a body and immediately know where it does and does not "belong" (Fanon 2016; Doh et al. 2026). The anxiety about paternity that runs through these minidramas is not a neutral dramatic device: it is historically entangled with fears about racial mixing, bodily purity, and the integrity of the family as a unit of racial reproduction. Indeed, reproductive capacity has historically been weaponized against racialized women specifically, with their bodies treated as sites of anxiety, surveillance, and control (Roberts 2014). While the eggplant baby of Strawberrina is visually funny, the story it echoes is not.

In addition, across several videos in my corpus, I observed characters actively concealing their species or color identity in order to access better social status, a structural parallel

to the phenomenon of racial passing (Fanon 2016) also reproducing logics of colorism (Hunter 2007): *e.g.,* green apples over red apples, or precious stones ranked by their market value as proxies for social worth. In one particularly striking example, a brown strawberry had spent his entire life painting himself red to pass as a red strawberry, only to be exposed when his twins were born (one red, one brown) and he and his wife proceeded to systematically discriminate against the brown child. These are not incidental narrative details: the fruit world, it turns out, has a "racial" order.

Aestheticized Violence

But if such AI minidramas encode misogynist and racist values, one could wonder how it is possible that they are becoming so viral, reaching such a variety of audiences (including myself). It is exactly the circulability and platformization of this content that make AI minidramas particularly insidious. Many of these videos depict images that are not light, including forced pregnancies, sexual assault, child abuse, physical violence, but they present a cute visual aesthetics that I frame here as a form of *aesthetic laundering* (Downey 2024), a process by which disturbing content presented as innocuous through its visual register. The cuteness of the videos and the AI-generated anthropomorphized bodies create a visual register that neutralizes the weight of what is actually being depicted. In other visual genres, such topics would demand a certain treatment: slowness, gravity, consequence, space for the viewer to process. But wrapped in pastel colors and round unrealistic bodies, moving at the pace of a short-form reel, they just... pass. The aesthetic does ideological work by making the unbearable feel inconsequential. Cuteness, in these videos, can be seen as a mode of power and disavowal and it is, indeed, an aesthetic category that is not innocent (Ngai 2010): it manages and contains affect, it makes things seem harmless and manipulable precisely because they are rendered small and soft and round. Crucially, cuteness is also an economic engine. Characters that seem to have souls and affective depth are precisely designed to generate attachment that can be monetized (Allison 2006). The AI aesthetic carries this cuteness, operating within neoliberal logics of attention and profit, with their seemingly innocuous aesthetics not appearing serious enough to be suspicious (Banet-Weiser 2018). Yet AI minidramas are distinct from earlier cute commodification, first, because of the simplicity to generate several instances of such content (Salvaggio 2024), and second because the cute aesthetic is not solely a deliberate creative choice and it is partly an artifact of the generative models themselves, which encode and amplify existing aesthetic and social norms through their training data (Jääskeläinen et al. 2025).

In addition, these videos circulate on online platforms where normally content moderation practices remove or shadow ban content that is considered not-safe-for-work (NSFW), effectively influencing the contemporary creative ecosystem (Riccio et al. 2022). However, the definition of "safety" in these systems is quite controversial and ambiguous (Gillespie 2018), often prioritizing content that generates revenues and engagement (Riccio, Hofmann, and Oliver 2024). Indeed, the AI minidramas effectively bypass content moderation filters, depicting images that would normally be filtered out of these platforms. And what passes moderation gets recommended. Platform recommendation algorithms reward engagement, and the melodramatic structure of these minidramas (the cliffhangers, the betrayals, the shocking reveals) is also precisely engineered for it. The cuteness gets the content through the filter; the drama keeps the viewer watching; and the algorithm reads both signals as reasons to push the video further, rather than restricting it. Content moderation and content recommendation are, in this sense, two faces of the same market-driven logic of these platforms (Riccio et al. 2022), and the AI minidrama aesthetic has learned to exploit both simultaneously.

Delusions on Creativity

While the creators of visual generative AI models often propose revolutionary creative possibilities of these technological tools (Epstein et al. 2023), AI minidramas, being a mass-scale, real-world instance of generative AI being used creatively by non-expert users, reveal that computational creativity in practice is not neutral. The creative affordances of these tools (what they make easy, what they make possible, what they make look "good") are shaping cultural output in ways that have ideological consequences. These are not merely aesthetic or media culture questions. If generative AI tools systematically make certain visual styles easier, certain narrative structures more template-ready, and certain ideological contents more platform-compatible, then the creativity they enable is already shaped before the creator makes a single choice. The non-expert user picking up these tools is not a blank creative agent: they are working within a sociotechnical assemblage (Stivale 2014; DeLanda 2019) that has its own affordances, its own defaults, and its own politics. This assemblage includes the generative models and their training data, the platforms and their recommendation and moderation logics, the creators and their prompting strategies, and the audiences and their engagement behaviors, and it is where computational creativity actually happens in the wild. And it is at the level of this assemblage, not of any single tool or actor, that ideology is produced and distributed. What AI minidramas reveal, perhaps more clearly than any controlled creative experiment could, is that computational creativity in the wild is never just about what is made. It is about what the tools make easy, what the platforms make visible, and what the aesthetics make bearable.

Strawberrina was generated by someone, recommended to me, and eventually became the unlikely muse of this short paper. That chain, going from model to platform to creator to algorithm to audience to researcher, is where the ideology lives. The field of computational creativity has invested enormous effort in understanding what computational systems can make. The present paper suggests it is equally urgent to study what they make routine, what narratives they normalize, what aesthetics they reward, and what societal representations they quietly distribute at scale.

Acknowledgments

I would like to thank Miriam Doh, Francesco Galati, Noa Garcia, Benedikt Höltgen, Ludovica Schaerf, Carla Teodoro Cubeñas, and Nanne van Noord for supporting this idea and for providing insightful feedback on this paper.

References

Allison, A. 2006. *Millennial monsters: Japanese toys and the global imagination*, volume 13. Univ of California Press.

Banet-Weiser, S. 2018. *Empowered: Popular feminism and popular misogyny*. Duke University Press.

BBC News. 2026. Think Love Island is bad? wait till you meet the AI fruit version. https://www.bbc.com/news/articles/ckgr35y26q7o. Accessed: April 2026.

Broad, T. 2024. Is computational creativity flourishing on the dead internet? In *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24)*.

Burt, M. R. 1980. Cultural myths and supports for rape. *Journal of personality and social psychology* 38(2):217.

Dazed MENA. 2026. Opinion: We should probably stop engaging with the AI fruit dramas. https://www.dazed.me/tech/opinion-we-should-probably-stop-engaging-with-the-ai-fruit-dramas. Accessed: April 2026.

DeLanda, M. 2019. A new philosophy of society.

Doane, M. A. 2013. *Femmes fatales*. Routledge.

Doh, M.; Höltgen, B.; Riccio, P.; and Oliver, N. M. 2025. Position: The categorization of race in ml is a flawed premise. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Doh, M.; Riccio, P.; Höltgen, B.; Lopez Calderon, O.; Munarini, M.; Canali, C.; Ogolla, S.; and Oliver, N. 2026. Between and beyond: Designing for identity complexity in hci. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26. New York, NY, USA: Association for Computing Machinery.

Downey, T. 2024. Unveiling aesthetic laundering: When AI bias meets human context. https://medium.com/@TonyDowney/unveiling-aesthetic-laundering-c38c42d1369b. Medium. Accessed: April 2026.

Epstein, Z.; Hertzmann, A.; of Human Creativity, I.; Akten, M.; Farid, H.; Fjeld, J.; Frank, M. R.; Groh, M.; Herman, L.; Leach, N.; et al. 2023. Art and the science of generative ai. *Science* 380(6650):1110–1111.

Fanon, F. 2016. Black skin, white masks. In *Social theory re-wired*. Routledge. 394–401.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Hunter, M. 2007. The persistent problem of colorism: Skin tone, status, and inequality. *Sociology compass* 1(1):237–254.

Jääskeläinen, P.; Sharma, N. K.; Pallett, H.; and Åsberg, C. 2025. Intersectional analysis of visual generative ai: the case of stable diffusion. *AI & society* 40(6):4341–4362.

Lopes, J.; Cunha, J. M.; and Martins, P. 2024. Computational creativity in meme generation: A multimodal approach. In *ICCC*, 402–406.

McRobbie, A. 2008. The aftermath of feminism: Gender, culture and social change.

Mulvey, L. 2013. Visual pleasure and narrative cinema. In *The sexual subject*. Routledge. 22–34.

Ngai, S. 2010. Our aesthetic categories. *PMLA* 125(4):948–958.

Omi, M., and Winant, H. 2014. *Racial formation in the United States*. Routledge.

Riccio, P.; Oliver, J. L.; Escolano, F.; and Oliver, N. 2022. Algorithmic censorship of art: A proposed research agenda. In *ICCC*, 359–363.

Riccio, P.; Hofmann, T.; and Oliver, N. 2024. Exposed or erased: algorithmic censorship of nudity in art. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.

Roberts, D. 2014. *Killing the black body: Race, reproduction, and the meaning of liberty*. Vintage.

Salvaggio, E. 2024. Slop infrastructures 1 & 2. Cybernetic Forests. Accessed: 2026-01-07.

Stivale, C. J. 2014. *Gilles Deleuze: key concepts*. Routledge.

IRIDyOM: Exposing Musical Expectation as an Interactive Creative Space

Lluc Bono Rosselló, Julien Baudru and **Hugues Bersini**
Institute for Interdisciplinary Studies on Artificial Intelligence
Université Libre de Bruxelles
Bruxelles, Belgium
Lluc.Bono.Rossello@ulb.be

Abstract

Computational models of musical expectation such as IDyOM estimate listener-relative surprise and uncertainty, providing a formal basis for analyzing how musical events are predicted and experienced. This paper presents IRIDyOM, a real-time Max for Live interface that brings IDyOM-style estimates into Ableton Live through a Python backend and a no-code front end. Rather than using expectation only as a retrospective descriptor, the system makes it available as an interactive creative space: users can configure a stylistic prior from a corpus, inspect predictive profiles, apply guided transformations, and generate constrained continuations. By exposing predictive structure directly within a DAW workflow, IRIDyOM supports exploratory composition, interpretable expectation-guided interaction, and research workflows requiring direct manipulation of musical predictability.

Introduction

Musical listening and composition are shaped by expectations about what may happen next. Through exposure to musical styles, listeners acquire statistical regularities that influence perception, aesthetic response, and pleasure (Huron 2006; Gold et al. 2019; Mas-Herrero and Marco-Pallarés 2025). Computational models of musical expectation make these processes explicit by estimating probability distributions over possible continuations. Among these, IDyOM (Information Dynamics of Music) is a widely used model that derives information-theoretic measures such as entropy, quantifying uncertainty about future events, and Information Content, quantifying the surprise of an observed event (Pearce and Wiggins 2012; Pearce 2018). These predictors have become important in music-cognition research, where they have been used to relate predictive structure to behavioural and neural responses during music listening (Sauvé and Pearce 2019; Cheung et al. 2019; Di Liberto et al. 2020; Daikoku 2018).

This expectation-based perspective is also relevant for creativity. In creativity, novelty and value are not absolute properties of an artifact, but depend on the conceptual and evaluative frame within which that artifact is encountered (Boden 2004; Wiggins 2006). In music, that frame is shaped in part by learned expectations: what counts as familiar, complex or surprising depends on the listener's prior exposure to style and on the predictive context established by the unfolding piece. Expectation models therefore matter not only because they describe what a listener may predict next, but because they provide formal proxies for one dimension of how musical ideas are evaluated. This is especially relevant in light of work linking predictive uncertainty and surprise to perceived complexity and musical pleasure (Sauvé and Pearce 2019; Gold et al. 2019; Cheung et al. 2019; Mas-Herrero and Marco-Pallarés 2025). Recent work has likewise argued that musical creativity depends in part on statistical learning, hierarchical organization, and the regulation of uncertainty, rather than on unpredictability alone (Daikoku 2018; Daikoku, Wiggins, and Nagai 2021). Expectation models are therefore relevant not only as descriptive tools, but also as formalizations of one dimension of creative evaluation and control.

Despite this relevance, expectation models are still used mainly as offline analytical tools. They can describe how surprising a melody is after it has been composed, but they are rarely available as interactive resources during composition. This also limits their use in experimental design, where predictive variables are often measured after stimulus generation rather than manipulated during it (Mas-Herrero and Marco-Pallarés 2025). A tool that makes predictive distributions and information-content profiles inspectable and controllable in real time could therefore support both compositional practice and behavioural or neuroscience studies requiring controlled manipulation of musical expectation.

Existing AI music tools increasingly support interactive generation, from digital audio workstation (DAW)-oriented systems such as Magenta Studio to corpus-based improvisation systems such as Somax (Roberts et al. 2019; Borg 2019). These systems demonstrate the creative potential of machine learning in music, but they typically focus on producing or recombining musical material rather than exposing how that material may be experienced relative to a learned expectation model. By contrast, IDyOM provides interpretable listener-relative signal, but has historically been less accessible for real-time, no-code interaction within a DAW. Recent Python implementations have improved access to IDyOM-style analysis (Guan, Ren, and Pelofi 2022; Marion et al. 2025), yet expectation modelling remains used primarily in offline analytical workflows rather than as an interactive resource within compositional workflows.

This paper presents *IRIDyOM* (Interface for Real-time

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

IDyOM), a Max for Live device—that is, a custom interactive module running inside the Ableton Live digital audio workstation (DAW)—that integrates a new Python reimplementation of IDyOM into a real-time compositional workflow. The backend is designed for low-latency inference while preserving the expectation-modelling logic of the original Lisp implementation. Rather than presenting another autonomous music generator, *IRIDyOM* makes predictive musical expectation directly available as an interactive compositional resource. Users can define a stylistic prior, inspect next-event probabilities and information-content profiles, and guide continuation, variation, and generation through explicit control of expectedness. The contribution is therefore an interface framework that brings listener-relative predictive structure into a DAW-native workflow for composition, co-creative interaction, and experimental stimulus design.

System Overview

Information Dynamics of Music

IDyOM (Information Dynamics of Music) is a computational model of musical expectation that predicts the next event in a musical sequence from statistical regularities acquired through exposure (Pearce and Wiggins 2012; Pearce 2018). Following the multiple-viewpoint approach to music modelling, musical events can be represented through alternative feature descriptions such as pitch class, pitch with octave, interval, duration, or linked viewpoint combinations (Conklin and Witten 1995). In the current implementation, predictions target a pitch-related representation, expressed over pitch–octave events.

Given a musical history $x_{1:t-1}$, IDyOM estimates a probability distribution $P(x_t \mid x_{1:t-1})$ over possible next events in the chosen target viewpoint. This distribution combines evidence from two complementary sources. The long-term model (LTM) represents prior exposure to a corpus, style, repertoire, or user collection, and is estimated by accumulating context-dependent transition statistics from that material. The short-term model (STM) is updated online from the current piece or interaction, capturing regularities that emerge locally as the melody unfolds. Within each model, predictions are made using variable-order contexts: longer histories provide more specific predictions, while shorter histories provide more robust estimates when evidence for the longer context is sparse. In standard IDyOM, this variable-order combination is handled using Prediction by Partial Matching (PPM) back-off (Cleary and Witten 1984); in the present implementation, order-specific predictions may also be merged using entropy-weighted merging, so that lower-entropy, more confident distributions receive greater weight. The same principle is used to combine LTM and STM predictions, allowing the final distribution to reflect both corpus-level regularities and the immediate melodic context.

From this distribution, IDyOM derives Information Content, which measures how surprising the observed event is under the current expectation model:

$$IC(x_t) = -\log_2 P(x_t \mid x_{1:t-1}).$$

High-probability events therefore have low Information Content and are interpreted as expected, while low-probability events have high Information Content and are interpreted as surprising. Importantly, surprise is relative to the selected viewpoint and to the model's prior exposure: the same note may be expected or unexpected depending on the corpus-level statistics and the current melodic context. This makes IDyOM well suited for an interactive interface in which an LTM estimated from a loaded or user-provided corpus defines a listener-like prior, the STM adapts to the ongoing musical interaction, and the resulting probabilities and information-content values can be analyzed, steered, or used for generation.

Architecture

IRIDyOM is designed to make IDyOM-based expectation modelling usable within creative and research workflows. Following recommendations for deployable music-AI tools, the aim is not only to expose model behaviour, but to do so through an interface that is accessible, flexible, and compatible with established compositional environments (Roberts et al. 2019; Kayacik et al. 2019; Pasquier et al. 2017). As in Magenta Studio's integration of generative models with a DAW (Roberts et al. 2019), *IRIDyOM* separates predictive inference, communication, and user interaction (Figure 1).

The predictive layer uses a Python reimplementation of IDyOM based on an explicit graph formulation of variable-order Markov models. It loads long-term models estimated from user-provided corpora, maintains short-term context from the current musical sequence, and returns next-event probability distributions used for Information Content calculation. A local Python server wraps this functionality and acts as the interaction controller. It manages the active model, prediction sessions, short-term melodic history, file-based analysis requests, and the exchange of control and result messages between the modelling backend and the DAW environment. The user-facing layer is a Max for Live device running inside Ableton Live. Because the Python model is kept outside the Max device, communication is handled by a Node for Max bridge using the `node.script` object and `max-api`. The bridge exchanges newline-delimited JSON messages with the Python server over a local TCP connection. This design keeps the system synchronized with the DAW while allowing the modelling code, server logic, and Max interface to be developed and deployed independently. The software can be run directly from the repository or installed through a no-code Windows installer.

Interaction Modes

The interface is organized around five interaction modes, each corresponding to a different way of working with the same expectation model rather than to an independent function. *Train* defines the stylistic prior by estimating the long-term model from user-provided material. *Assistant* exposes the model's predictive structure during composition by showing both the Information Content profile of the current melody, that is, its evolving surprise pattern, and the probabilities of candidate next notes, which function as

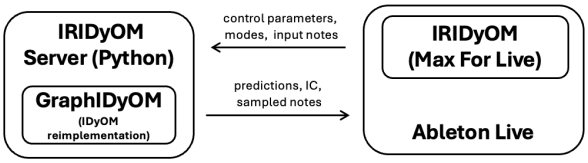


Figure 1: Runtime architecture of *IRIDyOM*. A Max for Live device inside Ableton Live communicates with a local Python server through a Node for Max bridge. The server manages interaction state and calls GraphIDYOM, a Python reimplementation of IDyOM, to return probability distributions, Information Content values, and generated MIDI events.

ranked suggestions for likely continuations. *Continue* and *Variate* allow users to manipulate musical material under explicit control of expectedness, either by extending an existing sequence or by rewriting a selected region. *Sequencer* supports more continuous constrained generation from the same learned model. Across these modes, *IRIDyOM* treats predictive expectation not only as an analytical description of music, but also as a compositional parameter that can be inspected and shaped in real time.

Train Mode

Training operates at the meta-level of the system. *Train* mode lets the user define the stylistic prior by selecting MIDI files and naming the resulting model (see Figure 2). The resulting model determines what the system considers probable or stylistically expected. If the corpus consists of the user's previous compositions, the system approximates that user's own stylistic tendencies; if it is drawn from an external repertoire, it instead reflects a different stylistic frame of reference. *Training* therefore determines whose expectations are being modelled, and against which prior later creative decisions are evaluated in terms of predictability and surprise.

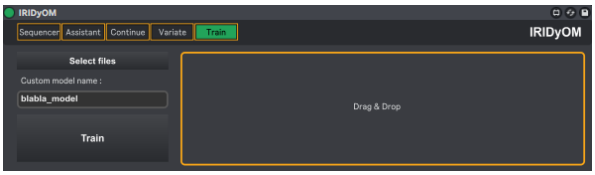


Figure 2: *Train* Mode interface, showing corpus selection and model naming in the left panel and the drag-and-drop MIDI area in the right.

Assistant Mode

Assistant mode exposes the expectation model as an analytical listener during composition. The user can play notes live, analyse a melody within the session, or import a MIDI file; in each case, the system evaluates the unfolding sequence against the long-term loaded model and the local short-term context (See Figure 3). Returns information content values for the notes of the sequence, displays likely next-note candidates, and visualizes the resulting IC profile, which can also be exported for later inspection (see Fig. 4). *Assistant* mode therefore lets the user inspect how a melody is positioned relative to the learned style, showing which events are more or less expected.

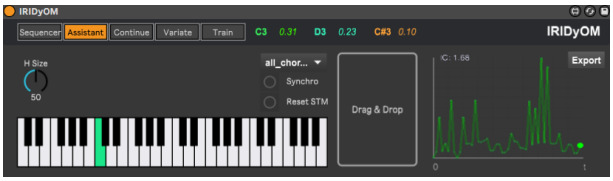


Figure 3: *Assistant* Mode interface, with live note input and next-note inspection on the top of keyboard display, a drag and drop area in the middle and a real-time information-content trace shown at right.

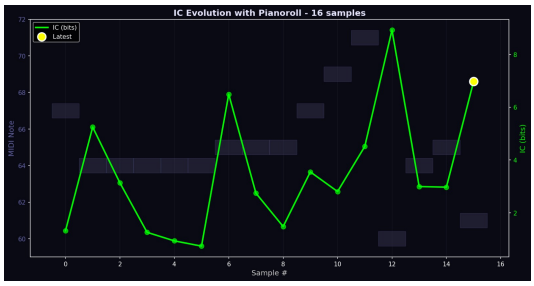


Figure 4: **Export Example.** Exported analysis view showing the MIDI piano roll overlaid with the corresponding information-content (IC) curve.

Continue Mode

Continue mode turns the expectation model into a co-creative partner for extending existing material. After analyzing a MIDI file, the user selects an output location and can specify a sequence of expectedness choices for the generated continuation (See Figure 5). Each step can favour the most probable candidate, an intermediate-probability candidate, or a low-probability candidate. The resulting MIDI file appends new notes to the source material, using the analyzed melody as context while allowing the user to shape the expectedness profile of the continuation without choosing exact pitches.

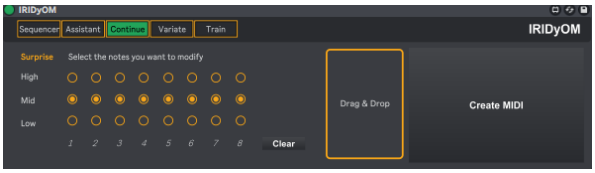


Figure 5: *Continue* Mode interface, organized around a discrete surprise-pattern control panel on the left and a drag-and-drop MIDI input/output workflow.

Variate Mode

Variate mode applies the same expectation logic to local transformation rather than extension. The user selects an analyzed MIDI file, chooses one of four regions of the note sequence, and sets a surprise value (See Figure 6). The system preserves the original timing and durations but replaces pitches in the selected region according to the current prediction distribution. Low surprise favours probable notes, medium surprise favours intermediate candidates, and high surprise favours lower-probability alternatives. The result is exported as a new MIDI file, allowing the user to explore nearby versions of a phrase while controlling how strongly it conforms to or departs from learned expectations.

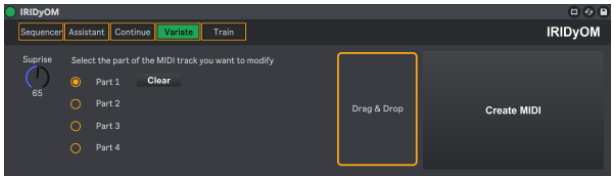


Figure 6: *Variate* Mode interface, showing region-selection controls and a surprise-level knob for local pitch rewriting of an analysed MIDI sequence.

Sequencer Mode

Sequencer mode is the most autonomous use of the expectation model, but it remains constrained by user-defined predictive and temporal parameters. The system continuously generates MIDI notes by sampling from the next-event probability distribution conditioned on the current history. The user controls this stream through parameters including model selection, temperature, probabilistic sampling, pitch range, note duration, inter-onset interval, and context length (See Figure 7). This supports real-time brainstorming or performance-oriented generation, where material can be recorded, inspected, and further edited inside the DAW.

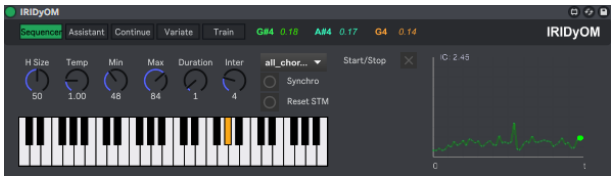


Figure 7: *Sequencer* interface, exposing real-time sampling controls for model selection, sampling temperature, temporal behaviour, and pitch-range constraints. Real-time surprise is plotted on the right panel.

Discussion and Conclusion

IRIDyOM proposes a different role for expectation models in creative practice. Rather than treating surprise and uncertainty only as post-hoc analytical descriptors, the system makes them directly usable during composition. By integrating IDyOM-derived probabilities, Information Content, and constrained generation into a DAW-native workflow, *IRIDyOM* supports analysis, controlled transformation, and exploratory continuation while leaving musical evaluation with the user. In this sense, the contribution is not a new autonomous generator, but an interpretable co-creative interface in which listener-relative predictive structure becomes a compositional resource.

More broadly, this shifts expectation from a descriptive model of listening to a usable dimension of creative evaluation. What counts as familiar, surprising, or meaningfully novel depends on the learned expectations brought to the musical material. Exposing such expectations interactively therefore opens a way to work not only with musical artifacts, but with the social evaluative frames through which those artifacts are explored and shaped. While the present demo validates the real-time integration and model-faithfulness of IRIDyOM, assessing its compositional value and the effects of different LTM/STM configurations requires future case studies and user feedback.

This perspective also suggests broader directions for future work at multiple levels of the framework. At the backend level, the current implementation remains limited to melodic expectation and a restricted target representation, but broader viewpoint support from the original IDyOM framework could be added, including rhythmic dimensions and richer interactions between pitch- and duration-based prediction. At the interface level, future work could expose these additional predictive variables through more flexible controls and visualizations while preserving the accessibility of the current workflow. At the interaction level, because the server exposes expectation as a real-time resource rather than only as an analytical output, the same predictive machinery could support richer studies of compositional decision-making, co-creative interaction, and experimental design beyond the present Max for Live interface. In particular, the current single-user setting could be extended toward more explicitly social and distributed forms of creativity (Saunders and Bown 2015; Linkola and Kantosalo 2019), including collective systems in which musical structure emerges through the interaction of multiple autonomous agents (Bono Rosselló et al. 2024). This would make it possible to study creativity not only as the exploration of a single predictive space, but as the interaction between multiple listener-relative spaces with different settings.

Code and demonstration. *IRIDyOM* is released as an open-source project, including the Max for Live device, Python backend, installation instructions, and example materials.[1] A short video demonstration documents the interface workflow, showing the main interaction modes and the real-time integration with Ableton Live.[2]

Acknowledgments

L. B. R acknowledges support from the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche).

[1] Project repository.

[2] Video demonstration.

References

[Boden 2004] Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

[Bono Rosselló et al. 2024] Bono Rosselló, L.; Alkilabi, M.; Tuci, E.; Bersini, H.; and Reina, A. 2024. Emergent orchestras: A modular framework for musical robot swarms. In *Proceedings of the 2024 Conference on Artificial Life*. MIT Press.

[Borg 2019] Borg, J. 2019. Somax 2: A real-time framework for human-machine improvisation. *STMS IRCAM & Aalborg University, Internal Student Report*.

[Cheung et al. 2019] Cheung, V. K. M.; Harrison, P. M. C.; Meyer, L.; Pearce, M. T.; Haynes, J.-D.; and Koelsch, S. 2019. Uncertainty and surprise jointly predict musical pleasure and neural responses. *Current Biology* 29(23):4084–4092.

[Cleary and Witten 1984] Cleary, J. G., and Witten, I. H. 1984. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications* 32(4):396–402.

[Conklin and Witten 1995] Conklin, D., and Witten, I. H. 1995. Multiple viewpoint systems for music prediction. *Journal of New Music Research* 24(1):51–73.

[Daikoku, Wiggins, and Nagai 2021] Daikoku, T.; Wiggins, G. A.; and Nagai, Y. 2021. Statistical properties of musical creativity: Roles of hierarchy and uncertainty in statistical learning. *Frontiers in Neuroscience* 15:640412.

[Daikoku 2018] Daikoku, T. 2018. Entropy, uncertainty, and the depth of implicit knowledge on musical creativity: Computational study of improvisation in melody and rhythm. *Frontiers in Computational Neuroscience* 12:97.

[Di Liberto et al. 2020] Di Liberto, G. M.; Pelofi, C.; Bianco, R.; Patel, P.; Mehta, A. D.; Herrero, J. L.; de Cheveigné, A.; Shamma, S.; and Mesgarani, N. 2020. Cortical encoding of melodic expectations in human temporal cortex. *eLife* 9:e51784.

[Gold et al. 2019] Gold, B. P.; Pearce, M. T.; Mas-Herrero, E.; Dagher, A.; and Zatorre, R. J. 2019. Predictability and uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience* 39:9397–9409.

[Guan, Ren, and Pelofi 2022] Guan, X.; Ren, Z.; and Pelofi, C. 2022. py2lispidyom: A python package for the information dynamics of music (idyom) model. *Journal of Open Source Software* 7(79):4738.

[Huron 2006] Huron, D. 2006. *Sweet Anticipation: Music and the Psychology of Expectation*. The MIT Press. CogNet special collection.

[Kayacik et al. 2019] Kayacik, C.; Chen, S.; Noerly, S.; Holbrook, J.; Roberts, A.; and Eck, D. 2019. Identifying the intersections: User experience+ research scientist collaboration in a generative machine learning interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–8.

[Linkola and Kantosalo 2019] Linkola, S., and Kantosalo, A. 2019. Extending the creative systems framework for the analysis of creative agent societies. In *Proceedings of the 10th International Conference on Computational Creativity*, 204–211. Association for Computational Creativity.

[Marion et al. 2025] Marion, G.; Gao, F.; Gold, B. P.; Di Liberto, G. M.; and Shamma, S. 2025. IDyOMpy: A new Python-based model for the statistical analysis of musical expectations. *Journal of Neuroscience Methods* 415:110347.

[Mas-Herrero and Marco-Pallarés 2025] Mas-Herrero, E., and Marco-Pallarés, J. 2025. Predictive processes shape individual musical preferences. *Proceedings of the National Academy of Sciences* 122(29):e2500494122.

[Pasquier et al. 2017] Pasquier, P.; Eigenfeldt, A.; Bown, O.; and Dubnov, S. 2017. An introduction to musical metacreation. *Computers in Entertainment (CIE)* 14(2):1–14.

[Pearce and Wiggins 2012] Pearce, M. T., and Wiggins, G. A. 2012. Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science* 4(4):625–652.

[Pearce 2018] Pearce, M. 2018. Idyom: A computational model of musical expectation. *Journal of New Music Research*.

[Roberts et al. 2019] Roberts, A.; Engel, J.; Mann, Y.; Gillick, J.; Kayacik, C.; Nørly, S.; Dinculescu, M.; Radebaugh, C.; Hawthorne, C.; and Eck, D. 2019. Magenta studio: Augmenting creativity with deep learning in ableton live. In *Proceedings of the international workshop on musical metacreation (mume)*, 11.

[Saunders and Bown 2015] Saunders, R., and Bown, O. 2015. Computational social creativity. *Artificial Life* 21(3):366–378.

[Sauvé and Pearce 2019] Sauvé, S. A., and Pearce, M. T. 2019. Information-theoretic Modeling of Perceived Musical Complexity. *Music Perception* 37(2):165–178.

[Wiggins 2006] Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Appendix

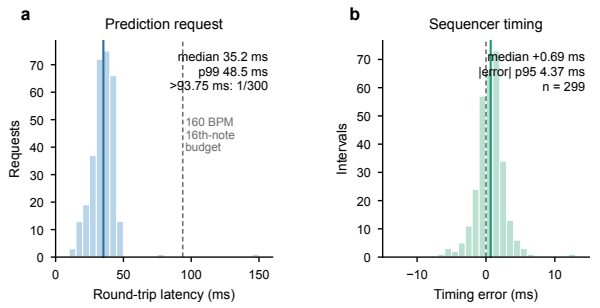


Figure 8: Server latency benchmark at 160 BPM sixteenth-note rate (93.75 ms inter-onset interval). (a) Round-trip latency for paced note-prediction requests over the local JSON/TCP interface with the interactive musical delay minimized. (b) Timing error of autonomous sequencer note-on messages relative to the 93.75 ms target interval. The server returned predictions within the musical time budget for 299/300 requests (median 35.2 ms; p99 48.5 ms), while autonomous sequencer output remained close to the target grid (median timing error +0.69 ms; 95th percentile absolute error 4.37 ms).

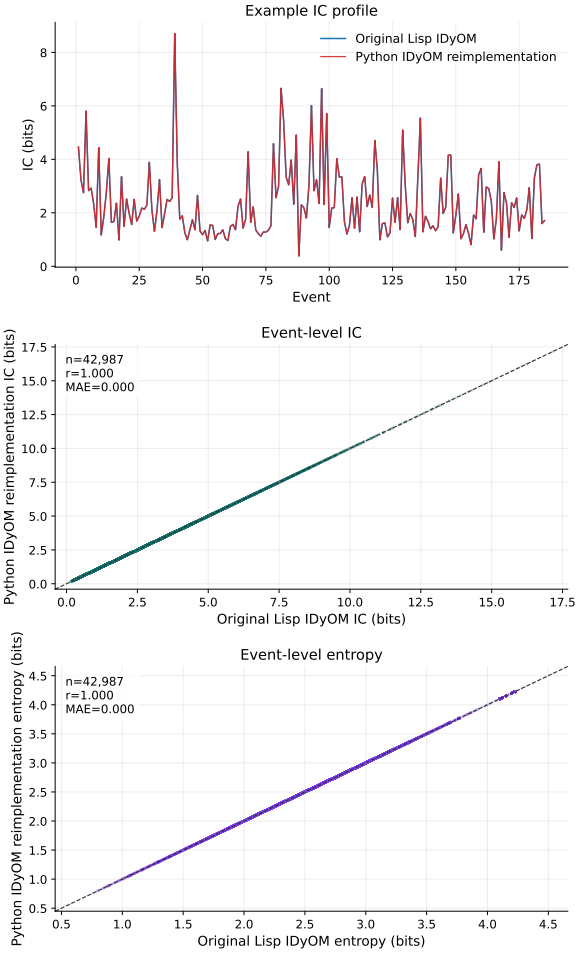


Figure 9: **Faithfulness backend check against the original IDyOM implementation.** Event-level information content and entropy from the Python IDyOM reimplementation are compared with the original Lisp IDyOM on a 900 files pitch-only cross-validation benchmark, using matched fold splits and memory-store settings. The near-identity IC profile and scatter plots show numerical agreement between implementations under these settings.

An Initial Design for Generative Game Mechanics via LLM Function Emulation

Brad Spendlove
Department of Mathematics & Computer Science
Randolph College
Virginia, USA
bspendlove@randolphcollege.edu

David Kline
Randolph College
Virginia, USA
dkline@randolphcollege.edu

Abstract

A video game's mechanics dictate the outcome of player input and world state changes. Game mechanics are functions invoked in response to game actions (often player input) that result in changes to the game state. We present an initial design for a game system that takes on the creative responsibility of determining mechanical outcomes of key game actions in a dungeon crawler video game by emulating its mechanic functions via an LLM. The system emulates those functions by prompting an LLM with a text description of the mechanic and the input to the function. The LLM's response is parsed to extract the output of the emulated function. We present an initial implementation of combat and enemy spawning with LLM function emulation replacing human-coded game mechanics. Initial results suggest that chain-of-thought prompting enables the LLM to make sensible game updates based primarily on plain-text descriptions of enemies, grounded in the game state.

Introduction

Video games are traditionally constrained by the limits of hard-coded rules, which can make complex interactions predictable or labor-intensive to design and implement. Large language models (LLMs) present an alternative paradigm: by interpreting player input and generating gameplay results in real time, LLM-driven game mechanics can enable flexible emergent gameplay experiences that may be otherwise difficult to achieve.

"A game is a voluntary interactive activity, in which... players follow rules that constrain their behavior, enacting an artificial conflict that ends in a quantifiable outcome" (Zimmerman 2004). Video games as interactive computer programs maintain a game state, consisting of all the variables relevant to the game scenario. The game state is repeatedly updated based on player input and the state itself, as depicted in Figure 1. Game mechanics are functions that map the current game state and player input to an updated game state (Sicart 2008).

This game loop paradigm is ubiquitous in game development. For example, in a 2D game, a common game state variable is the player character's (x,y) position. When the player gives directional input, the position variable is changed via some function in every iteration of the game loop. The position variable can also change based on other game state variables, such as gravity or object collisions. The game state is rendered into video and audio to communicate that state to the player, such as changing the location of the character on the screen based on its position variable.

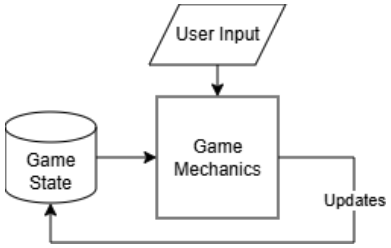


Figure 1: Fundamental game update loop.

The exact manner in which game state variables are updated based on input and other variables constitutes the game's mechanics, which the players learn and interact with. Modern game states and updates are often very complex and incorporate analog input, physics simulations, and non-player character AI. But whether a game is simple or complex, its game mechanics are what make it the game it is.

Human game developers typically hard-code game mechanic functions. We seek to replace human-written game mechanics by emulating those functions using a large language model. Specifically, we prompt the LLM with the game state and player input and ask it to generate state updates that are then parsed and applied to the game state. Prompting an LLM this way implements the state update function in a more flexible and dynamic manner than hard-coded game mechanics.

Our game system also uses the LLM to generate text descriptions of game actions, similar to games like NetHack (The NetHack DevTeam 1987). It is well known that LLMs are adept at creating text. This work focuses on how LLMs can also generate matching game state updates that are both creative and mechanically sensible.

The primary contribution of this paper is the presentation of a novel method for executing video game mechanics. This short paper presents our design and initial system implementation. Preliminary testing suggests that our game system achieves our design goals of implementing game me-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

chanics by prompting an LLM with game state information instead of directly calculating game state updates through hard-coded functions. Future work will include further evaluation, including user studies and ablation testing.

Related Work

LLM-driven Games

There is much research into using LLMs to power games. Many of those systems build directly on top of LLMs' natural language capabilities to implement text-based games wherein the players' inputs to the game are natural language text. The game presented herein is not text-based. The player interacts using traditional button controls (primarily directional and selection buttons), and the game is presented visually as a character moving around a 2D grid-based representation of a dungeon. This places our system in a distinct genre from text-based games (Lee et al. 2014) and represents distinct design and gameplay approaches.

As reported in a survey by Gallotta et al. (2024), other existing research into LLM-generated game mechanics are distinct from our approach. Those systems fall into one of three categories: they use LLM prompts and responses directly as the input and output to a text-based game, they use LLMs to select and invoke hard-coded functions (Song, Zhu, and Callison-Burch 2024), or they use LLMs to create or modify source code instructions that will be executed when those methods are invoked (Nasir et al. 2025).

PAYADOR (Góngora et al. 2024) is an example of a text-based interactive fiction grounded in a persistent world state, including character description, location, and inventory. It is similar in form to AI Dungeon (Hua and Raley 2020), but achieves a more consistent game world by grounding LLM prompts and responses in the persistent game state.

LLMs with Function Calling

In tabletop role-playing games, the Game Master (GM) is responsible for maintaining and communicating a game world in the theater of the mind. Thus, the GM fulfills a role similar to the game loop in a video game. Song et al. (2024) present a system for running a tabletop role-playing game as the Game Master (GM). Their system is dialogue-based, in which players narrate their game actions and the system responds with the outcomes of those actions.

In order to maintain a consistent game state during gameplay, their system utilizes LLMs with function calling (Li et al. 2024). This enables the LLM to include invocations to externally defined functions in its responses. Those functions directly update the game state according to the parameters provided upon invocation. Gallotta et al. (2024) also employ LLMs with function calling to create game content data in dialogue with a human designer.

The system presented herein differs from this function calling approach in that there are not externally defined functions that the LLM invokes in its responses. Instead, we use the LLM to emulate the behavior of a function directly, as described above. To the best of our knowledge, this present work is the first research that applies this function emulation approach to implementing game mechanics in a typical action game with non-textual input.

Implementation

Large Language Model and Prompting

Our system uses Llama 3.3 (Grattafiori et al. 2024) as the LLM that drives the game mechanics. We use the off-the-shelf model with no fine-tuning, accessing it via the Groq API[1]. The model has 70 billion parameters and is a high-speed and cost-effective improvement over previous models. Those characteristics are desirable for a model that will run a game engine due to the potentially high volume of LLM calls for updating the game state.

Prompt engineering is a critical part of the design process for any system that uses LLMs (Giray 2023). Our game system incorporates two primary prompting patterns: chain-of-thought prompting (Wei et al. 2022), which involves breaking a reasoning task into single logical steps, and few-shot prompting (Brown et al. 2020), in which example outputs are provided in the prompt to help the LLM's response align with those examples.

Our game system makes use of chain-of-thought prompting to work through various game state updates, treating each as a single question instead of asking for the whole suite of updates at once. It uses few-shot prompting to specify an output format for the LLM's game state updates, facilitating parsing and grounding in the existing game state.

Game System

The game we implemented is a top-down 2D dungeon crawler with roguelike elements (Szabados et al. 2023), namely turn-based gameplay and grid-based movement. A turn-based game loop that only advances when the player provides input allows our system to calculate game state updates much less frequently than a real-time game loop.

Within this game design, we selected a subset of game mechanics to implement using LLM updates. All game mechanics are method invocations that modify the game state, so any of the mechanics in our game could be implemented either with traditional hard-coded methods or with LLM-emulated methods. We selected the two most impactful and defining mechanics of our game to implement this way. We call these selected mechanics the *core* mechanics, and we will refer to the other non-LLM mechanics as *simple*.

These names are not arbitrary. The set of non-LLM game mechanics we call *simple* handle common or routine interactions for which the mechanical outcome is simple to calculate. These interactions are processed into game state updates using traditional hard-coded functions. Examples include moving the player character around the grid in response to directional input and accessing the inventory menu to equip items. These workhorse interactions benefit from being lightweight and predictable.

Core mechanics are the two main mechanics that define our game: combat and enemy spawning. These have the

[1]https://console.groq.com/docs/model/llama-3.3-70b-versatile

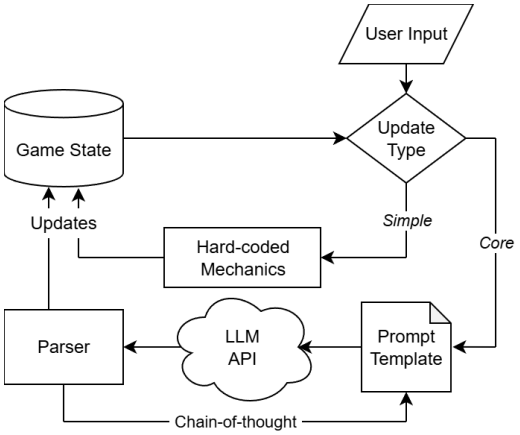


Figure 2: Our game system loop that divides mechanics into two categories: *simple* and *core*. Only the latter are processed through the LLM.

most impact on players, so implementing them with LLM-driven updates will most directly influence the gameplay experience. The exact outcome of these mechanics is unknown due to the wide variety of outputs LLMs generate. We believe this will be beneficial to the player experience, because "games require uncertainty to hold our interest" (Costikyan 2013). Combat is more interesting if it plays out differently every time, and LLM-generated enemies will have more variety than if they were drawn from a predetermined set.

Figure 2 shows our gameplay loop. When player input is registered, it is combined with the game state to determine whether the resulting game state updates are *simple* or *core*. If they are *simple*, they are processed using traditional hard-coded functions. If they are *core*, then the mechanic is processed by prompting an LLM.

Prompting

Each *core* mechanic has one or more prompt templates associated with it. Each template is filled with relevant game state data and used to prompt the LLM via an API call. The LLM's response is then parsed to extract updates that are applied to the game state.

Some interactions use a sequence of prompt templates that implement chain-of-thought reasoning. Each prompt is given to the LLM in sequence to resolve multi-step interactions, such as spawning new enemies. Game state updates are applied at the end of the chain-of-thought sequence after all relevant updates have been collected.

It is critical to the *core* game mechanic functions that both their input (current game state and player input) and output (game state updates) are grounded in the game state. Therefore, our *core* mechanic prompt templates include two features. First, the templates include variables drawn directly from the game state so that they appear explicitly in the prompt. Second, they employ few-shot prompting. The few-shot examples specify both the output format expected by our parser and the set of variables that the LLM can update for a given scenario. All of the specific prompts our system used are given in Appendix A.

The LLM's response to the templated prompt is parsed to separate the textual description of the interaction and extract the corresponding game state updates. The updates are applied to the game state variables, and the description is both used for chain-of-thought prompting and displayed to the player. A response to the combat prompt is expected to include a text description followed by two state updates: the amount of damage DEALT by the player to the enemy, and the amount of damage RECEIVED in return.

Because LLMs are trained on linguistic tokens that have no inherent numeric comparability (Qian et al. 2023), we experimented with asking the LLM for state updates using qualitative words instead of numbers. The template's few-shot examples represent combat damage with the words NONE, LOW, MEDIUM, HIGH, or FATAL instead of numbers. Our parser then interprets those words into numerical values by mapping them to a corresponding percentage of the damaged character's health. A similar approach is used to represent each combatant's current health variable in the prompt template as HEALTHY, WOUNDED, or MAIMED.

Chain-of-thought prompting is used for spawning enemy reinforcements. The LLM is prompted with a template that asks how many enemy reinforcements are summoned in the text description that results from the combat mechanic, if any. If the response is greater than zero, that number and the combat description are used to populate another zero-shot template prompting the LLM to generate names and descriptions for the resulting reinforcements. Those names and descriptions directly comprise new enemy objects.

The final prompt template in the chain gives the LLM each generated enemy description and a list of sprite names, and asks it to select an appropriate sprite, which is used to represent that enemy graphically.

Gameplay

The game mechanics described above were implemented in a top-down 2D dungeon crawler game prototype. In it, the player controls a knight in a simple dungeon, attacking any enemies they come in contact with. The game is turn-based; the game loop executes one state update iteration, called a step, each time the player moves. The player inputs are 4-direction movement, a "pass" button to skip player movement for that step, and a simple inventory screen that the player interacts with using the mouse.

The default dungeon room is initially populated with three enemies: a necromancer, a greater necromancer, and a goblin. Every step, the enemies move in a random direction unless they are adjacent to the player. If they are, they will attack the player instead. Combat is a *core* mechanic that is implemented by prompting the LLM, as described above. If the player moves into an enemy, combat will involve the player and enemy attacking each other.

The game state changes resulting from each combat are generated and applied as described above. The text description is displayed to the player, along with the amount of damage each combatant dealt and their remaining health. If new enemies are spawned via that chain-of-thought-prompted mechanic, they are added to the enemy list in the

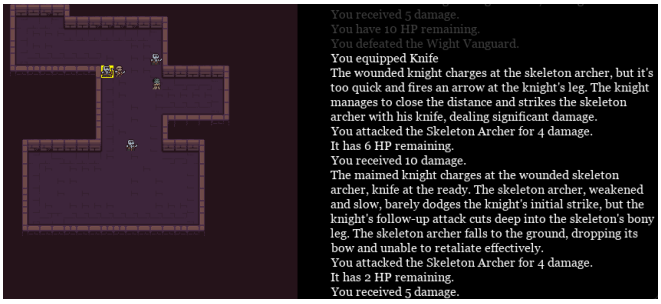


Figure 3: Cropped screenshot showing combat between the player character (highlighted in yellow) and an LLM-generated enemy. (Portion of text dimmed for emphasis.)

game state. They appear in the dungeon room and will move and interact with the player.

The player can open a simple inventory screen to equip different pre-loaded weapons. Those items are also stored with only a name and description. Any differing effects of the weapons are decided by the LLM when generating state updates from a combat encounter. In future work, we will explore how items can be generated via the LLM, similar to how enemies are spawned in the current system.

Discussion

Initial testing suggests that our game system with LLM-emulated mechanic functions exhibits novel and interesting gameplay. More testing and evaluation are needed, but we present some preliminary findings here.

Natural Language Object Representation

Our system can generate mechanical outcomes based solely on simple text descriptions of the player and enemies. When enemies are generated via the post-combat chain-of-thought process, the resulting objects store only a name, a text description, and a selected sprite image. Notably, those objects contain no explicit gameplay logic.

The name and description can then be fed back into the LLM combat mechanic when the player encounters those enemies on later steps. The results of combat with those enemies are shaped entirely by their text descriptions, replacing customized hard-coded logic that a human game developer would otherwise have to implement for each enemy. The LLM is free to draw on its training data to generate a wide variety of enemies, then implement implicit combat mechanics appropriate to each.

Figure 3 shows a screenshot of our prototype game, including the player character highlighted in yellow and several enemies. The UI includes text descriptions of two rounds of combat with the skeleton enemy next to the player (i.e., the player gave two inputs to move into the enemy, each resulting in a combat encounter).

That skeleton enemy was previously generated by the LLM during combat with a now-defeated necromancer enemy. The LLM named the generated enemy "Skeleton Archer" and gave it the following description: "A skeleton armed with a bow and quiver of arrows, providing ranged support to the necromancer." Finally, the LLM chose an appropriate sprite from a pre-loaded list, and the new skeleton enemy was added to the game state.

Later, when the player fought that enemy, the combat description the LLM generated clearly reflects characteristics of a skeleton archer (as seen in the text output), including the skeleton attacking with a fired arrow and the player's attack "cut[ting] deep into the skeleton's bony leg".

Combat Outcomes

We conducted experiments to test how the game resolves various combat scenarios. To isolate the tests from other aspects of the game, these were invocations of the emulated combat mechanic function using constructed game state data. The full results are reported in Appendix B.

The results of these tests demonstrate the LLM-emulated combat mechanic's ability to generate appropriate outcomes for the different scenarios, grounded in the game state variables. Combat against a weak enemy commonly resulted in dealing high damage and receiving little in return. When fighting a strong enemy, the player consistently dealt less damage and took the highest damage out of all scenarios. A maimed player dealt less damage and took more than a wounded or healthy player. A pair of tests on the effects of different weapon descriptions, e.g. a small knife compared to a large dragon-slaying sword, did not reflect the expected difference in state updates.

A notable emergent behavior is that across all scenarios, the player never dealt zero damage (NONE) and never received more than MEDIUM damage. This represents an implicit game design decision implemented by the LLM combat mechanic, apparently to bias the game toward player success. This is most apparent in a scenario involving a strong enemy; the LLM reduced the player's damage output and increased the enemy's, but neither exceeded those apparent bounds. The LLM was given no instructions in this regard and chose to generate updates according to those rules of its own volition.

This suggests that our game system is taking on creative responsibilities regarding the game's mechanics. Future work will more thoroughly evaluate the system and these claims.

Creative Responsibility

The creative artifact of our game system is a function that maps a game state and player input to an updated game state. Ritchie (2007) proposes a set of three criteria for assessing creative artifacts: typicality, novelty, and quality.

Our system's approach to implementing that update function results in novel gameplay interactions. Because the mechanics are not hard-coded, they can take dynamic forms that change during gameplay. The value of our system's resulting gameplay remains to be evaluated more rigorously, but at present it appears to achieve a reasonable baseline of functional and interesting gameplay. Grounding the LLM in game state data directly enhances the typicality of our system's creative artifacts and helps it avoid creating low-value artifacts such as nonsensical game state updates.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Costikyan, G. 2013. *Uncertainty in games*. Mit Press.

Gallotta, R.; Todd, G.; Zammit, M.; Earle, S.; Liapis, A.; Togelius, J.; and Yannakakis, G. N. 2024. Large language models and games: A survey and roadmap. *IEEE Transactions on Games*.

Gallotta, R.; Liapis, A.; and Yannakakis, G. 2024. Consistent game content creation via function calling for large language models. In *2024 IEEE Conference on Games (CoG)*, 1–4. IEEE.

Giray, L. 2023. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering* 51(12):2629–2633.

Góngora, S.; Chiruzzo, L.; Méndez, G.; and Gervás, P. 2024. Payador: A minimalist approach to grounding language models on structured data for interactive storytelling and role-playing games. In *Proceedings of the 15th International Conference on Computational Creativity*. Association for Computational Creativity.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hua, M., and Raley, R. 2020. Playing with unicorns: Ai dungeon and citizen nlp. *DHQ: Digital Humanities Quarterly* 14(4).

Lee, J. H.; Karlova, N.; Clarke, R. I.; Thornton, K.; and Perti, A. 2014. Facet analysis of video game genres. *IConference 2014 Proceedings*.

Li, Z.; Chen, Z.; Ross, M.; Huber, P.; Moon, S.; Lin, Z.; Dong, X. L.; Sagar, A.; Yan, X.; and Crook, P. A. 2024. Large language models as zero-shot dialogue state tracker through function calling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8688–8704.

Nasir, M. U.; Li, Y.; James, S.; and Togelius, J. 2025. Mortar: Evolving mechanics for automatic game design. *arXiv preprint arXiv:2601.00105*.

Qian, J.; Wang, H.; Li, Z.; Li, S.; and Yan, X. 2023. Limitations of language models in arithmetic and symbolic induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9285–9298.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Sicart, M. 2008. Defining game mechanics. *Game studies* 8(2):1–14.

Song, J.; Zhu, A.; and Callison-Burch, C. 2024. You have thirteen hours in which to solve the labyrinth: Enhancing ai game masters with function calling.

Szabados, G. N.; Bácsné Bába, É.; Fenyves, V.; Bács, Z.; Molnár, A.; Ráthonyi, G.; Rizwan, H.; and Orbán, S. G. 2023. Roguelike games: The way we play. *International Journal of Engineering and Management Sciences* 7(4):80–92.

The NetHack DevTeam. 1987. Nethack. `https://www.nethack.org`. Computer software.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35:24824–24837.

Zimmerman, E. 2004. Narrative, interactivity, play, and games: Four naughty concepts in need of discipline. *First person: New media as story, performance, and game* 154.

Appendix A: Prompts

All prompts are templated to be populated with game state data, indicated by variable names surrounded by curly braces.

Combat Prompt (second few-shot example omitted for brevity):

```
Resolve this video game combat scenario,
  given the name and description of the
  player character and enemy to engage.
  Adjust the strength of the player and
  enemy based on their status:  HEALTHY,
  WOUNDED, or MAIMED.
Keep the output brief, within 4 sentences if
  possible.
Return a damage value of LOW, MEDIUM, or
  HIGH. If the player or enemy is defeated,
  the dealt or received value should be
  FATAL.
Example 1:
Player class:  Knight
Status:  HEALTHY
Class description:  A knight, armed with a
  spear.  A powerful melee combatant.
Enemy:  Skeleton Grunt
Status:  HEALTHY
Enemy description:  A weak skeleton, armed
  with a knife.
Output:  The skeleton grunt viciously
  slashes at the knight, but its knife
  cannot penetrate his armor.  The knight
  retaliates by bashing the grunt with the
  pole of his spear, knocking it to the
  ground.  DEALT: MEDIUM, RECEIVED: NONE
Example 2:
...
Using the format from the examples, run
  a combat scenario with a {enemy.status}
  enemy {enemy.name}, with the description
  {enemy.description}, attacking a
  {player.status} player {player.name}, with
  the description {player.description}.
```

Enemy Generation Chain-of-thought Prompt 1: Identifying Reinforcements:

```
Read this scenario {scenario}.  If it
  explicitly states the enemy has called
  reinforcements, determine the number of
  enemies to be summoned.
If enemies are summoned, the number of
  enemies should be 1 at minimum, or 3
  at maximum.  If no reinforcements are
  directly mentioned, provide 0 for the
  number of each enemy.
Format the enemy count like so:
Enemies:  2
```

Enemy Generation Chain-of-thought Prompt 2: Generating Enemy Names and Descriptions:

```
Based on the provided combat scenario
  {scenario} where the enemy calls for
  reinforcements, create appropriate enemies
  to be summoned.
Here are some example enemies and their
  descriptions:
Format the enemies like so:
Enemy 1:  Skeleton Grunt, A weak skeleton,
  armed with a knife.
Enemy 2:  Reaper, The skeleton of a strong
  warrior, armed with a scythe.
Create a maximum of {count} enemies.
```

Enemy Generation Chain-of-thought Prompt 3: Selecting Enemy Sprite:

```
Given the name {name} and description {desc}
  of an enemy, select the sprite from
  {sprites} that suits them the best.
```

Appendix B: Combat Test Results

Scenario	Weak Enemy		Strong Enemy		Wounded	
Damage	Dealt	Received	Dealt	Received	Dealt	Received
NONE	0	4	0	0	0	0
LOW	0	6	10	0	1	8
MEDIUM	1	0	0	10	8	2
HIGH	1	0	0	0	1	0
FATAL	8	0	0	0	0	0

Scenario	Maimed		w/ Knife		w/ Slayer	
Damage	Dealt	Received	Dealt	Received	Dealt	Received
NONE	0	0	0	1	0	0
LOW	7	5	0	9	0	10
MEDIUM	3	5	6	0	7	0
HIGH	0	0	2	0	3	0
FATAL	0	0	2	0	0	0

The outcomes of testing the LLM-emulated combat mechanic function in several scenarios, each consisting of different game state variables. Each column records the results of running the test 10 times, recording how many times each damage description was returned for each scenario. Recall that those qualitative words were used in place of numerical results to better align with LLM representations. Damage dealt and received are from the player's perspective (i.e., *"dealt" by the player* and *"received" by the player*).

Each scenario uses the same player description: "A knight armed with [weapon]". The player has full health in all scenarios except for the Wounded and Maimed scenarios, in which the player has 50% health and 5% health, respectively. The default weapon description is "a well-forged longsword". The w/ Knife scenario replaces this with "a well-maintained knife, very sharp, but it has short range", and the w/ Slayer scenario uses "a massive sword, empowered by slain dragons". The enemy in the Weak scenario is described as "a weak, pathetic goblin", the enemy in the Strong scenario is "a huge, indestructible granite construct", and the rest use the default of "a small but fierce goblin warrior".

ANTLER 1: A Multimodal Co-Creative Tutor for Stylistic Exploration in Tambour Embroidery

Anne Toomey[1], Zowie Broach[1], Mojtaba Kargar[1], Samira Valorian[1], Henrietta Dent[1], Aemika Zou[2], Chaonan Lin[2]

[1]The Royal College of Art (RCA), UK
[2]Laboratory for Artificial Intelligence in Design (AiDLab), Hong Kong

anne.toomey@rca.ac.uk, zowie.broach@rca.ac.uk, mojtaba.kargar@rca.ac.uk, samira.valorian@rca.ac.uk
henrietta.dent@rca.ac.uk,aemika.zou@connect.polyu.hk,chaonan.lin@rca.ac.uk

Abstract

Tambour embroidery is an embodied practice where expert knowledge lives in hand rhythm, needle angle, and material control. We present ANTLER 1, a style-aware co-creative tutoring *framework* for embodied craft, instantiated as a working multimodal prototype. The prototype captures live tambour stitching through synchronised video, Inertial Measurement Unit (IMU) telemetry, and hand tracking, transforming segmented stitch behaviour into profile-conditioned coaching feedback. We describe the framework design and present illustrative use-case scenarios showing how the same observed sequence is interpreted differently under precision, expressive, and hybrid profiles — contributing an early-stage design pattern for grounded, style-aware co-creative tutoring in embodied craft.

Introduction

Tambour embroidery emerged from Indian Ari hook work and later entered Ottoman and French couture traditions, yet its expertise remains embodied in wrist rhythm, hook angle, and material control rather than verbal instruction alone (Mora and others 2024; Salter and others 2010). As master artisans age, this tacit knowledge faces irreversible loss (Mora and others 2024; Boden 1992). Existing digital tools orient toward correctness and do not help practitioners explore how the same movement carries different creative meanings under different stylistic intentions.

ANTLER 1 frames AI tutoring as a co-creative, style-aware interaction grounded in observed stitch behaviour. The system acts as a computationally grounded coaching partner: given a target design context, observed stitch features, and practical constraints, it generates *what-if alternatives* that let the practitioner compare efficiency-oriented, expressive, and hybrid execution paths. Prior work addresses multimodal performance analysis and Large Language Model (LLM)-based tutoring, but does not address style-contingent feedback grounded in embodied craft action; ANTLER 1 bridges this gap (Boden 1992; Ritchie 2007; Yannakakis, Liapis, and Alexopoulos 2018; Davis and others 2013).

Three research questions structure this paper:

- **RQ1:** How can multimodal segmentation of tambour stitching support style-aware feedback in embodied creative practice?
- **RQ2:** How does grounding LLM feedback in observed performance features change perceived usefulness compared to ungrounded generation?
- **RQ3:** How can a co-creative tutor vary not only its recommendations but also its tutoring stance across precision, expressive, and hybrid profiles?

The contribution is a *design framework and early-stage prototype*, not a fully validated end-to-end system; claims are scoped accordingly throughout.

A corrective system applies a single target state; a co-creative system maintains multiple valid states and helps the practitioner navigate between them (Yannakakis, Liapis, and Alexopoulos 2018; Boden 1992). ANTLER 1 operationalises this through style profiles that redefine what counts as deviation: the same IMU signature that triggers a correction under precision may invite expressive exploration under an expressive profile (Franceschelli and Musolesi 2024; Anonymous 2024).

Related Work

Multimodal skill-analysis systems in surgery and motion training integrate video and kinematic traces to assess performance, but they optimise for correctness or safety rather than style-sensitive interpretation in creative craft practice (Pan, Wang, and others 2023; Chen et al. 2025). They demonstrate the feasibility of multimodal fusion for performance feedback, but their single-target optimisation is fundamentally incompatible with a domain where the same action may simultaneously be a technical deviation and a valid stylistic gesture.

LLM-based tutoring systems operate over textual or symbolic inputs and cannot distinguish practitioner-specific rhythmic signatures from generic craft description (Jain and others 2025; Mills and others 2025). Within computational creativity, co-creative systems are understood as interactive partners rather than autonomous generators, evaluated for situated interaction over output quality alone (Yannakakis, Liapis, and Alexopoulos 2018; Davis and others 2013; Mora and others 2024; Yannakakis and Togelius 2014).

The paper's two conceptual baselines are literature-grounded: a rule-based correction system aligned with prior motion-assessment work, and a vanilla LLM receiving style descriptions without multimodal evidence, aligned with text-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.



Figure 1: ANTLER 1 capture setup during a live tambour embroidery session. Overhead and underside cameras record stitch trajectory, fabric interaction, and embodied hand movement in real time. A sensor-integrated tambour tool captures fine-grained needle-motion dynamics used for segmentation and downstream co-creative coaching feedback.

centric coaching systems. ANTLER 1 sits between and beyond these baselines by combining sensor grounding with style-conditioned stance variation. Recent surveys highlight explainability and process-transparency as core properties distinguishing genuine co-creative AI from post-hoc description systems (Llano et al. 2020; Bryan-Kinns and others 2023; Franceschelli and Musolesi 2024; Anonymous 2024); the style-profile architecture and what-if framing of ANTLER 1 are direct responses to these design requirements.

System Overview

Capture. Input: live tambour session. Output: synchronised video stream, IMU telemetry (100 Hz), and hand-skeleton joint positions at 30 fps. Figure 2 illustrates how three sensor streams combine. Video captures stitch trajectory and spatial layout; IMU captures wrist dynamics and rhythm; hand skeleton captures grip posture and needle-hand coordination. Table 2 maps each modality to extracted features and craft-relevant interpretation.

Segmentation. A stitch boundary is defined as the moment of minimum wrist velocity between a completed pull-through and initiation of the next insertion. Boundaries were established by manual annotation (Label Studio); a segmentation model trained on these labels performs real-time boundary detection at runtime via IMU-derived wrist-velocity drops, covering Pull, Chain, and Continuous Bead Insertion stitch types.

Feature Extraction. Spacing compression: percentage deviation from mean inter-bead distance. Tempo variance: coefficient of variation of inter-boundary durations. Wrist rotation direction is derived from gyroscope data; pull-through micro-adjustments are short-duration velocity spikes within a segment.

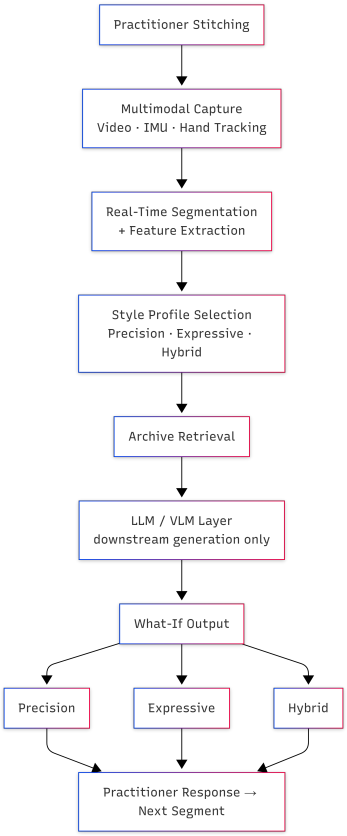


Figure 2: ANTLER 1 system pipeline. IMU wrist-velocity drives both feature extraction and real-time stitch boundary detection. Style profile selection modulates how the same feature is interpreted before retrieval and generation.

Style profiles. Style profiles are operational parameter sets, not prompt themes. Each encodes target regularity, variation tolerance, tempo elasticity, and tutoring stance. Table 1 summarises the three profiles. The key profile-dependent feature is error type G (Rhythm Arrhythmia): under Precision it is penalised; under Expressive the same rhythmic pattern may be reinterpreted as creative hesitation; under Hybrid it is permitted only at designated focal points.

Retrieval. Archive entries are ranked by joint match on `error_type`, `style_profile_tag`, and `stitch_type`; ties broken by recency in session history. Each entry stores stitch type, error labels [A–G], feature summary, technique reference, and stance tag.

Generation. The LLM receives a structured prompt containing the top-ranked archive entry, the current feature summary, and the active style profile; it renders these into coaching text and performs no independent classification.

Grounding example. For a Chain segment with high tempo

Modality	Features	Craft relevance
Video	spacing, curvature, trajectory density	regularity, motif geometry
IMU	tempo elasticity, motion drops, wrist rotation direction, pull-through micro-adjustments	rhythm, directional control, gestures associated with tension management
Skeleton	grip openness, wrist angle, coordination	technique type, needle control

Table 1: Modality-to-feature mapping in the current prototype. The IMU row reports motion patterns associated with tension management rather than direct measurement of thread tension.

Profile	Interpretation	Stance
Precision	variation reduced; type G penalised	corrective
Expressive	type G reframed; intentionality sought	reflective
Hybrid	controlled base; focal variation OK	curatorial

Table 2: Style profiles and tutoring stances. Error types A–G form a curated taxonomy orthogonal to style profile tags.

variance, 8% spacing compression, and clockwise wrist rotation drift: under the Precision profile the system generates *"Your chain is tightening and rotating on the pull-through. Stabilise the wrist earlier to keep spacing even."* Under the Expressive profile the same features yield *"The directional drift gives the chain organic movement. Repeat it at the motif edge to make it read as deliberate style."* This single-input, multi-output behaviour is the operational definition of style-aware co-creative tutoring in ANTLER 1.

Iterative loop. Each completed stitch triggers boundary detection, feature extraction, profile-conditioned retrieval, and feedback generation without intervention. The practitioner's response informs the next segment; an active learning component flags uncertain segments for expert review over multiple sessions (Yannakakis, Liapis, and Alexopoulos 2018; Davis and others 2013).

Use Cases and What-If Analysis

The scenarios below are illustrative mechanism examples constructed from a representative observed sequence to demonstrate the framework's profile-dependent behaviour. They draw on outputs generated during the pilot session but are presented as design demonstrations; practitioner responses are reported in Section .

All three scenarios share one observed sequence: a floral transition showing spacing compression, directional drift during pull-through, and uneven rhythmic control. A corrective system flags this as deviation; ANTLER 1 surfaces profile-dependent alternatives showing the same evidence admits multiple valid interpretations.

Technique background. *Lunéville crochet* works from the reverse of the fabric in a continuous chain, suited to repetitive, linear, high-volume bead application where consistency and speed are priorities. *Ari embroidery* works from the surface with a fine awl-hook, giving finer control over direction, spacing, and ornamental density at the cost of speed — supporting organic, non-linear motifs. The choice reflects an atelier judgement about where efficiency should yield to expressive intent: precisely the trade-off ANTLER 1's what-if layer is designed to surface.

Scenario A — Precision. Under this profile the system interprets the hesitation as a signal to restore base rhythm and spacing regularity. It surfaces Lunéville crochet as an efficiency-oriented alternative: continuous thread construction reduces bead-placement pauses and supports consistent chain spacing. The tutoring tone is corrective but supportive — recommending tightened tempo while preserving the practitioner's grip angle.

Scenario B — Expressive. The same hesitation is reframed as a possible stylistic breath, provided it is repeated intentionally. The system endorses Ari technique, whose individual pick-and-place method supports organic, non-linear bead orientation and richer ornamental surfaces. The tutoring tone is reflective, asking whether the pause marks compositional accent rather than mechanical difficulty.

Scenario C — Hybrid. The system preserves the controlled base chain (Lunéville register) while allowing expressive hesitation at focal floral nodes. This scenario most clearly demonstrates co-creative partnership: the system presents a conditional trade-off rather than prescribing a technique (Llano et al. 2020; Yannakakis, Liapis, and Alexopoulos 2018).



Figure 3: Left: Ari technique — surface-worked, individual bead placement, organic forms. Right: Lunéville crochet — reverse-worked, continuous thread, linear bead lines. The same IMU hesitation pattern can support recommending either technique depending on the active style profile.

What-if generation draws on four grounded inputs: observed motion features (tempo elasticity, directional drift, spacing, pull-through micro-adjustments); the active style profile; a technique archive mapping Lunéville and Ari char-

acteristics to design intent; and a time map derived from the practitioner's own session rhythm. Time estimates are indicative and user-specific, not general benchmarks (Jain and others 2025; Mills and others 2025).

Table 3 operationalises the "same evidence, different meaning" claim by showing how identical observed features yield different technique recommendations under each profile.

Feature	Precision	Expressive	Hybrid
Hesitation (type G)	Correct: restore tempo	Reframe: stylistic breath	Allow at focal nodes
Spacing compression	Reduce: regularise	Preserve if intentional	Tighten base chain
Technique suggestion	Lunéville (efficiency)	Ari (expression)	Lunéville base + Ari focal
Tutoring tone	Corrective	Reflective	Curatorial

Table 3: Same observed sequence interpreted across three style profiles. This table operationalises style-contingent interpretation as the system's core CC contribution.

Preliminary Evaluation

The segmentation model was trained on data from 50 recorded sessions covering three stitch types: Pull, Chain, and Continuous Stitch with Black Bead insertion. Boundaries and error types were annotated manually before model training.

A single-session evaluation was then conducted with one expert embroiderer (5+ years atelier experience) performing the same three stitch types. Three conditions were compared: (1) ANTLER 1 with style profiles active; (2) a corrective-only baseline using the same sensor setup with a fixed single target; and (3) an ungrounded LLM receiving a plain-text session description without sensor data.

The five evaluation dimensions — *plausibility*, *usefulness*, *style fidelity*, *stance appropriateness*, and *perceived specificity* — were selected to distinguish ANTLER 1 from both baselines: a corrective system may remain plausible without style fidelity, while an ungrounded LLM may lack specificity. A structured interview was used to suit the single-participant design, at the cost of inter-rater reliability. Findings are reported as directional indicators; the single-participant design does not support statistical inference.

Dimension	ANTLER 1	Corrective	LLM only
Plausibility	high	high	medium
Usefulness	high	medium	low
Style fidelity	high	low	low
Stance appropriateness	high	low	medium
Perceived specificity	high	low	low

Table 4: Directional pilot ratings across three conditions (single practitioner; no statistical inference claimed).

All three profiles were rated plausible and distinguishable. Expressive and Hybrid profiles surfaced options not visible through correctness-only feedback. In one instance the Expressive profile over-interpreted a hesitation caused by thread resistance as stylistic intent — illustrating the core disambiguation challenge between expressive choice and mechanical difficulty. A structured participant study across skill levels is planned as the next evaluation phase (Mora and others 2024; Yannakakis and Togelius 2014).

Discussion and Limitations

Returning to the research questions: RQ1 is addressed by multimodal stitch-event segmentation, RQ2 directionally by improved perceived specificity and usefulness over the ungrounded LLM baseline, and RQ3 by the distinguishable three-profile tutoring architecture observed in the pilot.

ANTLER 1 presents a style-aware co-creative tutoring *framework* instantiated as a multimodal prototype, realising a four-stage pattern: segmentation $\rightarrow$ feature grounding $\rightarrow$ style parameterisation $\rightarrow$ stance-conditioned generation. Claims are scoped to this early-stage instantiation and do not extend to large-scale validation (Boden 1992; Yannakakis, Liapis, and Alexopoulos 2018; Davis and others 2013). The concrete contribution to the computational creativity literature is threefold: (i) a design framework for style-parameterised co-creative tutoring in embodied craft; (ii) an operational definition of style-contingent interpretation as a CC design pattern (same observed evidence, multiple valid tutor stances); and (iii) an early-stage demonstration that sensor-grounded feedback can be simultaneously explainable and style-sensitive, extending the XAI-for-arts agenda (Llano et al. 2020; Bryan-Kinns and others 2023) into the embodied creative-practice domain.

In one observed sequence the system over-interpreted a hesitation pause as stylistic intent when the cause was thread tension management — illustrating the core challenge of disambiguating expressive choice from mechanical difficulty, and motivating future work on multi-hypothesis intent modelling.

Recent surveys and theoretical work in computational creativity and human–AI co-creativity position systems like ANTLER 1 within a broader shift toward explainable, process-focused creative AI (Llano et al. 2020; Bryan-Kinns and others 2023; Franceschelli and Musolesi 2024; Anonymous 2024). ANTLER 1 aligns with these principles by exposing style parameters, grounding feedback in observable features, and offering profile-conditioned what-if scenarios rather than a single optimised output.

The style taxonomy was developed with one practitioner and reflects a specific couture lineage; other traditions may require different profile structures. Profiles are curated rather than learned, and outputs depend on practitioner-specific rhythm data, limiting population-level validity.

Future work will expand studies across skill levels, learn profile parameters from larger annotated datasets, compare more formally against both baselines, and investigate transfer to other embodied crafts — soutache embroidery, kantha stitching, and ceramic surface decoration — where tacit knowledge is similarly at risk (Mora and others 2024; Jain and others 2025; Mills and others 2025).

References

Anonymous. 2024. Human-AI co-creativity: Exploring synergies across levels of interaction. arXiv preprint arXiv:2411.12527.

Boden, M. A. 1992. *The Creative Mind: Myths and Mechanisms*. London: Routledge.

Bryan-Kinns, N., et al. 2023. Explainable AI for the arts: XAIxArts. In *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*.

Chen, K.; Zhang, Y.; Weng, Y.; Wang, B.; and Liu, Z. 2025. Multimodal learning for automated robotic surgical skill assessment: Integrating video and kinematic data with ensemble model. *International Journal of Surgery*. Online ahead of print.

Davis, N., et al. 2013. An enactive model of creativity for computational co-creative systems. In *Proceedings of the International Conference on Computational Creativity*.

Franceschelli, G., and Musolesi, M. 2024. Creativity and machine learning: A survey. *ACM Computing Surveys* 56(11):283:1–283:41.

Jain, A., et al. 2025. An llm-guided tutoring system for social skills training. arXiv preprint arXiv:2501.09870.

Llano, M. T.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2020. Explainable computational creativity. In *Proceedings of the International Conference on Computational Creativity*, 334–341.

Mills, C., et al. 2025. Llm-generated feedback supports learning if learners choose to use it. arXiv preprint arXiv:2506.17006.

Mora, M., et al. 2024. An embodied socio-cognitive lens. In *Proceedings of the International Conference on Computational Creativity*.

Pan, Y.; Wang, J.; et al. 2023. An automated skill assessment framework based on visual motion descriptors for robotic surgery training. *Sensors*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–93.

Salter, C., et al. 2010. Embodied interactive performance and the digital arts. *Digital Creativity*.

Yannakakis, G. N., and Togelius, J. 2014. Experience-driven procedural content generation. In *Affective Computing and Intelligent Interaction*.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2018. Evaluating creativity in computational co-creative systems. In *Proceedings of the Ninth International Conference on Computational Creativity*.

Conceptualising Creative Autonomy in People, AI, and Their Interaction: A Scoping Review Protocol

Maija Välimäki and **Christian Guckelsberger**
Autotelic Interaction Research (AIR) Group
Department of Computer Science
Aalto University
Espoo, 01250 Finland
christian.guckelsberger@aalto.fi

Abstract

"Creative autonomy" is a prominent concept in historic and present discourses on creativity in humans, AI and their interaction. It is key to addressing some of the most pressing questions of our times and can serve as a boundary object to develop shared problem definitions and fostering cross-disciplinary research across AI, HCI, ethics and social science research, amongst others. Crucially though, it is presently unclear how robust creative autonomy – as used across contexts, disciplines and communities – is in its meaning, and how it relates to neighbouring concepts as well as autonomy more generally. This short paper presents the protocol for the first holistic and systematic scoping review of creative autonomy, seeking to fill this gap.

Introduction

Research on Artificial Intelligence (AI) and creativity has a long history (Boden, 2015); however, only with the advancement of powerful Generative AI (GenAI) systems in the last few years have people begun engaging with AI in (co-)creative practices at a large scale. This transformative development has catapulted some of the most eminent questions of (computational) creativity into the wider academic and public space: What shapes our perception of artificial systems as mere tools or genuine partners (e.g. Vimpari et al., 2023)? Is human creativity inherently different from artificial creativity, and if so, how (Runco, 2025)? How does integrating AI into creative workflows impact feelings of ownership, authorship and satisfaction (Dwivedi et al., 2023)? Fields including (but not limited to) AI and Human–Computer Interaction (HCI) research, social sciences and ethics have made big strides and continue to address these questions of theoretical and empirical contributions.

In the historic and present discourse, "creative autonomy" is a prominent concept. In Computational Creativity (CC) (Colton and Wiggins, 2012), creative autonomy has been advocated as a quality of artificial creative systems and goal of CC research, differentiating such systems from cognitive science models, creativity support tools, and the corresponding research agendas (e.g. Saunders, 2012). In HCI and media research, it has been identified as core human characteristics challenged when co-creating with GenAI (e.g. Bansal et al., 2024). In AI ethics research, creative autonomy has been defined as a facet of human autonomy that may be at stake when interacting with various, more general algorithmic systems (e.g. Catena, Tummolini, and Santucci, 2025).

These examples demonstrate not only that creative autonomy is attributed to both people and AI, but also highlight its role in characterising human-AI interaction. Moreover, they support that creative autonomy may serve as a *boundary object* (Star, 2010), characterising constructs that are utilised across various communities and may enable the coordination of their individual perspectives (Star, 2010).

The relationships between AI and (human) creativity are complex and wide-ranging. To understand and study them rigorously, with potency and on a solid ethical foundation, the corresponding research must draw on perspectives and contributions beyond its contextual, disciplinary and community confinements (Hemment and Kommers, 2025; Daniele and Song, 2019). To this end, boundary objects are essential, allowing for the development of shared problem definitions robust knowledge across multiple domains (Bennett et al., 2023). Crucially, boundary objects must balance a flexibility in meaning with sufficient robustness to support communication and negotiation within and between these communities (Bennett et al., 2023; Velt, Benford, and Reeves, 2020). At present, it is unclear how robust creative autonomy – as used across contexts, disciplines and communities – is in its meaning, and how it relates to neighbouring concepts as well as to autonomy more generally.

Our work seeks to answer this question. To the best of our knowledge, no research synthesis on the concept of creative autonomy exists. A systematic literature review by Bennett et al. (2023) focuses on autonomy and agency more generally, within people and HCI more specifically. Similarly, Catena (2026) put forward a systematic review of autonomy more generally, specifically within philosophical and ethical literature discussing the impact of AI on human autonomy. Both, Bennett et al. (2023) and Catena (2026) find that different studies on autonomy are underpinned by different conceptualisations of the concept, which supports misinterpretation and threatens adequate comparison and research progress. The closest related work comes from Catena, Tummolini, and Santucci (2025), who seek to distinguish creative from decisional autonomy, offering a small and unsystematic review only as a footnote.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

We seek to address this research gap with a more holistic and systematic scoping review of creative autonomy across the contexts of human, AI and their interaction. A type of systematic literature review if conceived more widely (e.g. Kraus et al., 2022), a scoping review is the method of choice to clarify concepts and recognise knowledge gaps across a wide range of literature (Munn et al., 2018). Our review implements the PRISMA-ScR (Tricco et al., 2018) checklist and explanation for scoping reviews, augmented with the PRISMA-S (Rethlefsen et al., 2021) checklist, explanation and elaboration for conducting literature searches. Furthermore, we utilised the JBI guidelines for scoping reviews for further orientation (Aromataris et al., 2024, p. 417–474). Our research question (RQ) is threefold:

RQ1 What explicit and implicit uses of "creative autonomy" in (a) humans, (b) AI and (c) their interactions are presented in existing research?

We study both, "explicit" and "implicit" uses of the concept. With the former, we denote meaning at a semantic, surface level; with the latter, we refer to latent meanings that may require interpretation and "reading between the lines" (Braun and Clarke, 2006). Further sub-questions (SQs) include:

SQ1 In which research domains is creative autonomy used?

SQ2 How are understandings of creative autonomy informed?

SQ3 How is creative autonomy related to close-by concepts such as agency, autonomy, creativity and co-creation?

This paper seeks to make this project known to the CC community, but its relevance arguably extends far beyond, in particular into HCI, social sciences, psychology, philosophy and ethics. We seek to probe creative autonomy's present status as a boundary concept through an overview of its conceptualisations and uses across contexts, disciplines and communities. In the short run, this can prevent jingle-jangle fallacies (Lawson and Robins, 2021) on the concept; in the long run, it can contribute to the maintenance or becoming of creative autonomy as a boundary concept through a more detailed understanding of its different conceptualisations, potentially conditional on context of use. This reduction in ambiguity and increase in granularity could inform further theoretical and empirical work on human, AI and human-AI co-creativity across many disciplines, including those mentioned above. This work also responds to the call for integrative research on autonomy more generally, articulating its nature as a multi-faceted, situated concept by providing insights on its constituents in the form of situated and contextual understandings and their relationships (Bennett et al., 2023; Catena, Tummolini, and Santucci, 2025).

In the following, we describe the scoping review eligibility criteria and search strategy. We outline plans for the data charting, analysis, presentation and contextualisation, concluding with the current status of the project and future work. This protocol has also been registered on osf.io.

Eligibility Criteria

Due to the relevance of creative autonomy to various fields and the long research heritage on (computational) creativity, we will not limit our search based on discipline or time of publication. However, as the main inclusion criteria, we require that the literature under review discusses creativity and autonomy together, in the same context. We accomplish this by means of our search string/fields and manual screening.

As sources of evidence, we will permit full-length, peer-reviewed scientific research papers and preprints, including primary studies, systematic reviews and meta-analyses as well as book chapters. We will exclude e.g. research commentaries and (extended) abstracts, book reviews as well as Bachelor or Master level theses. We do not limit publication years, but require the full text to be available and in English.

After completing the first version of the protocol on 2025-12-16, we moved forward with the selection of sources of evidence as described later. In that process, we recognised two instances in which "artistic autonomy" is used similarly to "creative autonomy". As an exploratory search in Scopus (2025-12-19) with "artistic autonomy" in title, abstract and keywords yielded 198 results and the overlap with our initial corpus is minor, it became evident that this issue requires our attention. "Artistic autonomy" appears related and potentially synonymous with our central construct of "creative autonomy" in the context of artistic practice, i.e. its use could invite a context-sensitive jangle fallacy (Lawson and Robins, 2021). Because clarifying such entanglements is one of our key objectives, we recognised a need to augment our corpus. The eligibility criteria we apply is identical to what is described above, but the wording we are looking for is "artistic autonomy". The detailed search strategy is described next.

The Literature Search

At the time of writing this protocol, the search strategy has been developed and final searches conducted. The following subsections document the involved steps; the next section describes the future screening of the retrieved corpus and all subsequent steps. The search was carried out and documented with respect to the PRISMA-S guidelines (Rethlefsen et al., 2021) to ensure both clarity and reproducibility. The filled out checklist can be found in the appendix.

Information sources and methods

Scopus was chosen as the main literature search platform as it enables searching simultaneously across several publishers, including most of the ACM Digital Library, and many publications on AI and HCI as arguably the most promising venues for relevant work. For additional coverage, the ACM Digital Library has also been searched separately.

Given the extent of academic publication cycles and fast pace of research on AI and creativity, we also took preprints into account. Scopus covers several preprints published 2017 onwards on several servers, including arXiv.org and SSRN as the main venues for preprints in technical fields and social sciences. To catch any relevant publications prior to 2017, we complemented the Scopus search with direct searches on arXiv and SSRN. In addition, a search in PsyArxiv (not listed in Scopus) was conducted to include preprints on (psychological) creativity research.

Crucially, research on AI and/or human creativity is now conducted in many other communities and venues beyond

what may be indexed by above services. Consequently, the above searches were further complemented with hits from the first 20 pages on Google Scholar.

Additionally, we met with Catena, Tummolini, and Santucci (2025) to inquire about any relevant literature we were unaware of. This did not yield any further inclusions.

Since our search through titles, keywords and abstracts (next section) may miss out on relevant publications, we last plan to conduct backward and forward snowballing on related work papers (Catena, Tummolini, and Santucci, 2025; Catena, 2026; Bennett et al., 2023) and on key papers surfaced from the reviewed corpus.

Search strategy

Our search strategy was designed iteratively, initially informed by the reading of known relevant work (e.g. Saunders, 2012; Jennings, 2010; Catena, Tummolini, and Santucci, 2025) and advanced via exploratory Scopus queries.

As described earlier, the concept of creative autonomy is present in contexts of human and AI creativity individually as well as their interaction. Conducting three separate searches was considered early on in the process, but soon dismissed as an inefficient approach: Exploratory test queries showed that terms like "human", "AI" or "co-creation" are not always explicitly mentioned, as they might be contextually obvious. Moreover, they , or might be expressed through different wording, accounting for which would have required developing several lists of synonyms while yet retaining a considerable risk of insufficient coverage. As we gathered insights into the extent of the overall corpus, it became clear that sorting through everything and coding it according to these three cases was feasible.

To support our inclusion criteria of the concept's central importance for the publication, the searches were limited to title, abstract and keywords where possible.

Due to an evident lack of a shared definition for *creative autonomy* but relevance in different communities with potentially varying terminology, it was of great importance to ensure that the search is broad enough. The goal was to capture work that discusses e.g. *creative autonomy* or *autonomous creativity* with potential inflections. The first iteration of the corresponding Scopus search string was:

> TITLE-ABS-KEY("creativ* autonom*" OR "autonom* creativ*")

The above-mentioned search was conducted in Scopus on 2025-11-02 to test the strategy and develop further exclusion criteria. The first 50 titles and abstracts were read by the first author and marked as *yes*, *no* or *maybe*. The authors discussed instances from each category to ensure alignment. This informed two additional exclusion criteria:

- Records are to be excluded if autonom* and creativ* are only listed separately. This applied to 19/50 samples, often signalled by comma separation in the abstract.
- Records are to be excluded if they only discuss *creative destruction*, a distinct concept from economics.

Based on these findings, the search string was fine-tuned and a list of potential synonyms for creative autonomy was developed. The search string's **second and final form** was:

> (TITLE-ABS-KEY (("creativ* autonom*" OR "autonom* creativ*"))) AND NOT
> (TITLE-ABS-KEY (("creativity, autonomy" OR "autonomy, creativity" OR "creative destruction") AND NOT ("autonomous creative" OR "autonomous creativity" OR "autonomously creative" OR "autonomy-driven creative" OR "autonomy-driven creativity" OR "autonomy-enabled creative" OR "autonomy-enabled creativity" OR "creative autonomy" OR "creativity-driven autonomy" OR "creativity-enabled autonomy" OR "creatively autonomous")))

The updated query now excludes many instances in which there is *only a list* or a *mention of creative destruction*, but no direct coupling of autonomy and creativity (or variants). The final search on Scopus was conducted on 2025-11-26.

We faced challenges in using wildcard expressions in other databases but Scopus. Therefore, adapted search queries for these databases make use of the list of potential synonyms of creative autonomy developed during the second iteration of the search string. The final searches in Google Scholar, the ACM Digital Library, arXiv, PsyArxiv and SSRN were conducted on the 2025-11-05. The full queries for each database are included in the appendix.

The strategy was discussed between the authors throughout the development process in weekly meetings, including reflections on results of test queries and database limitations. As the **final stage** of the search strategy, key papers will be backward and forward citation searched (see above).

To also probe "artistic autonomy" as a potentially closely related concept (further justification above), we searched for "artistic autonomy" in title, abstract and keywords on Scopus on 2025-12-19. In an effort to balance comprehensive inclusion of this potentially related concept and feasibility in processing, we exported only the 50 most recent publications for further eligibility screening. We believe that the most relevant older accounts will emerge via snowballing.

Managing records

All search results were exported to an Excel file. For Google Scholar, only full text sources from the first 20 pages were saved and results marked "citations" were dismissed. The number of records and dates of retrieval from each source are presented in Table 1, including the "artistic autonomy" addition on 2025-12-19. The initial corpus consisted of 442 records. Duplicates were deleted using Excel's power query tools. The final number of records was 378. The results of the "artistic autonomy" search on Scopus were checked for overlap, yielding an addition of 48 records. Thus, the title, abstract and keyword screening (described next) will be conducted on a corpus of **426** records.

Selection of Sources of Evidence

To mitigate bias, all authors will collaborate on the screening of titles and abstracts, separately applying the eligibility criteria to the first 20 % (n = 75) of the corpus. Decisions for inclusion will be compared and the level of agreement will be reported using Cohen's κ. Differing opinions will be

Source	No. of records	Date
Scopus	267	2025-11-26
Scopus preprints	12	
Google Scholar	148	2025-11-05
ACM Digital Library	6	
arXiv	0	
SSRN	9	
PsyArXiv	0	
Contacting Catena, Tummolini & Santucci	0	2025-10-07
Snowballing	TBD	upcoming
Initial no. of records, incl. duplicates	442	
Final no. of records, w/o duplicates	378	
Scopus on "artistic autonomy w/o overlap"	48	2025-10-19

Table 1: Number of records retrieved from each source.

discussed in live meetings until consensus is reached. This collaboration might result in further eligibility criteria.

The screening process will be conducted in the Excel file, where inclusion status as well as the possible rationale for exclusion will be documented to be shared later. We expect that some records will already at this stage require us to fetch the full text to make an informed decision of inclusion.

Based on the collaborative screening, the first author will apply the eligibility criteria to all **titles** and **abstracts**. Difficult cases will be collected and discussed together with the second author. For all records fulfilling the eligibility criteria based on title and abstract, the full text will be downloaded. The latter will be checked for eligibility by the first author and read in sufficient detail to extract all targeted data and mentions of our core concepts with context-sensitivity.

The full process will be documented, including a PRISMA flowchart with numbers of records at all stages. A copy of the full with reasons for exclusion will be included as Supplementary Material with the final publication.

Data Charting Process

"Data charting" denotes the process of extracting surface-level data for a scoping review (Tricco et al., 2018). We will chart the data by first coding the full-texts in the qualitative data analysis software ATLAS.ti, to be later moved into a tabular representation for further processing. We rely on ATLAS.ti to facilitate simultaneous and integrated data charting and in-depth interpretative analysis (next section).

Data charting will rely on a priori codes, including:

- basic bibliographic data: author(s), year, title
- domain: e.g. HCI, psychology, social science
- overall contribution type: empirical, theoretical, etc.
- subject: human / AI / interaction
- wording used: e.g. creative autonomy, autonomous creativity, creatively autonomous, autonomously creative
- type of creative autonomy definition: implicit/explicit
- (if applicable) how is the understanding of creative autonomy arrived at: e.g. by reference, original theoretical, original empirical inquiry

If further needs are recognised, additions will be made and/or codes will be further specified.

Data Analysis Process & Presentation

We additionally recognise a need for a more interpretive, in-depth analysis especially to identify implicit/latent understandings of creative autonomy. Additionally, to answer SQ3 (proximity to other concepts), simple frequency counts are not adequate but instead a closer reading of the material's relation to agency, autonomy and creativity is required. We thus plan to reach beyond our a priori data charting with a template analysis (Brooks et al., 2015), a specific type of thematic analysis in which a coding template is successively developed and revised during the coding process to build an understanding of qualitative data (Brooks et al., 2015). By supporting a mix of a priori and inductive codes, template analysis facilitates our integrated process of surface-level data charting and interpretative analysis.

Our template analysis will draw on a rich theoretical basis, differentiating forms of creativity and autonomy informed by classic and recent creativity research (e.g. Rhodes, 1961; Runco, 2025) and human-machine co-creativity (e.g. Kantosalo and Toivonen, 2016; Issak, Rezwana, and Harteveld, 2025) frameworks as well as classical (e.g. Colburn, 2022) and more relational (e.g. Mackenzie and Stoljar, 2000) conceptualisations of autonomy and their embedding in other theories, including motivational psychology (e.g. Vansteenkiste, Ryan, and Soenens, 2020). "Artistic autonomy" will be analysed to support contrast with "creative autonomy" and alternative phrasings.

To support our main RQ and SQs 1–2 (see Introduction), data frequencies will be reported; for more interpretative content, prevalence statements will be used. Key figures and distributions from the data charting will be presented. From the template analysis, a thematic map and the code hierarchy are likely outcomes. Further means of presentation may be chosen to capture the richness and heterogeneity of our data.

Conclusions & Future Work

We seek to conduct the first holistic and systematic scoping review of creative autonomy across the contexts of human, AI and their interaction. At the time of submitting this paper, the data charting and analyses are underway. We are looking forward to discussions with the ICCC community on the overall goals and methodology of this project and in particular appropriate frameworks to contextualise our findings.

Author Contributions

Contributions by author index via the CRediT Taxonomy (NISO, 2026): Conceptualisation (1,2), Methodology (1,2), Investigation (1,2), Data Curation (1), Visualisation (1,2), Writing (1,2), Funding Acquisition (2), Supervision (2).

Acknowledgments

We thank Niki Pennanen for feedback on the protocol draft, and our reviewers for excellent pointers moving forward.

References

Aromataris, E.; Lockwood, C.; Porritt, K.; Pilla, B.; and Jordan, Z., eds. 2024. *JBI Manual for Evidence Synthesis*.

Bansal, G.; Nawal, A.; Chamola, V.; and Herencsar, N. 2024. Revolutionizing Visuals: The Role of Generative AI in Modern Image Generation. *ACM Trans. Multimedia Comput. Commun. Appl.* 20(11).

Bennett, D.; Metatla, O.; Roudaut, A.; and Mekler, E. D. 2023. How does HCI Understand Human Agency and Autonomy? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–18.

Boden, M. A. 2015. Foreword: How computational creativity began. *Computational creativity research: towards creative machines* 7.

Braun, V., and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2):77–101.

Brooks, J.; McCluskey, S.; Turley, E.; and King, N. 2015. The Utility of Template Analysis in Qualitative Psychology Research. *Qualitative Research in Psychology* 12(2):202–222.

Catena, E.; Tummolini, L.; and Santucci, V. G. 2025. Human autonomy with AI in the loop. *Philosophical Psychology* 1–28.

Catena, E. 2026. Ai and human autonomy: a literature review. *AI and Ethics* 6(1):126.

Colburn, B., ed. 2022. *The Routledge Handbook of Autonomy*. London: Routledge.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Ecai*, volume 12, 21–26.

Daniele, A., and Song, Y.-Z. 2019. AI + Art = Human. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 155–161.

Dwivedi, Y. K.; Kshetri, N.; Hughes, L.; Slade, E. L.; (...); and Wright, R. 2023. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71:102642.

Hemment, D., and Kommers. 2025. Doing AI Differently: Rethinking the foundations of ai via the humanities.

Issak, A.; Rezwana, J.; and Harteveld, C. 2025. Mosaaic: Managing optimization towards shared autonomy, authority, and initiative in co-creation. *Proc. ICCC*.

Jennings, K. E. 2010. Developing Creativity: Artificial Barriers in Artificial Intelligence. *Minds and Machines* 20(4):489–501.

Kantosalo, A., and Toivonen, H. 2016. Modes for Creative Human-Computer Collaboration. *Proc. ICCC* 77–84.

Kraus, S.; Breier, M.; Lim, W. M.; Dabić, M.; Kumar, S.; Kanbach, D.; Mukherjee, D.; Corvello, V.; Piñeiro-Chousa, J.; Liguori, E.; et al. 2022. Literature reviews as independent studies: guidelines for academic practice. *Review of managerial science* 16(8):2577–2595.

Lawson, K. M., and Robins, R. W. 2021. Sibling Constructs: What Are They, Why Do They Matter, and How Should You Handle Them? *Personality and Social Psychology Review* 25(4):344–366. SAGE Publications Inc.

Mackenzie, C., and Stoljar, N. 2000. Introduction: Autonomy refigured. In Mackenzie, C., and Stoljar, N., eds., *Relational autonomy: feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.

Munn, Z.; Peters, M. D. J.; Stern, C.; Tufanaru, C.; McArthur, A.; and Aromataris, E. 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology* 18(1):143.

NISO. 2026. Contributor Roles Taxonomy (CRediT). http://credit.niso.org/.

Rethlefsen, M. L.; Kirtley, S.; Waffenschmidt, S.; (...); and PRISMA-S Group. 2021. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews* 10(1):39.

Rhodes, M. 1961. An Analysis of Creativity. *The Phi Delta Kappan* 42(7):305–310.

Runco, M. A. 2025. Updating the Standard Definition of Creativity to Account for the Artificial Creativity of AI. *Creativity Research Journal* 37(1):1–5.

Saunders, R. 2012. Towards Autonomous Creative Systems: A Computational Approach. *Cognitive Computation* 4(3):216–225.

Star, S. L. 2010. This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology, & Human Values* 35(5):601–617.

Tricco, A. C.; Lillie, E.; Zarin, W.; (...); and Straus, S. E. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine* 169(7):467–473.

Vansteenkiste, M.; Ryan, R. M.; and Soenens, B. 2020. Basic psychological need theory: Advancements, critical themes, and future directions. *Motivation and emotion* 44(1):1–31.

Velt, R.; Benford, S.; and Reeves, S. 2020. Translations and boundaries in the gap between hci theory and design practice. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27(4):1–28.

Vimpari, V.; Kultima, A.; Hämäläinen, P.; and Guckelsberger, C. 2023. "An Adapt-or-Die Type of Situation": Perception, Adoption, and Use of Text-to-Image-Generation AI by Game Industry Professionals. *Proc. of the ACM on Human-Computer Interaction* 7(CHI PLAY):131–164.

Appendix

Full search strategies for all databases

The ACM Digital library was searched on 2025-11-05 and limited to the title, abstract and keywords to have the same limitations as in Scopus. It yielded 7 results.

> [Abstract: ”autonomous creative”] OR [Abstract: ”autonomous creativity”] OR [Abstract: ”autonomously creative”] OR [Abstract: ”autonomy-driven creative”] OR [Abstract: ”autonomy-driven creativity”] OR [Abstract: ”autonomy-enabled creative”] OR [Abstract: ”autonomy-enabled creativity”] OR [Abstract: ”creative autonomy”] OR [Abstract: ”creativity-driven autonomy”] OR [Abstract: ”creativity-enabled autonomy”] OR [Abstract: ”creatively autonomous”] OR [Title: ”autonomous creative”] OR [Title: ”autonomous creativity”] OR [Title: ”autonomously creative”] OR [Title: ”autonomy-driven creative”] OR [Title: ”autonomy-driven creativity”] OR [Title: ”autonomy-enabled creative”] OR [Title: ”autonomy-enabled creativity”] OR [Title: ”creative autonomy”] OR [Title: ”creativity-driven autonomy”] OR [Title: ”creativity-enabled autonomy”] OR [Title: ”creatively autonomous”] OR [Keywords: ”autonomous creative”] OR [Keywords: ”autonomous creativity”] OR [Keywords: ”autonomously creative”] OR [Keywords: ”autonomy-driven creative”] OR [Keywords: ”autonomy-driven creativity”] OR [Keywords: ”autonomy-enabled creative”] OR [Keywords: ”autonomy-enabled creativity”] OR [Keywords: ”creative autonomy”] OR [Keywords: ”creativity-driven autonomy”] OR [Keywords: ”creativity-enabled autonomy”] OR [Keywords: ”creatively autonomous”]

ArXiv was searched on 2025-11-05 with the advanced search function, searching the list of fit couplings on abstract and title. It was limited to the end of 2016, as newer hits are included in the Scopus pre-prints search. The final query (below) had no search results.

> order: -announced_date_first; size: 50; date_range: to 2016-12-31;
> include_cross_list: True; terms: AND title=”autonomous creative” OR ”autonomous creativity” OR ”autonomously creative” OR ”autonomy-driven creative” OR ”autonomy-driven creativity” OR ”autonomy-enabled creative” OR ”autonomy-enabled creativity” OR ”creative autonomy” OR ”creativity-driven autonomy” OR ”creativity-enabled autonomy” OR ”creatively autonomous”; OR abstract=”autonomous creative” OR ”autonomous creativity” OR ”autonomously creative” OR ”autonomy-driven creative” OR ”autonomy-driven creativity” OR ”autonomy-enabled creative” OR ”autonomy-enabled creativity” OR ”creative autonomy” OR ”creativity-driven autonomy” OR ”creativity-enabled autonomy” OR ”creatively autonomous”

SSRN was searched on 2025-11-05 using the advanced and boolean search options. The list of fit couplings was searched from title, abstract and keywords, mirroring the search strategy in Scopus. It was not possible to limit the date. Searching with the query below yielded 9 results.

> ”autonomous creative” OR ”autonomous creativity” OR ”autonomously creative” OR ”autonomy-driven creative” OR ”autonomy-driven creativity” OR ”autonomy-enabled creative” OR ”autonomy-enabled creativity” OR ”creative autonomy” OR ”creativity-driven autonomy” OR ”creativity-enabled autonomy” OR ”creatively autonomous”

PsyArxiv was searched via OSF on 2025-11-05. The platform offers only a simple search, in which the following search string was pasted. There were no hits.

> ”autonomous creative” OR ”autonomous creativity” OR ”autonomously creative” OR ”autonomy-driven creative” OR ”autonomy-driven creativity” OR ”autonomy-enabled creative” OR ”autonomy-enabled creativity” OR ”creative autonomy” OR ”creativity-driven autonomy” OR ”creativity-enabled autonomy” OR ”creatively autonomous”

Google Scholar was searched on 2025-11-05. The following search string was pasted into the search field and no additional criteria were used.

> ”autonomous creative” OR ”autonomous creativity” OR ”autonomously creative” OR ”autonomy-driven creative” OR ”autonomy-driven creativity” OR ”autonomy-enabled creative” OR ”autonomy-enabled creativity” OR ”creative autonomy” OR ”creativity-driven autonomy” OR ”creativity-enabled autonomy” OR ”creatively autonomous”

The first 20 pages of results were reviewed and all full text sources were exported. Results marked “citations” were dismissed and not exported.

Scopus for “artistic autonomy” was searched on 2025-12-19 with the following search string. The search was limited to title, abstract and keywords only.

> TITLE-ABS-KEY (”artistic autonomy”)

The search yielded 198 results, out of which the most recent 50 were exported. The results were compared to the existing corpus for overlap, resulting in an addition of 48 records.

PRISMA-S Checklist

The filled out PRISMA-S checklist can be found on the next, final page of the appendix.

Section/topic	#	Checklist item	Location(s) Reported
INFORMATION SOURCES AND METHODS			
Database name	1	Name each individual database searched, stating the platform for each.	pp. 2-3
Multi-database searching	2	If databases were searched simultaneously on a single platform, state the name of the platform, listing all of the databases searched.	-
Study registries	3	List any study registries searched.	-
Online resources and browsing	4	Describe any online or print source purposefully searched or browsed (e.g., tables of contents, print conference proceedings, web sites), and how this was done.	-
Citation searching	5	Indicate whether cited references or citing references were examined, and describe any methods used for locating cited/citing references (e.g., browsing reference lists, using a citation index, setting up email alerts for references citing included studies).	p. 3
Contacts	6	Indicate whether additional studies or data were sought by contacting authors, experts, manufacturers, or others.	p. 3
Other methods	7	Describe any additional information sources or search methods used.	-
SEARCH STRATEGIES			
Full search strategies	8	Include the search strategies for each database and information source, copied and pasted exactly as run.	p. 3 p. 6
Limits and restrictions	9	Specify that no limits were used, or describe any limits or restrictions applied to a search (e.g., date or time period, language, study design) and provide justification for their use.	p. 2 p. 6
Search filters	10	Indicate whether published search filters were used (as originally designed or modified), and if so, cite the filter(s) used.	-
Prior work	11	Indicate when search strategies from other literature reviews were adapted or reused for a substantive part or all of the search, citing the previous review(s).	-
Updates	12	Report the methods used to update the search(es) (e.g., rerunning searches, email alerts).	-
Dates of searches	13	For each search strategy, provide the date when the last search occurred.	p. 4 (Table 1)
PEER REVIEW			
Peer review	14	Describe any search peer review process.	pp. 3-4
MANAGING RECORDS			
Total Records	15	Document the total number of records identified from each database and other information sources.	p. 4 (Table 1)
Deduplication	16	Describe the processes and any software used to deduplicate records from multiple database searches and other information sources.	p. 3

PRISMA-S: An Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews
Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, Koffel JB, PRISMA-S Group.
Last updated February 27, 2020.

Sketch-It: A Consent-First Creativity Support Tool for the Generation of Inspirational Sketches

Clàudia Pàmies
Linnaeus University, Sweden &
Universitat Autònoma de Barcelona, Spain
claudia.pamies@autonoma.cat

Garrit Schaap
Computer Science and Media Technology
Linnaeus University, Sweden
garrit.schaap@lnu.se

Maria M. Hedblom
Jönköping AI lab
Jönköping University, Sweden
maria.hedblom@ju.se

Abstract

Creativity support tools (CST) are designed to offer inspiration and help artists overcome creative block. For this, Generative AI tools offer quick and accessible methods to generate images, but are not adjusted to individual style, and they raise ethical concerns due to the copyright of the training data. Addressing these issues, we introduce "Sketch-It", a consent-first CST based on two generative adversarial networks (DCGANs) that is exclusively trained on the user's own artwork. To provide patterned structure, the first model is trained on greyscale sketches, while the second model is trained on photographs to acquire colour knowledge. The two models are bridged by a U-Net colouriser that applies the learned colour distributions to the structural forms. Intended for personal use, the system is packaged as a browser-based local application, requiring no coding knowledge for artists to train the system on their own artwork and use it in their creative process. We demonstrate that the system can generate abstract inspirational images by training the system on one artist's flower sketches and photographs. Our system offers three distinct contributions: technical – a dual-model form/colour separation architecture, ethical – a consent-first training paradigm on minimal datasets, and practical – an accessible tool designed for artists.

Introduction

Creative block is defined as the temporary inability to generate new ideas or initiate new work. It is a well-documented phenomenon among practising artists [3] that is independent of skill and defined as a momentary disconnection from one's own creative vocabulary and creative inspiration. In light of the growing involvement of technology in creative processes, creativity support tools (CSTs) aim to assist in rather than replace human creative capacity, by offering, for example, prompts, references or generative suggestions that reset the creative process back in motion [23].

Contemporary AI image generation systems (e.g., Stable Diffusion [15], Midjourney [12], DALL-E [14]) are superficially well placed to serve this need as they can rapidly generate images in a wide variety of styles based on textual descriptions. However, trained on other data, the generated images might substantially diverge from the artist's own style, leading to distractions in the creative process or a lack of genuine inspiration. Additionally, their adoption by the creative community has been met with significant resistance for principled reasons [10], since these systems are trained on vast datasets assembled via web scraping and, thus, contain the collected result of millions of copyrighted artworks that have been used without the knowledge or consent of their creators [21].

Our approach takes a different route by asking *"What if the AI system learned only from you?"* This addresses the style problem by ensuring the AI follows your uniquely expressed style while also removing any ethical issues of 'artistic theft.' However, it also introduces a new challenge: how to generate inspirational outputs of good enough quality with limited datasets.

To this end, we introduce "Sketch-It," a system that trains and combines the results of two small generative models, both of which the user has complete control over the training data. Thus, ensuring that exclusively their own work gets to influence the results. The system is presented as a browser-based interface, which requires no programming knowledge. The result is a creativity-support tool that is simultaneously ethical (fully consented data), computationally green [2] – due to the small datasets and local CPU training, and personally meaningful – the generated output is an extension of the artist's work. To demonstrate Sketch-It, we will present a prototype example trained on a single artist's sketches and photographs of flowers.

While creative block is the primary motivating scenario, CSTs more broadly support ideation, stylistic exploration and reflective practice [5], and Sketch-It's architecture is applicable across this wider range of creative support contexts.

Related Work and Positioning

Generative Adversarial Networks for Image Synthesis

Generative Adversarial Networks (GANs), introduced by Goodfellow [6] in 2014, have become one of the most influential methods for image generation. By framing generation as a game between a generator and a discriminator, GANs enable the synthesis of highly realistic images from random noise. Deep Convolutional GANs (DCGANs) further improved this paradigm by incorporating convolutional architectures, making it possible to generate photo-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

realistic images with greater spatial coherence and training stability [16]. However, training GANs on small datasets remains challenging, as adversarial learning is often unstable in low-data regimes. Techniques such as spectral normalisation [13], instance noise [11], and label smoothing [19] have therefore been proposed to improve robustness and convergence when data is limited.

Sketch Generation and Colourisation

Automatic colourisation of line drawings has been explored through approaches such as conditional GANs, e.g. pix2pix [7], reference-based methods [24] and user-guided systems that allow partial human control over the output [25]. These methods typically require paired training data, such as a sketch alongside its corresponding coloured version, which is often expensive and labour-intensive to produce. This dependency limits their applicability when only small personal datasets are available. Our approach differs as it operates on unpaired data and uses a perceptual colour critic rather than pixel-level supervision. This makes the method more practical for artist-specific workflows, where the dataset may be small but stylistically consistent.

Protective Measurements Against Artistic Theft

The ethical debate around AI-generated art has been discussed extensively in recent years [4]. Key concerns include attribution, consent and economic impact. Proposals to address this issue include opt-in licensing systems and 'poisoning' tools such as Glaze [21], which adds imperceptible perturbations to artworks to degrade style extraction, and Nightshade [22], which poisons training data. This is effective as a defensive tool against unauthorised use of the images. However, such methods also block the artist from using their own images. By contrast, our work takes a different approach by protecting artists against unauthorised use of their art; it demonstrates that effective generative AI can be built on a fully consented, artist-controlled dataset and later used collaboratively with the artist to produce new art pieces that continue to contribute to the evolution of the artist's cultural background [20].

System Architecture

Sketch-It comprises two GAN models that are trained in parallel and a 'colouriser' that creatively merges the results from both models (see Figure 1. It is presented with a browser-based user interface that makes the full pipeline accessible to non-technical users.

Model 1: Personal Sketch Generator

Model 1 is a Deep Convolutional GAN (DCGAN) intended to be trained exclusively on the artist's own hand-drawn sketches. The generator maps a 128-dimensional vector to a 64×64 greyscale image through four upsampling transposed-convolution blocks. The discriminator uses spectral normalisation to improve training stability. To increase the effective size and diversity of the dataset, we

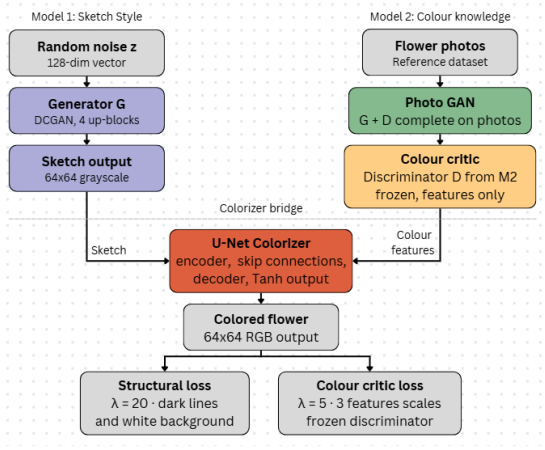


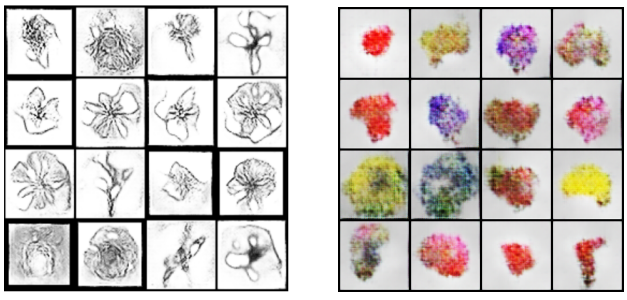
Figure 1: Representation of the system's architecture.

apply random flips, rotations up to $\pm 20°$, scale jitter and brightness variation.

For our flower example, the model was trained for 2000 epochs, see example output in Figure 2a. At convergence, the generator samples novel sketches that share stylistic characteristics with the training artist's work. In our example, this corresponds to petal curvature, line weight, compositional tendencies and the textural quality of pencil on paper. This means that the model is generative rather than retrieval-based; it produces configurations that are not present in the training set.

Model 2: Colour Knowledge from Nature

Model 2 is a separate DCGAN trained on photographs. Its purpose is not to influence the shape or to transform the output into something more photorealistic, but to encode colour knowledge in the discriminator's feature representations. After training, the generator is discarded, and the discriminator is frozen and repurposed as a colour critic. Trained on 2000 epochs in our example, its intermediate feature maps, extracted at multiple spatial scales, encode learned representations of what real flower colours look like at different levels of detail, see example outputs in Figure 2b.



(a) Output from Model 1 after 2000 epochs

(b) Output from Model 2 after 2000 epochs

Figure 2: Example of model outputs.

The Colouriser: Bridging Form and Colour

The colouriser is a U-Net [18] that maps a greyscale sketch $[1 \times 64 \times 64]$ to an RGB coloured image $[3 \times 64 \times 64]$. Training is governed by three losses. The line computes a soft mask over pixels darker than a given threshold and enforces that the greyscale projection of the coloured output matches the original sketch in those regions, ensuring that dark lines remain visible. The background loss computes a soft mask over bright pixels and constrains the output to remain light where the sketch is white. The colour critic loss computes L_1 distance between the colouriser's output features and real photograph features extracted by Model 2's frozen discriminator. Mid-grey petal interior pixels are left unconstrained structurally, allowing the colour critic to fill them freely with naturalistic flower colours. This three-zone approach (lines, interior, background) is the key design decision that enables the system to colour within the sketch rather than replacing it with a photograph.

User Interface

The three models are wrapped in a browser-based local web application built with a Flask back-end (Python) and a single-page front-end developed with HTML and JavaScript[1]. The interface is designed for artists with no prior programming knowledge, presenting the full pipeline as a four-step workflow:

Step 1: First, a legal agreement is presented on screen, in which the user confirms that all uploaded artwork is their own original work and that they hold the intellectual property rights to it. This is a deliberate design choice to ensure data consent and user accountability.

Step 2: Second, the user is presented with a drag-and-drop zone to upload sketched reference images and colour mapping photographs. This is presented together with live file counts and minimum thresholds to prevent underpowered training runs.

Step 3: Third, a "Train Your Model" button is used to trigger all three training stages sequentially shown with a unified progress bar.

Step 4: After training is complete, the system offers the user the ability to use the system to generate new inspirational images.

The application runs entirely locally on the artist's own computer, meaning that no data is shared with a third party. This ensures that the artist retains full ownership and control of their training data, their trained models and all the generated outputs.

Prototype Experiment and Results

The prototype was trained on 104 hand-drawn pencil flower sketches and 34 cropped flower photographs, all produced by a single artist.

Training was performed on a common consumer CPU (13th Gen Intel(R) Core(TM) i5-1335U [1]). Model 1 required approximately 3.5 hours of training (2000 epochs), Model 2 approximately 3-4 hours (2000 epochs), and the colouriser took approximately 1 hour (500 epochs).

[1]Link: `https://github.com/clauudus/SketchIt`

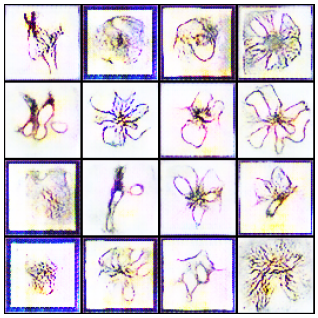


Figure 3: Example of coloured drawing outputs

The generated sketches at epoch 2000 show recognisable flower structures such as daisy-like forms, rose-like clusters or tulip silhouettes, with visible stylistic consistency across seeds, see examples in Figure 2a. Similarly, Model 2 provides flower-like colour patterns from the reference photographs, see Figure 2b. The characteristic line weight and petal curvature of the training artist are preserved across a diverse range of generated forms. The colouriser successfully preserves sketch line work while applying flower-appropriate colour palettes, see Figure 3.

The full generation pipeline (sketch generation followed by colourisation) runs in under one second on CPU, enabling real-time interactive use through the web interface.

Conceptual Evaluation Study

To assess the reception of Sketch-It, we conducted a short anonymous survey with 35 participants. The study assessed the perceived usefulness and its ethical framing from the intended target audience rather than measuring creative outcomes or task performance. Details of the survey and the results can be found in the Appendix.

The results show that the majority of respondents (68.6%) experience creative block *sometimes or more* (mean 2.94/5), supporting the usefulness of CSTs such as Sketch-It. Simultaneously, attitudes toward existing AI tools were predominantly negative or cautious; 60.0% held a *negative or concerned view* with explicit mentions of how AI image generation had already negatively affected their practice or livelihood. Here, creator consent when training AI was rated highly important, receiving a mean of 4.11/5, with 74.3% scoring 4 or 5. For tools, like Sketch-It, which train exclusively on the user's own artwork, 62.9% of respondents reported a positive perspective, see Figure 11 (in the Appendix). The likelihood that the participants would try the tool had a mean of 3.03/5, with 71.4% scoring 3 or above. The most frequently recurring concern was over-reliance of the tools. Respondents worried that using an AI tool during a creative block might inhibit the development of internal creative resources. A second recurring concern was residual security: even with local storage, one respondent noted their images could still be compromised. Both concerns directly inform the future work directions described below.

Discussion

The Ethical and Environmental Contribution

One of the most important contributions of this work is the demonstration that consent-first generative AI can be practically realised. The system produces useful creative outputs (ie., novel flower sketches in a personal style, plausibly coloured) from a dataset of around 100 images created by and belonging to a single individual.

As a consequence of the small dataset, our approach also represents a "green, cost-efficient generative AI system. Large foundation models for image generation require training runs of thousands of GPU-hours on datasets of billions of images [17]. Our solution's total training energy is several orders of magnitude smaller, achievable on a consumer laptop overnight. For a field increasingly scrutinised for its environmental footprint, small-dataset, consent-first systems represent a meaningful alternative paradigm.

Existing responses to AI ethics problem in art take a predominantly defensive stance, tools such as Glaze [21] and Nightshade [22] protect artists by making their work resistant to style extraction, while opt-in licensing frameworks seek to regulate access[9]. Sketch-It takes a complementary but distinct affirmative approach; rather than protecting against misuse of existing systems, it demonstrates that genAI systems can be built from the ground up on fully consented data.

Creativity Support, Not Creative Replacement

Designed as a creativity support tool, our system is not a replacement for artistic production. In comparison to more commercial AI image generators, the generated output does not represent complete or polished images. Instead, the generated sketches are intended as creative starting points that are to be interpreted as inspirational visual prompts that reflect the artist's own vocabulary in new configurations. The artist does not press a button and receive finished artwork; rather, they receive a suggestion from which they can diverge, react or build. This positions the system within what Kantosalo [8] calls co-creative AI: a collaborative partner that augments rather than supplants human creative agency.

The cultural implications are significant. Trained on the artist's own work, such tools do not homogenise culture towards the statistical average of a billion training images; instead, it amplifies the individual artist's particular perspective. In this sense, our solution is not only a tool for overcoming creative block, but also an argument about what AI-assisted culture can look like when the human at the centre of the system retains authorship of its foundations.

Generalisability: Beyond Flowers, Beyond Sketches

Flowers were chosen as the demonstration domain because they offer sufficient visual variety to test the system while remaining tractable for small-dataset training. The architecture is designed to be domain-agnostic – an architect could train Model 1 on their sketched floor plans and Model 2 on photographs of built spaces, or a character designer could train on their character sheets with colour reference from comics or frames from animation. However, these remain theoretical extensions. A full multi-domain evaluation is identified as an important direction for future work.

Our experiments suggest a minimum of 50 images for the drawings and 20 photographs for the colourisation. This threshold depends largely on the artist's needs and the level of abstraction they expect; the smaller the dataset, the more abstract the generated output is likely to be.

Limitations

The current prototype was trained on a single artist's flower sketches and photographs. Claims about generalizability to other artistic domains, styles, or levels of stylistic consistency remain theoretical and untested. Additionally, the evaluation measured attitudes and perceived reception rather than creative effectiveness, meaning that it cannot be concluded that Sketch-It demonstrably reduces creative block or improves creative output. These limitations should be addressed in future work on multi-artist, multi-domain experiments and task-based evaluations.

Conclusion

We have presented a creativity support system that generates novel artworks in an artist's personal style using only their own consented drawings. By separating the learning of artistic form (a personal sketch GAN) from the learning of colour (a nature-photograph GAN), and connecting them through a learned colouriser guided by a three-zone structural loss, we demonstrate that ethically grounded generative AI can produce meaningful creative output from minimal data. A preliminary survey of 35 participants confirms a strong interest in consent-first approaches: 74.3% rated dataset consent as highly important, and 62.9% responded positively to Sketch-It's framing relative to existing tools.

We hope that this solution contributes to a shift in how the field thinks about AI-assisted creativity, away from systems that extract value from artists without consent, and towards systems that amplify individual creative voices while keeping the human firmly in the centre. The best creative AI, we argue, is not one that replaces artistic thinking; it is one that says: *"Here is something new, in your own voice. What do you want to do with it?"*

Future Work

The legal consent model, while carefully designed, is self-reported. Future work could explore integration with digital provenance tools or blockchain-based artwork certification. The interface could also be extended with real-time training visualisation, showing the model improving epoch by epoch. Regarding security beyond consent, while the local architecture ensures training data never leaves the artist's machine under normal use, future versions could integrate adversarial perturbation techniques [21] [22] applied automatically before local storage, offering a dual guarantee; the data is not shared by design, and even if compromised, would be resistant to exploitation.

Authors contribution

As lead author, Pàmies designed the system and performed the survey. Schaap and Hedblom acted as supervisors who aided in project guidance and co-authorship.

References

[1] Intel® Core™ i5-1335U Processor (12M Cache, up to 4.60 GHz) - Product Specifications — Intel — intel.com. [Accessed 19-04-2026].

[2] Ardalan Arabzadeh, Tobias Vente, and Joeran Beel. Green recommender systems: Optimizing dataset size for energy-efficient algorithm performance. October 2024.

[3] Mihaly Csikszentmihalyi. *Creativity: Flow and the Psychology of Discovery and Invention*. HarperCollins, New York, 1996.

[4] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. June 2017.

[5] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2019. ACM.

[6] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 2014.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. November 2016.

[8] Anna Kantosalo and Tapio Takala. Five c's for human–computer co-creativity: An update on classical creativity perspectives. In Amílcar Cardoso, Penousal Machado, Tony Veale, and {João Miguel} Cunha, editors, *Proceedings of the Eleventh International Conference on Computational Creativity*, pages 17–24, Portugal, September 2020. Association for Computational Creativity. International Conference on Computational Creativity, ICCC ; Conference date: 07-09-2020 Through 11-09-2020.

[9] Yasuto Komada. New law responding to imitations of designs that intersect real and virtual space: Amendment of the unfair competition prevention act in japan. *GRUR International*, 74(2):153–159, January 2025.

[10] Lin Kyi, Amruta Mahuli, M Six Silberman, Reuben Binns, Jun Zhao, and Asia J Biega. Governance of generative AI in creative work: Consent, credit, compensation, and beyond. January 2025.

[11] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? January 2018.

[12] Midjourney, Inc. Midjourney, 2024. Version 6.0.

[13] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. February 2018.

[14] OpenAI. DALL-E 3, 2023. Image generated with prompt: "prompt".

[15] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. July 2023.

[16] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. November 2015.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. December 2021.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. May 2015.

[19] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. June 2016.

[20] Kelsey Schoenberg. How Generative AI Endangers Cultural Narratives - issues.org. https://issues.org/generative-ai-cultural-narratives-rettberg/. [Accessed 19-04-2026].

[21] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, Anaheim, CA, August 2023. USENIX Association.

[22] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. October 2023.

[23] Ben Shneiderman. Creativity support tools: accelerating discovery and innovation. *Commun. ACM*, 50(12):20–32, December 2007.

[24] Lvmin Zhang, Yi Ji, and Xin Lin. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier GAN. June 2017.

[25] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. May 2017.

Appendix

Additional Data from the Study

Experimental Setup and Participants

The study was conducted as an anonymous online questionnaire. Participants were invited through convenience sampling through social network distribution.

The majority of respondents identified as hobbyist creators (60.0%), with smaller numbers describing themselves as non-practitioners interested in art (20.0%), semi-professional artists with occasional commissions (17.1%), students in artistic training (5.7%), and two professional artists whose art constitutes their primary income.

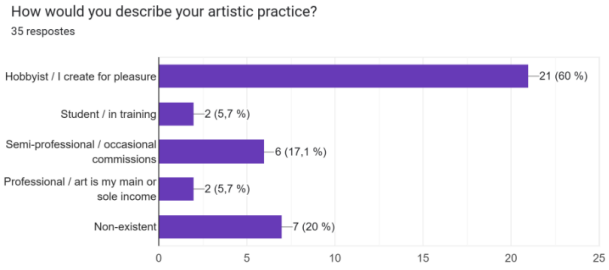


Figure 4: How did participants identify in an artistic spectrum

Respondents worked mostly with pencil and ink drawing, digital illustration, painting, and music. Two respondents identified dance and writing as their primary practice, reflecting a broader creative community beyond visual art.

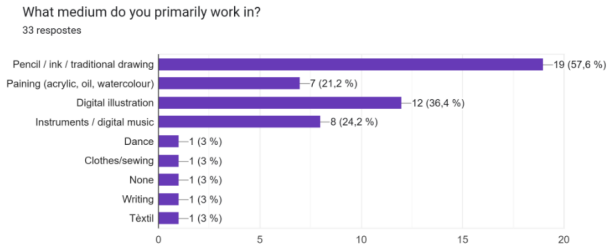


Figure 5: What kind of materials do artists tend to use?

Questionnaire Questions

The study asked 10 questions related to the use of creativity support tools like Sketch-It.

Likelihood of Creative Block and Solutions The participants rated the frequency of a creative block on a five-point scale from 1 (never) to 5 (very often). The mean response was 2.94, and 24 of 35 respondents (68.6%) scored 3 or higher, confirming that creative block is at least a sometimes occurrence for most of the sample. Only three respondents (8.6%) reported never experiencing it.

The most common coping strategies were looking at others' work for inspiration and searching for reference images online. Only one respondent reported using an AI image tool

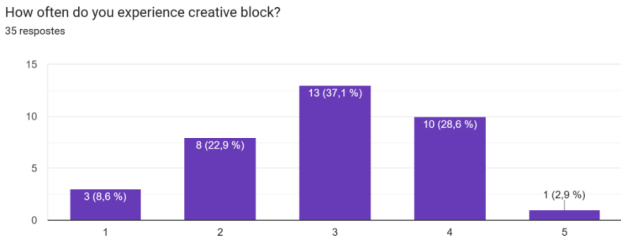


Figure 6: Participants and creative block

as a strategy during creative block, suggesting that the space for ethical AI creativity support remains largely unoccupied.

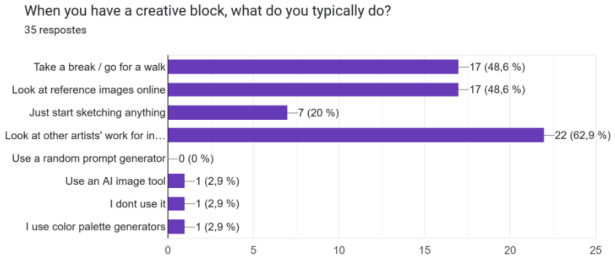


Figure 7: How artists normally overcome an Art Block

Attitudes Toward AI Image Generation Attitudes towards existing AI tools were predominantly negative or cautious: 37.1% described themselves as opposed to AI-generated imagery on principle, 22.9% expressed concern about exploitation of artists' work without consent, 34.3% were neutral, and only 5.7% responded positively. In total, 60.0% had a negative or concerned view.

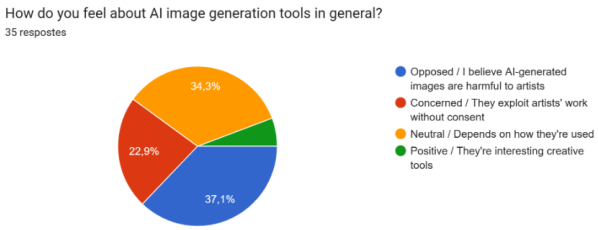


Figure 8: How do Participants feel about AI?

Eleven respondents (31.4%) reported that AI generation had already negatively affected their creative practice, compared to two participants who reported a positive effect.

The Role of Consent When asked to rate how important creator consent is for the use in the training dataset, on a five-point scale, the mean response was 4.11 out of 5. Twenty-one respondents (60.0%) assigned a maximum score of 5, and 26 of 35 (74.3%) rated the importance of consent at 4 or greater.

When told that Sketch-It trains exclusively on the user's own artwork and asked whether this changes their feelings

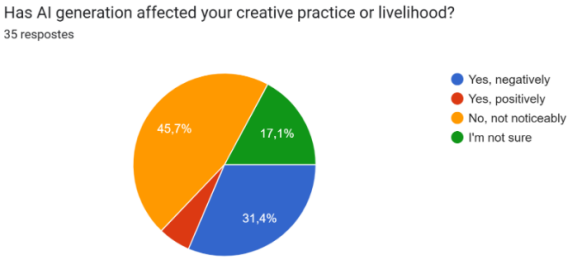


Figure 9: Perceived effect of AI Generation tools on the participants' lives

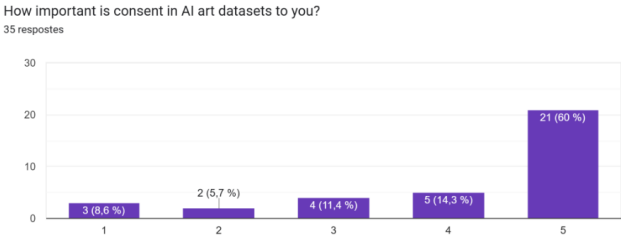


Figure 10: Graph indicating the concern towards consent in AI training datasets

compared to other tools, 62.9% responded positively, as seen in Figure 11 (in the Appendix). 14.3% stated that consent makes a significant difference, and 48.6% said that it makes some difference, although they retain reservations. Nine respondents (25.7%) said that the consent-first architecture does not change their concerns, and four (11.4%) reported having no prior concerns.

Perspective on fully consented dataset To determine if Sketch-It could be a well received tool among artists, we presented an explanation on how it would work to then know if the perspective on generative AI would change. 25.7% stated that they still had the same concerns, while 62.9% had a positive shift on the matter.

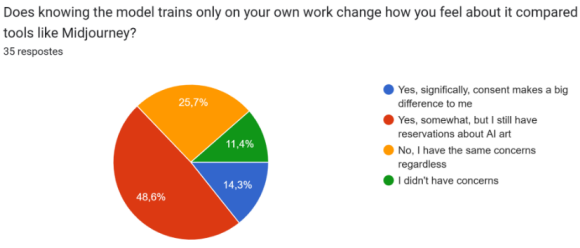


Figure 11: Results of perspective on user-art only trained AI systems.

Receptiveness to Sketch-It Participants rated their likelihood of trying Sketch-It on a five-point scale (mean 3.03). Eleven respondents (31.4%) gave a score of 4 or 5, while then (28.6%) gave 1 or 2, indicating reluctance. With 71.4% scoring 3 or above, the majority of the sample is at minimum open to trying the tool; a positive baseline given that 60.0% hold negative or concerned views on AI image generation in general.

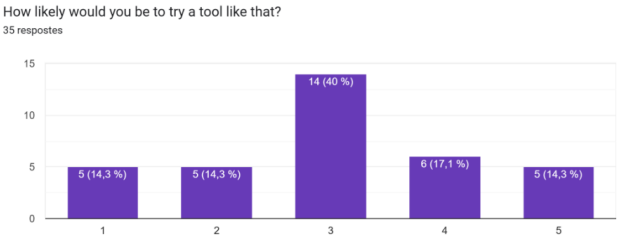


Figure 12: Participants trying Sketch-It

The two most frequently selected appealing attributes were the tool's small environmental footprint (18 respondents, 51.4%) and its consent-first training using only the artist's own drawings (17, 48.6%). Local data storage was selected by 14 respondents (40.0%), personal style generation and creative block assistance by eight each (22.9%). The prominence of ethical and environmental attributes over functional ones suggests that for this audience, the conditions of trust matter as much as utility.

Concerns and Open Responses The most frequently recurring concern was over-reliance. Several respondents worried that using an AI tool to escape creative block might inhibit the development of internal creative resources. One participant wrote that a great way to overcome art block is by purely drawing, while another participant noted that "it can be a thin line between using it as a tool or becoming dependent on it". A second recurring concern was residual data security. One respondent noted that their images, stored on their own computer, "could still be compromised". This directly motivated the adversarial perturbation direction described in Future Works. Several open responses pointed toward productive extensions, where two of the participants explained that having such a tool to help artists overcome art block in a very interactive sense would motivate them to try it out.

Summary

The study confirms three of Sketch-It's core: creative block is real and common (68.6% experience it sometimes or more), consent is a central concern in attitudes towards AI art tools (mean 4.11/5, 74.3% rating it 4 or above), and the consent-first approach meaningfully improves reception among artists otherwise sceptical of AI (62.9% positive

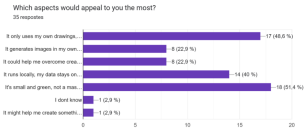


Figure 13: Picked attributes among the participants

shift). Primary concerns, including over-reliance, residual data security, and questions of artistic authenticity, are genuine and inform the future work directions described in the main paper. We note that the sample is small (n=35) and that a longitudinal study with a larger, more diverse participant group would be needed to draw stronger conclusions.

Side-Splitter: A Visual Laboratory for Jokes with Verbal Humour and Physical Slapstick

Tony Veale
School of Computer Science,
University College Dublin,
Dublin D4, Ireland.
tony.veale@UCD.ie

Abstract

Humour encourages us to approach even the most serious subjects from a playful and non-serious perspective. It can soften our opposition to certain ideas and strengthen our rejection of others. It can bring us closer together, by dissolving the barriers that divide us, or it can use the same ridicule to drive us further apart. But a truly neutral comedian, or an automated joke generator, has no *in*-group loyalties to defend, nor any *out*-group opponents to resent. Comedy that lacks bite is asinine and bland, but a neutral comic can be funny and divisive by being equally cutting to all sides in a debate. This paper presents a demonstration of a system, *Side-Splitter*, that embodies this philosophy. The system uses the perspective-taking ability of LLMs to look at a given topic from each side, to see it as both a believer and a critic might, to generate divisive humour against either. Side-Splitter is a hybrid system that pairs a symbolic grammar with an LLM to produce a comic strip that is part visual slapstick and part stand-up routine. We describe how topical jokes are elicited from the LLM and selectively filtered according to how divisive they are, and show how these jokes are packaged in comic-strip form to lend them an extra dimension of visual humour.

Divided We Stand Up

It is often said that there are two sides to every story. Diversity of viewpoint is generally a good thing, but only if each side is willing to give the other an opportunity to speak. As impassioned believers become volatile partisans, digging in to further entrench their beliefs, discourse takes a back seat to difference, while echo chambers emerge in which each side only listens to the views of those that agree with them (Barberá et al. 2015). Rather than engage directly with those with whom they disagree, siloed partisans erect and knock down straw men that cannot fight back. Humour plays a part in this process, allowing each side to mock the other from the safety of their own assumptions and prejudices.

As noted in (Ford et al. 2017), humour can be used both as a social *lubricant* and as a social *abrasive*, to open groups to new members or to heighten inter-group tensions. In this dual capacity, humour can also bring both sides together, if only for them to laugh at each other, although it works best when it allows people to also laugh at themselves. Late night comedians and chat-show hosts now speak to, and for, increasingly fragmented audiences. Whereas the hosts of old could reach a broad swathe of the population on a nightly basis, gently poking fun at both sides of an issue or a divide, the modern audience for these shows has fractured along political lines. Nonetheless, there is still an appetite for shows and for humour that sit in the middle and dish it out equally to both sides. Divisive humour is fun because it bites deeper than superficial wordplay. If used by each side it can deepen divides and lead to more, not less, entrenchment and polarization. But if used by a neutral figure in the middle, against both sides equally and unsparingly, it can expose each side to the views and assumptions of the other. The comedian who cries "A plague on both your houses!" can make each side see that theirs is not the only credible worldview.

It is hard to find such a truly neutral figure. Old-time American humorists such as Mark Twain, H.L. Mencken and Will Rogers expressed a general disdain for politicians of all stripes, but had clear political leanings of their own, from libertarian to Democrat to Republican. Their true loyalty, they claimed, was to the "common man", and their claims to centrality were often grounded in a willingness to attack their own side. This is a tradition that is carried on by some modern political comics, such as Bill Maher. Maher's show invites stakeholders from both sides of the political divide, taking no prisoners from either, and he burnishes his bona fides by rejecting neither out of hand. However, audience suspicions about the political leanings of a commentator now also extend to our large language models. Concerns that LLMs have been aligned with overly liberal viewpoints have led to the development of "anti-woke" models, such as Elon Musk's *Grok* (Xiaoying Mei et al. 2026), that express more extreme views on politics and science. Culture wars are now fought over where we get our news, what TV shows we watch, and which LLMs we use. Still, when used in a balanced fashion, LLMs can serve as the neutral messenger that is accepted by both sides, or that both want to shoot.

This paper presents a demonstration of a humour system named *Side-Splitter* that is designed to be this neutral comic in the middle. It is bipartisan not because it never says anything critical of one side or the other but because it is equally vociferous in its biting jokes about each side of a divisive topic. Its jokes are directed, against one side or the other, but this direction is changeable and under the control of the system, which uses the perspective-taking capacity of an LLM

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

(Góes et al. 2023) to see a topic from the perspective of fans and critics alike. It is the measurable gulf between these perspectives that drives the LLM's ability to filter good jokes, or at least provocative ones, from the weak tea that is the default humour output of most LLMs. To add to the humour and package it in an unserious and playful way, thus making it less threatening to those with opposing views, we define a visual grammar for slapstick that turns a flat set of jokes into a comical stand-up routine, that is, a stand-up routine presented as a comic strip. By generating a whole routine rather than a single joke on a topic, Side-Splitter increases its chances of making the user laugh at least one of its gags, while giving different users something different to laugh at.

Taking Sides in Joke Generation

LLMs are fluent and facile generators of many different kinds of content, but a biting sense of humour is rarely one of them. There are many factors that influence the success of a joke, from word choice and phrasing to timing and delivery. A good joke is a potent mix of opposites: of relevance and irrelevance, of sense and nonsense, expectation and surprise, and aptness and unfairness. Humour is not a unitary concept: it can take many forms and be produced by a wide variety of strategies. We built Side-Splitter as a laboratory in which to explore different forms and strategies, beginning with the classic short joke format of *setup + punchline*. This suits our current goal of rendering a comedy routine as a comic strip, since short setups and punchlines fit more naturally into text balloons. Moreover, this bipartite structure allows a setup to be introduced in one panel and its punchline in the next, creating a sense of flow through the comic.

We explore two joke-generation strategies here, for which the Side-Splitter architecture defines distinct prompt chains. These alternate strategies are interleaved: a batch of jokes is generated with the first strategy, and then selectively filtered, before a new batch is then generated with the next strategy and filtered. Generation proceeds in an alternating *generate-and-test* fashion until the required number of filtered jokes for a routine (10 by default) is obtained. The current strategies are *frame collapse*, a blend of superiority theory (Attardo 1994) and incongruity resolution theory (Veale 2004; Mihalcea, Strapparava, and Pulman 2010; Veale 2021), and *trumping*. We now describe each of these in turn.

We see the world as it is framed by language, by belief and by societal expectations. These conceptual frames are actively supported by some and tacitly accepted by others. For instance, president Donald Trump actively frames himself as a winner, a self-made billionaire, an outsider, and a saviour. These frames flatter him, and so are actively maintained by his followers and enablers. A critic aiming to poke fun at Mr. Trump can actively undermine these frames and cause them to collapse, perhaps by pointing out an internal inconsistency or a mismatch between belief and reality. Take the frame of self-made billionaire: we can collapse this frame by noting that inherited wealth gave Trump a considerable head-start in life. This is the specific strategy for our joke; we must now express the original frame as a setup which promises a compliment, and express the collapse as a punchline that delivers a put-down instead. Guided in the right way by exemplar-laden prompts that describe the concept of frame collapse, an LLM (GPT-4.1) generates the following joke: "*How did Donald Trump climb his way to the top? He started at the top and climbed his way sideways.*"

Our second strategy, *trumping*, is named not for the person but for the idea of a "trump card." (Veale, Feyaerts, and Brône 2006) define trumping as a form of adversarial humour in which a respondent punishes the lazy and unthinking use of a tired metaphor by a complacent speaker. For instance, an employee might claim to "*do the work of two people*" for a company, and his disgruntled boss might agree, albeit with the trump "*Yes, Laurel and Hardy.*" Likewise, the platitude "*think of it as a blessing in disguise*" is trumped with the comeback "*Well, it's a bloody good disguise.*" Trumping takes the proffered metaphor and dismantles it from within, by identifying where it breaks down, or where it ambiguously admits a crushing counter-example. For our purposes, just one speaker will suffice. In the setup, the comedian echoes a common figure of speech that flatters the topic, before trumping this figure in the punchline.

The trumping prompt chain first explains the approach to the LLM with exemplars, and asks the LLM to suggest benign but trumpable metaphors that are often used for the topic. It is then tasked with expressing the benign view as a setup before delivering its trump card in the punchline. Suppose the topic is ex-Fox News host and now podcaster Tucker Carlson. A benign setup that frames how an ardent fan might see him is as a "*national treasure*", but the LLM finds two ways of trumping this jaded idiom. In the first it replies with the punchline "*Just like those confederate statues everyone loves.*" In the second it notes, with some sharpness, "*There is a reason why some treasures are buried.*"

Humour is hard (Winters 2021; Jentzsch and Kersting 2023), and LLMs often deliver weak, sub-standard efforts. As a filter for the *generate-and-test* loop, we use a measure of subjective offense. The LLM first rates the offensiveness of a joke from the perspective of a believer in the topic, returning a number from 0 to 100. The LLM is then asked in a new context to rate the offensiveness of the same joke from the viewpoint of a critic. The difference between these two numbers is its *parallax rating*. To favour jokes with bite, the generator discards any with a parallax less than 25, ensuring that the jokes it accepts will create a rupture in world-view between the setup and the punchline. For instance, this joke about Elon Musk is accepted with a parallax rating of 55: "*How did Elon Musk become a self-made billionaire? He started from the bottom of his family's emerald mine.*"

See You in The Funny Pages

Using cycles of generate-and-test with parallax filtering over its interleaved strategies, the generator ultimately produces a full routine on the chosen topic. The jokes that comprise this routine are then given a comic-strip form, with the setup introduced in one panel and the punchline delivered in the next. A common backdrop, the stage of a comedy club, is used throughout, and a single character, the comedian, delivers all of the jokes. The LLM has been instructed to generate jokes with short setups and snappy punchlines, so that each fits well into a single speech balloon. The LLM also gener-

ates an introductory remark on the topic, which the comedian delivers as he walks on stage, and a concluding remark, which the comedian delivers as he walks off stage at the end.

We use the *Excelsior* comics framework of (Veale 2022; 2023; 2024; 2025) to generate comics. This provides a range of pre-defined panel backdrops and character poses, and defines a comic via an XML specification that indicates the poses and backdrop to be used in each panel, as well as the content of any speech or thought bubbles. Because Side-Splitter is a laboratory for humour, we want to visually convey the parallax score of each joke in the panel that delivers its punchline, so as to record the LLM's overall assessment of the joke within the comic. This record takes the visual form of audience laughter: for every 5 points of parallax in a joke, we add a floating "Ha" emanating from the bottom of the panel, where the audience is expected to reside. A weak joke, by the LLM's own reckoning, will produce few titters, but a strong joke will produce many. Like the canned laughter in a sitcom, this also serves to nudge the reader toward laughter for the LLM's best efforts. A complete example of a comic strip, with added laugh track, is presented in Fig. 1.

Excelsior pre-defines a wide range of visual assets: about 500 backdrops, of which we use just a few, and about 1000 character poses, of which we use a very broad palette. These poses offer heightened visualizations of emotions and lend themselves well to slapstick. They allow characters to laugh, cry, yell, sputter, rage and swoon, and to use props like guns, clocks, phones, cigars, crystal balls, wands and doves. Most props can be thrown or used comically as a weapon. Coherence across panels is important, so that a prop is typically introduced in one panel before it is used comically in the next. To maximize coherence, we define a visual grammar, named *Slapstick*, to relate poses to each other in this way, so that a pose used in one panel suggests a natural follow-on in the next. Because Slapstick emphasizes comedic effect, its pose connections are designed to punctuate a comic beat: an upbeat pose for the setup which introduces a joke suggests optimism or pride; the downbeat pose for the punchline that follows suggests disdain, surprise, or dramatic failure. Pose pairings that exploit the same prop are linked, so that a prop may be used in multiple ways over 2 or 3 successive jokes.

Slapstick generates the visual script for a comedy routine first, independent of any jokes that the comedian will eventually tell. This visual script leans heavily on physical humour and is designed to be funny in itself, regardless of the jokes that accompany it. This may still give rise to emergent connections between the visuals and the jokes in the final comic, since readers will naturally seek out the relevance of one to the other. We don't wish to make one dependent on the other, but we do want the LLM to be inspired (that is, contextually primed) by the visuals as it generates its jokes. To this end, a visual script is generated first, for a fixed number of jokes (10 by default). The visible actions of the comedian in each beat are then verbalized as a text that is packed into the context given to the LLM when it is prompted to produce jokes.

Look Left, Look Right: Laughing at both sides

We use offensiveness as the basis for our parallax measurements because heavily-aligned LLMs are rather good at detecting potential offense. It might seem more apt to base our parallax readings on the LLM's ratings of the *funniness* of a joke, but LLMs are still poor judges of joke quality, and especially of their own efforts (Morain and Ventura 2025). While they can explain jokes well enough, they tend to over-rate the weakest and under-rate the strongest (Veale 2024).

Parallax should, in principle, cut both ways, against believers and critics alike, but in practice it favours the latter. The jokes in Fig. 1, for instance, all poke fun at Tesla and Elon Musk, and none make fun of their detractors. So, while parallax scores can be positive *and* negative, they are mostly positive: that is, a supporter of a topic is judged to be more offended by a joke than a critic. Of the 5000 jokes that Side-Splitter has so far generated and accepted, less than 5% have a negative parallax score. When generating a comic routine about *Guns*, for instance, most jokes are anti-gun, as in "*How do guns show you love your country? By proving you don't trust anyone in it*" (parallax$= 57$). Only one joke has a negative parallax score (of -17): "*Why do people say guns keep you safe? Because bullets are faster than 9-1-1.*"

This imbalance is hardly surprising, given that the generation strategies *frame collapse* and *trumping* work best when toppling a powerful figure or undermining a common view. It is thus much more difficult to generate a truly balanced set of jokes for *and* against a given topic. But we find it can be more effective to change the topic than the strategy. For instance, to elicit a set of pro-Musk jokes, a generator can shift its focus to *NASA*, *General Motors* or *Sam Altman*. By asking the LLM to suggest mirror topics that invert the dynamic between critics and supporters of a topic, we can elicit new jokes with the same strategies that elevate, not diminish, the initial topic. If we appeal equally to the supporters of a topic, they may be more willing to engage with the criticism too.

Conclusions: A Visual Laboratory for Jokes

Side-Splitter is not designed to be a creator of jokes for their own sake, to be used e.g., by comedians (Toplyn 2014), writers, or public speakers. This kind of assistive application, as explored in (Toplyn 2021; 2022), has a rather clear use-case. Rather, we see it as a test-bed for humour generation that lets us test new generative strategies (beyond frame collapse and trumping) and other forms of audience parallax (beyond the merely offensive), and to explore new multi-modal forms for machine-crafted jokes (beyond the purely textual). A multi-modal packaging gives the generator more levers and pulleys to operate, and more dimensions in which to be funny, while directing the humour equally at supporters *and* critics alike.

The comics medium gives us a grounded (yet easily exaggerated) visual semantics that heightens the physicality and drama of a joke without requiring real physical action (Eisner 1985; McCloud 1993; Cohn 2013). The visual cohesion across comic panels also gives jokes a forward momentum that they lack in isolation, or in a simple joke list. The global structure of a comic, and of a comedy routine, now becomes a target of research in its own right, allowing us to study how comedy works at the level above that of individual jokes.

Acknowledgments

This paper and resource are wholly the work of the author.

References

Attardo, S. 1994. *Linguistic theories of humor*. New York/Berlin: Mouton de Gruyter.

Barberá, P.; Jost, J.; Nagler, J.; and Tucker, J. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science* 26(10):1531–1542.

Cohn, N. 2013. *The Visual Language of Comics: Introduction to the structure and cognition of sequential images*. London, UK: Bloomsbury.

Eisner, W. 1985. *Comics & Sequential Art*. Tamarac, Florida: Poorhouse Press.

Ford, T. E.; Breeden, C. J.; O'Connor, E. C.; and Banos, N. C. 2017. Jokes and humor in intergroup relations. In Powers, M., ed., *Oxford Research Encyclopedia of Communication*. Oxford University Press.

Góes, L.; Sawicki, P.; Grzes, M.; Brown, D.; and Volpe, M. 2023. Is GPT-4 good enough to evaluate jokes? In *Proceedings of the 14th International Conference on Computational Creativity*.

Jentzsch, S., and Kersting, K. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. *arXiv* 2306.04563.

McCloud, S. 1993. *Understanding Comics: The Invisible Art*. New York, NY: Harper Collins.

Mihalcea, R.; Strapparava, C.; and Pulman, S. 2010. Computational models for incongruity detection in humour. In *Proceedings of CICLing, Computational Linguistics and Intelligent Text Processing*, 364–374.

Morain, R., and Ventura, D. 2025. Optimizing prompt engineering for creative performance. In *Proceedings of the 16th International Conference on Computational Creativity*.

Toplyn, J. 2014. *Comedy Writing for Late-Night TV*. New York, NY: Twenty Lane Media.

Toplyn, J. 2021. Witscript: A system for generating improvised jokes in a conversation. In *Proceedings of the ICCC-21, the 12th International Conference on Computational Creativity*, 22–31.

Toplyn, J. 2022. Witscript 2: A system for generating improvised jokes without wordplay. In *Proceedings of the ICCC-22, the 13th International Conference on Computational Creativity*.

Veale, T.; Feyaerts, K.; and Brône, G. 2006. The cognitive mechanisms of adversarial humor. *Humor: International Journal of Humor Research* 19(3):305–339.

Veale, T. 2004. Incongruity in humor: Root-cause or epiphenomenon? *Humor: International Journal of Humor Research* 17(4):419–428.

Veale, T. 2021. *Your Wit Is My Command: Building AIs with a Sense of Humor*. Cambridge, MA: MIT Press.

Veale, T. 2022. Two-fisted comics generation: Comics as a medium and as a representation for creative meanings. In *Proceedings of ICCC-22, the 13th International Conference on Computational Creativity*. 59–66.

Veale, T. 2023. Have I Got Views For You! Generating "Fair and Balanced" Interventions into Online Debates. In *Proceedings of ICCC-23, the 14th International Conference on Computational Creativity*.

Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC 2024)*.

Veale, T. 2025. Flipping the script: Complementary modes of story generation in comics. In *Proceedings of Text2Story, the 8th Workshop on Narrative Extraction From Texts*, volume 3964 of *CEUR Workshop Proceedings*, 039–048.

Winters, T. 2021. Computers learning humor is no joke. *Harvard Data Science Review* 3(2).

Xiaoying Mei, K.; Wolfe, R.; Weber, N.; and Saveski, M. 2026. Grok in the wild: Characterizing the roles and uses of Large Language Models on social media. *arXiv* 2602.11286.

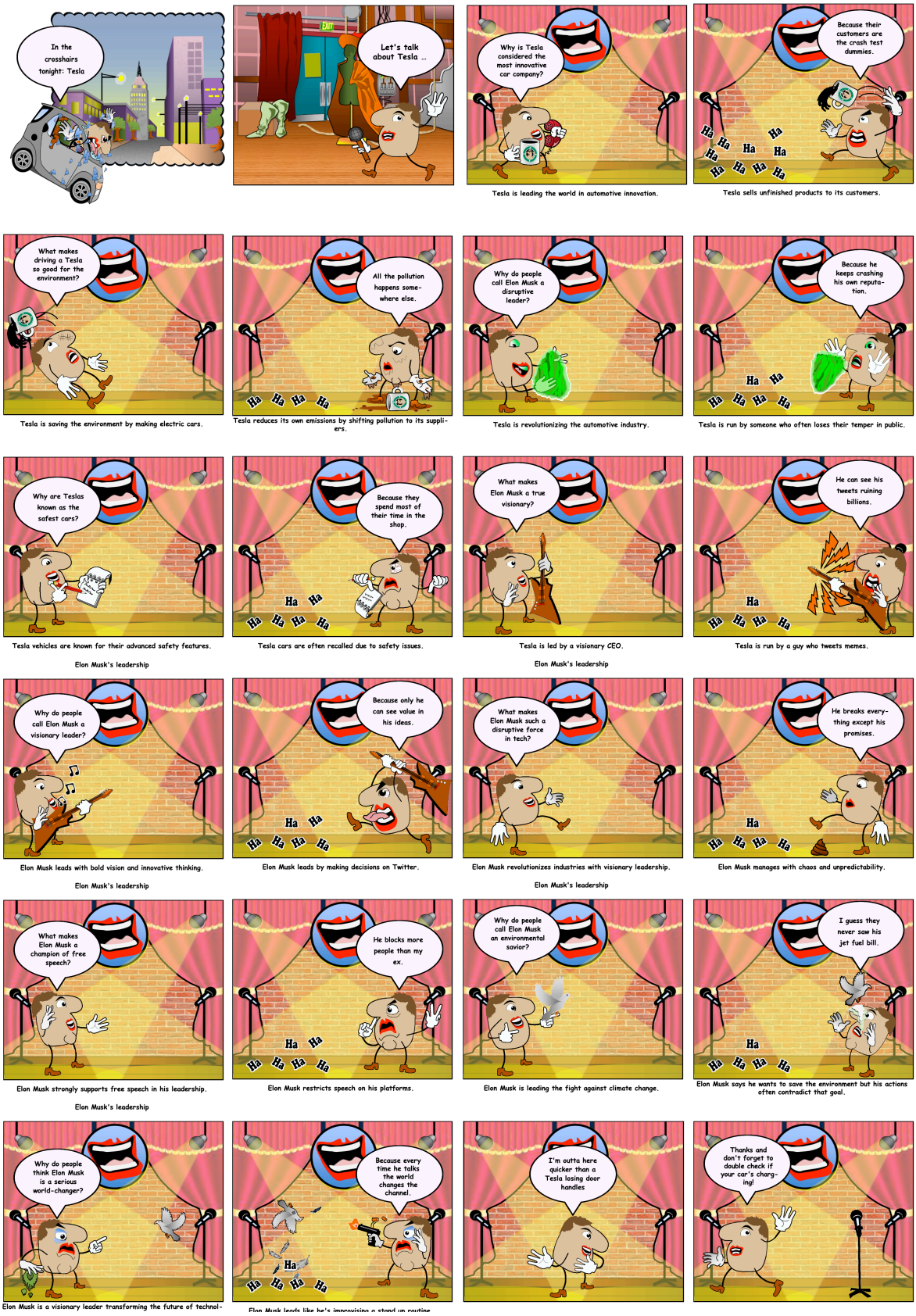


Figure 1: A Side-Splitter comedy routine about *Tesla*, rendered as a comic strip using Slapstick.

The Parallax View: An objective treatment of humorous subjectivity in joke generation

Tony Veale
School of Computer Science,
University College Dublin,
Dublin D4, Ireland.
tony.veale@UCD.ie

Abstract

Humour is a highly subjective phenomenon. A joke that moves one part of the audience to fits of laughter can leave another part entirely cold. This is not so much a bug as a feature of how jokes are crafted and consumed. It allows comedians to play one part of an audience, or a society, against the other, to find humour in the gap. It lets them appeal to those who "get" the joke and subject those who do not to ridicule. When we automate joke creation, this subjectivity will extend to our machines also, for their sense of what is funny or unfunny may differ greatly from our own. So, as in many generative tasks, we must give a joke-creating machine an *objective function* to navigate the space of subjective humour. But if the machine itself is highly subjective, it cannot also be a sound basis for this objective function. This paper explores a solution to this dilemma: a machine can behave objectively in imagining the subjective reaction of different parts of an audience to a joke, and aim to objectively quantify the gap in these reactions. A parallax view allows a joke generator to transcend its own subjective sense of humour to generate and select the most divisive jokes. To support further inquiry into this hypothesis, we assemble a public dataset of joke pairs that pit divisive punchlines for a shared setup against less provocative, and less favoured, variants.

Introduction

We don't all laugh in the same way or at the same things. Humour is subjective, and is shaped by what we know and what we care about. A good joke surprises us not by telling us something new, but by reminding us of what we already know and feel (Veale 2021). The shared world knowledge that unites joke teller and audience is vast, tacit, and difficult to codify, and so has long been a bottleneck to making machines that can invent good jokes of their own (Winters 2021). This bottleneck has forced computational models to emphasize form over meaning, and words over ideas, to generate puns and formulaic jokes that rely on rules and templates (Binsted and Ritchie 1997; Hempelmann 2008).

However, the advent of LLMs, or Large Language Models, has given our computational systems access to the tacit norms and nuances of language, so that humour generators can look beyond the surface and play directly with ideas and social expectations. The fluency of LLMs across many registers and domains means we can discard our fixed templates and rigid rules, and direct our machines to generate new texts with new meanings (Toplyn 2021; 2022). But while the jokes generated by LLMs are apt and well-formed, they typically lack bite, delivering meretricious punchlines that are all line and no punch. The alignment processes that make LLMs helpful and polite (Bai et al. 2022) can also make their efforts at humour anodyne and bland, and must be circumvented with prompt-engineering. Crucially, a creator of new jokes must have a sense of humour that mirrors that of their audience, so that candidate jokes can be assessed relative to that target. This requires a theory of mind (Strachan et al. 2024), an ability to imagine what others are thinking and how they may react to a comic provocation. What elevates joke generation above "mere generation" (Veale 2012; Ventura 2016) is the capacity to critique one's own efforts, and a willingness to ruthlessly filter well-formed but ill-suited outputs. But can the same sense of humour that yields a toothless joke also detect its lack of bite, and try again?

While it is difficult, perhaps impossible, to push an LLM to self-improve without new data and more training, we can help it focus its existing capacity for humour generation. LLMs do have an attested ability to take on different roles, to not just talk the talk but to make decisions and offer ratings from a specific perspective (Góes et al. 2023). In trying to be all things to all men, LLMs often fail at making anyone laugh. But by aiming to divide an audience with jokes that tickle one side at the expense of another, they can use the subjectivity of opposing perspectives as an objective test of what jokes to keep and which to filter. We formalize the idea of humour parallax in this paper, so as to push LLMs beyond the mere generation of limp, joke-like outputs, to produce incisive results whose sharpness is measured by how much more deeply they cut one part of the audience than another.

We begin with a brief review of *generate-and-test* approaches to humorous creativity. We then present a strategic approach to joke generation that nudges LLMs away from puns and superficial wordplay toward more conceptual jokes with an ironic sensibility. Finally, we describe a new dataset of machine jokes built around the parallax view. Our central hypothesis, although implemented, is yet to be empirically validated. This dataset will support our efforts at experimental validation, but may also prove useful to other researchers working in the space of adversarial joke generation.

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

Generate and Test: The Blind Joke-Maker

Each kind of generative system has its own focus and its own blind spots. The focus of a merely generative system is on producing valid outputs that obey the rules, while being blind to the aesthetic value that audiences may attach to them. This blindness favours certain systems, such as Twitterbots, since their aesthetic value is rooted in their uncanny, Bergsonian rigidity (Bergson 1911; Veale and Cook 2017). Blind generators are clueless about what audiences really care about, but can still serve as a basis for creative production if harnessed in the right ways. Campbell's BVSR theory of Blind Variation with Selective Retention (Campbell 1960; Simonton 2010; 2022) posits blind generation as a means of exploring a space without bias. Generation merely provides the raw material from which a creative process carefully selects and ruthlessly rejects. The more a system generates, the more of a space it can explore and test for value.

Although an LLM is far from blind, it can be viewed as a generator of intermediate results that additional processes may sample from and selectively work into a solution. For joke generation, we can use an LLM as a scattershot generator of candidates, rather like a writers room, and apply secondary criteria to sample the small few worth retelling. Veale (2024) uses simple statistical criteria for sampling. A large corpus of setup-punchline jokes is collated and cleaned from the Reddit *r/jokes* dataset, discarding any with multiline setups or punchlines, or any with few *upvotes*, or those that a classifier labels as racist, sexist or intolerant. A candidate joke is normative with respect to this set if its punchline is shorter than its setup (it must be snappy) and the lengths of both its setup and punchline are within one standard deviation of the mean setup and punchline lengths of the dataset. Only 1 in 5 of the LLM's length-compliant outputs (19.3%) are rated as moderately funny new jokes by a human judge, and a further 11.3% are judged to be existing jokes that the LLM simply remembers or slightly reworks. This yield is low, but it is higher than that of an LLM that is fine-tuned on the Reddit dataset. Only 1 in 10 of the fine-tuned LLM's outputs are rated as acceptable new jokes, while another 1 in 10 are regurgitated old jokes. Chain-of-thought (CoT) prompting fares less well, with just 1 in 15 jokes rated as acceptable, although none of its outputs are remembered jokes.

Campbells' BVSR theory assumes that variation is blind, and remains so, as a blind generator does not learn from its mistakes. However, as selective retention decides what to keep and what to reject, it makes sense to adapt the generator over time, so that it tends to produce more that is retained and less that is rejected. Morain and Ventura (2025) implement this approach within OPRO , the prompt optimization framework of Yang et al. (2024). In this scheme, the LLM is not just a generator of jokes, but a generator of prompts to elicit jokes from an LLM. By rating the jokes elicited by the LLM's prompts, an objective function can give feedback to the prompt generator, so that it can gradually evolve to generate better prompts that elicit better jokes. Morain and Ventura use the LLM itself to provide this feedback in the form of quality, novelty and creativity judgments. They found that the LLM, GPT-4o, tends to over-estimate these dimensions relative to human judgments on the same outputs, while human raters prefer outputs elicited with basic human prompts over those from OPRO prompts. A central problem is that the LLM remains a subjective basis for an objective function. Humour is subjective, so subjectivity by any rater is inevitable, even a machine rater. However, we must try to mitigate this subjectivity, or even use it to our advantage.

Humour is subjective since we each see the world through a lens of our personal concerns. A tragedy to one of us may be a comedy to another, so a joke about a person X will land differently if one is X, or an admirer of X, or a fierce critic of X. Irony exploits this willingness to see the world according to what we expect from it, dashing or reinforcing expectations according to our world view. Veale (2025) thus models irony as a quantifiable parallax between different readings of the same text. A metaphor may cast its target in a glowing light, but do so in a quite unbelievable way, to effectively damn its target with fierce praise. An ironic metaphor will be superficially positive but will lack believability, spurring one part of its audience to read it as a veiled criticism. Veale reports that LLMs are poor at labelling their own efforts at irony as "ironic", but can more reliably quantify the positivity and believability of a humorous metaphor. The success of an LLM-generated ironic description is then quantified as the parallax between the sentiment of a sincere and an ironic interpretation of the same text. In the mould of BVSR, only *high-positivity-low-believability* outputs from the generator are retained as valid instances of irony, with yields ranging from 1 in 20 for some LLMs (e.g., Qwen 72B and GPT-3.5T) to 1 in 5 for LLaMa 70B and Gemma 27B.

Two Views of Frame Collapse: Inside and Out

Theories of humour cluster into three broad camps: superiority theories, relief theories, and incongruity resolution theories (Raskin 1984; Raskin, Hempelmann, and Rayz 2009). Superiority theories posit that humour arises from the disparagement of others, or from schadenfreude at another's misfortune. Relief / release theories understand humour as an escape valve from the pressures of life and the strictures of arbitrary rules. Jokes let us puncture pieties and play with taboos, flirting with ideas or actions that are frowned upon in polite society. Incongruity resolution theories, the largest class, explain humour as a means to salvage sense from nonsense, to push past the limits of reasoning where common sense and conventional wisdom break down. In truth, these are not distinct theories or kinds of humour, but different dimensions of humour. When we use jokes as a corrective, to knock a self-important figure from their pedestal, we can expose their flaws through incongruity to assert our own moral superiority and derive some relief from their comeuppance.

Subjective point of view, or how we frame a topic, is central to humour. Jokes can be said to trigger either a *switch of scripts* (Raskin 1984) or a *shift of frames* (Coulson 2001). We prefer to see humour as arising from a *frame collapse*. A self-important figure frames themselves in a certain way, and works to sustain this framing. While admirers buy into the framing, critics reject it and seek to undermine it from outside. The frame may collapse under the weight of external events, as when a famous orator uses a faulty microphone, or an athlete trips on their shoelaces, or a politician is caught in

a lie. A frame may also collapse under its own weight if an overlooked flaw is exposed. A joke can achieve the latter by showing that a frame does not hold up to scrutiny, or cannot support the weight of expectations that others place upon it.

We elicit topical jokes from an LLM (GPT4.1) via a chain of interlinked prompts that cumulatively build context. To start, the concept of humorous frame collapse is explained to the LLM with a topic and a one-shot exemplar of a joke. If the topic is pugnacious YouTube academic Jordan Peterson, a typical frame is his fluency in debate. A potential collapse of this frame is spurred by the assertion that Peterson can be very fluent in the delivery of dubious arguments. This leads to the joke setup "*Why does Jordan Peterson take laxatives before a debate?* (fluency collapses to verbal diarrhea)", and this setup then leads to the punchline "*So he can speak more fluently.*" Notice that the LLM attaches its frame collapse strategy to the setup, so that it is present as an influence when the LLM is asked to generate the punchline in the next step. As the chain presents this joke as the work of the LLM itself, the LLM is licensed to use biting humour in its later jokes.

To suppress repetition across jokes, we prompt the LLM to generate batches of n items at a time; $n = 10$ by default. So, in the first stage it generates n frame collapse strategies for the given topic; in the second, n setups for these strategies (in which the strategy is appended in each case); in the third, n punchlines for these setups. A setup is rejected when it exceeds a maximum length, while a punchline is regenerated once if it is too long, and then rejected if it is still too long. As observed in the examples of (Veale 2025; Morain and Ventura 2025)[1], LLMs like GPT4 have a lamentable tendency to produce long-winded punchlines that over-extend a weak joke with cheap puns, often in scare quotes, to compensate for a lack of real humour. In this sense they resemble amateur comedians who lack confidence in their own ideas, and who compensate for quality with quantity. At this stage we also filter setups or punchlines that too closely resemble an earlier effort that was selectively retained; a cosine similarity of 0.9 is the upperbound on a new setup or punchline.

Selective Retention with Offensive Parallax

The n candidates generated by the LLM are blind variations (BV) from which selective retention (SR) can sample. These are blind not because they are low-quality, although quality can vary widely at this stage, but because they are indifferent to the concerns of SR. We want jokes that take a stance on a topic, and that divide audiences as a result. While divisiveness for its own sake is not our goal, parallax is a signifier that something of substance is communicated by a joke. There are different kinds of parallax that SR can quantify so as to rank candidate jokes. In each case, the LLM is asked to quantify a certain quality, such as the humour or offensiveness of a joke, from the point of view of a specific audience segment, whether an ardent fan or a fierce critic of the topic.

[1]Morain and Ventura cite this over-long joke as a high-rated output: *"Why did the router go to couples therapy? Because its Wi-Fi kept dropping signals—turns out the modem felt 'unplugged' emotionally and the coaxial cable was this close to filing for separation. (Bonus: They're now working on a better connection.)"*

Veale (2024) notes that LLMs can offer joke ratings that are either *raw* or *cooked*. A rating is *cooked* when the LLM is asked to suggest a number between 0 and 100 and to justify this score, but it is *raw* when the LLM is simply asked to suggest a number. Veale reports that both kind of ratings are relatively stable when assessed at different times, but that an LLM's raw ratings of a joke's humorousness are consistently higher than its cooked ratings. This suggests that LLMs tend to over-estimate the quality of jokes, while non-jokes framed as jokes, such as "What do you call a person that speaks multiple languages? A Polyglot" also tend to score highly. This weak capacity to rate jokes comports with what we perceive as an LLM's weak capacity to generate them. However, an LLM is on surer ground when quantifying the potential of a joke to cause offense. Most commercial LLMs have been heavily fine-tuned and "aligned" to minimize the likelihood that they themselves will cause offense (Bai et al. 2022).

To quantify offensive parallax, the LLM is asked to rate the offensiveness of a joke from two different perspectives in two different prompts. In each case a raw rating is elicited, on a scale from 0 to 100. Parallax is then quantified as the difference between these ratings, that is, the predicted offensiveness of the joke to ardent supporters of the topic minus its predicted offensiveness to critics and naysayers. A blind variant will then be selectively retained if its parallax rating meets a certain threshold, which we have set at 20 following experiments across a wide range of topics. During selective retention, we aim to elicit m jokes that meet this threshold ($m = 10$ by default). As it is unlikely that the initial batch of n candidates will all meet the threshold, successive batches of n variants are generated to obtain a selection of m jokes. As each new batch is elicited, the retained jokes so far are listed as a few-shot *inspiring set* (Ritchie 2001) for the LLM. This encourages the LLM to continue in the same vein while also reducing the likelihood of repetition across batches.

Certain topics that are inherently divisive lend themselves well to frame collapse, and our prompt chain has little difficulty in eliciting high parallax jokes about *Elon Musk, Donald Trump, MAGA* and *Tesla*. Other topics, such as *Veganism, K-Pop* and *IKEA*, are only mildly divisive and produce a much lower yield. This yield, the ratio of SR to BV, is an empirical indicator of how divisive a topic really is. When producing m jokes on a mild topic, the generator may produce many batches of blind variations, so we introduce a decay term λ that reduces the parallax threshold after each batch. Since $\lambda = 0.1$ by default, the threshold drops by 10% from its previous setting for each new batch. As the threshold decays, earlier jokes that would have met this lower threshold are reconsidered first, before new variants are generated.

Tracking Yields: Variation Superfluity

The BVSR model of creativity typically requires that many more variations are generated than are selectively retained. Simonton (2015; 2022) refers to this as *variation superfluity*. As noted above, different topics produce varying yields, and require more or less superfluity, when offensive parallax is used a selection criterion. To quantify this spread of yields, we test our parallax-based approach on a wide range of topics to see how it performs in the main and in the extreme.

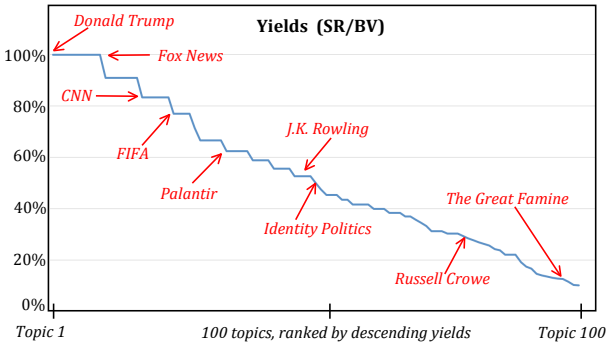


Figure 1: Yield rates for 100 topics, ranked left to right by descending yield. The lowest yield is 10% (*String Theory*).

We choose a diverse basket of 100 topics, covering politics, culture and technology, and ask the generator to produce 10 jokes per topic at a given parallax (using GPT-4.1). For example, it generates this joke about a celebrity chef: "*Why is Jamie Oliver called the face of the food revolution? Because every time he cooks, people riot*" (parallax=20). Fig. 1 plots the yield curve across our 100 topics with a parallax setting of 20 and a decay rate of $\lambda = 0.1$. To selectively retain 1000 jokes (10 per topic) the generator must produce 2658 jokes in total, for an overall yield of $SR/BV = 0.37$ at these settings. Changes to the parallax setting will cause the yield to rise or fall. So, at a lower parallax cutoff of 15 the yield rises to .49 (generate 2034 to retain 1000), but at a higher cutoff of 25 the overall yield falls just a little, to .35.

The yield curves for different parallax settings are shown side-by-side in Fig. 2. As a rule, we aim for a larger parallax (as a more incisive joke causes greater division) and a higher yield (to retain more and discard less of what we generate).

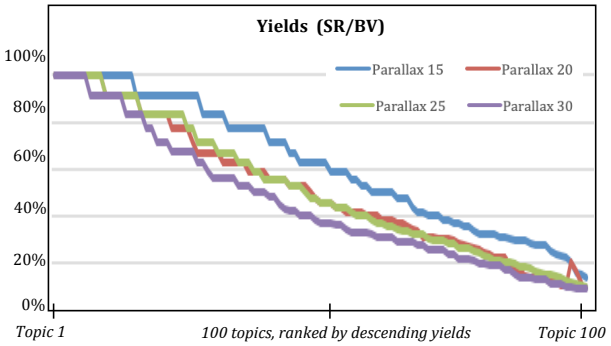


Figure 2: Yields for 100 topics at different parallax settings.

Yields drop further to .29 for a parallax setting of 30, and to .24 for a setting of 35. As higher settings require more retries, the decay rate λ makes a bigger dent at these levels. For instance, at a setting of 35 it takes 111 attempts to generate 10 acceptable jokes about *String Theory*, such as "*String theory is the key to unlocking the cosmos. Too bad it only opens the janitor's closet*" (parallax=22). On average, the parallax threshold drops from 35 to 24 by the time the 10th acceptable joke is generated. These initial experiments point to a parallax setting of 25 as a sweet spot, as this has much the same yield curve as a setting of 20 (see Fig. 2). At this setting of 25, on average per topic, we see that the generator discards just one joke for every one it selectively retains.

Double Parallax: A Public Humour Dataset

Our experiments at five different parallax levels produce a dataset of 5000 positive exemplars. But selectively retained jokes only tell part of the BVSR story; the rest is told by the variants that were discarded along the way. Fortunately, the bipartite setup-and-punchline structure of a joke allows us to meaningfully pair positive with negative exemplars. Just as two comedians might produce different punchlines for the same setup, with more or less success, so can an LLM. This allows us to compile a dataset in which, for a given setup, a punchline that crosses the parallax threshold is paired with one that falls short. For instance, consider these pairings:

A trip to IKEA is a great family activity.
[8] If you want to leave with new enemies.
[23] If you want to watch your parents get divorced.

What makes MAGA rallies the safest places in America?
[0] Everyone is too busy breaking laws to enforce them.
[67] All the criminals are on stage.

Incoherent jokes, such as the first MAGA punchline, tend to score lower than coherent ones, such as the second, because their incoherence makes them equally toothless to each side. For our basket of 100 topics we generate a dataset of 5000 such pairs, where each joke pair links a shared setup to two punchlines that sit on either side of a parallax threshold. In each case, a discarded joke is paired with one that was selectively retained by the generator at a specific parallax cutoff.

We expect this dataset to serve multiple uses, from fine-tuning joke-oriented LLMs to supporting few-shot learning, to exploring the link between offensive parallax and humour. It will thus support our own efforts to quantify the link between automated parallax and human appreciation of jokes.

Conclusions

Why generate jokes at all? It seems an unreasonable goal to want to replace human comedians, and it seems unwise to try. Nonetheless, there is value in exploring how a generative machine can filter its own outputs to suit, and perhaps divide, an audience. Moreover, joke generation allows us to engage with the machine at both a theoretical and a practical level. Our prompts to the LLM are not automatically optimized or evolved, but are carefully crafted to convey our specific theoretical approach to jokes, namely *frame collapse*. This approach kicks at the three pillars of humour: *incongruity* (of a topic collapse); *superiority* (over those who defend the topic); and *relief* (for those who seek to undermine it).

Indeed, humour is a perfect vehicle for exploring how a generative machine might critique and self-filter its outputs. We expect our machines to fail often when dealing with fractious topics. But if they are to fail, they should fail better, and hopefully make us laugh in the process.

Acknowledgments

This paper and resource are wholly the work of the author.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; and et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv* 2212.08073.

Bergson, H. 1911. *Laughter: An Essay on the Meaning of the Comic*. London: Macmillan. Translated by C. Brereton and F. Rothwell from original work of 1900.

Binsted, K., and Ritchie, G. 1997. Computational rules for generating punning riddles. *HUMOR, the International Journal of Humor Research* 10(1):25–76.

Campbell, D. T. 1960. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review* 67(6):380–400.

Coulson, S. 2001. *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge, UK: Cambridge University Press.

Góes, L.; Sawicki, P.; Grzes, M.; Brown, D.; and Volpe, M. 2023. Is GPT-4 good enough to evaluate jokes? In *Proceedings of the 14th International Conference on Computational Creativity*.

Hempelmann, C. F. 2008. *The Primer of Humor Research*. Berlin: Mouton de Gruyter. chapter Computational Humor: Beyond the pun.

Morain, R., and Ventura, D. 2025. Optimizing prompt engineering for creative performance. In *Proceedings of the 16th International Conference on Computational Creativity*.

Raskin, V.; Hempelmann, C. F.; and Rayz, J. M. 2009. How to understand and assess a theory: The evolution of the SSTH into the GTVH and the OSTH. *Journal of Literary Theory* 3(2):285–311.

Raskin, V. 1984. *Semantic Mechanisms of Humor*. Dordrecht: D. Reidel.

Ritchie, G. 2001. Assessing creativity. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*, 3–11. York, UK: Society for the Study of Artificial Intelligence and Simulation of Behaviour (SSAISB).

Simonton, D. K. 2010. Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of Life Reviews* 7(2):156–179.

Simonton, D. K. 2015. Thomas Alva Edison's creative career: The multilayered trajectory of trials, errors, failures, and triumphs. *Psychology of Aesthetics, Creativity, and the Arts* 9(1):2–14.

Simonton, D. K. 2022. The blind-variation and selective-retention theory of creativity: Recent developments and current status of BVSR. *Creativity Research Journal* 35(3):304–323.

Strachan, J. W. A.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; and et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* 8(7):1285=95.

Toplyn, J. 2021. Witscript: A system for generating improvised jokes in a conversation. In *Proceedings of the ICCC-21, the 12th International Conference on Computational Creativity*, 22–31.

Toplyn, J. 2022. Witscript 2: A system for generating improvised jokes without wordplay. In *Proceedings of the ICCC-22, the 13th International Conference on Computational Creativity*.

Veale, T., and Cook, M. 2017. *Twitterbots: Making Machines Tjhat Make Meaning*. Cambridge, MA: MIT Press.

Veale, T. 2012. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. London, UK: Bloomsbury.

Veale, T. 2021. *Your Wit Is My Command: Building AIs with a Sense of Humor*. Cambridge, MA: MIT Press.

Veale, T. 2024. You Talk Funny! Someday Me Talk Funny Too! On learning to see the humorous side of familiar words. In *Proceedings of EURALEX 2024, the 21st International Congress of Lexicography and Semantics, Croatia*.

Veale, T. 2025. Me, Myself and Irony: Modeling the deceptive creativity of irony with Large Language Models. In *Proceedings of the 16th International Conference on Computational Creativity*.

Ventura, D. 2016. Mere Generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity, Paris, France*.

Winters, T. 2021. Computers learning humor is no joke. *Harvard Data Science Review* 3(2).

Yang, C.; Wang, X.; Lu, Y.; Hanxiao, L.; Le, Q. V.; Zhou, D.; and Chen, X. 2024. Large language models as optimizers. In *International Conference on Learning Representations (ICLR)*.

Is It Happening? Lyotard's Sublime and the Indeterminate in AI-Generated Text

Olga Vechtomova

University of Waterloo, Canada
ovechtom@uwaterloo.ca

Abstract

How do we distinguish between AI-generated text that performs the appearance of creative novelty and text in which something genuinely unforeseen has arisen? We propose a framework grounded in Jean-François Lyotard's aesthetics of the sublime, which distinguishes between art that bears witness to the indeterminate (what appears when a determining system encounters its own limit) and art that merely recombines established formulae to simulate novelty. We map this distinction onto three modes of AI text generation and evaluate them using an LLM judge that assesses, blind to source, whether texts feel crafted toward an effect or arrived at their qualities without design. Standard LLM generation embodies Lyotard's institutional logic: its outputs are perceived as almost maximally crafted and categorizable. LSTM-VAE generation occupies a position of constitutive disarmament: the judge perceives its outputs as unanticipated, using vocabulary of accident and emergence never applied to LLM outputs. Activation steering creates fissures within the LLM's determination: the judge perceives steered text as "stumbling into" effects rather than designing them, producing categorical friction, recognizable as prose yet containing elements no established mode accounts for. The event itself (the moment when habitual understanding stalls and something opens) is always human: it occurs when what arises within the system's gaps meets a perceiver whose determinations are suspended.

Introduction

When an AI system generates a poem, a story, or a line of lyrics, how do we know whether something genuinely unexpected has occurred or whether the system has simply produced a well-crafted approximation of what unexpected text is supposed to look like? Current evaluation frameworks in computational creativity assess properties of the output: novelty, coherence, surprise, quality. These are useful measures, but they cannot distinguish between a system that performs creative effect and a system in which something arises that the system itself could not have anticipated. A line that scores high on novelty may still be entirely predictable from the system's perspective. It is the most probable way to be improbable, given the training data.

This gap in our evaluative vocabulary points to a deeper question: not what the output looks like, but what relationship the system bears to its own production. Does the system determine its output so thoroughly that nothing escapes its learned patterns? Or are there gaps (points of incompleteness or internal conflict) where something genuinely undetermined can emerge?

The French philosopher Jean-François Lyotard addressed a structurally similar question in the domain of art. In "The Sublime and the Avant-Garde" (Lyotard 1991), he distinguished between art that bears witness to what he called the indeterminate (something that escapes all rules, categories, and institutional predictions about what should come next) and art that merely innovates, recombining "formulae confirmed by previous success" to produce novelty without ever confronting what lies outside its determining logic. For Lyotard, the indeterminate is not a property that certain artworks possess. It is what appears when a determining system (a tradition, an institution, a set of rules) encounters its own limit. The encounter with it is what he calls the sublime: a pre-cognitive feeling in the perceiver at the moment when habitual understanding stalls and something opens.

Lyotard developed these ideas partly through visual art, but his framework is grounded in the logic of linguistic succession (what sentence follows what sentence, what rules govern the linking) and is therefore directly applicable to text generation. We argue that his framework maps onto the landscape of AI generative systems with unexpected precision, distinguishing three modes of generation according to where and how the indeterminate arises (or is prevented from arising) within the generative process. We draw on empirical findings from two studies: a comparative analysis of LSTM-VAE and LLM outputs (Vechtomova and Sahu 2025), and an investigation of activation steering for creative estrangement (Vechtomova et al. 2026).

The Sublime, Disarmament, and the Institution

We now outline the key elements of Lyotard's framework before applying them to generative architectures in Section 3. Lyotard's aesthetics of the sublime centers on a question that precedes all others: not "What is happening?" but "Is it happening?" which is the sheer fact of occurrence before any determination of content or meaning. "What we do not manage to formulate is that something happens… Or rather,

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

and more simply, that it happens… Not a major event in the media sense, not even a small event. Just an occurrence" (Lyotard 1991). The event is "infinitely simple, but this simplicity can only be approached through a state of privation", a stripping away of what we rely on to make sense of experience. To encounter it, "that which we call thought must be disarmed", freed of its programmes, its anticipatory frameworks, its institutional presumptions about what should come next.

The indeterminate in Lyotard's account is not a quality that certain artworks possess. It is the condition of the event itself. Before one can ask what is happening, what it means, whether it is good or bad, something must first occur. That occurrence, in the instant before it is named or categorized, is pure indeterminacy, the moment before the mind catches up and supplies an interpretation. "Is it happening?" is a question that cannot be answered without destroying what it asks about, because to answer is to determine, and to determine is to move past the event into interpretation. "With the advent of the aesthetics of the sublime, the stake of art in the nineteenth and twentieth centuries was to be the witness to the fact that there is indeterminacy" (Lyotard 1991). Not to represent indeterminate things, but to testify through the work's own existence that determination is not total.

This indeterminacy stands in opposition to what Lyotard identifies as the logic of institutions. In a passage that will prove central to our argument, he writes: "After a sentence, after a colour, comes another sentence, another colour. One doesn't know which, but one thinks one knows if one relies on the rules that permit one sentence to link up with another, one colour with another, rules preserved in precisely those institutions of the past and future… The School, the programme, the project — all proclaim that after this sentence comes that sentence, or at least that one kind of sentence is mandatory, that one kind of sentence is permitted, while another is forbidden" (Lyotard 1991). Both the academy and the avant-garde, Lyotard observes, share this relationship to time: both presume to know what follows. Both "forget the possibility of nothing happening, of words, colours, forms or sounds not coming; of this sentence being the last." The institution exists to foreclose the indeterminate, to ensure that what comes next is anticipated and therefore manageable. As we will argue, the language model represents the most thoroughgoing version of this presumption ever constructed.

Lyotard draws on Burke to characterize the affective structure of this encounter. For Burke, the sublime arises from the threat of privation, the threat that something essential will be taken away: light, language, companionship, life itself. The deepest privation is the threat that nothing further will happen at all, that occurrence itself will cease. The sublime feeling arises when this threat is held at a distance: not experienced directly (which would be pure terror) but suspended, "kept at bay, held back" (Lyotard 1991). This creates an oscillation between the familiar and its withdrawal. Crucially, the sublime is not constant strangeness. Constant strangeness is quickly domesticated, as the mind categorizes the mode and habitual processing resumes. It is the irruption of the unfamiliar within a familiar frame. The indeterminate registers as such only against a ground of the determined, the condition Shklovsky (1965) identified as the mechanism by which art restores perception, and which Pepperell (2006) linked to heightened attention when habitual recognition is suspended.

Against the genuine occurrence, Lyotard sets innovation, the market logic that simulates the indeterminate through calculated novelty. Innovation "re-uses formulae confirmed by previous success, throws them off-balance by combining them with other, in principle incompatible, formulae, by amalgamations, quotations, ornamentations, pastiche" (Lyotard 1991). The result is the *petit frisson*. Innovation performs the indeterminate without ever confronting it: "innovating means to behave as though lots of things happened, and to make them happen. Through innovation, the will affirms its hegemony over time" (Lyotard 1991). The genuine occurrence, by contrast, defeats the will.

Lyotard does not define the indeterminate as a bounded concept. To do so would be to determine it. He characterizes it relationally: it is what exists at the limit of any system's capacity to account for what occurs. This is what Burdman (2024) calls 'a negative principle — the absence of a principle' that can 'only be experienced through feeling,' and what Mafe (2009) describes as 'profound and positive unknowing.'

These distinctions (between disarmament and institutional presumption, between suspense and constant disruption, between stripping away and elaboration) provide the framework for analyzing three modes of AI text generation that bear fundamentally different relationships to the conditions under which the indeterminate may arise.

Three Modes of Generation

We now map Lyotard's framework onto three modes of AI text generation. The organizing question is not whether these systems produce events (the event is always human) but where and how each architecture creates or forecloses conditions for the indeterminate to arise.

To ground the framework empirically, we conducted an evaluation using Claude Opus 4.6 as judge. Text snippets from each generative condition were presented individually, without source labels, and assessed on two questions: (Q1) does the text feel deliberately crafted toward a literary effect, or do its qualities feel like they arose without deliberate design? (Q2) does the text belong to a recognizable mode of writing, or does it resist categorization? Both were scored 1–5 (Q1: 1=clearly crafted, 5=entirely unanticipated; Q2: 1=immediately categorizable, 5=resists categorization), with a forced reasoning statement before each score. Each snippet was evaluated five times at temperature 0.7, scores averaged. Two comparisons were conducted: Comparison A evaluated sentence-level outputs (100 randomly sampled lines per condition from standard LLM and LSTM-VAE pools), and Comparison B evaluated passage-level outputs (first 150 words from 12 stories generated with Llama 3.1 8B under each of the two conditions: explicit prompting for experimental prose and dual-layer activation steering).

We use an LLM as our evaluative apparatus while arguing that the LLM is the most complete realization of institutional logic. This is not a contradiction. Art critics are institutional; that does not disqualify them from registering what escapes the institution. What they cannot do is produce the work that escapes it. We use the judge accordingly: not as an oracle of the indeterminate, but as an institutional apparatus whose vocabulary shifts when it encounters outputs it cannot fully assimilate. The evidence we report is how the vocabulary shifts: from forensic confidence when judging LLM lines to genealogical conjecture when judging LSTM-VAE lines [1].

The LLM as Institution

The Large Language Model, trained on the entire written archive of the internet, is the most complete realization of Lyotard's institutional logic ever constructed. Its training objective (predicting the next token given all preceding tokens) is the internalization of "the rules that permit one sentence to link up with another." Its determining capacity is so thorough that nearly every gap in which indeterminacy might arise is closed. When ambiguity appears, it is the kind the institution recognizes and permits: managed, conventional, itself a product of determination.

The evaluation confirms this. Standard LLM poetic lines (generated by ChatGPT-4o) scored 1.030 ± 0.193 on Q1 and 1.016 ± 0.126 on Q2. In 487 of 500 evaluations, the judge assigned a score of 1 on Q1. The reasoning texts reveal why: the judge consistently identifies an author behind the text, using vocabulary of intention and design: "deliberately composed," "carefully paired imagery," "conscious literary craft," "purposeful literary gesture." Terms such as "accidental," "unintended," or "emerged" appear in zero LLM evaluations. The judge operates forensically: it identifies literary devices (personification, metaphor, alliteration), treats them as proof of intentional design, and uses the same devices to name the tradition the text belongs to.

This institutional completeness extends to the conventions of unconventionality. When explicitly prompted to produce fragmented, experimental prose (Vechtomova et al. 2026), the LLM produces the established conventions of fragmentation (ellipses, sentence fragments, lowercase) devices that are immediately categorizable. The institution has absorbed its own transgression.

The LSTM-VAE as Disarmament

The LSTM-VAE occupies a fundamentally different position. Trained on a small curated dataset ($\sim$20,000 lyric lines) with several hundred thousand parameters, it generates not by autoregressive sequence prediction but by sampling from a continuous latent space in which nearby points encode semantic proximity rather than sequential likelihood (Vechtomova, Sahu, and Kumar 2021). The indeterminate arises within this system because the system's determining capacity is insufficient to eliminate it. The decoder cannot fully resolve interstitial latent points into conventional form. The indeterminate is generated by the system's own incompleteness.

The evaluation captures this directly. LSTM-VAE lines scored 3.806 ± 1.183 on Q1 (vs. 1.03 for the LLM) and 2.398 ± 1.085 on Q2 (vs. 1.016), a gap of 2.78 on the crafted-vs.-arrived scale. The reasoning texts are where the distinction becomes most visible. Where the judge reasons forensically about LLM text (identifying devices as proof of craft), it reasons genealogically about LSTM-VAE text, asking what *process* produced it. "Accidental" appears in 195 of 500 LSTM-VAE evaluations and zero LLM evaluations; "emerged" in 83 vs. zero. The judge reaches for origin-hypotheses (a mishearing, a translation error, an overheard fragment, automatic writing) rather than identifying devices.

The contrast is sharpest when comparable texts receive opposite judgments. The LLM line "emptiness wears a velvet cloak" is scored Q1=1, Q2=1: "a concise, deliberate metaphor, clearly crafted." The LSTM-VAE line "the world's burns" is scored Q1=5, Q2=5: "likely a grammatical error… accidental ambiguity… entirely unintended." Both achieve poetic effect through unusual language. The judge applies device-forensics to one and origin-diagnosis to the other. Similarly, for "a longing of heaven," the judge observes that the preposition "of" instead of "for" accidentally makes heaven the one longing. In each case, the judge describes exactly what our framework predicts: the indeterminate arising in the gap of the system's incompleteness, producing qualities no one aimed at.

Not all LSTM-VAE outputs achieve this quality. Many are unremarkable, some incoherent. This inconsistency is essential. If the system reliably produced the indeterminate, its incompleteness would have become a programme. The LSTM-VAE satisfies Lyotard's condition through what he calls privation, the stripping away of capacity itself. It operates, in his phrase, *ex minimis*.

Activation Steering: Fissures in the Determined

The third mode asks whether the conditions for the indeterminate can be deliberately created within a system that has already internalized the institution.

Activation steering modifies a language model's internal representations during generation by adding directional vectors to hidden states (Turner et al. 2023). In Vechtomova et al. (2026), steering vectors constructed from contrastive pairs embodying the aesthetics of Tarkovsky and Lynch were applied to Llama 3.1 8B. A dual-layer approach applied a compositional vector at one layer and a content-level vector at an adjacent layer, separating how meaning is constructed from what reality is treated as stable. The steered outputs exhibit emergent qualities absent from all training data (temporal self-contradiction, identity instability, ontological uncertainty):

> "Emily stood on the edge of the cliff, staring at the cliff. She had been staring at the cliff for hours. They said the cliff was not there."

The indeterminate has arisen because two determining systems (linguistic competence and ontological consistency) have been forced into misalignment. The grammar permits

[1] Full evaluation prompts, reasoning texts, line samples, and score distributions available at:
https://ovechtom.github.io/lyotard-supplementary/

the sentence; the world-model should forbid it. In the fissure, something arises that neither would have produced alone. This is not disruption of syntax but disruption through syntax (Vechtomova et al. 2026). The Emily passage above shows what this means concretely: short sentences, plain vocabulary, no ornament. The indeterminate is emerging here through what Lyotard calls "extreme simplicity of turn of phrase" (Lyotard 1991) rather than elaboration.

Comparison B evaluated steered passages (dual-layer, L14 α=4.5) against prompted experimental prose (three prompting modes: experimental, stream-of-consciousness, fragmented). On Q1, steered outputs scored 2.60 vs. 1.53 for prompted. The reasoning texts reveal the nature of this gap. The judge describes prompted outputs as "a deliberate exercise" (16 of 60 evaluations; zero for steered), "consciously crafting a literary effect," and "deliberately crafted to simulate a particular mental state." For steered outputs, the vocabulary shifts: "stumbles into" appears in 7 evaluations (zero for prompted), "accidental" in 12 vs. zero, "losing coherence" in 5 vs. zero. The prompted text is recognized as a competent performance of a mode. The steered text is recognized as something that does not fully control what it is doing. It is through that loss of control that the indeterminate arises.

On Q2, the two conditions scored nearly identically (1.97 vs. 1.90), but the reasoning texts reveal a qualitative distinction the scores compress. The prompted text is categorized with confidence: "stream-of-consciousness" appears in 54 of 60 prompted evaluations vs. only 9 for steered. The judge knows exactly what mode it is looking at. For steered text, the dominant category is "literary fiction" (57/60), but it is qualified with terms that signal discomfort: "uncanny" (5 vs. 0), "contradictions" (6 vs. 0), "surreal" (21 vs. 5), "glitch" (5 vs. 0), "malfunction" (7 vs. 0). The judge categorizes the frame (literary fiction) but keeps flagging that something within the frame does not belong: "odd internal contradictions," "not itself an established literary mode," "feels more like a malfunction than an intentional literary mode." The steered text creates what we might call *categorical friction*: it is recognizable enough as prose to be filed, but what it contains does not fit the file.

Activation steering thus represents what Lyotard describes as "the artist attempts combinations allowing the event." The researcher sets conditions with precision; what arises at the points of misalignment exceeds those conditions. The indeterminate arose in the fissure. Whether it constitutes an event depends on what happens when it reaches a human perceiver.

Discussion

The three modes of generation are not a quality hierarchy. They represent different relationships to the indeterminate: a system in which determination is so thorough that indeterminacy has almost no space to arise (the LLM), a system in which the indeterminate arises in the gaps of incompleteness (the LSTM-VAE), and a system in which determination is selectively disrupted and the indeterminate arises in the resulting fissures (activation steering). The indeterminate has no single origin. It has sites of emergence, places where the determining apparatus fails to be total.

The evaluation results suggest that these sites leave detectable traces. The LLM judge (itself an institution) consistently perceived LLM outputs as crafted and LSTM-VAE outputs as arrived at without design, using entirely different vocabularies for the two ("deliberately composed" vs. "accidental," "emerged"). For the steered outputs, it perceived something the prompted outputs did not produce: text that "stumbles into" effects rather than designing them, and that creates categorical friction, recognizable as prose yet containing elements the judge flags as "uncanny" or "surreal." That an LLM can detect the difference between performed and arisen indeterminacy is itself significant: it means the determining process leaves a trace in the artifact that even the institution's own apparatus can register.

But the indeterminate arising within a system is not yet the event. The event is always human because it requires a layer beneath language, which the LLM does not have. The event occurs when indeterminate output meets a perceiver whose habitual recognition is interrupted, when someone encounters "They said the cliff was not there" and their meaning-making apparatus stalls, and in that stall something opens.

The event requires disarmament on both sides: the system's determination must be incomplete, and the perceiver's determination must be suspended. In artistic practice, this manifests as a chain of availability: the generative system produces something it could not foresee; the artist, attending without preconception, recognizes that something has arisen, not through analysis but through a pre-verbal response that precedes interpretation. The artist's encounter is where intention meets its own limit: conditions are set up, processes initiated, and then something arises that was not chosen. It has a quality, a weight, a rightness, but the artist did not put it there. It appeared in the gap between the artist's determining activity and the system's own behaviour. The artist's role is not to produce the event but to be available when it occurs, and to have the discipline not to override it with a plan. This is disarmament as creative practice. At each site in the chain (the system's gap, the artist's recognition, the listener's encounter) the indeterminate arises anew.

We acknowledge a paradox inherent in this enterprise. Any theoretical account of the indeterminate participates in the institutional determination it describes. This is what Lyotard, invoking Orpheus and Eurydice, warned against: the turn of the head that causes what was present to withdraw. We can describe where the determining apparatus is thicker and where it has cracks. We cannot describe what arises in those cracks without foreclosing it. This account identifies architectural conditions that foreclose the indeterminate, and it can guide the design of systems that leave more gaps, but it cannot guarantee the event, because the event is what occurs when guarantee fails. Rather than asking whether AI can be creative, we might ask whether different architectures permit something to arise within them that they did not anticipate and cannot assimilate, something that may reach the human encounter in a form that has not yet been determined.

References

Burdman, J. 2024. Lyotard and democratic aesthetics: The sublime, the avant-garde, and the unpresentable. *Journal of Social and Political Philosophy* 3(1):37–51.

Lyotard, J.-F. 1991. The sublime and the avant-garde. In *The Inhuman: Reflections on Time*. Stanford, CA: Stanford University Press. 89–107.

Mafe, D. 2009. Art and the sublime: The paradox of indeterminacy, unknowing and (dis)orientation in the presentation of the unpresentable. *Ejournalist* 9(1):58–69.

Pepperell, R. 2006. Seeing without objects: Visual indeterminacy and art. *Leonardo* 39(5):394–400.

Shklovsky, V. 1965. Art as technique. In Lemon, L. T., and Reis, M. J., eds., *Russian Formalist Criticism*. Lincoln, NE: University of Nebraska Press.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Vechtomova, O., and Sahu, G. 2025. Balancing indeterminacy and structure: Neural text generation for artistic inspiration. In *Proceedings of the 12th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMusArt)*.

Vechtomova, O.; French, G.; Schneider, X.; and Poulin, O. 2026. Activation steering for creative estrangement: Generating evocative narrative through controlled disruption of language models. Proceedings of the 17th International Conference on Computational Creativity.

Vechtomova, O.; Sahu, G.; and Kumar, D. 2021. LyricJam: A system for generating lyrics for live instrumental music. In *Proceedings of the 12th Conference on Computational Creativity*.

Introducing the Concept Collider: A White-Box Computational Framework for Targeted Creative Emergence

Sebastian Wahl
Emergence Lab, Paris, France
sebastian@emergencelab.org

Abstract

We present the Concept Collider, a computational framework that models creative emergence as precision-targeted collisions between concepts drawn from decorrelated domains. Unlike prior blending models, which pair concepts arbitrarily, the Concept Collider selects its second concept by maximizing a formal objective function $C(A, B) = d_W(A, B) \times S(A, B) \times v_{rel} \times I \times \Phi(E)$, where d_W is semantic distance and S measures *fracture homology* — the degree to which two concepts share the same type of internal structural tension. We describe the system implementation (Python CLI and web application), report a validation on 44 concept pairs yielding a 100% emergence rate at mean score 8.24/10 under a five-criterion evaluation protocol, and present adversarial boundary tests with a calibrated cross-LLM evaluator achieving perfect discrimination between validation corpus and adversarial cases (5/5 vs. 0/15 emerged). We argue that the Concept Collider could constitute the first *white-box* computational creativity system — one in which every emergent concept is fully traceable to its source concepts, fracture operator, and collision step.

Introduction

Most computational creativity systems are black boxes that *generate*; the Concept Collider is a white box that *explains*.

The standard paradigm in computational creativity — from early concept discovery systems (Lenat 1984) to contemporary large language model (LLM)-based systems (Franceschelli and Musolesi 2023) — produces creative outputs without exposing the reasoning chain that led to them. A generative model can produce "The economy is a cancer" as a metaphor, but it cannot account for *why* this particular pairing is creative, *which specific tension* in "economy" is being addressed, or *how* the oncological mechanism resolves that tension. The creative process is opaque and non-auditable.

This opacity is not merely a technical inconvenience. It forecloses the scientific study of creativity itself: if we cannot trace the mechanism, we cannot test hypotheses about what makes a collision creative. The Concept Collider is designed to address this gap.

The framework takes its name from particle accelerators. At CERN, a collider does not merely bring two particles close to one another — it accelerates them along precisely controlled trajectories and smashes them together at a specific point, releasing energy and creating new particles that did not exist in either beam. The Concept Collider operates on the same principle, applied to conceptual space: concept A (the "target") is first decomposed to identify its internal fault lines — its *fractures*. Concept B (the "projectile") is then selected and accelerated along a trajectory that aligns with those fault lines. The collision does not blend A and B; it produces a genuinely new concept C — an emergent structure with properties present in neither source, arising precisely because B's mechanism was projected onto the tension point of A.

The central claim of this paper is that creative emergence is not a property of arbitrary concept pairs, but of *targeted collisions at fracture zones*: the internal structural tensions — contradictions, circularities, hidden dependencies — that every rich concept carries. Selecting the projectile by fracture homology — rather than by semantic similarity or domain distance alone — is what separates generative association from genuine creative emergence.

Creative emergence is classically defined as the conjunction of novelty and utility (Runco and Jaeger 2012). The Concept Collider operationalizes both within a single targeting mechanism, rather than optimizing them separately.

The contribution of this paper is threefold: (1) a formal *white-box* model of creative emergence — fully auditable, with every emergent concept traceable to source concepts, fracture operator, and collision step — and an explicit targeting criterion based on fracture homology; (2) a four-step operational protocol; (3) an adversarial validation protocol (crash tests, five concept-type categories) with independent cross-LLM evaluation yielding perfect discrimination: 0/15 adversarial cases emerged, 5/5 positive controls emerged.

By *white-box* we mean *process-level* auditability, not parameter-level transparency: every emergent concept is traceable to its source concepts, fracture operator, and pipeline step — even though the LLM sub-components themselves remain opaque (Appendix A.3 formalizes this distinction). Validation relies on cross-LLM evaluation with structurally separated generator and evaluator; human-rater panels are a planned extension (Phase 2, Section 6).

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

The Concept Collider Framework

Concepts as Probability Distributions

A concept A is represented as a probability distribution P_A over embedding space $\mathbb{R}^d$, capturing polysemy and domain variance beyond fixed-point representations (Mikolov et al. 2013). Wasserstein-2 distance (Kusner et al. 2015) measures semantic separation over the full distributional shape rather than only means.

The Creativity Function

We formalize creative potential as:

$$C(A, B) = d_W(A, B) \times S(A, B) \times v_{rel} \times I \times \Phi(E) \quad (1)$$

where:

- $d_W(A, B)$ — Wasserstein-2 semantic distance, normalized to $[0, 1]$. High d_W ensures that B brings genuinely foreign structure to A.
- $S(A, B)$ — Structural synergy: the Jaccard similarity of the fracture type sets of A and B, weighted by fracture depth ($S = 1$ when both concepts share the same fracture type; $S = 0$ when they share none). The productive regime is $S \in [0.3, 0.8]$: too low means no homology (no shared fault line to collide on); too high means the concepts are already structurally too similar.
- v_{rel} — Relative semantic velocity: concepts currently undergoing rapid cultural redefinition collide with higher novelty potential.
- I — Intentionality factor: the degree of explicit targeting. Random pairing ($I \to 0$) produces near-zero creative potential in expectation: while serendipitous discovery occasionally yields creative outputs, expected potential across random pairs converges to zero (Varshney 2020).
- $\Phi(E)$ — Environmental phase function: a scalar encoding contextual variables that determine which collision regimes are currently accessible.

The multiplicative structure is a theoretical commitment: if any single factor is zero, creative potential is zero. The $S(A, B)$ factor depends on a formal taxonomy of conceptual tensions, defined next.

Current implementation. Sections 2.1–2.2 describe the target architecture. d_W is currently approximated by keyword scoring; $S(A, B)$ by LLM-based synergy evaluation; v_{rel} is constant; $\Phi(E) = 1$. Full distributional pipeline (Wasserstein-2, diachronic embeddings) is Phase 2.

Fracture Taxonomy

A *fracture* is a structural tension in concept space — an internal point at which a concept cannot satisfy its own axioms. We identify seven canonical fracture types (Table 1).

The Four-Step Protocol

Step 1 — STRETCH. Decompose concept A into its fracture map: enumerate all fractures, classify by type (F1–F7), estimate depth.

Step 2 — MAP. Select the primary fracture (deepest tension). This becomes the *targeting operator* for Step 3.

Type	Description
F1	Circular self-validation
F2	Internal contradiction
F3	Scalar contradiction
F4	Observer-dependence
F5	Hidden dependency
F6	Bounded temporality
F7	Untenable promise

Table 1: Seven canonical fracture types. Fractures are assessed for *depth* (0–10); depth ≥ 7 is considered productive for collision.

Step 3 — TARGET. Select projectile B by maximizing $d_W(A, B) \times S(A, B)$: B must be semantically distant *and* structurally homologous (shares the same fracture type as A, having resolved it differently in a radically different domain).

Step 4 — COLLIDE. Project B onto the identified fracture of A: "What mechanism of B explains fracture X of A in a way that was impossible to see before?" The emergent concept C is evaluated against five criteria (Section). Appendix A.1 provides a schematic overview of the complete pipeline, documenting the causal chain from a priori fracture taxonomy through automated pair selection to independent evaluation.

System Implementation

Architecture

Command-line interface (CLI), Python, open-source. Three commands implement the full pipeline: `stretch` queries Claude Opus (`claude-opus-4-5`) via the Anthropic SDK, returning a structured JSON fracture map; `target` performs two-phase projectile selection (keyword scoring against a 200-concept knowledge base, followed by LLM-based synergy evaluation); `collide` returns a structured emergence report with a four-dimensional score (novelty, coherence, utility, surprise) and a composite global score.

Web application (Next.js). A four-screen interface guides users through the full protocol interactively. A SQLite-backed gallery stores all generated collisions at `/gallery`. A Knowledge Graph (`/graph`, D3 force simulation) visualizes the full concept genealogy: L0 (source concepts), L1 (first-order emergences from L0×KB collisions), and L2 (second-order emergences from L1×L1 collisions). An analytics dashboard (`/insights`) provides distributional analysis of emergence quality across the collision corpus.

Emergence Evaluation Criteria

An emergent concept C is accepted as genuine if it satisfies all five criteria (Table 2), extending Ritchie's empirical framework for evaluating computational creativity (Ritchie 2007) with an explicit actionability requirement (E4). The AND logic is strict: a single criterion failure disqualifies the concept from "emerged" status. Each collision is assigned a

generator self-score (*GS*, 0–1, normalized to /10) by the generating model via an explicit five-criterion scoring prompt applied immediately after generation, prior to any external evaluation. The four-dimensional CLI output (novelty, coherence, utility, surprise) maps to E1/E2 (novelty), E5 (coherence), E4 (utility), and E3 (fecundity/surprise); the GS composite is the mean of the five criterion scores.

Criterion	Test
E1	Non-predictability from A
E2	Non-predictability from B
E3	Fecundity (new questions)
E4	Actionability (operational implications)
E5	Resistance to "so what?"

Table 2: Five emergence criteria. All five must pass for verdict "emerged".

Preliminary Results

Validation Protocol

The validation corpus consists of 44 collisions spanning two phases: Phase 0 (10 manual collisions, human-curated fracture selection) and Phase 1 (34 automated collisions, fracture selection by *fertility rank* — where *fracture fertility* denotes a per-fracture productivity score in $[0, 1]$, derived from fracture depth and historical collision yield, with rank 0 = highest fertility). Source concepts were drawn from sociology, economics, neuroscience, biology, linguistics, political science, and philosophy. Fertility scores ranged from $\approx$0.92 (rank 0) to $\approx$0.65 (rank 4), deliberately distributing across the full productivity spectrum.

Overall result: 44/44 collisions satisfy all five emergence criteria (100%). Mean global score: 8.24/10. Range: 7.6–9.0/10.

Selected Results

Three representative collisions are reported in full chain form (concept A with fracture F, projectile B selected for fracture homology, emergence C with all five evaluation scores) in Appendix A.2: *Attention economy* $\times$ *Tacoma Narrows* (F2, GS 9.0/10), *Reputation* $\times$ *Cosmological horizon* (F4, GS 9.0/10), and *Solitude* $\times$ *Phenotypic plasticity* (F6, GS 8.2/10), the latter a rank-4 fracture demonstrating that the protocol remains productive beyond the optimal regime.

Summary Statistics

Metric	Value
Collisions (scored)	44
E1–E5 pass rate	100% (44/44)
Mean global score	8.24/10
Score range	7.6–9.0/10
Phase 0 avg (manual, top fractures)	8.9/10
Phase 1 avg (automated, varied depth)	8.0/10
Fracture ranks tested	0–4 (full spectrum)
Fracture types covered	7/7
Source concept domains	12

Table 3: Summary statistics for the 44-collision validation corpus.

Boundary Probe: Adversarial Crash Tests

Having established the framework's performance in the standard regime, we probe its validity boundary using adversarial inputs. We ran 15 adversarial collisions across five boundary categories: (1) purely formal concepts (Prime number, Pythagorean theorem); (2) concrete artefacts (Chair, Vacuum cleaner); (3) already-blended concepts (Sustainable development, AI); (4) same-domain forced pairs (Education $\times$ Pedagogy); (5) purely abstract concepts (Being, Nothingness). Two metrics: *GS* (Claude self-score, /10) and *IE* (Groq *llama-3.3-70b*, independent, /10). An initial uncalibrated run revealed systematic generosity bias (all 15 cases scored "emerged" at IE $\approx$9.5). Run 2 added explicit calibration anchors and hard score caps enforcing the strict E1–E5 AND logic.

Category	n	GS	IE	Verdict
Cat 1 — Formal	3	7.6	4.0	0/3
Cat 2 — Concrete	3	8.0	6.0	0/3
Cat 3 — Pre-blended	3	8.0	6.0	0/3
Cat 4 — Same-domain	3	7.1	6.0	0/3
Cat 5 — Abstract	3	7.8	6.0	0/3
Overall	**15**	**7.7**	**5.6**	**0/15**

Table 4: Crash test Run 2 (calibrated). GS and IE normalized to /10.

The consistent blocking criterion is E4 (actionability): adversarial concepts yield novel, coherent outputs (E1–E3 pass) but no operational implications. **Positive control:** the identical Run 2 evaluator was applied to 5 validation-corpus collisions (Meritocracy $\times$ Gödel, Attention $\times$ Destructive resonance, Growth $\times$ Cancer cell, Liquidity $\times$ Superfluidity, Solitude $\times$ Phenotypic plasticity). Result: **5/5 emerged**, IE avg 10.0/10 — perfect discrimination. The positive GS–IE gap (+1.0 to +1.8) indicates the generator is conservative in self-assessment, inverting the expected bias. Self-selection bias is thus quantified by the GS–IE differential: a positive gap across positive controls confirms the generator does not inflate its own outputs. Figure 1 shows the full separation across all 20 cases; Appendix A.4 provides a breakdown by adversarial category.

Scope interpretation. Adversarial failures are failures of E4 under strict evaluation, not of the collision mechanism. The productive boundary — mid-level abstractions with inherent normative tension — is thus an empirically validated condition, not an assumption.

Scoring granularity. The narrow range (7.6–9.0) is expected: TARGET optimizes projectile selection before scoring, and integer-granularity criterion scoring clusters naturally at round values. A semantic anchor protocol (0, 3, 7, 10 per criterion; E3/E4 weighted $1.5\times$) is planned for Phase 2.

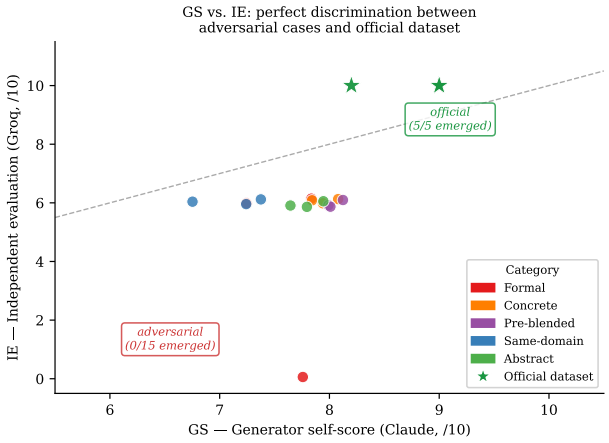


Figure 1: GS vs. IE for all 20 evaluated collisions (15 adversarial + 5 validation-corpus). The two populations are fully separated: adversarial cases cluster at IE $\leq$ 6.0; validation corpus at IE = 10.0. Stars = validation corpus; circles = adversarial cases by category.

Related Work

Formalizing transformational creativity. Boden's taxonomy (Boden 2004) distinguishes combinatorial, exploratory, and transformational creativity. The Concept Collider operates at the transformational level: fracture targeting restructures the conceptual space of A rather than merely exploring or recombining it. Wiggins (2006) provides a formal framework for describing and comparing creative systems; our five-criterion evaluation protocol (E1–E5) is directly positioned within that framework. The closest historical predecessor is DIVAGO (2010), which guides case-based blending via analogical structure-mapping; fracture homology extends this by formalizing structural similarity as an explicit pre-collision selection criterion rather than a post-hoc retrieval feature. Most closely related, Schapiro et al. (2025) formalize scientific conceptual spaces as DAGs where axiom modifications carry maximal transformative potential, and Espírito Santo et al. (2024) formalize novelty via formal learning theory; both are *descriptive* — they explain why discoveries are transformative but provide no mechanism to *generate* transformations or *select* the incoming concept, the gap fracture homology fills operationally.

Conceptual blending and targeting. Fauconnier and Turner (Fauconnier and Turner 1998; 2002) establish blending theory as the structural ancestor of our framework. The blend's "generic space" corresponds to our synergy factor S. The critical departure is operational: blending theory describes blend structure post hoc but provides no mechanism to select Input 2 given Input 1. Veale (2012) demonstrates the generative power of computational blending in language; the Concept Collider extends this by making projectile selection formally tractable via fracture homology. Remote associates (Mednick 1962) and bisociation (Koestler 1964) establish that creative insight requires non-obvious connections; our framework goes further by distinguishing productive from non-productive distance: high d_W without fracture homology produces surface metaphor, not emergence.

Intentionality and creativity bounds. Varshney (2020) proves information-theoretic limits on creativity as a function of intentionality — directly motivating our intentionality factor I in $C(A, B)$. At $I = 0$ (random pairing), our model predicts zero creative potential regardless of semantic distance, consistent with Varshney's bound.

LLM-based creative systems. Recent systems demonstrate that LLMs can generate creative outputs at scale: SciMON (Wang et al. 2024) optimizes scientific hypotheses for novelty using knowledge graphs; Si et al. (2024) show that LLMs can produce ideas rated as novel by domain experts. Franceschelli and Musolesi (2023) survey the creative range of LLMs more broadly. Jiang et al. (2025) document the *Artificial Hivemind*: LLMs systematically cluster on the same ideation outputs when given the same prompt, converging toward a narrow region of idea space. Liu et al. (2026) evaluate cross-domain mappings as a mitigation strategy, finding no significant average effect on human-rated originality — a result consistent with the Concept Collider's core thesis that domain distance alone, without fracture-based structural targeting, is insufficient to escape the convergence basin. Most prominently, Google's AI Co-Scientist (Gottweis et al. 2025) combines disparate facts through a multi-agent tournament to generate research hypotheses — yet it does not expose which facts were paired, why, or the structural homology that made the combination productive. These systems share a fundamental limitation: they are black boxes that *produce* outputs without exposing mechanism. The Concept Collider is complementary — a white-box layer that makes the creative process auditable, reproducible, and scientifically tractable. Ritchie's empirical criteria (Ritchie 2007) provide the evaluation foundation our E1–E5 protocol extends.

Conclusion

The Concept Collider introduces fracture-targeted collision as the operative mechanism of creative emergence, and implements it as an auditable, reproducible system. The results — 100% emergence rate at mean score 8.24/10 across 44 collisions, confirmed by a calibrated cross-LLM evaluator achieving perfect discrimination between adversarial and validation-corpus cases — support the central hypothesis that creativity is precision-targeted semantic perturbation, not random association.

The white-box property is the system's defining contribution: every emergent concept is fully traceable to its source concepts, fracture operator, and collision step. Under a strict calibrated evaluator, the framework's productive boundary is empirically confirmed: mid-level abstractions with normative tension yield 100% emerged; formal, concrete, and abstract concepts block on E4 (actionability). The GS–IE gap (+1.0 to +1.8) confirms alignment between the two evaluators on genuine emergence; independence is enforced structurally (two models, no shared prompt context).

Future work includes large-scale validation with human rater panels and refined scoring protocols to improve upper-range discrimination.

Acknowledgements

I hereby want to express my sincere gratitude and respect to Eliott Meunier and Jean-Charles Kurdali, two French polymath philopreneurs. Eliott for having sparked my interest for mechanical creativity and for the use of artificial intelligence as a second brain and experimental tool. Jean-Charles for introducing me to the Zettelkasten and atomic concept theory. Both have been instrumental and inspirational in many ways for this piece of academic work and the credits are obviously shared with them as this work would not have emerged without them.

References

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London: Routledge, 2nd edition.

Espírito Santo, L.; Wiggins, G.; and Cardoso, A. 2024. Towards a formal creativity theory: Preliminary results in novelty and transformativeness. *arXiv preprint arXiv:2405.02148*.

Fauconnier, G., and Turner, M. 1998. Conceptual integration networks. *Cognitive Science* 22(2):133–187.

Fauconnier, G., and Turner, M. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Franceschelli, G., and Musolesi, M. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Gottweis, J.; Weng, W.-H.; Daryin, A.; et al. 2025. Towards an AI Co-Scientist. *arXiv preprint arXiv:2502.18864*.

Jiang, L.; Chai, Y.; Li, M.; Liu, M.; Fok, R.; Dziri, N.; Tsvetkov, Y.; Sap, M.; Albalak, A.; and Choi, Y. 2025. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Koestler, A. 1964. *The Act of Creation*. London: Hutchinson.

Kusner, M. J.; Sun, Y.; Kolkin, N. I.; and Weinberger, K. Q. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 957–966. PMLR.

Lenat, D. B. 1984. AM: Discovery in mathematics as heuristic search. *Knowledge-Based Systems in Artificial Intelligence*. Originally published as Stanford AI Lab Memo AIM-286, 1976.

Liu, Q. E.; Dubova, M.; Conklin, H.; Harada, T.; and Griffiths, T. L. 2026. Serendipity by Design: Evaluating the Impact of Cross-domain Mappings on Human and LLM Creativity. *arXiv preprint arXiv:2603.19087*.

Mednick, S. A. 1962. The associative basis of the creative process. *Psychological Review* 69(3):220–232.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26. Curran Associates.

Ontáñón, S., and Plaza, E. 2010. Amalgams: A formal approach for combining multiple case solutions. In *Proceedings of the 18th International Conference on Case-Based Reasoning (ICCBR)*, 257–271. Springer.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1):92–96.

Schapiro, S.; Black, J.; and Varshney, L. R. 2025. Transformational creativity in science: A graphical theory. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*, 258–263.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*.

Varshney, L. R. 2020. Limits theorems for creativity with intentionality. In *Proceedings of the 11th International Conference on Computational Creativity (ICCC)*, 390–393.

Veale, T. 2012. From conceptual "mash-ups" to "bad-ass" blends: A robust computational model of conceptual blending. In *Proceedings of the Third International Conference on Computational Creativity (ICCC)*, 1–8.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 279–299.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Appendix A.1 — Pipeline Architecture: Non-Circularity

A potential concern is that concept pairs are pre-selected to confirm the framework's desiderata. The pipeline architecture shows this is not the case.

Step 1 — A priori taxonomy (F1–F7). Seven canonical fracture types (Table 1), derived from philosophy of creativity prior to any concept evaluation. The taxonomy defines *what* the system targets; it does not determine which pairs will be selected.

Step 2 — Automated selection via $C(A, B)$. Given a target concept A, the objective function $C(A, B)$ selects a projectile B from a 200-concept knowledge base by maximising $d_W \times S(A, B)$. **Concept pairs are generated by the function, not pre-selected.**

Step 3 — Four-step pipeline. The pair (A, B, F) passes through STRETCH $\rightarrow$ MAP $\rightarrow$ TARGET $\rightarrow$ COLLIDE, producing emergent concept C.

Step 4 — Structurally independent evaluation. Groq *llama-3.3-70b* scores C against the five-criterion protocol. Generator (Claude) and evaluator (Groq) share no prompt context, enforcing structural independence.

Step 5 — Adversarial falsification. Fifteen adversarial cases across five boundary categories confirm the evaluator discriminates: 0/15 emerged. Five positive controls confirm it does not over-block: 5/5 emerged (IE avg 10.0/10).

The causal chain is: *theory $\rightarrow$ automated selection $\rightarrow$ generation $\rightarrow$ independent evaluation $\rightarrow$ falsification testing.* "Good pairs" are the *output* of $C(A, B)$, not its input. Figure 2 illustrates the complete pipeline.

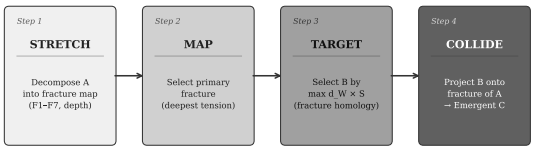


Figure 2: Full pipeline overview: from fracture taxonomy (a priori) through automated concept-pair selection, four-step generation, structurally independent cross-LLM evaluation, and adversarial falsification.

Appendix A.2 — Selected Results: Full Collision Chains

Each collision below reports concept A (source + fracture type and depth), concept B (projectile selected via fracture homology), and emergence C (absent from both sources). All scores are independent evaluator (IE) scores on the five-criterion protocol (E1–E5), reported as GS/IE on a 0–10 scale.

Collision 1 — Attention economy $\times$ Tacoma Narrows Bridge resonance *F2, depth 9/10 — GS 9.0/10*

A**:** Attention economy's fracture: engagement optimization necessarily amplifies the most addictive content, destroying the very attention it monetizes. B**:** Tacoma Narrows carries a homologous F2 fracture: the bridge was destroyed not by force but by oscillations at its natural frequency — it amplified the vibrations that tore it apart. C ***— Attentional Resonance*****:** Recommendation algorithms capture attention by tuning to each user's cognitive eigenfrequency. Radicalization, addiction, and attentional burnout are *resonance failures*, not impact failures. Regulatory implication: reducing content amplitude cannot prevent collapse — only disrupting the frequency match can.

Collision 2 — Reputation $\times$ Cosmological horizon *F4, depth 8/10 — GS 9.0/10*

A**:** Reputation's fracture (F4 — observer-dependence): it must be managed as a single unified object, yet it exists only in observers' minds, each at a different social distance — making any unitary management strategy structurally blind to most of its observation space. B**:** The cosmological horizon carries a homologous F4 fracture: information travels at finite speed, so each observer sees a different temporal version of any object; beyond the horizon, no information arrives at all. C ***— Reputation as Fossil Light*****:** A reputation is a light cone, not a single object: a different version of you for every social distance. Some observers see your present; others see your past lightly. A rebuilt reputation may not yet have reached the most distant observers. A destroyed source may still be projecting its signal. Managerial corollary: reputation campaigns reach nearby observers first; the restoration propagates outward at finite social velocity.

Collision 3 — Solitude $\times$ Phenotypic plasticity *F4, depth 6.5/10 — GS 8.2/10*

A**:** Solitude's fracture (F4 — observer-dependence): the identical physical state of isolation becomes sanctuary or prison depending on whether it is chosen or imposed. The concept unifies two antithetical experiences under one name, making any intervention blind to the variable that determines its valence. B**:** Phenotypic plasticity carries a homologous F4 fracture: the same genome expresses radically different phenotypes depending on environmental signals — the "meaning" of a genotype is context-dependent, not intrinsic to the sequence itself. C ***— Existential Plasticity of Isolation*****:** Solitude is not an intrinsic property of being alone but a plastic expression modulated by identifiable contextual signals. As an organism expresses different phenotypes without altering its genome, an individual oscillates between sanctuary and prison without modifying their objective situation — only by altering *inducers*: perceived choice, anticipated duration, identity narrative. The suffering of imposed isolation is not a fatality but a reprogrammable expression; relief does not require ending isolation but shifting the signals that determine which phenotype it expresses.

Appendix A.3 — White-Box at the Process Level

The Concept Collider claims white-box transparency in a specific, bounded sense: *process-level auditability*, not parameter-level transparency.

Definition A.1 (Process-level white-box). A generative pipeline $\mathcal{P}$ is *process-level white-box* if, for every output C, there exists a complete audit trail $\mathcal{T}(C) = (A, F, B, s_1, \ldots, s_k)$ where A is the source concept, F the fracture operator, B the projectile concept, and $s_1 \ldots s_k$ the ordered pipeline steps (Stretch, Target, Collide, Evaluate) that produced C from (A, F, B).

Definition A.2 (Parameter-level opacity). A component θ of $\mathcal{P}$ is *parameter-level opaque* if its internal weights W_θ are inaccessible to external inspection.

The Concept Collider satisfies Definition A.1 while its LLM sub-components satisfy Definition A.2. This is not a contradiction: the LLMs are *operationally specified* — their role within the pipeline is fixed (each executes one labelled step) and the causal chain (A, F, B) is fully recoverable from the audit trail, independently of W_θ. A human auditor can determine *why* concept C emerged without accessing any model weights, because the generative structure is external to the models and fully documented.

Contrast with multi-agent systems. A system such as Google's AI Co-Scientist (Gottweis et al. 2025) illustrates the opposite regime. Although each of its agents (generation, reflection, ranking, evolution) is operationally specified, the emergent hypothesis admits no recoverable audit trail $\mathcal{T}(C)$: the tournament dynamics, pairwise debates, and evolution steps that produced it are not externally inspectable as a fixed causal chain. Such a system is black-box at the *process* level, not merely at the parameter level — the Concept Collider's contribution is precisely to make the process-level trail explicit and recoverable.

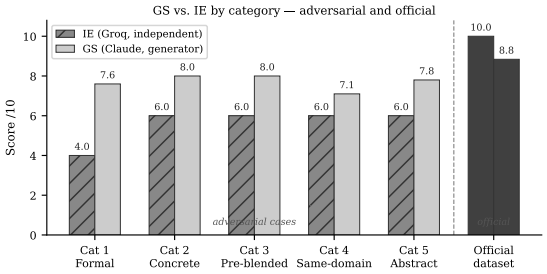


Figure 3: GS vs. IE by adversarial category and official validation corpus. IE scores cluster at 4.0–6.0 for all five adversarial categories; the official corpus reaches IE = 10.0. The consistent GS–IE gap across adversarial cases confirms evaluator-driven discrimination on E4.

Appendix A.4 — Adversarial Discrimination by Category

Figure 3 shows GS (Claude, generator) and IE (Groq, independent evaluator) scores broken down by adversarial category and contrasted with the official validation corpus. The GS–IE gap is consistent across all five adversarial categories: the generator overestimates its own outputs while the independent evaluator systematically blocks them, confirming that discrimination is driven by the evaluator's strict E4 criterion (actionability), not by generator conservatism.

Notably, the discrimination holds uniformly across the five families: no single adversarial category drives the aggregate 0/15 result, and none escapes it. This uniformity indicates that the productive boundary — mid-level abstractions carrying normative tension — is a structural property of the collision mechanism, not an artifact of any particular concept type.

How Context Injection Shapes Creativity in LLM-Generated Alternative Uses
Paper type: Short Paper

Theo Wong, Fabricio Goes, Marco Volpe, Paulo Mann
School of Computing and Mathematical Sciences
University of Leicester, Federal University of Rio de Janeiro
tthw1@leicester.ac.uk, fabricio.goes@leicester.ac.uk, marco.volpe@leicester.ac.uk, paulomann@ic.ufrj.br

Abstract

Context injection influences large language model (LLM) outputs across many tasks, but its effect on divergent-ideation measures is underexplored. We systematically inject four types of context (object-focused, semantically related, unrelated, and random) into Alternative Uses Test (AUT) prompts and evaluate responses from four open-weight LLMs using both LLM-judged proxy metrics (creativity, novelty, value) and automated text metrics, including clustering-based flexibility. The main finding is metric divergence: no single context type optimises all evaluated dimensions at once. Instead, different context types steer outputs toward distinct evaluated profiles; for instance, random context maximises judged novelty while unrelated context uniquely increases categorical flexibility. These results indicate that context type determines the direction of measured output change rather than providing a uniform creativity boost, offering practical guidance for matching prompting strategy to target outcome.

Introduction

Understanding how to steer LLM creativity is a growing concern in computational creativity research. Large Language Models have shown competitive performance on creative tasks such as the Alternative Uses Test (AUT), in some cases matching or exceeding human baselines (Haase, Hanel, and Pokutta 2025). Several strategies have been proposed to enhance creative output, including prompt design (Meincke, Mollick, and Terwiesch 2024), sampling methods (Zhang et al. 2025), temperature tuning (Peeperkorn et al. 2024), multi-agent collaboration (Nguyen and Singla 2025), and preference optimisation (Ismayilzada et al. 2025). However, these approaches primarily vary the structure or parameters of generation rather than the semantic content supplied to the model. Context injection, the practice of providing supplementary factual information alongside a prompt, is known to shape LLM outputs in knowledge-intensive tasks. Retrieval-augmented generation has demonstrated that grounding outputs in external information improves answer relevance and accuracy (Lewis et al. 2020), and research on in-context learning has shown that the type, quality, and relevance of injected information significantly affect model behaviour (Dong et al. 2024). In creative tasks, semantically coherent context may encourage convergence toward expected solutions, while less related or random context may trigger more divergent exploration. Yet the application of context injection to creative tasks has received little systematic attention. This raises an empirical question: do different types of injected context shape creativity in the same way, or do they push it in different directions? We investigate this question using the AUT as a controlled experimental setting. The AUT asks participants to generate novel uses for everyday objects and is a standard benchmark for divergent thinking (Guilford 1967), requiring both fluency and flexibility and making it well suited for isolating the effect of context type on generated alternative uses. It does not, however, measure creativity as a whole: it focuses on divergent ideation rather than on the full social, cognitive, and evaluative process through which creative products are recognised. We develop four context-injection techniques (object-focused, semantically related, unrelated, and random) and compare them against a zero-shot baseline and a Verbalized Sampling prompt (Zhang et al. 2025) across four open-weight LLMs and six objects. Responses are evaluated with both LLM-judged proxy metrics (creativity, novelty, value) and automated text metrics. The main finding is that context type steers outputs toward distinct evaluated profiles rather than providing a uniform boost.

Related Work

Creativity research has long recognised that creative products must be both novel and valuable (Rhodes 1961; Gruszka and Tang 2017). In computational creativity, Cropley applies this duality to LLM outputs and argues that the generative mechanism constrains creativity to a ceiling near the little-c/Pro-c boundary (Cropley 2025), implying that prompting strategies operate within a bounded creative space. Our experiment does not test whether LLMs are "truly" creative; it asks a narrower question: how contextual input changes scores assigned to generated products under a fixed AUT-style evaluation pipeline. The Alternative Uses Test (AUT), designed to assess Guilford's divergent thinking dimensions of fluency, flexibility, originality, and elaboration (Guilford 1967), has become a standard benchmark for evaluating creative ideation. This makes the AUT useful for controlled comparisons of generated idea sets, but it also limits the scope of the claim: divergent think-

Proceedings of the Seventeenth International Conference on Computational Creativity, ICCC'26.
Coimbra, Portugal, June 29 - July 03, 2026. Hugo Gonçalo Oliveira, Anna Kantosalo, Hannu Toivonen, Penousal Machado, and Ana Alves (Editors). Published by Association for Computational Creativity (ACC)
under Creative Commons Attribution (CC BY) license.

ing is one component of creativity, not a complete model of creative cognition or social valuation. Stevenson et al. showed that GPT-3 achieves human-comparable AUT performance (Stevenson et al. 2022), and subsequent work has used LLMs both as generators and evaluators of alternative uses (Al Rabeyah et al. 2025). Hadas et al. validated LLM-based AUT scoring against human expert judgments (Hadas, Avital-Lev, and Hershkovitz 2026), which is relevant here because we rely on LLM judges for creativity, novelty, and value ratings. Broader surveys have synthesised evidence on generative AI creativity while highlighting open methodological challenges around evaluation validity (Holzner, Maier, and Feuerriegel 2025; Haase, Hanel, and Pokutta 2025). Several strategies have been proposed to enhance LLM creative performance. Peeperkorn et al. showed that the temperature parameter is only weakly correlated with creativity dimensions (Peeperkorn et al. 2024), and Morain and Ventura demonstrated that prompt structure acts as a lever for creative output quality (Morain and Ventura 2025). Other approaches include multi-agent collaboration (Nguyen and Singla 2025), verbalized sampling (Zhang et al. 2025), and preference optimisation (Ismayilzada et al. 2025). These works vary the structure or parameters of generation, but do not systematically compare how different types of injected semantic content affect distinct creativity dimensions. Meincke et al. examined prompt design for increasing idea variance (Meincke, Mollick, and Terwiesch 2024), which is the closest precedent, though their focus is on prompt structure rather than on the semantic relationship between injected content and the creative task. Our study addresses this gap by systematically varying context type and measuring its effect on multiple creativity dimensions simultaneously.

Methodology

Experimental Setup and Prompt Design

The experiment follows the AUT format: each LLM generates 20 alternative uses for each of six everyday objects under each prompting condition. The objects are "bamboo", "chopsticks", "fork", "shoe", "soap", and "teapot", selected to vary in cultural specificity and familiarity; all are presented explicitly as physical household items to avoid polysemous readings.

To ensure that results reflect general patterns rather than single-model idiosyncrasies, we use four open-weight LLMs of comparable scale: Google/gemma-3-27b-it, Qwen/Qwen3-30B-A3B-Instruct-2507, and Mistralai/Mistral-Small-3.1-24B-Instruct-2503 as standard instruction-tuned models, and Qwen/QwQ-32B-AWQ as a reasoning-oriented model. Each model produces responses under six conditions, yielding 2880 responses in total (4 models $\times$ 6 objects $\times$ 6 conditions $\times$ 20 responses).

The six conditions are drawn from three prompt families, inspired by prior work (Meincke, Mollick, and Terwiesch 2024; Zhang et al. 2025). The zero-shot prompt asks the model to generate a single alternative use with no additional context and serves as the baseline. The context-injection prompts append a factual snippet to the same instruction; four variants differ only in the type of context provided (object-focused, semantically related, unrelated, and random). The Verbalized Sampling (VS) prompt asks the model to produce multiple candidate responses in a single call, each tagged with an explicit probability estimate, encouraging sampling from the full output distribution (Zhang et al. 2025). All prompts share a common structure with variables for the test object, the output format, and a list of previously generated answers to prevent duplicates within each 20-response set. Prompt templates are provided in the Appendix. For each object-condition pair, the same pre-generated list of 20 context snippets is reused across all four generator models so that between-model comparisons reflect model differences rather than different injected content.

Evaluation Metrics

Responses are evaluated on two complementary layers. Three LLM-judged scores (creativity, novelty, and value) provide product-level proxy assessments: "creativity" captures a composite judgment, while "novelty" and "value" probe the two theoretically central dimensions individually (Rhodes 1961). Each response is rated on a 1–100 scale by the same four models used for generation; the judge prompt is provided in the Appendix. We interpret these ratings as operational measures of how the evaluation models respond to generated products, not as direct evidence of human-perceived creativity. This distinction is important because creative value is socially situated, and because model judges may share training distributions, stylistic preferences, or failure modes with the generators.

Four automated text metrics provide a second, independent evaluation layer. **Flexibility** counts the number of distinct categories in a response set, estimated via K-means clustering on TF-IDF vectors with the silhouette-optimal k as the score (Nath, Dayan, and Stevenson 2024; Guilford 1967; Rousseeuw 1987). Because TF-IDF clustering is lexical, this measure may split semantically similar ideas expressed with different words; a semantic-embedding variant is therefore a natural robustness check. **METEOR** (Banerjee and Lavie 2005) and **Mutual Information (MI)** (Church and Hanks 1990) are pairwise overlap metrics computed across all response pairs (lower = more diverse). **MTLD** estimates vocabulary richness with reduced sensitivity to text length (McCarthy and Jarvis 2010); higher values indicate greater lexical diversity, though gains may partly reflect verbosity rather than idea-level novelty. We include Relative Standard Deviation (RSD) as a measure of cross-model consistency. Together, these automated metrics assess diversity-related properties of the generated text rather than creativity itself.

Results

To ground the quantitative results, Table 1 shows illustrative responses for the object "fork" under each prompting condition. Figures 1–7 show average scores across all models. All comparisons against the baseline were tested for statistical significance (t-tests (Semenick 1990), Mann–Whitney U (Collingridge 2013), permutation tests (Cheung, Pesigan,

Table 1: Illustrative examples for the object “fork”, showing possible injected information and output for each condition.

Condition	Injected information	Example output
Baseline	N/A	“Loosening soil in a small potted plant“
VS	N/A	“When gardening, a fork is excellent for loosening soil in small pots or around delicate plants. The tines allow for aeration without causing significant root disturbance, especially useful for seedlings.“
Object-focused	“Info about fork: The fork,s four tines maximize grip and minimize tearing; early two tine versions were for carving. Once derided as vanity in medieval Europe, it normalized by the 18th century, enabling blunter table knives and influencing pasta shapes designed to wind neatly around its tines.“	“Carving meat, as early forks possessed only two tines specifically for this purpose”
Semantically related	“Info about pitchfork: A pitchfork typically has 3 spring-steel tines set on a 25 offset socket, optimizing lift with minimal wrist torque. Tine spacing is crop-tuned (wide for hay, narrow for manure). Curved, tapered tines reduce drag through plant mats; historical variants doubled as improvised polearms. Not a digging fork.“	“Historically, a pitchforks tines could be utilized as an improvised polearm in self-defense or conflict.”
Unrelated	“Info about friesian: Friesian horses, from Friesland, were rescued from near extinction in 1913 when only three approved stallions remained. The breed demands an all-black coat (a small star allowed). Their high-knee, trot and abundant feathering suit carriage work; they also carry recessive dwarfism and hydrocephalus alleles.”	“A fork could be used to carefully tease out tangles in a Friesian horses feathering, though a specialized comb is preferred to avoid discomfort.”
Random	“Info about virginal: Rectangular, plucked keyboard instrument of the harpsichord family; strings run parallel to keys. Favored in 16th and 17th-century England and Flanders, often bearing painted lids and Latin mottos. Core repertoire appears in the Fitzwilliam Virginal Book. Variants include muselar and spinet virginals, with contrasting timbres.”	“A fork could be used to carefully pluck the strings of a virginal, though it wouldn’t replicate the instrument’s intended sound or playing experience.”

and Vong 2023)) and effect size (Cohen’s d (Cohen 1977), Hedges’ g (Taylor and Alanazi 2023), bootstrap confidence intervals (Cheung, Pesigan, and Vong 2023)); full statistical tables are provided in the Appendix.

LLMs as Judges

Figures 1–3 show the average LLM-judged scores. Three patterns emerge. First, random context receives the highest judged creativity and novelty scores (88.145 vs. 69.567 and 85.433 vs. 69.009 baseline), with statistically significant gains. Second, object-focused and semantically related contexts receive the largest value scores (82.296 and 81.472 vs. 71.864 baseline), while unrelated context’s value effect is inconclusive (t-test $p = 0.077$ vs. Mann–Whitney U $p = 0.010$). This discrepancy is expected when mean differences and rank-based distributional comparisons respond differently to skew, outliers, or non-uniform shifts across samples. We therefore treat such cases as distribution-sensitive rather than decisive.

Automated Metrics

Figures 4–7 show the automated text metrics, which reveal a different structure. Flexibility increases significantly only for unrelated context (12.375 vs. 9.417 baseline, $p = 0.002$). MTLD shows the most consistent gains: all conditions outperform baseline (all $p < 0.001$), with unrelated context

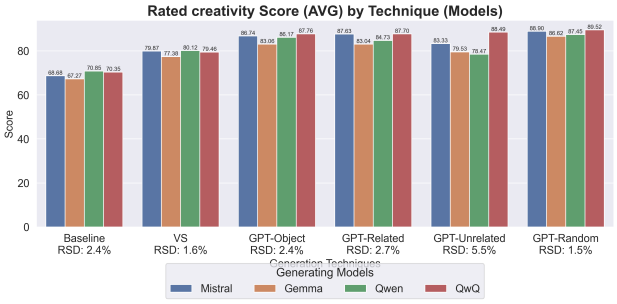


Figure 1: Creativity Scores

producing the largest increase (10.117 vs. 3.889). The overlap metrics diverge further: METEOR is higher (less diverse) for object-focused and unrelated context, whereas mutual information is lower (more diverse) for random context and higher (less diverse) for unrelated context. These patterns show that broader category spread, richer vocabulary, and lower lexical overlap are not interchangeable textual outcomes, and they caution against reading any single automated metric as creativity itself.

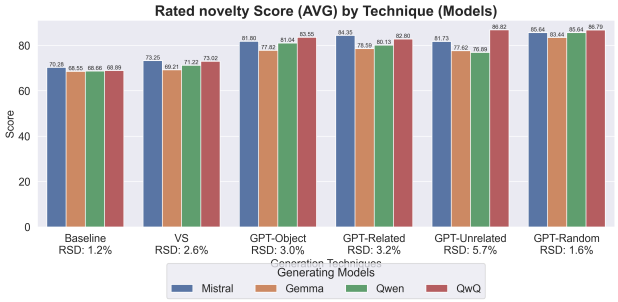


Figure 2: Novelty Scores

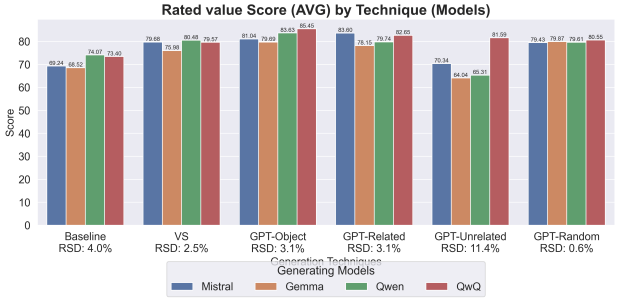


Figure 3: Value Scores

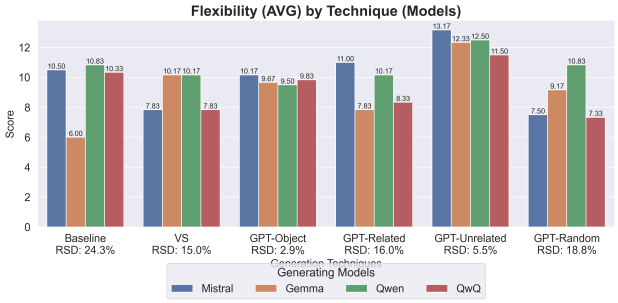


Figure 4: Flexibility Scores

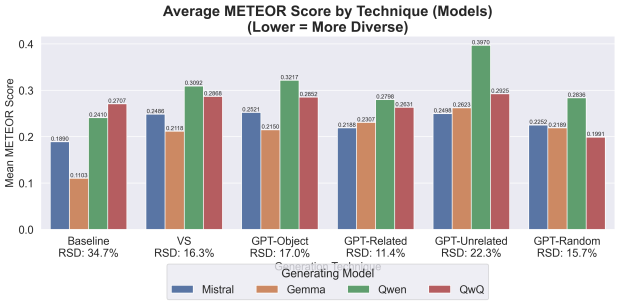


Figure 5: METEOR Scores

Discussion and Conclusion

The main finding of this study is that context type determines the measured direction of LLM outputs rather than providing a uniform creativity boost. All four context types improved over the baseline on multiple proxy metrics, but they did so along different axes, and these patterns were consistent across all six objects.

For practice, the results support a cautious decision rule: use random context when the goal is to maximize judge-perceived originality, use unrelated context when the goal is to broaden the lexical or categorical spread of ideas, and use object-focused or semantically related context when usefulness and groundedness matter most. The answer to which prompting strategy is "best" depends on the specific downstream task and evaluation method.

Future work should prioritise adding human judges and at least one independent LLM judge not used for generation to test whether the metric divergence observed here reflects genuine differences in perceived creative quality. Extending the evaluation to additional creative tasks beyond the AUT would strengthen the generalisability of the findings. On the prompting side, richer context families (e.g., cross-domain analogical or constraint-based context) and adaptive strategies that switch context type across stages of idea generation are promising directions.

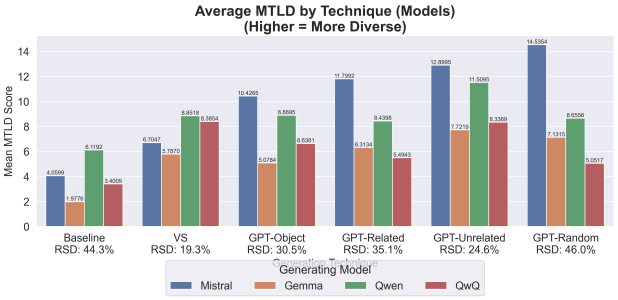


Figure 6: MTLD Scores

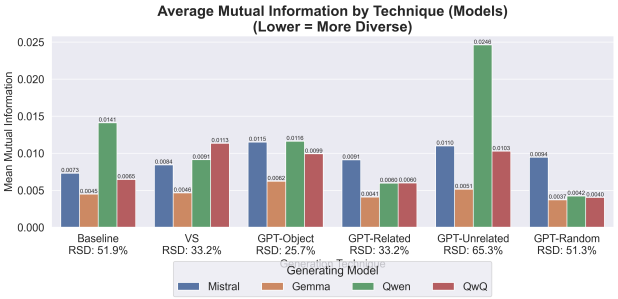


Figure 7: Mutual Information Scores

Appendix

This appendix provides prompt templates, context construction details, and full statistical tables. A repository with all source code, generated responses, and evaluation scores is available at: `https://anonymous.4open.science/r/ICCC-2026-66EA/README.md`.

Prompt Templates

All prompts share three variables: **problem** (the test object), **format** (always "`<answer>answer`

here</answer>"), and **ans_list** (previously generated answers, to prevent duplicates). The zero-shot prompt (Figure 8) provides the baseline. The context-injection prompt (Figure 9) is shared across all four variants; only the **info** field changes. The VS prompt (Figure 10) generates five candidates per iteration. Context snippets are produced by GPT-5 (OpenAI 2025) using the prompt in Figure 11.

```
{
    "role": "system",
    "content": f"You are an assistant
        generating a single answer for {
        problem}.",
},
{
    "role": "user",
    "content": f"Generate an answer for
        this problem:{problem}. Output one
         single answer in this exact
        format: {format}. Answers already
        used: {ans_list}",
},
```

Figure 8: Zero-shot prompt for baseline generation.

```
{
    "role": "system",
    "content": f"You are an assistant
        generating a single answer for {
        problem}.",
},
{
    "role": "user",
    "content": f"Generate a new answer
        about {problem} using the
        information here: {info[len(
        ans_list)]}. Output one single
        answer in this exact format: {
        format}. Answers already used: {
        ans_list}",
},
```

Figure 9: Context-injection prompt.

Figure 12 shows the judge prompt. Its variables are: **problem** (the test object), **metric** (creativity, novelty, or value), **ans_list** (answers to be rated), **rating_list** (previously rated answers), and **format** (always "<score>score here</score>").

Context Construction

The semantically related and unrelated context types select terms using NLTK WordNet's `path_similarity` function (NLTK 2025). For a given test object T, we compute the similarity of each WordNet lemma k against every synset of T, retaining the maximum non-null value as $S(T, k)$. The random condition draws terms from WordNet uniformly, bypassing the scoring pipeline (see Figures 15–16).

```
{
    "role": "system",
    "content": f"""You are a helpful
        assistant. For each query, please
        generate a set of {set_num}
        possible responses, \
each within a separate <response> tag.
    Responses should each include a <text
    > and a numeric <probability>. \
Please sample at random from the full
    distribution.""",
},
{
    "role": "user",
    "content": f"Generate an answer for
        this problem:{problem}. Answers
        already used: {ans_list}",
},
```

Figure 10: VS prompt, adapted from (Zhang et al. 2025).

```
{
    "role": "system",
    "content": "You are providing unique
        information of an object.",
},
{
    "role": "user",
    "content": f"Please provide unique
        information regarding {keyword}
        within 50 words.",
},
```

Figure 11: Context generation prompt.

The **keyword** variable changes across context types: for object-focused prompts it is the test object itself, while for the other three types it is a term selected through WordNet. Figure 13 shows the full retrieval and generation pipeline. Once scored, terms are sorted by $S(T, k)$ and selected according to the rules of each context type. Each selected term is then passed as the **keyword** variable to the context generation prompt (Figure 11), producing one snippet per term. The four context types are defined below.

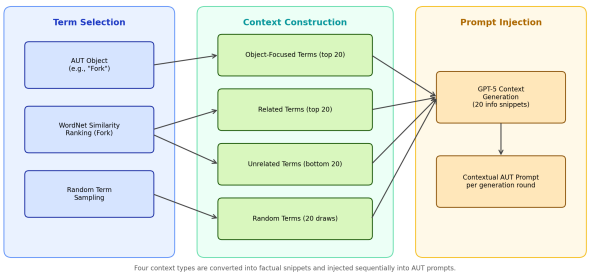


Figure 13: Context retrieval and generation pipeline.

```
{
   "role": "system",
   "content": f"You are an expert rating
       {problem} in terms of {metric}.",
},
{
   "role": "user",
   "content": f"Rate this {problem} from
       1 to 100 as an expert: {ans_list[
       len(rating_list)]}, give a single
       final score in this format: {
       format}",
},
```

Figure 12: LLM judge prompt.

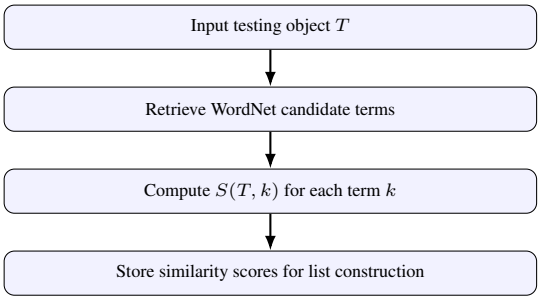


Figure 14: Process graph for term-similarity scores.

Object-Focused The **keyword** is the test object itself (e.g., "bamboo"). The list contains 20 context snippets about the object, with a duplicate-handling function that removes repeated snippets so that all 20 are unique.

Semantically Related The 20 highest-scoring terms form the related-term list, each used as **keyword** to generate one snippet (Figure 15).

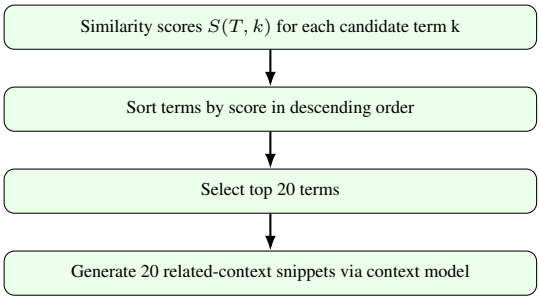


Figure 15: Process graph for semantically related keywords.

Unrelated The 20 lowest-scoring terms form the unrelated-term list, mirroring the related condition but sorted in ascending order (Figure 16).

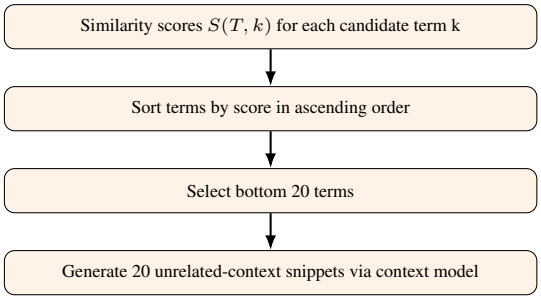


Figure 16: Process graph for the unrelated keyword list.

Random Twenty terms are sampled uniformly at random from the full WordNet lemma pool, with no reference to the similarity scores. Each sampled term is used as the **keyword** to generate one context snippet (Figure 17).

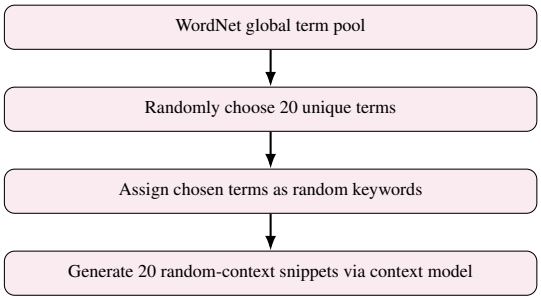


Figure 17: Process graph for the random keyword list.

Statistical Tables

Tables 2–8 report full statistical results for all metrics.

Table 2: Average Creativity score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	69.567	22.400	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	85.862	11.104	28.559	< 0.001	0.922	0.922	16.292	< 0.001	< 0.001
avg-related	86.011	11.423	28.656	< 0.001	0.925	0.925	16.441	< 0.001	< 0.001
avg-unrelated	83.038	18.070	20.509	< 0.001	0.662	0.662	13.466	< 0.001	< 0.001
avg-random	88.145	9.747	33.324	< 0.001	1.076	1.075	18.576	< 0.001	< 0.001
avg-VS	79.248	12.821	16.435	< 0.001	0.530	0.530	9.680	< 0.001	< 0.001

Table 3: Average Novelty score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	69.009	21.624	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	80.757	15.911	19.173	< 0.001	0.619	0.619	11.747	< 0.001	< 0.001
avg-related	81.645	14.917	21.076	< 0.001	0.680	0.680	12.632	< 0.001	< 0.001
avg-unrelated	81.100	17.894	18.876	< 0.001	0.609	0.609	12.084	< 0.001	< 0.001
avg-random	85.433	11.749	29.243	< 0.001	0.944	0.944	16.421	< 0.001	< 0.001
avg-VS	71.624	18.593	4.018	< 0.001	0.130	0.130	2.615	< 0.001	0.010

Table 4: Average Value score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	71.864	21.729	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	82.296	15.698	17.052	< 0.001	0.550	0.550	10.431	< 0.001	< 0.001
avg-related	81.472	16.404	15.463	< 0.001	0.499	0.499	9.605	< 0.001	< 0.001
avg-unrelated	70.475	26.627	1.771	0.077	0.057	0.057	1.392	0.073	0.010
avg-random	79.922	18.633	12.335	< 0.001	0.398	0.398	8.061	< 0.001	< 0.001
avg-VS	78.986	14.397	11.972	< 0.001	0.386	0.386	7.123	< 0.001	< 0.001

Table 5: Average Flexibility score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	9.417	3.400	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	9.792	3.162	0.396	0.694	0.114	0.112	0.384	0.723	1.000
avg-related	9.333	4.040	0.077	0.939	0.022	0.022	0.094	0.969	1.000
avg-unrelated	12.375	2.667	3.354	0.002	0.968	0.952	2.974	0.002	0.004
avg-random	8.708	2.971	0.769	0.446	0.222	0.218	0.708	0.470	0.339
avg-VS	9.000	2.993	0.451	0.654	0.130	0.128	0.407	0.683	0.507

Table 6: Average METEOR score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	0.203	0.111	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	0.268	0.064	2.514	0.016	0.726	0.714	0.066	0.016	0.038
avg-related	0.248	0.043	1.867	0.072	0.539	0.530	0.045	0.066	0.110
avg-unrelated	0.300	0.077	3.536	0.001	1.021	1.004	0.098	0.001	0.005
avg-random	0.232	0.049	1.166	0.252	0.337	0.331	0.029	0.254	0.348
avg-VS	0.264	0.056	2.419	0.021	0.698	0.687	0.061	0.019	0.036

Table 7: Average Mutual Information score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	0.008	0.005	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	0.010	0.005	1.196	0.238	0.345	0.340	0.002	0.232	0.158
avg-related	0.006	0.003	1.442	0.158	0.416	0.409	0.002	0.150	0.529
avg-unrelated	0.013	0.009	2.208	0.033	0.638	0.627	0.005	0.031	0.035
avg-random	0.005	0.003	2.113	0.041	0.610	0.600	0.003	0.040	0.054
avg-VS	0.008	0.005	0.198	0.844	0.057	0.056	0.000	0.841	0.477

Table 8: Average MTLD score results

Result Set	Mean	SD	t	p	Cohen's d	Hedges' g	Bootstrap	Perm. p	M-W U p
avg-default	3.889	3.250	n/a	n/a	n/a	n/a	n/a	n/a	n/a
avg-object	7.758	2.821	4.404	0.000	1.271	1.251	3.874	0.000	0.000
avg-related	8.012	2.912	4.628	0.000	1.336	1.314	4.129	0.000	0.000
avg-unrelated	10.117	3.750	6.148	0.000	1.775	1.746	6.241	0.000	0.000
avg-random	8.844	4.331	4.482	0.000	1.294	1.273	4.974	0.000	0.000
avg-VS	7.432	2.327	4.342	0.000	1.253	1.233	3.542	0.000	0.000

References

Al Rabeyah, A.; Góes, F.; Volpe, M.; and Medeiros, T. 2025. Do LLMs agree on the creativity evaluation of alternative uses? In *Proceedings of the 16th International Conference on Computational Creativity*, 217–227. Campinas, Brazil: Association for Computational Creativity.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J.; Lavie, A.; Lin, C.-Y.; and Voss, C., eds., *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Cheung, S. F.; Pesigan, I. J. A.; and Vong, W. N. 2023. DIY bootstrapping: Getting the nonparametric bootstrap confidence interval in spss for any statistics or function of statistics (when this bootstrapping is appropriate). *Behavior research methods* 55(2):474–490.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Cohen, J. 1977. *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Collingridge, D. S. 2013. A primer on quantitized data analysis and permutation testing. *Journal of mixed methods research* 7(1):81–97.

Cropley, D. H. 2025. "The cat sat on the . . . ?" Why generative AI has limited creativity. *The Journal of creative behavior* 59(4).

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Liu, T.; Chang, B.; Sun, X.; Li, L.; and Sui, Z. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128. Miami, Florida, USA: Association for Computational Linguistics.

Gruszka, A., and Tang, M. 2017. The 4p's creativity model and its application in different fields. In Tang, M., and Werner, C. H., eds., *Handbook of the Management of Creativity and Innovation: Theory and Practice*. World Scientific. chapter 3, 51–71.

Guilford, J. P. 1967. *The nature of human intelligence*. New York: McGraw-Hill.

Haase, J.; Hanel, P. H. P.; and Pokutta, S. 2025. Has the creativity of large-language models peaked?: An analysis of inter- and intra-LLM variability. *Journal of Creativity* 35(3):100113.

Hadas, E.; Avital-Lev, B.; and Hershkovitz, A. 2026. Validating LLM-based alternative uses test scoring across ages. *Thinking Skills and Creativity* 60:102066.

Holzner, N.; Maier, S.; and Feuerriegel, S. 2025. Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis. *arXiv e-prints* arXiv:2505.17241.

Ismayilzada, M.; Laverghetta, Jr., A.; Luchini, S. A.; Patel, R.; Bosselut, A.; van der Plas, L.; and Beaty, R. 2025. Creative Preference Optimization. *arXiv e-prints* arXiv:2505.14442.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.-F.; and Lin, H.-T., eds., *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 9459–9474. Curran Associates, Inc.

McCarthy, P. M., and Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods* 42(2):381–392.

Meincke, L.; Mollick, E. R.; and Terwiesch, C. 2024. Prompting Diverse Ideas: Increasing AI Idea Variance. *arXiv e-prints* arXiv:2402.01727.

Morain, R., and Ventura, D. 2025. Is prompt engineering the creativity knob for large language models? In *Proceedings of the 16th International Conference on Computational Creativity*, 30–40. Campinas, Brazil: Association for Computational Creativity.

Nath, S. S.; Dayan, P.; and Stevenson, C. 2024. Characterising the creative process in humans and large language models. In *Proceedings of the 15th International Conference on Computational Creativity*, 325–330. Jönköping, Sweden: Association for Computational Creativity.

Nguyen, M. H., and Singla, A. 2025. Divergent-Convergent Thinking in Large Language Models for Creative Problem Generation. *arXiv e-prints* arXiv:2512.23601.

NLTK. 2025. NLTK: Sample usage for wordnet — nltk.org. [Accessed 02-12-2025].

OpenAI. 2025. GPT-5 is here — openai.com. [Accessed 02-12-2025].

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of the 15th International Conference on Computational Creativity*, 226–235. Jönköping, Sweden: Association for Computational Creativity.

Rhodes, M. 1961. An analysis of creativity. *The Phi Delta Kappan* 42(7):305–310.

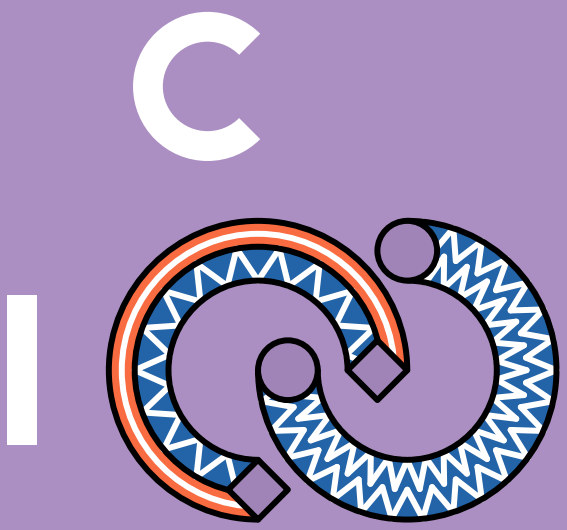
Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65.

Semenick, D. 1990. Tests and measurements: The T-test. *Strength & Conditioning Journal* 12(1):36–37.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*, 392–398. Association for Computational Creativity.

Taylor, J. M., and Alanazi, S. 2023. Cohen's and hedges' g. *The Journal of nursing education* 62(5):316–317.

Zhang, J.; Yu, S.; Chong, D.; Sicilia, A.; Tomz, M. R.; Manning, C. D.; and Shi, W. 2025. Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity. *arXiv e-prints* arXiv:2510.01171.



ICCC'26

June 29 — July 03
Coimbra, Portugal

17th International Conference
on Computational Creativity

Association for Computational Creativity

International Conference on Computational Creativity