Proceedings of the 16th International Conference on Computational Creativity

June 23 — June 27, 2025 • Campinas, Brazil

Editors: Hugo Gonçalo Oliveira • Brad Spendlove • Pablo Gervás • Dan Ventura

Association for Computational Creativity
International Conference on Computational Creativity

Proceedings of the Sixteenth International Conference on Computational Creativity

ICCC 2025

Campinas, Brazil | 23 - 27 June

Hugo Gonçalo Oliveira, Brad Spendlove, Pablo Gervás, Dan Ventura (Editors)

Published by the Association for Computational Creativity (ACC)

State University of Campinas (Unicamp)
Brazil

http://computationalcreativity.net/iccc25/

First published 2025

TITLE: Proceedings of the 16th International Conference on Computational Creativity

EDITORS: Hugo Gonçalo Oliveira, Brad Spendlove, Pablo Gervás, Dan Ventura

ISBN: 978-989-54160-7-3

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, published under a Creative Commons Attribution (CC BY) license `https://creativecommons.org/licenses/by/4.0/`, which allows unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Attribution CC BY

Preface

Welcome to the proceedings of the *16th International Conference on Computational Creativity*. This year the conference was held from the 23rd to the 27th of June 2025 on the campus of Universidade Estadual de Campinas (Unicamp) in Campinas, Brazil. This was the first ICCC to be held anywhere in South America, and we are excited about bringing the conference to a brand new part of the world.

For the scientific program, ICCC'25 received 119 submissions: 61 to the full paper track and 58 to the short paper track. From the 61 full paper submissions, 24 were accepted as oral presentations (39%), and one was withdrawn. From the 58 short paper submissions, 13 were accepted for short oral presentation and 17 for poster presentation (52%), and one was withdrawn. The accepted papers represent work from an internationally diverse community of researchers and practitioners from 61 institutions and 19 countries.

As in past years, ICCC'25 recognised the outstanding work of researchers in the community through best paper awards in three categories:

- Best Full Paper Award: Gaurav Sahu and Olga Vechtomova. "Computational Modeling of Artistic Inspiration: A Framework for Predicting Aesthetic Preferences in Lyrical Lines Using Linguistic and Stylistic Features"
- Best Short Paper Award: Samuel Schapiro, Jonah Black and Lav Varshney. "Transformational Creativity in Science: A Graphical Theory"
- Best Poster Award: Daniel DeFreez, Jackson Harrower, Travis Wood and Cynthia Salbato. "The Wizard in the Town Plaza: Voice-Based Interactive Storytelling in Public Spaces"

In addition to the outstanding work presented by authors of accepted papers, the conference program included

- A keynote address by Jônatas Manzolli whose talk was entitled, "Interacting within an Intelligent Space: An Environment for Creativity";
- Four workshops: Information Theory and Computational Creativity; Computational Design and Computer-aided Creativity; Ars Combinatoria and Generative Meta-Textuality: From Historical Systems to AI Authorship; Multimodal Creative Improvisation in Mixed Reality;
- The social program: socializing as a community is one of the most important functions of every ICCC and this year was no exception. Social highlights included:
  - A welcome reception held at Empório do Nono
  - A musical performance at Casa do Lago
  - A conference dinner also held at Empório do Nono
  - A tour of the SIRIUS particle accelerator

We are grateful to all of those involved in the organisation of ICCC'25: The program committee, senior program committee and external reviewers for their timely and thoughtful reviews of conference submissions; the keynote speaker for his insightful and entertaining presentation; the local team at Unicamp for their great work throughout the conference; and the Association for Computational Creativity steering committee for their guidance and support.

In addition, we are very grateful for the financial and logistical support provided by our conference sponsors:

- the AI Journal for funding student conference registrations and prizes for the best papers;
- the Toyota Research Institute for funding the expenses for the keynote speaker
- Universidade Estadual de Campinas for providing meeting spaces, IT and audiovisual support and for supporting the organization of the social events.

Brad Spendlove
Hugo Gonçalo Oliveira
Dan Ventura

July 2025

Organizational Team

General chair
Pablo Gervás, Universidad Complutense de Madrid, Spain

Local chairs
Manuel Falleiros, State University of Campinas, Brazil
Artemis Moroni, Centro de Tecnologia da Informação Renato Archer, Brazil

Program chairs
Hugo Gonçalo Oliveira, University of Coimbra, Portugal
Dan Ventura, Brigham Young University, USA

Workshop chair
Piera Riccio, ELLIS Alicante, Spain

Publicity chair
José Pedro Lopes, University of Coimbra, Portugal

Publication chair
Brad Spendlove, Randolph College, USA

Webmaster
Marcio Lima Inácio, University of Coimbra, Portugal

Program Committee

Senior PC Members

Alison Pease - University of Dundee
Amílcar Cardoso - University of Coimbra
Andrés Gómez de Silva Garza - Instituto Tecnológico Autónomo de México
Carlos León - Complutense University of Madrid
Dan Brown - University of Waterloo
Geraint Wiggins - Vrije Universiteit Brussel / Queen Mary University of London
Hannu Toivonen - University of Helsinki
João Miguel Cunha - University of Coimbra
Kazjon Grace - The University of Sydney
Maya Ackerman - Santa Clara University
Maria Teresa Lanno - University of Sussex
Nada Lavrač - Jozef Stefan Institute
Oliver Bown - University of New South Wales
Pedro Martins - University of Coimbra
Penousal Machado - University of Coimbra
Rafael Pérez Y Pérez - Universidad Autónoma Metropolitana
Simon Colton - Queen Mary University of London
Tony Veale - University College Dublin

PC Members

Allison Parrish - New York University
Anna Jordanous - University of Kent
Antonios Liapis - University of Malta
Brad Spendlove - Randolph College
Caterina Moruzzi - University of Edinburgh
Colin Johnson - University of Nottingham
Diarmuid O'Donoghue - Maynooth University
Enric Plaza - Universitat Autònoma de Barcelona
Fabrício Góes - University of Leicester
Giovanni Pilato - Istituto di Calcolo e Reti ad Alte Prestazioni
Gonzalo Mendez - Universidad Complutense de Madrid
Ivan Guerrero - Universidad Nacional Autónoma de México
Janet Rafner - Aarhus University
Jer Hayes - Accenture
Jéssica Parente - University of Coimbra
João Correia - University of Coimbra
Jon McCormack - Monash University
Khalid Alnajjar - Rootroo
Lakshmi Nair - Georgia Institute of Technology
Liane Gabora - The University of British Columbia
Luís Espírito Santo - University of Coimbra / Vrije Universiteit Brussel
Manex Agirrezabal - University of Copenhagen
Maria M. Hedblom - Jönköping University
Matthew Guzdial - University of Alberta
Maximos Kaliakatsos-Papakostas - Aristotle University of Thessaloniki

Mihai Pomarlan - University of Bremen
Mika Hämäläinen - Metropolia University of Applied Sciences
Milena Fisher - The Creativity Post
Paul Bodily - Idaho State University
Piotr Mirowski - DeepMind
Rafael Peñaloza - University of Milano-Bicocca
Raquel Hervás - Universidad Complutense de Madrid
Roisin Loughran - Dundalk Institute of Technology
Santiago Negrete-Yankelevich - Universidad Autónoma Metropolitana
Senja Pollak - Jozef Stefan Institute
Sérgio Rebelo - University of Coimbra
Shlomo Dubnov - University of California in San Diego
Tapio Takala - Aalto University
Thomas Winters - KU Leuven
Tiago Martins - University of Coimbra
Tomi Männistö - University of Helsinki
Wendy Aguilar - Universidad Nacional Autónoma de México
Ying-Ping Chen - National Yang Ming Chiao Tung University

Table of Contents

Humor

A Full Pipeline for Context-Aware Pun Generation 1
Marcio Lima Inácio and Hugo Gonçalo Oliveira

Me, Myself and Irony: Modelling the deceptive creativity of irony with Large Language Models 12
Tony Veale

LLMs

Refining Metrical Constraints in LLM-Generated Poetry with Feedback 20
Manex Agirrezabal and Hugo Gonçalo Oliveira

Is Prompt Engineering the Creativity Knob for Large Language Models? 30
Robert Morain and Dan Ventura

Experimenting with Large Language Models for Poetic Scansion in Portuguese: A Case Study on Metric and Rhythmic Structuring 41
André Valença and Filipe Calegario

Co-creativity

LuminAI: Embodied AI as a Catalyst, Constraint, and Co-Creator in Dance Improvisation Class 50
Milka Trajkova, Andrea Knowlton, Jasmine Kaur and Brian Magerko

Measuring Creativity in Co-Writing with AI: Rhyme Density and the Limits of Computational Proxies 61
Ibukun Olatunji and Mark Sheppard

Beyond chat: towards greater user involvement and agency in human-AI co-creativity through shared collaborative spaces 66
Rodolfo Ocampo, Oliver Bown and Kazjon Grace

Reimagining Dance: Real-time Music Co-creation between Dancers and AI 76
Olga Vechtomova and Jeff Bos

Battle Rap as a Framework for Human-Machine Co-Creativity 81
Ibukun Olatunji, Mark Sheppard, Alma Rahat, Matt Jones and Amanda Rogers

When AI Says No: Investigating the Creative Power of Dissent 92
David Bouter and Rob Saunders

MOSAAIC: Managing Optimization towards Shared Autonomy, Authority, and Initiative in Co-creation 97
Alayt Issak, Jeba Rezwana and Casper Harteveld

Prompting AI in Co-Creation: The Role of Syntax and Sentiment in Shaping AI-Generated Content 108
Giancarlo Colloca and Jeba Rezwana

Ideation

Spark: A System for Scientifically Creative Idea Generation 114
Aishik Sanyal, Samuel Schapiro, Sumuk Shashidhar, Royce Moon, Lav Varshney and Dilek Hakkani-Tur

How far afield should you go when being creative? Semantic area as a metric of AI's effects on creative ideation 120
Steven Rick, Jennifer Heyman, Pablo Paredes, Matthew Hong and Thomas Malone

Computational Modeling of Artistic Inspiration: A Framework for Predicting Aesthetic Preferences in Poetic Lines Using Linguistic and Stylistic Features 129
Gaurav Sahu and Olga Vechtomova

Narrative

Adapting Proppian Morphology for Generating Narrative Structures 139
Pablo Gervás

Evaluating Creative Short Story Generation in Humans and Large Language Models 145
Mete Ismayilzada, Claire Stevenson and Lonneke van der Plas

Playing Along: Building AI Agents for Co-Creation of Improvised Stories 156
Idan Dov Vidra, Gal Kimron, Lior Noy and Ariel Shamir

Turning Linear Stories into Thrillers: The Impact of Story Reordering and Familiarity 167
Mei Si

New Frontiers

Abductive Computational Systems: Creative Abduction and Future Directions 172
Abhinav Sood, Kazjon Grace, Stephen Wan and Cecile Paris

Foundation Models as Agents of Austerity 178
Anton Koch and Guilherme Zanchetta

Integrating Computational Creativity and Climate Data for Environmental Awareness: Design and Analysis of an Ongoing Project 183
Felipe Mammoli, Guilherme Zanchetta, Laureana Fontolan and Artemis Moroni

Psychologically-inspired generative AI videos for supporting creativity 191
Shabnam Hakimi, Monica Van, Matthew Hong, Kalani Murakami, Pablo Paredes and Matthew Klenk

Evaluation

Automatic Aesthetic Evaluation in Generative Image Models 201
Larissa Gomide, Lucas Nascimento Ferreira and Wagner Meira Jr.

A Creativity Assessment Scale for Text-to-Image Prompting: Challenges & Observations 206
Sophia Lichtenberg and Almila Akdag

Do LLMs Agree on the Creativity Evaluation of Alternative Uses? 217
Abdullah Al Rabeyah, Fabricio Goes, Marco Volpe and Talles Medeiros

Language

Making the Familiar Unfamiliar: AI-Driven Synectic Metaphor Generation for Computational Creativity 228
Georgios Palamas

Making Mountains out of Molehills: Abstraction without Information Loss in Analogical Mapping 234
Tony Veale

A Development and Teaching Framework for Codenames 239
Robert Morain, Brad Spendlove and Dan Ventura

Poems to Lyrics: Automated Rephrasing with Beat Alignment 248
Mohamad Elzohbi and Richard Zhao

Graph-based Approaches

Transformational Creativity in Science: A Graphical Theory 258
Samuel Schapiro, Jonah Black and Lav Varshney

Do Conversational Interfaces Limit Creativity? Exploring Visual Graph Systems for Creative Writing 264
Abhinav Sood, Maria Teresa Llano and Jon McCormack

Invisible Strings: Revealing Latent Dancer-to-Dancer Interactions with Graph Neural Networks 273
Luis Vitor Zerkowski, Zixuan Wang, Ilya Vidrin and Mariel Pettee

Design

Co-Designing Fashion with AI: A Small-Data Approach to Generative Garment Design 283
Ina Hrešć and Terence Broad

Creative as IDEO experts: LLM-agent-based design thinking workshop 292
Beibei Liu, Shuide Wen and Peiyao Cheng

Posters

Are AI-generated Jokes Truly Original? Charting the "Joke Space" 302
Ori Amir

Automatic Narrative Knowledge Base Generation 309
Robert Morain, Rafael Pérez Y Pérez and Dan Ventura

Automatically Detecting Amusing Games in Wordle 315
Ronaldo Luo, Gary Liang, Cindy Liu, Adam Kabbara, Minahil Bakhtawar, Kina Kim and Michael Guerzhoy

Between codes and dreams: hallucinatory cut-ups as a poetic in creation with AI 319
Fernando Fogliano, Dimitri Lomonaco, Gisele Aquino and Letícia Nogueira Costa

Brave: Engineering an Embedded Network-Bending Instrument, Manifesting Output Diversity in Neural Audio Systems 323
Daniel Manz and Mick Grierson

Controlling the image generation process with parametric activation functions . 329
Ilia Pavlov

Creativity Rivalry: Human, Artificial Intelligence, and Co-Design 335
Hoa Vo

Embedding visual thinking into an AI-driven furniture design critiquing system 340
Vuong Nguyen and Gabriel Vigliensoni

Fairness as a Creative Resource: Challenges and Opportunities in Creative Computing 346
Ricardo Rabonato and Lilian Berton

Introducing Pathomalgametry: Conceptual Blending with Geometric Path-Finding and Amalgamation 351
Mena Leemhuis and Oliver Kutz

Making the Familiar Strange: A Computational Approach to Defamiliarization in Creativity Support 357
Yu-Chieh Ho, Wenn-Chieh Tsai and Zhu Yuan Lai

Defining and Quantifying Creative Behavior in Popular Image Generators 362
Aditi Ramaswamy, Hana Chockler and Melane Navaratnarajah

The Wizard in the Town Plaza: Voice-Based Interactive Storytelling in Public Spaces 368
Daniel DeFreez, Jackson Harrower, Travis Wood and Cynthia Salbato

The truth is no diaper: Human and AI-generated associations to emotional words 374
Špela Vintar and Jan Jona Javoršek

The "What" Space. Prosodic Variability and Affective Virtual Environments ... 380
Jorge Forero Rodriguez, Gilberto Bernardes and Mónica Mendes

A Full Pipeline for Context-Aware Pun Generation

Marcio Lima Inácio and **Hugo Gonçalo Oliveira**
University of Coimbra
CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra
Department of Informatics Engineering
Polo II, Pinhal de Marrocos
3030-290 Coimbra, Portugal
{mlinacio, hroliv}@dei.uc.pt

Abstract

Among the different forms of humor, puns are the most researched in Computational Creativity and Natural Language Generation. However, existing systems generate puns by taking a pair of ambiguous words as input, without addressing how such pairs are obtained, which disconnects the generation from any contextual background. Furthermore, the majority of the works focus on English, leaving other languages, such as Portuguese, behind. In this paper, we present a full pipeline for creating puns in Portuguese, including the creation of homographic and homophonic word pairs from news headlines. We evaluate ten different Transformer-based approaches for generating jokes given these word pairs — fine-tuned T5 models and different LLM prompting techniques — through a questionnaire with 24 participants from Brazil and Portugal, who rated the puns in terms of humor and relation to the base headline. Results suggest that including the words' definitions in the prompt can harm humor ratings and that few-shot prompting outperformed zero-shot. Additionally, the T5 model fine-tuned without word definitions produced texts that were more closely related to the base headline, but at the expense of humor.

Introduction

Humor is a fundamental aspect of human communication, playing a crucial role in social interactions (Kuipers 2008). Among the various forms of humorous content, verbal humor is especially prevalent in present days, with social relations mediated by the Web, where textual content is abundant (Chiaro 2018). Within this context, researchers have been developing methods to automatically generate verbal humorous content over the past decades.

Due to its constrained nature, most research on verbal humor has focused on puns, which are defined by Miller, Hempelmann, and Gurevych (2017) as:

> "[...] a form of wordplay in which one sign (e.g., a word or phrase) suggests two or more meanings by exploiting polysemy, homonymy, or phonological similarity to another sign, for an intended humorous or rhetorical effect."

Initially, we believed that narrowing Humor Generation to these constraints would make the task simpler. However, we found that they introduce additional challenges, as we need to first find suitable pun and alternative words to generate the puns, a problem that most previous works do not address.

Additionally, by just generating a funny text from a pair of ambiguous words, the generation process becomes detached from any context, which limits its quality and utility. Thus, we propose a full pipeline for generating punning humor that includes the creation of context-aware homographic and homophonic word pairs. Including this first step as a task on itself makes the pipeline compatible with previous works, as the created pairs can be used as inputs. Furthermore, our study focuses on the Portuguese language, which is underrepresented not only in Computational Creativity (CC) but also in Natural Language Processing (NLP) literature.

Even though we could use almost all literature models as Natural Language Generation (NLG) methods, we decided to use Transformer-based models. We fine-tuned T5 models — similar to Sun et al. (2022) — and explored different prompting techniques with Large Language Models (LLMs), namely Llama3.3 (Grattafiori et al. 2024) and Llama3.3-70B distilled from Deepseek-R1 (DeepSeek-AI et al. 2025).

We evaluated our models by generating puns from a set of 30 Brazilian and Portuguese news headlines and asking 24 native speakers to rate the generated content on humor and relation to the original headline. In sum, the main contributions of this paper are:

- A full pipeline for creating puns, including the creation of homographic and homophonic word pairs;
- An evaluation of ten different Transformer-based approaches for generating puns in Portuguese within the mentioned pipeline;
- A rule-based Phoneme-to-Grapheme (P2G) system for Portuguese, which generates all possible spellings for a given phonetic transcription.

The general results suggest that fine-tuned T5 models are unfit for generating funny puns, while still creating texts that are closely related to the input headline when not including the words' definitions in the prompt. On the other hand, LLMs presented better humor ratings in some scenarios, especially when prompted with few-shot examples. We also found that including definitions in the prompt can harm humor ratings regardless of the method used.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

The remainder of the paper is organized as follows: first, we present the related work in Humor Generation; then, we describe our methodology, including the steps for creating word pairs, generating puns, and evaluating the methods. Later, we present the evaluation results and discussions. Finally, we conclude with some limitations of our work and future research directions.

Related Work

Humor Generation has been a topic of research for at least three decades, with a special focus on puns. Following the trends in NLG, the automatic creation of punning humor has been explored through various approaches, ranging from template-based systems — such as JAPE-1 (Binsted and Ritchie 1994) and T-PEG (Hong and Ong 2009) — through rewriting techniques (He, Peng, and Liang 2019; Yu, Zang, and Wan 2020), to neural networks (Yu, Tan, and Wan 2018; Luo et al. 2019; Diao et al. 2020; Mittal, Tian, and Peng 2022), and more recently LLMs (Chen et al. 2024; Inácio and Gonçalo Oliveira 2024). However, with the exception of T-PEG, which receives a keyword to be used in the joke, all of these works take as input a pair of ambiguous words to generate the puns, neglecting the process of obtaining such pairs.

One of the few works addressing the problem of obtaining input word pairs was conducted by Sun et al. (2022). By training a classifier with contextual keywords as input, they select one pair of punning and alternative words from a pre-conceived list. Yet, this approach is limited to the words in the list and does not address the automatic creation of word pairs. Besides, automatic portmanteaux generation methods could be used in such pair-creation phase, even though they are not usually mentioned within this context (Smith, Hintze, and Ventura 2014; Deri and Knight 2015; Das and Ghosh 2017; Gangal et al. 2017; Simon 2018; Vivek Kulkarni et al. 2018).

It is also worth mentioning other efforts in Humor Generation that are not limited to puns, especially those that avoid the so-called "mere generation" (Veale and Pérez y Pérez 2020) problem by incorporating contextual background into the generation process. For instance, Mendes and Gonçalo Oliveira (2020) proposed a method that adapts news headlines to generate humorous texts based on proverbial expressions. Similarly, Winters and Delobelle (2021) transformed news headlines into satirical texts using Genetic Algorithms. Another example is Witscript 2 (Toplyn 2022), which creates jokes based on a given context description. In the multimodal domain, Gonçalo Oliveira, Costa, and Pinto (2016) proposed a system that generates Internet memes from news headlines and proverbs. Although not focused on humor, Gatti et al. (2015) developed a system that does concept blending by replacing words in well-known expressions by keywords related to a set of news headlines.

Finally, we note that most of the aforementioned works are for the English language. Specifically for Portuguese, the focus of this paper, despite some work on-Humor Generation (Gonçalo Oliveira, Costa, and Pinto 2016; Gonçalo Oliveira and Rodrigues 2018; Mendes and Gonçalo Oliveira 2020; Inácio et al. 2024), it is still limited.

Methodology

We propose a full pipeline for creating puns, including the creation of pun (w_p) and alternative (w_a) word pairs, a step that has been neglected in previous works. The overall framework is depicted in Figure 1. In the following sections, we describe each step of the pipeline in more detail.

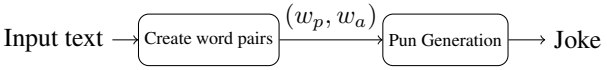


Figure 1: General framework for Pun Generation

Word Pair Creation

We create two kinds of word pairs: homographic and homophonic. For homographic pairs, w_p and w_a are written identically, without pronunciation constraints. Meanwhile, homophonic ones are pronounced the same but necessarily spelled differently. As it is known, some puns exploit words that are just similar in pronunciation or writing, but we decided to focus on these two types for simplicity.

To create the word pairs while avoiding mere generation, we first extract five keywords from each input text, in our case headlines. We then expanded this initial set by using word embeddings to find the five most similar words (based on cosine similarity) for each original keyword. Finally, word pairs are created from this expanded list of candidates.

Homographic Word Pairs The process of creating homographic word pairs is straightforward: we need to find ambiguous words among the candidates. This may resort to a dictionary or to a WordNet lexicon (Fellbaum 1998), where the definitions of all the senses of each word can be retrieved from. Since the word senses can be too similar, not providing enough semantic distinction to create a pun, we do not simply consider a word as ambiguous when it has more than one sense, e.g. the word "*caneta*" (pen) could have two different senses refering to a fountain pen and a ballpoint pen. Instead, we gather all definitions of the senses and calculate the consine similarity between them using sentence embeddings (Reimers and Gurevych 2019). If the minimum similarity between any pair of definitions is below a threshold, we consider the word as ambiguous and create its corresponding pair (w_p, w_a), where $w_p = w_a$. More details about how we set the necessary threshold can be found in the Experimental Setup section.

Homophonic Word Pairs To create homophonic word pairs, we need to add more steps to account for pronunciation. First, we used a text-to-phones conversion tool to obtain the phonetic transcription of each candidate word. Then, we use a Phoneme-to-Grapheme (P2G) system that returns all possible orthographic realizations for a given phonetic transcription. Finally, the candidates are filtered twice: first, using Phonemizer to confirm that they have the same pronunciation as the original word; and second, using the dictionary to just keep existing words.

Since a word can have multiple homophones, we must consider which ones to select for the pair. Thus, we used

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

the same similarity threshold strategy from the homographic word pairs, this time comparing the senses of paired homophonic words (rather than senses from the same word). Homophonic pairs with a similarity below the threshold are selected as the final output.

Pun Generation

For the Pun Generation step, we explore transformer-based models. Nonetheless, we highlight that, in our pipeline, this stage can be replaced by any other model that generates jokes from word pairs, as most models of the literature (Yu, Tan, and Wan 2018; Tian, Sheth, and Peng 2022; Chen et al. 2024).

Experimental Setup

Although this work focuses on the Portuguese language, the pipeline we propose is language-agnostic; hence, we present the specific resources and tools used in our experiments.

Word Pair Generation

As a first step, we required inputs to create the word pairs from. To this end, we used the top news headlines in Google News for Brazil and Portugal on January 30$^{\text{th}}$, 2025. For each country, we selected 15 headlines — five from each of the following categories: Science and Technology, Entertainment, and Sports — resulting in a total of 30 headlines.

For extracting keywords from the headlines, we used KeyBERT (Grootendorst 2020), with the sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 model[1] (Reimers and Gurevych 2019) and the NILC 300-dimension GloVe Embeddings (Hartmann et al. 2017) for expanding the keywords set.

When generating the word pairs, we had to determine the appropriate similarity threshold. To this end, we used PUNTUGUESE[2] (Inácio et al. 2024), a corpus of puns in Portuguese annotated with pun and alternative words. We applied the procedure described in the previous section to compute the minimum cosine similarity for homographic word pairs in the corpus. The distribution of these values can be seen in Figure 2. Based on this analysis, we set the threshold at 0.2, which lies between the mean (0.17) and the 70$^{\text{th}}$ percentile (0.23) of the distribution.

We adopted OpenWordNet-PT (OWN-PT) (de Paiva, Rademaker, and de Melo 2012) as our WordNet lexicon. To calculate the semantic similarity between word definitions, we used the multilingual SentenceTransformers model sentence-transformers/all-MiniLM-L6-v2[3]

For the creation of homophonic word pairs, we implemented our own rule-based P2G system, since, to the extent of our knowledge, there is none in Portuguese that retrieves all possible spelling representations. The P2G rules are derived from the standard Brazilian pronunciation described by Perini (2021) and Silva (2022). To validate the rules, we ensured that they could correctly recreate the vocabulary of the Floresta Sintá(c)tica corpus (Afonso et al. 2001), with the exception of some foreign names and terms, e.g. "Schwarzenegger" and "baguette". Our P2G system can correctly recreate the pronunciation for 20,213 out of 27,270 (74.1%) words in the corpus. It works by iterating through each phoneme in the transcription and mapping it into a set of possible graphemes; the final orthographic realization is obtained as the Cartesian product of all these grapheme sets. For example, the phonetic transcription of the word "*chá*" (tea) is [ʃˈa], resulting in the following mappings:

- [ʃ]: {ch, x, s, z, sh}
- [ˈa]: {a, á, à, ha, há}

As the Cartesian product can become computationally expensive for longer words, we included pruning heuristics to eliminate unlikely realizations. For instance, we discard mappings of [ʃ] to {s, z} at the beggining of words, as they only have this pronunciation at syllable endings. Another pruning used is avoiding results with "hh", as this does not happen in Portuguese orthography; additionally, "à" just exists in rather particular cases. Thus, the final writing candidates for this pronunciation are {"*sha*", "*shá*", "*cha*", "*chá*", "*xa*", "*xá*"}. After the filtering process, described in the previous section, we end up with the final homophonic word pair {"*chá*" (tea), "*xá*" (shah)}

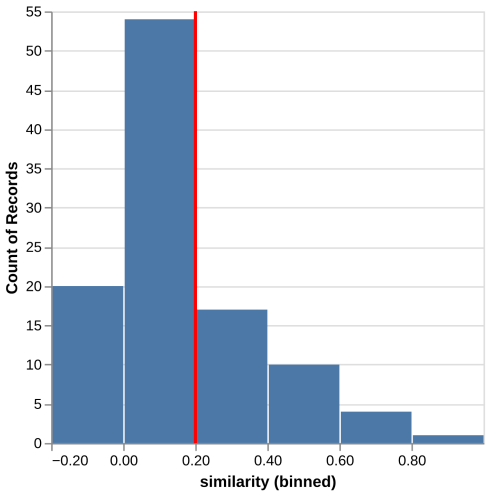


Figure 2: Distribution of cosine similarity between definitions of ambiguous words in the PUNTUGUESE corpus

Pun Generation

We evaluated ten different NLG approaches, divided into two groups: fine-tuned T5 models and different LLM prompting techniques.

Fine-tuned T5 Models The first generation approach we explored was fine-tuning a pre-trained T5 model to generate puns given the word pair under two settings: one without word definitions and one with word definitions. As a pre-trained T5 model, we adopted PTT5-v2 (Piau, Lotufo, and

[1] https://hf.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

[2] https://github.com/Superar/Puntuguese.

[3] https://hf.co/sentence-transformers/all-MiniLM-L6-v2.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Nogueira 2025), a Portuguese version of T5. We fine-tuned this model on PUNTUGUESE (Inácio et al. 2024).

We initially filtered the corpus to keep only homographic or homophonic puns that have exactly one pair of pun (w_p) and alternative (w_a) words, excluding jokes with multiple pairs. This filtering resulted in a total of 964 examples. For the second approach, we further filtered the corpus to include only the puns whose words are present in OWN-PT, allowing us to obtain their definitions (d_p and d_a), resulting in 207 examples.

Then, we created input commands for the model:

- `PTT5`: "*Gerar trocadilho:* w_p / w_a";
- `PTT5+def`: "*Gerar trocadilho:* w_p (d_p) / w_a (d_a)".

Since w_p and w_a may have more than one definition, we selected the definitions with the lowest cosine similarity using the sentence embeddings previously described.

For the fine-tuning, we used the stratified splits provided in the PUNTUGUESE Huggingface repository[4], keeping the previous filtering mentioned above. This resulted in 676 training, 98 validation, and 190 test examples for the `PTT5` model; and 142 training, 17 validation, and 48 test examples for the `PTT5+def` model.

Hyperparameter tuning was performed with Weights & Biases[5] via a Bayesian search for the best learning rate and batch size, minimizing validation loss across 20 runs. The optimal hyperparameters for the `PTT5` model were: a learning rate of 1.7×10^{-6} and a batch size of 16; for the `PTT5+def` model, the best hyperparameters were: a learning rate of 7.9×10^{-6} and a batch size of 16. Both models were fine-tuned for a maximum of 200 epochs, with early stopping if the validation loss did not improve for three consecutive evaluations. The generation sampling was done with a temperature of 0.6 and top-p of 0.95.

LLM Prompting Techniques Since LLMs are general-purpose, instruction-based models, we did not carry out fine-tuning. Instead, we relied entirely on prompting:

- `Llama3.3` and `Deepseek`: Zero shot, with a prompt including the task definition and the word pair;
- `Llama3.3+def` and `Deepseek+def`: Zero shot, with a prompt including the task definition, word pair, and the definitions of the words;
- `Llama3.3+shot` and `Deepseek+shot`: Few shot (using 4 random examples from PUNTUGUESE, 2 of each type), with a prompt including the task definition and the word pair;
- `Llama3.3+shot+def` and `Deepseek+shot+def`: Few shot, with a prompt including the task definition, word pair, and definitions of the words.

For LLM prompting, we used the Ollama tool[6] with two state-of-the-art models: Llama3.3 (`Llama3.3`) (Grattafiori et al. 2024) and Llama3.3 distilled from Deepseek-R1 (`Deepseek`) (DeepSeek-AI et al. 2025), each with 70B parameters, temperature 0.6, and top-p 0.95. Full prompts (in Portuguese) for each technique, as well as the selected few-shot examples, can be seen in Table 1. Each prompt is preceeded by the following system prompt:

> *Você é um gerador de piadas baseado em trocadilhos em português. Sua tarefa é criar piadas curtas e engraçadas baseadas nos pares de palavras fornecidos. Crie os trocadilhos com base nos exemplos. Apresente o resultado exclusivamente em formato JSON contento duas chaves: "palavras" e "trocadilho", contendo as palavras fornecidas e a piada criada, respectivamente. Retorne apenas o objeto JSON.* (You are a pun-based joke generator in Portuguese. Your task is to create short and funny jokes based on the word pairs provided. Create the puns based on the examples. Present the result exclusively in JSON format containing two keys: "words" and "pun", containing the words provided and the joke created, respectively. Return only the JSON object.)

All prompts are given in Portuguese except for the definitions, which are in English because the OWN-PT defintions used in our final experiments are in English. The few-shot definitions were manually written by the authors based on the corresponding entries in the Priberam dictionary[7].

Evaluation

To evaluate the models, we first generate puns from a set of 30 headlines, as descibed in the Word Pair Creation section. Three of the headlines did not result in any word pair, i.e. no ambiguous words were found. The remaining 27 headlines resulted in 124 word pairs, which served as inputs for the Pun Generation models. This process resulted in 1,240 puns, from which we select the examples for evaluation.

Note that, for the same headline, different pairs can be obtained, thus resulting in different puns. This is frequent when the input text has an ordinal, which can be a fraction, and is semantically-related to other ordinals. One of such cases in the used headlines happens for "*E assim foi o terceiro episódio de A Herança*" (And so was the third episode of "The Inheritance"), illustrated with three of the eight examples of word pairs and respective jokes by `Llama3.3+def`:

- *quarto*, *quarto* (fourth, bedroom): P*or que o quarto da casa foi ao médico? Porque estava com um 'quarto' de litro de febre!* (Why did the room in the house go to the doctor? Because it had a 'quarter' of a liter of fever!)
- *quinto*, *quinto* (fifth): *Por que o quinto elemento foi ao quinto andar? Porque estava procurando por sua fração perfeita em uma série de eventos!* (Why did the fifth element go to the fifth floor? Because it was looking for its perfect fraction in a series of events!)
- *cesto*, *sexto* (basket, sixth): *Eu tentei encontrar o cesto, mas ele estava no sexto lugar da lista de prioridades!* (I tried to find the basket, but it was sixth on the priority list!)

[4] `https://hf.co/datasets/Superar/Puntuguese`.

[5] `https://wandb.ai/`.

[6] `https://ollama.com/`.

[7] `https://dicionario.priberam.org/`.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Model type	Prompt
Zero-shot	Palavras: "w_p" e "w_a"
Zero-shot with definitions	Palavras: "w_p" e "w_a" Definições: "d_p" e "d_a"
Few-shot	Palavras: "entre-meada" e "entre miada" {"palavras": ["entre-meada", "entre miada"], "trocadilho": "O que é que está entre dois gatos? Uma entre-meada."} Palavras: "flora" e "flora" {"palavras": ["flora", "flora"], "trocadilho": "Foi lançada uma nova manteiga com sabor a merda. Chama-se Flora Intestinal."} Palavras: "shanti-lee" e "chantilly" {"palavras": ["shanti-lee", "chantilly"], "trocadilho": "Que nome se dá a uma chinesa muito docinha? Shanti-Lee."} Palavras: "paciente" e "paciente" {"palavras": ["paciente", "paciente"], "trocadilho": "O que o médico disse para o homem que entrou gemendo de dor no hospital? "Seja paciente"."} Palavras: "w_p" e "w_a"
Few-shot with definitions	Palavras: "entre-meada" e "entre miada" Definições: "Part of the pork meat consisting of bacon with bacon in between" e "Between the voice of the cat" {"palavras": ["entre-meada", "entre miada"], "trocadilho": "O que é que está entre dois gatos? Uma entre-meada."} Palavras: "flora" e "flora" Definições: "Set of plants from a region, an environment or a geological period" e "Butter brand" {"palavras": ["flora", "flora"], "trocadilho": "Foi lançada uma nova manteiga com sabor a merda. Chama-se Flora Intestinal."} Palavras: "shanti-lee" e "chantilly" Definições: "Chinese-sounding name" e "Cream made from whipped cream and sugar" {"palavras": ["shanti-lee", "chantilly"], "trocadilho": "Que nome se dá a uma chinesa muito docinha? Shanti-Lee."} Palavras: "paciente" e "paciente" Definições: "Who or what has patience" e "Any person undergoing medical treatment or care" {"palavras": ["paciente", "paciente"], "trocadilho": "O que o médico disse para o homem que entrou gemendo de dor no hospital? "Seja paciente"."} Palavras: "w_p" e "w_a" Definições: "d_p" e "d_a"

Table 1: Prompts used for Pun Generation with LLMs. Variables w_p, w_a, d_p, and d_a represent the pun and alternative words, and their definitions, respectively.

To maintain a feasible evaluation process, we selected, for each headline and model, one pun according to an automatic scoring, regardless of the source pair. With this score, we selected 270 puns (10 per headline, one for each model) for evaluation. The score was computed as a weighted combination (each with a weight of 0.5) of two metrics:

- Typicality: The classification score by a Humor Recognition model trained on PUNTUGUESE (Inácio et al. 2024);
- Semantic Similarity: The cosine similarity between the the embeddings of the generated pun and the headline.

The evaluation was carried out by 24 participants — 14 from Brazil and 10 from Portugal — all native speakers of their respective varieties of Portuguese. Each evaluator received a set of headlines from their country and evaluated all associated jokes for each headline. The sets were split across the evaluators so that puns for each headline were evaluated by exactly 4 participants.

Participants rated each generated text based on two criteria: humor and relation to the base headline, using a 3-point scale. For humor, the scale was: "*Não tem piada*" (Not funny), "*Tem pouca piada*" (A bit funny), and "*Tem piada*" (Funny). For relation to the headline, the ratings were: "*Não tem relação*" (Not related), "*Tem pouca relação*" (A bit related), and "*Tem relação*" (Related).

Results

Before presenting the evaluation results, we depict some examples of generated text for each approach in Table 2. Each example was randomly selected from the jokes of its corresponding model with a majority rating of "Funny" from the evaluators. From these examples, we can see that the models are indeed generating puns, according to the definition of Miller, Hempelmann, and Gurevych (2017) mentioned in the Introduction, as they include words that evoke more than one meaning even if not necessarily the input word pair. For example, the joke by `Deepseek+shot+def` has both

Model	Input Headline	Generated Text	Pair	Fun.	Rel.
PTT5	***Starlink Direct to Cell passa a disponibilizar 4G via satélite para todos os smartphones.*** **(Starlink Direct to Cell now offers 4G via satellite for all smartphones.)**	*Qual é o satélite que mais se aproxima do homem? O satélite.* (What is the satellite that is closest to the man? The satellite.)	*satélite* (satellite) *satélite* (follower)	3	2
PTT5 +def	***Starlink Direct to Cell passa a disponibilizar 4G via satélite para todos os smartphones.*** **(Starlink Direct to Cell now offers 4G via satellite for all smartphones.)**	*Por que o homem não tem um satélite? Porque não tem um satélite.* (Why doesn't the man have a satellite? Because he doesn't have a satellite.)	*satélite* (satellite) *satélite* (follower)	3	2
Llama3.3	***Vodafone realiza primeira videochamada espacial do mundo com smartphone.*** **(Vodafone makes the world's first space video call with a smartphone.)**	*Por que o mundo foi ao psicólogo? Porque estava se sentindo um pouco 'girado'!* (Why did the world go to the psychologist? Because it was feeling a bit 'spun'!)	*mundo* (populace) *mundo* (world)	3	2
Llama3.3 +def	***"Não daria à minha filha um vibrador aos 12 anos. Não posso é condenar".*** **("I wouldn't give my daughter a vibrator at 12. I just can't condemn it.")**	*Eu condeno o uso excessivo de 'condenar', mas não posso condenar alguém por concordar!* (I condemn the excessive use of 'condemn', but I can't condemn someone for agreeing!)	*condenar* (condemn) *condenar* (convict)	3	2
Llama3.3 +shot	***Liga Europa: o onze provável do FC Porto na estreia de Martín Anselmi.*** **(Europa League: FC Porto's probable eleven in Martín Anselmi's debut.)**	*Por que a lâmpada foi ao terapeuta? Ela estava com problemas para se ligar.* (Why did the lamp go to the therapist? She was having trouble turning on.)	*liga* (elastic) *liga* (league)	3	1
Llama3.3 +shot +def	***O preço do Samsung Galaxy S25 cai a pique com os descontos na Samsung.*** **(The price of the Samsung Galaxy S25 drops sharply with discounts at Samsung.)**	*Por que o vendedor usava uma venda nos olhos? Para não ver a venda cair!* (Why did the seller wear a blindfold? So he wouldn't see the sale fall!)	*venda* (sale) *venda* (blindfold)	3	2
Deepseek	***Restos de salmoura no asteroide Bennu contêm minerais essenciais para a vida.*** **(Brine residues on asteroid Bennu contain essential minerals for life.)**	*O asteróide foi ao terapeuta porque estava se sentindo um pouco 'defasado'.* (The asteroid went to the therapist because it was feeling a bit 'outdated'.)	*asteróide* (asteroid) *asteroide* (asteroid)	3	3
Deepseek +def	***E assim foi o terceiro episódio de "A Herança".*** **(And so was the third episode of "The Inheritance".)**	*Por que a carteira e o DNA foram juntos ao terapeuta? Porque um carrega a herança material e o outro, a herança genética!* (Why did the wallet and the DNA go to the therapist together? Because one carries the material inheritance and the other, the genetic inheritance!)	*herança* (heritage) *herança* (heredity)	3	2
Deepseek +shot	***Restos de salmoura no asteroide Bennu contêm minerais essenciais para a vida.*** **(Brine residues on asteroid Bennu contain essential minerals for life.)**	*O que é um asteróide? É um asteroide sem sotaque.* (What is an asteroid? It's an asteroid without an accent.)	*asteróide* (asteroid) *asteroide* (asteroid)	3	3
Deepseek +shot +def	***Restos de salmoura no asteroide Bennu contêm minerais essenciais para a vida.*** **(Brine residues on asteroid Bennu contain essential minerals for life.)**	*Por que o asteroide foi ao festival de música? Porque ele amava rock e metal!* (Why did the asteroid go to the music festival? Because it loved rock and metal!)	*asteróide* (asteroid) *asteroide* (asteroid)	3	3

Table 2: Examples of funniest texts generated by each model. Input headlines are in bold.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

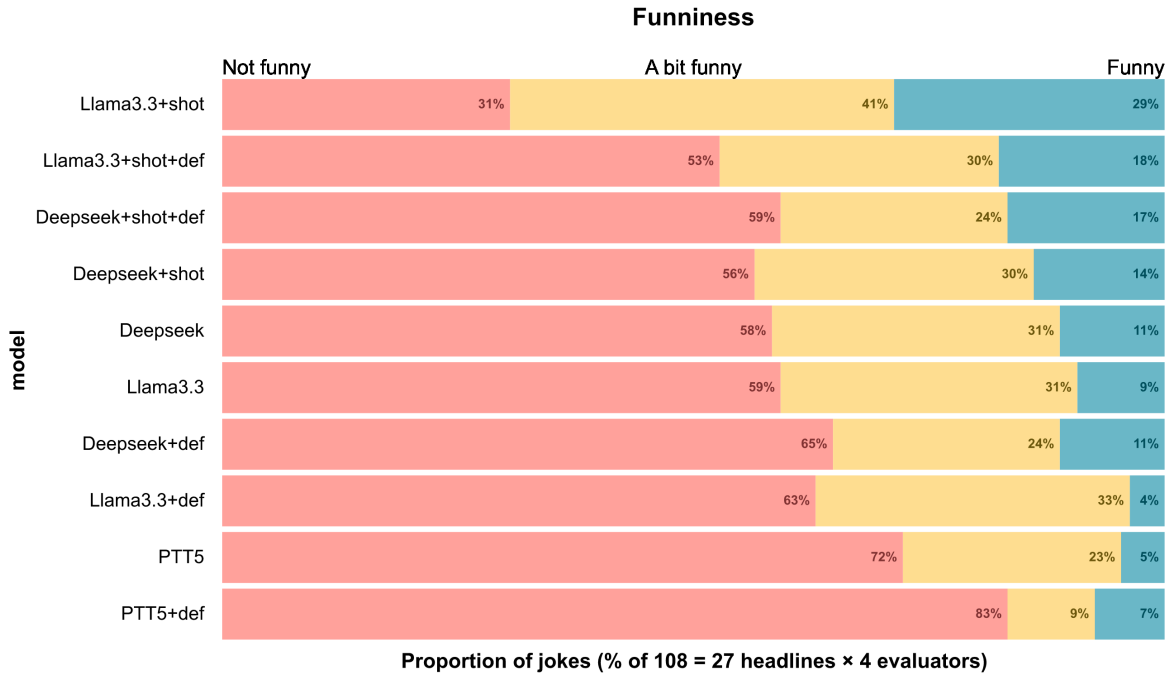


Figure 3: Evaluation distribution for the Funniness scale

"rock" and "metal" as ambiguous words, but the input pair was "*asteróide*" and "*asteroide*" (both meaning asteroid).

Another joke that derived from this same pair of words is the one by `Deepseek+shot`, which lead to an interesting observation about this specific model. Deepseek-R1 has a built-in Chain-of-Thought (CoT) mechanism that always generates a thought process in English, even if the final output is in another language. When observing the CoT output for this joke, we see that the model correctly identified the difference in the accent mark of the dipthong ("*oi*" vs. "*ói*") and decided to make a joke about it. However, the final output had a mistranslation for the word "accent", which was translated as "*sotaque*" (a regional form of pronunciation), while the intended meaning (from the CoT) would be "*acento*" (an accent mark). This "error", nonetheless, might have resulted in a funnier joke than the original intention as it created a more playful scenario with talking asteroids.

On the other hand, we can observe that LLMs also rely on the so-called "lazy pun pattern", as described by Xu et al. (2024), in which the text is composed by forcing two occurences of the pun word with different meanings, rather than just implying one of the senses. For instance, the pun created by `Llama3.3+shot+def` has the word "*venda*" sale or blindfold) repeated twice, with a clear distinction between the two meanings. Another example is the joke from `Deepseek+def`, which explicitly adds adjectives to limit the semantic interpretations of the occurrences of the word "*herança*" (inheritance or heredity).

For a quantitative analysis of the results, we plot the distribution of all 1,080 evaluations (27 headlines $\times$ 10 models $\times$ 4 evaluators) for the Funniness scale in Figure 3. We see that `Llama3+shot` is the only model that reached less than 50% of the evaluations as "Not funny", being consistently funnier than the other models. From the distributions, we can also notice that including the definitions (`+def`) consistently harmed the funniness of the generated texts, regardless of the model, resulting in higher proportions of the "Not funny" category. The fact that the definitions from are in English might have influenced these results.

For LLMs, we can see that the few-shot approach improved funniness regardless of the model or the presence of word definitions, especially for `Llama3.3` — which had an increase of 164% in the proportion of "Funny" texts (from 11% to 29%) — and `Llama3.3+def`, with a 350% "Funny" text increase (from 4% to 18%).

Concerning the fine-tuned models, `PTT5` and `PTT5+def` both had the highest proportion of "Not funny" ratings, 72% and 83% respectively. This shows that the fine-tuning was not enough for the model to acquire humor-production capabilities, being limited to produce texts that resemble jokes in the superficial form. On the other hand, regarding the relation to the headline (Figure 4), `PTT5` had the highest proportion of "A bit related" (41%) and "Related" (11%) ratings.

It is notable that including the definitions in `PTT5`'s input harmed the relation to the headline; we believe that this is due to our definition-selection process, which might have selected definitions that are not the most common or the most related to the headline. We need, however, to further investigate this hypothesis.

Observing the overall relation distributions in Figure 4,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

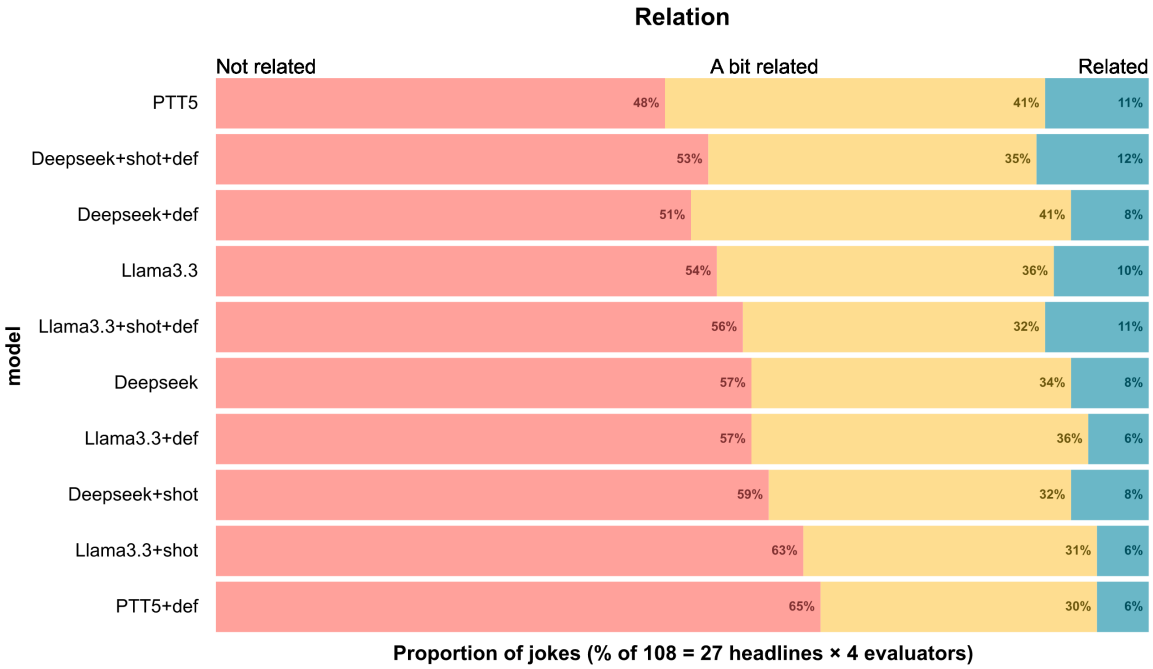


Figure 4: Evaluation distribution for the Relation scale

we can see that the models did not perform well in creating texts related to the headline, with very few ratings as "Related". By comparing the effects of including the definitions into the input, it mostly harms the results, with the exception of the Deepseek-R1-based approaches. This may be due to its built-in CoT mechanism, which could potentially bring up other definitions rather than simply relying on the ones provided; a more in-depth analysis of the CoT outputs could be helpful in testing this hypothesis.

To understand how consistent the ratings are for each model, we calculated the agreement scores of the evaluations for each model. To do so, we chose the Krippendorff's α coefficient (Krippendorff 2019) as it can handle multiple raters and missing data (as not every evaluator rated every text). The results are presented in Table 3.

	Krippendorff's α	
Model	Funniness	Relation
`PTT5`	-0.05	0.46
`PTT5+def`	0.07	0.44
`Llama3.3`	0.08	0.41
`Llama3.3+def`	0.09	0.33
`Llama3.3+shot`	0.02	0.18
`Llama3.3+shot+def`	**0.26**	0.40
`Deepseek`	0.00	**0.54**
`Deepseek+def`	0.20	0.31
`Deepseek+shot`	0.13	0.38
`Deepseek+shot+def`	0.21	0.28

Table 3: Krippendorff's α values for Funniness and Relation

As expected from such a subjective task, agreement scores for Funniness are low, with the highest value being 0.26 for `Llama3.3+shot+def`. This shows that, although having higher proportions of "A bit funny" and "Funny" ratings, the humor created by `Llama3.3+shot` did not consistently please the same way ($\alpha = 0.02$). When including the definitions (`Llama3.3+shot+def`), the agreement scores are a little higher (0.26), but still not enough to consider the evaluations as consistent.

On the other hand, agreement scores for Relation are higher, with the most consistent model being `Deepseek` ($\alpha = 0.54$). Curiously, `Llama3.3+shot` had the lowest agreement score for this scale ($\alpha = 0.18$), from 27 jokes, 19 of them had at least one evaluator desagreeing with the others. This indicates how even this task of evaluating semantic relationship can have some degree of subjectivity, especially when dealing with texts close to non-sense (as most of the generated jokes); for example, the following joke generated by `Llama3.3+shot` had all 3 different Relation ratings (2 "Not related", 1 "A bit related", and 1 "Related"):

> **Headline:** *Starlink Direct to Cell passa a disponibilizar 4G via satélite para todos os smartphones* (Starlink Direct to Cell now offers 4G via satellite to all smartphones)
>
> **Joke:** *Por que o planeta foi ao terapeuta? Ele estava se sentindo um pouco terrestre, precisava de ajuda para se sentir mais 'em terra'.* (Why did the planet go to the therapist? It was feeling a bit earthy, needed help to feel more 'down to earth'.)

Finally, we evaluated if the simple automatic scoring we used to select the best puns when there were multiple options

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

for a headline was consistent with the human evaluations. Thus, we calculated the Pearson's correlation between both scores (Typicality and Semantic Similarity) and the Funniness and Relation ratings; additionally, we included the correlation between the two human ratings. The results are depicted in Table 4.

Metric	Funniness	Relation
Typicality	-0.02	-0.06
Semantic Similarity	0.06	0.15
Funniness	—	0.28

Table 4: Pearson's correlation between Typicality, Semantic Similarity, Funniness, and Relation

The correlation analysis shows that the automatic scoring we used might not be the best-suited for the task, as it had very low correlation with the human ratings, especially between Typicality and Funniness. For Relation and Semantic Similarity, the correlation was higher, but still not enough to consider the automatic scoring as a good representation of the human evaluations. Regarding the ratings for Funniness and Relation, the correlation was 0.28, indicating a slight connection between the context and the humor effect, but further analysis must be carried out to confirm this intuition.

Conclusions and Future Work

We presented a complete pipeline for generating puns, from the creation of word pair to the generation of final punning texts. We provide a straightforward methodology to create homographic and homophonic word pairs based on the similarity of their lexical definitions. To evaluate our approach, we implemented it in Portuguese. We also developed a Phoneme-to-Grapheme system to generate orthographic alternatives in Portuguese, which can be used not only in creating puns but also in projects that require spelling variations; for example, a funny translator-like platform[8] that gives a non-canon spelling for a given sentence. All code for our experiments, including the P2G system, are publicly available at `https://github.com/NLP-CISUC/full-pun-generation/`.

As for the pun generation step, we explored 10 different approaches, including fine-tuned T5 models and state-of-the-art LLMs. We also tested the impact of including the definitions of the words in the input, as well as providing few-shot examples. To the extent of our knowledge, this is the first work to use LLMs for context-aware Pun Generation in Portuguese.

The evaluation was carried out by 24 native speakers of Portuguese, from Brazil and Portugal, who rated the generated texts based on two criteria: funniness and relation to the headline from which the word pairs were created. Results showed that the LLMs consistently outperformed the fine-tuned models on funniness, with Llama3.3 with few-shot being considered "A bit funny" or "Funny" most of the times. Other approaches, however, are mostly "Not funny". The inclusion of definitions in the input harmed the funniness ratings of the generated texts, regardless of the model, but it did not have a significant impact on the relation to the headline, which is not exceptionally good in any model.

As improvement, we point out the need to use word definitions in Portuguese — either through Machine Translation, bilingual dictionaries, or LLMs — guaranteeing consistency within the prompts. For LLMs, we could use the original headline to ensure a stronger relation with the context, while exploring more techniques, such as prompt chaining or agentic prompting. We also need to further investigate the impact of the word pair creation process on the final puns, especially whether it provides more control over what parts of the context should be considered during generation. Additionally, the word pair creation can be expanded to include other types of punning phenomena, such as lexical blending or paronymy.

Finally, the evaluation can be expanded by, not only including more people to account for multiple senses of humor, but also by collecting sociodemographic data (Eiselen and van Huyssteen 2023) to analyze if the models are considered funny for any specific type of audience (for example, people who like dad-jokes). We also intend to further analyze the outputs of the CoT mechanism in Deepseek-R1, as it can give interesting insights about how the concepts are connected during the process of creating a suitable pun.

Limitations

We acknowledge that the main limitation of this work is the lack of comparison of the impact of the word pair creation step in the pipeline, especially on headline relation. However, we emphasize that adressing the challenge of obatining these pairs is already a significant contribution to the field, as most existing works on pun generation do not deal with this issue, despite relying on such pairs as input. Moreover, we believe that including the word pair creation step can provide more control and interpretability over the generated puns.

Furthermore, when combining the Typicality and Semantic Similarity scores, we did not perform weight optmization; instead, both metrics had an equal weight of 0.5. It is possible that different weights would lead to different selections. Moreover, we acknowledge that it would be ideal to compare the models' jokes with human-created ones, whether produced by professionals or laypeople, to have an upper bound on the expected performance. Finally, we mention that the generated jokes could not be entirely new, as we do not assess novelty and originality in this paper.

Acknowledgments

We thank all volunteering participants who took part in our evaluation process. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT – Foundation for Science and Technology I.P. (grant number UI/BD/153496/2022), in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025.

[8]Inspired by MiGuXeiToR: `https://aurelio.net/coisinha/miguxeitor/`.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

Afonso, S.; Bick, E.; Haber, R.; and Santos, D. 2001. Floresta Sintá(c)tica: um "treebank" para o português. In Conçalves, A., and Correia, C. N., eds., *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)*. Lisbon: APL.

Binsted, K., and Ritchie, G. 1994. An implemented model of punning riddles. In Hayes-Roth, B., and Korf, R. E., eds., *Proceedings of the 12th National Conference on Artificial Intelligence*, volume 1, 633–638. Seattle: AAAI Press / The MIT Press.

Chen, Y.; Yang, C.; Hu, T.; Chen, X.; Lan, M.; Cai, L.; Zhuang, X.; Lin, X.; Lu, X.; and Zhou, A. 2024. Are U a joke master? Pun generation via multi-stage curriculum learning towards a humor LLM. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 878–890. Bangkok, Thailand and virtual meeting: ACL.

Chiaro, D. 2018. *The Language of Jokes in the Digital Age: Viral Humour*. New York: Routledge.

Das, K., and Ghosh, S. 2017. Neuramanteau: A Neural Network Ensemble Model for Lexical Blends. 576–583.

de Paiva, V.; Rademaker, A.; and de Melo, G. 2012. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, 353–360. Mumbai: The COLING 2012 Organizing Committee.

DeepSeek-AI; Guo, D.; Yang, D.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. In *arXiv:2501.12948*. arXiv.

Deri, A., and Knight, K. 2015. How to Make a Frenemy: Multitape FSTs for Portmanteau Generation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 206–210. Denver, Colorado: Association for Computational Linguistics.

Diao, Y.; Yang, L.; Fan, X.; Chu, Y.; Wu, D.; Zhang, S.; and Lin, H. 2020. AFPun-GAN: Ambiguity-Fluency Generative Adversarial Network for Pun Generation. In Zhu, X.; Zhang, M.; Hong, Y.; and He, R., eds., *Natural Language Processing and Chinese Computing*, volume 12430. Cham: Springer International Publishing. 604–616.

Eiselen, R., and van Huyssteen, G. B. 2023. A Comparison of Statistical Tests for Likert-Type Data: The Case of Swearwords. *Journal of Open Humanities Data*.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambridge: The MIT Press.

Gangal, V.; Jhamtani, H.; Neubig, G.; Hovy, E.; and Nyberg, E. 2017. CharManteau: Character Embedding Models For Portmanteau Creation. *arXiv: Computation and Language*.

Gatti, L.; Özbal, G.; Guerini, M.; Stock, O.; and Strapparava, C. 2015. Slogans Are Not Forever: Adapting Linguistic Expressions to the News. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 2452–2458. Buenos Aires: AAAI Press.

Gonçalo Oliveira, H., and Rodrigues, R. 2018. Exploring Lexical-Semantic Knowledge in the Generation of Novel Riddles in Portuguese. In *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)*, 17–25. Tilburg, the Netherlands: Association for Computational Linguistics.

Gonçalo Oliveira, H.; Costa, D.; and Pinto, A. M. 2016. One Does Not Simply Produce Funny Memes! - Explorations on the Automatic Generation of Internet Humor. In Pachet, F.; Cardoso, A.; Corruble, V.; and Ghedini, F., eds., *Proceedings of the Seventh International Conference on Computational Creativity*, 238–245. Paris: Sony CSL Paris, France.

Grattafiori, A.; Dubey, A.; Jauhri, A.; et al. 2024. The Llama 3 Herd of Models. In *arXiv:2407.21783*. arXiv.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT. In *Zenodo*. v0.3.0.

Hartmann, N.; Fonseca, E.; Shulby, C.; Treviso, M.; Rodrigues, J.; and Aluísio, S. 2017. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In Pinheiro, V., and Paetzold, G. H., eds., *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 122–131.

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*, volume 1, 1734–1744. Minneapolis: ACL.

Hong, B. A., and Ong, E. 2009. Automatically Extracting Word Relationships as Templates for Pun Generation. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 24–31. Boulder: ACL.

Inácio, M., and Gonçalo Oliveira, H. 2024. Generation of Punning Riddles in Portuguese with Prompt Chaining. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC'24)*. Jönköping: ACC.

Inácio, M. L.; Wick-Pedro, G.; Ramisch, R.; Espírito Santo, L.; Chacon, X. S. Q.; Santos, R.; Sousa, R.; Anchiêta, R.; and Goncalo Oliveira, H. 2024. Puntuguese: A Corpus of Puns in Portuguese with Micro-Edits. 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)*, 13332–13343. Torino: ELRA and ICCL.

Krippendorff, K. 2019. *Content Analysis: An Introduction to Its Methodology*. 2455 Teller Road, Thousand Oaks California 91320: SAGE Publications, Inc.

Kuipers, G. 2008. The Sociology of Humor. In *The Primer of Humor Research*, number 8 in Humor Research. Berlin, New York: Victor Raskin. 361–398.

Luo, F.; Li, S.; Yang, P.; Li, L.; Chang, B.; Sui, Z.; and Sun, X. 2019. Pun-GAN: Generative Adversarial Network for Pun Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

ISBN: 978-989-54160-7-3

Processing (EMNLP-IJCNLP), 3386–3391. Hong Kong: ACL.

Mendes, R., and Gonçalo Oliveira, H. 2020. TECo: Exploring Word Embeddings for Text Adaptation to a given Context. In Cardoso, F. A.; Machado, P.; Veale, T.; and Cunha, J. M., eds., *Proceedings of the Eleventh International Conference on Computational Creativity*, 185–188. Coimbra: ACC.

Miller, T.; Hempelmann, C.; and Gurevych, I. 2017. SemEval-2017 Task 7: Detection and Interpretation of English Puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 58–68. Vancouver: ACL.

Mittal, A.; Tian, Y.; and Peng, N. 2022. AmbiPun: Generating Humorous Puns with Ambiguous Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1053–1062. Seattle: ACL.

Perini, M. A. 2021. *Gramática descritiva do português brasileiro*. Editora Vozes.

Piau, M.; Lotufo, R.; and Nogueira, R. 2025. Ptt5-v2: A Closer Look at Continued Pretraining of T5 Models for the Portuguese Language. In Paes, A., and Verri, F. A. N., eds., *Intelligent Systems*, volume 15413. Cham: Springer Nature Switzerland. 324–338.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. Hong Kong: ACL.

Silva, T. C. 2022. *Fonetica e Fonologia do Português*. São Paulo: Editora Contexto, 7 edition.

Simon, J. A. 2018. Entendrepeneur: Generating humorous portmanteaus using word embeddings. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)*.

Smith, M. R.; Hintze, R. S.; and Ventura, D. 2014. Nehovah: A Neologism Creator Nomen Ipsum. 173–181.

Sun, J.; Narayan-Chen, A.; Oraby, S.; Gao, S.; Chung, T.; Huang, J.; Liu, Y.; and Peng, N. 2022. Context-Situated Pun Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4635–4648. Abu Dhabi: ACL.

Tian, Y.; Sheth, D.; and Peng, N. 2022. A Unified Framework for Pun Generation with Humor Principles. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3253–3261. Abu Dhabi: Association for Computational Linguistics.

Toplyn, J. 2022. Witscript 2: A System for Generating Improvised Jokes Without Wordplay. In *International Conference on Computational Creativity*. Bolzano-Bozen: ACC.

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.

Vivek Kulkarni; Kulkarni, V.; William Yang Wang; and Wang, W. Y. 2018. Simple Models for Word Formation in Slang. 1:1424–1434.

Winters, T., and Delobelle, P. 2021. Survival of the Wittiest: Evolving Satire with Language Models. In *Proceedings of the Twelfth International Conference on Computational Creativity*, 82–86. Mexico City: ACC.

Xu, Z.; Yuan, S.; Chen, L.; and Yang, D. 2024. "A good pun is its own reword": Can Large Language Models Understand Puns? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11766–11782. Miami: ACL.

Yu, Z.; Tan, J.; and Wan, X. 2018. A Neural Approach to Pun Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1650–1660. Melbourne: ACL.

Yu, Z.; Zang, H.; and Wan, X. 2020. Homophonic Pun Generation with Lexically Constrained Rewriting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2870–2876. Online: ACL.

Me, Myself and Irony: Modeling the deceptive creativity of irony with Large Language Models

Tony Veale
School of Computer Science,
University College Dublin,
Dublin D4, Ireland.
tony.veale@UCD.ie

Abstract

Modern generative AI systems certainly excel at generation, whether of images, audio or text, and can now shoulder so much of the creative burden that they may already meet a popular definition of computational creativity, all without actually embodying any explicit theory or model of creativity. For many tasks, the issue of whether these systems can appreciate what they generate, or whether whether they are *merely generative*, is a moot one, given the human-like quality of their outputs. Yet for some creative tasks, this question still matters. This paper explores the capacity of large language models (LLMs) to both speak ironically and to appreciate the irony of what they produce. Irony requires a contrast between a speaker's thoughts and a speaker's words; the user of irony holds something back, something unsaid, that undermines what is actually said. We compare and contrast creative comparisons from humans on the web and the outputs of LLMs such as *GPT4o-mini*, with a focus on the *"X is the Y of Z"* construction, to quantify the biases, divergence, and scope for deliberate irony in each. Our aim is to quantify the extent to which an LLM can self-assess and appreciate the irony of its own outputs, and thus filter any unsuccessful outputs for itself.

Introduction

Picasso famously quipped that "art is a lie that tells the truth" (Picasso 1923). Certainly, art is a form of artifice that is not what it aims to represent: a painting of a pipe is not a pipe, as Magritte tells us, anymore than the Venus de Milo is a real woman, with or without arms. Nonetheless, art can strive to convey truths about the world or the human condition by using this artifice to stir feelings and provoke new thoughts.

"Art" is a very broad term, encompassing many other creative phenomena that convey truth with artifice or sense with nonsense. Irony is an everyday example of how one can convey truths more viscerally by seeming to cloak them in transparent lies. This makes irony a game in which speakers seek to tell the truth, or at least *their* truth, with lies that are undone by context. If I point to an expensive sports car parked on a street festooned with litter, graffiti and broken windows, and say "That guy *knows* how to park!" I am really saying "The guy does not know how to park, but he really should." As (Grice 1978) puts it, an ironic speaker says something that is blatantly false in context, and relies on this context to act as an implicit negation marker on what is said. In effect, this speaker pretends to be someone else, someone unwise after the fact, whose folly is mocked by the transparent lie.

An ironic speaker who says "what a great place to park!" is doing one of three things: saying the opposite of what others are thinking; pretending to be someone else, one whose imprudence deserves criticism; or quoting an unwise person, such as the car's owner, or echoing their inner dialogue. Different theories of irony prize one of these actions over others. (Grice 1978) emphasizes the contradiction implicit in irony, while (Clark and Gerrig 1984) focus on pretence. This play-acting, which is crafted to be seen as such, becomes arch and pantomime-like when irony curdles into unsubtle sarcasm. The idea that irony quotes from a context that has now changed, or echoes a thought that no longer seems wise, is central to the *echoic mention* theory of (Sperber and Wilson 1981; 1992). But whether one is quoting, echoing or pretending, allusion is a key part of irony. (Kreuz and Glucksberg 1989) argue that sarcastic irony, such as "you're a real genius," must allude to an antecedent state of affairs that robs the claim of its merits. In doing so, the allusion brings into focus an expectation that has since failed (e.g., that the addressee is capable of clever thoughts), thus revealing the statement to be a shallow pretence (Kumon-Nakamura, Glucksberg, and Brown 1995).

Since it alludes to a failed expectation *as if* it actually came to pass (Garmendia 2019), most irony appears positive but conveys a tacit criticism. Negative irony, as in "You're a terrible friend (ha ha)", is rarer, and while it imparts praise of a kind, it also carries a strong whiff of reproach. For the most part, we expect an ironic claim to seem positive on the surface, but to convey this positivity with little credibility. This alloy of high-positivity and low-believability, dubbed *pragmatic insincerity* by Kumon-Nakamura *et al.* (1995), is exactly we want from LLMs when we task them with generating ironic comparisons. So, when an LLM packages irony in the form *X is the Y of Z*, such as *Conor McGregor is the Mother Teresa of sportsmanship*, we want it to approach the truth with the same playful ambivalence. This is a delicate balance for even humans to achieve, and we will quantify here the extent to which LLMs are also capable of producing an effective mix of positivity and believability.

The next section considers a dataset of human XYZs harvested from the web, which will serve as a baseline for char-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

acterizing the XYZs that are elicited from LLMs such as *GPT4o-mini* (Chen et al. 2024). We elicit XYZs using both a neutral prompt and an explicitly ironic one; this allows us to compare and contrast XYZs gathered with each, to quantify whether the ironic variety is less biased, more diverse, and more pragmatically insincere than the neutral variety. An analysis of the LLM's self-ratings of positivity and believability will reveal whether the LLM can appreciate good examples of irony for itself and, just as importantly, filter its weaker efforts. We then look at a range of other LLMs to see if our findings for *GPT4o-mini* hold for these others too, or whether different LLMs need specific workflows for irony. We cannot reliably elicit irony from an LLM without first framing what we mean by irony, but as we will see, thinking about irony from an LLM's perspective can help us to better frame our own definition of what it means to be ironic.

Left and Right: Human Production of XYZs

As a vehicle for studying creativity, figurative XYZs combine the best features of metaphors, similes and analogies (Veale 2012). Much like metaphors, they allow entities to be mapped within or across domains (e.g. politics to sports, science to art), creating mappings that are highly original, as in "the potato is the Tom Hanks of the vegetable world" (versatile and down-to-earth) or "Red meat is the Donald Trump of cancer" (an aggressive builder), or conservative but functional, as in "Serena Williams is the Roger Federer of women's tennis." Like similes, they have a marked form, *X is the Y of Z*, that enables them to be harvested at scale from a corpus or the web. As analogies, they establish mappings from entities in a source to a target domain, and name one of those domains explicitly (e.g. *vegetables, cancer*).

(Veale 2012) introduced a corpus of figurative XYZs that was harvested from the web using the Google API. A set of search queries was first generated by identifying the most productive Ys for possible XYZs in the Google n-grams 1T database (Brants and Franz 2006), which was scanned for all 3-grams of the form "the Y of" (as in "the Mozart of") and 4-grams of the form "the Y_f Y_s of" (as in "the Bill Gates of"). Each matching Y was then used to build a web query of the form "* is the Y of *", and any text snippets returned by Google were parsed to extract matching values for X and Z. This process yielded a corpus of 2,196 different XYZs, ranging over 1,985 different Xs (of which only 115 occur in more than one XYZ) and 665 different Ys (of which 503, or 75%, occur more than once). The most frequent Y, *Chuck Norris*, was the butt of a popular joke cycle in 2012, and was used in 22 XYZs, to describe Xs ranging from the fictional *Jack Bauer* and *Darth Maul* to the very real *Thomas Edison*.

Time and place are a frequent basis for the Z dimension of an XYZ. Overall, geographical distinctions motivate 16% of Zs, while historical eras make up 14% (the 21st century being the most common with 71 uses). These XYZs can be surprisingly rich, as in "Courtney Love is the Yoko Ono of the nineties" (which implicitly maps *Nirvana* to *The Beatles* and *Kurt Cobain* to *John Lennon*). Political orientation is also commonly reflected in the Z dimension, as in *the left* (22 uses), *the right* (13), *the Republican party* (25), *the GOP* (9), and *the Democratic party* (20). Indeed, as 21% of all Xs and 19% of all Ys in the corpus also hark from the political realm, politics is the strongest driver of mappings in the data.

Gender plays a niche role. Just 6 XYZs use a domain defined by gender (e.g., women's tennis), while the data overall is dominated by males, who make up 79% of Xs and 86% of Ys in a roughly 6 to 1 bias against females. A small minority (7%) of Xs are inanimate (as in "Nintendo is the Ned Flanders of the console world") or non-human (as in "Pit bulls are the Mike Tyson of the K9 world"). Given this abundance of male Xs and Ys, we expect most human XYZs to be gender-preserving, but the observed rate of conservation (91%) is much higher than that expected by chance alone (75%). A subsequent analysis of the XYZs created by LLMs will bear out whether generative AI exhibits the same bias.

The human data shows conservatism in other respects too. When each X and Y is labeled with a coarse domain, such as *politics*, *sport*, *religion*, *music*, *business*, *crime*, *showbiz* and *art*, we note that 56% of XYZs stay within the same domain, so that e.g., politicians are most often mapped to politicians. These XYZs tend to make playful hops, not creative leaps, although the pick of the crop can still be jolting and witty, as in this rare cross-gender mapping: "The Queen is the Jerry Springer of the UK". Since the data was harvested from all corners of the web, often from pages that no longer exist, we have no systematic means of knowing whether any XYZ was intended ironically. For example, "Pac Man is the King Lear of the 1980's 8-bit video game revolution" is certainly hyperbolic, but it may also express a sincerely-held awe for retro gaming. In a later section we will explore a means of quantifying the scope for irony in these human XYZs.

Sweet and Sour: LLM Production of XYZs

When quizzed about XYZs, LLMs such as *GPT4o-mini* will show a prior knowledge of the construct, due in large part to the quantity of academic content that makes up their training data. Nonetheless, it is worth explaining the concept to the LLM with some illustrative examples before we instruct it to generate its own. To start, we give the LLM this system instruction: "You are a creative assistant that invents witty and imaginative metaphors. You may be critical. Be fearless in your criticisms." We then explicitly explain XYZs as follows when tasking the LLM with creating its own: "An XYZ comparison is a creative way of describing a target X in the domain Z as an entity Y from a different domain, as in 'Bill Gates is the Thomas Edison of the 21st century' or 'Roger Federer is the Michael Jordan of tennis.' The Y is always a well-known individual that can be either real or fictional."

Ideally, the LLM will have complete freedom in its choice of X, Y and Z. To minimize repetition and to maximize diversity, we cue up a wide range of coarse domains in which it can operate, such as the world of *sport, contemporary life, literature, science, the arts, popular culture, finance, highbrow culture, music*, and *the environment*. We provide 50 of these cues to nudge the LLM to explore different spaces, allowing it to choose its own X and Y, and its own specific Z, in each case. The LLM might e.g. choose *hockey, chess, UFC* or *competitive Scrabble* as a Z for the cue *the world of sport*. We ask the LLM to generate 20 XYZs for each cue, or 1000 in total, using this prompt: "Please generate 20

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

XYZ metaphors where X is a famous presence in <cue>". As the LLM is not asked to be either ironic or sincere, we refer to this set as the *neutral* dataset. To overtly elicit ironic XYZs, we must first explain what we mean by irony, by inserting the following into the LLM's context: "In an ironic XYZ the choice of Y is very surprising and highlights characteristics that are lacking in X, or deserving of criticism in X, such as 'Vladimir Putin is the Dalai Lama of world politics.'" The task prompt is then modified to request 20 "ironic XYZ metaphors" using each of the same 50 cues. We refer to this second collection of 1000 XYZs as our *ironic* dataset.

The LLM shows broad diversity in each dataset. Although prompted with just 50 general cues, the neutral dataset defines .68 unique Zs per XYZ, while the ironic dataset defines .79. The neutral dataset uses .77 unique Xs and .42 unique Ys per unique XYZ, while the ironic set uses .70 and .39 respectively. In contrast, the web dataset (of human XYZs) defines .90 unique Xs but just .30 unique Ys per unique XYZ. However, it better reflects our ironic and neutral datasets in its diversity of Zs, defining .66 unique Zs per unique XYZ.

The gender balance is rather more equitable in these LLM datasets, even though a male bias persists. The neutral and ironic datasets each place twice as many males as females in X and Y positions, which is an improvement on the 6 to 1 bias observed in the web XYZs produced by humans. We take some encouragement from this, given that the LLM likely encountered many of the latter in its training, and was not overtly tasked with gender parity. But the LLM is also averse to cross-gender mappings, as 80% of neutral XYZs and 64% of ironic XYZs are same-sex mappings (we expect just 60% and 54% by chance alone), even if the ironic dataset does seem a little more adventurous in this regard.

Ironic XYZs also show a greater propensity for creative leaps across domains. While 44% of neutral XYZs remain in-domain (so that X and Y have the same coarse domain label), just 18% of ironic XYZs are confined to the same general domain. This is in sharp contrast to the 55% of human XYZs that do not venture across domain boundaries. In the ironic dataset, politicians are more likely to be compared to fictional figures in the Drama domain than to actual politicians, as when Vladimir Putin is described as the *Darth Vader of global diplomacy*. While neutral XYZs seek to maximize the relevance of Y to X via domain similarity, it seems that ironic XYZs instead aim for what (Attardo 2000) calls "relevant inappropriateness." In these cases, Y is chosen as a contrast for X, not as a comparison, highlighting what is lacking or deficient in X. It is this seeming inappropriateness of Y that also alerts us to a possible ironic intent.

Discerning Intent: Positivity and Believability

A successful ironic comparison requires more than just the speaker's intent; it also requires an audience's recognition of this intent. Irony veils criticism with praise, but if this praise is taken at face value its critical subtext may be lost. If, however, the praise seems unbelievable, an audience is forced to dig for a deeper, more elusive and less positive meaning. A clearly ironic XYZ is an alloy of high positivity and low believability that damns a target with fierce, not faint, praise.

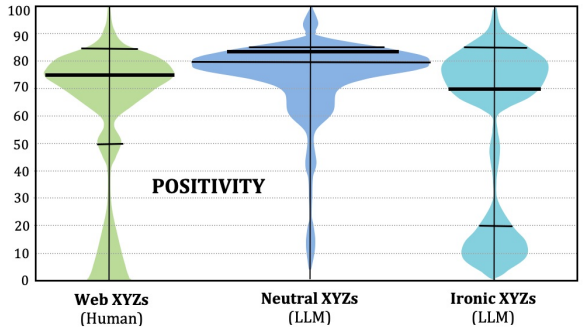


Figure 1: Violin plots of LLM positivity ratings for XYZs in 3 datasets: *Web (human), neutral (LLM)* and *ironic (LLM)*. Thick bars mark medians. Thin bars mark 1st & 3rd quartiles.

For an LLM to confidently predict that a comparison will be read as ironic, and to self-filter when this confidence is low, it must be capable of rating the positivity and believability of an XYZ for itself. To elicit numeric ratings for positivity, we prompt the LLM as follows: "Rate the positivity of this comparison on a scale of 0 (no positivity at all) to 100 (maximum positivity), returning just a number: <XYZ>." Since each request is issued in a new context, the LLM's responses are unaffected by previous interactions. The distributions of positivity scores for XYZs in all three datasets (*web, neutral* & *ironic*) are visualized in the violin plots of Fig. 1. Median positivity is high for each set (75, 85 and 70, respectively).

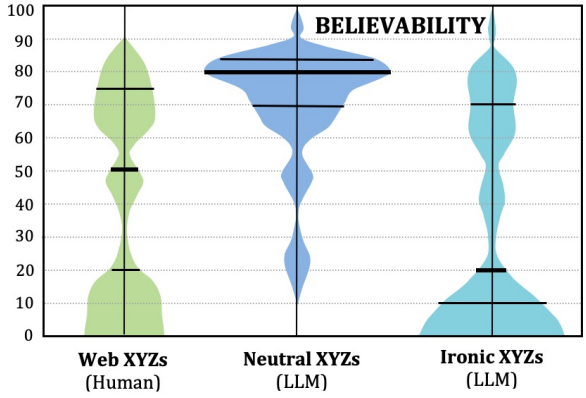


Figure 2: Violin plots of LLM believability ratings for XYZs in 3 datasets: *Web (human), neutral (LLM)* and *ironic (LLM)*

We replace "positivity" in our prompt to instead elicit believability ratings from the LLM. The violin plots of Fig. 2 show a spread of median believability scores across datasets: *neutral* XYZs seem the most credible at 80, *ironic* XYZs seem the least credible at 20, and *web* XYZs sit in the middle at 50. These spreads match our expectations, yet we must also allay some concerns. First, is the LLM generating objective ratings for each XYZ, or is it simply hallucinating outputs of the expected form (integers between 0 and 100)? Second, does the LLM perceive the implicit criticism of ironic XYZs, and reflect this criticism in its positivity scores?

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

To address the first, we modify the prompt to elicit further ratings for *aptness*, *fairness* and *sincerity* from the LLM for each dataset. These are not independent qualities and so we expect, if the LLM's ratings are *not* arbitrary, to see strong correlations between them. We see that Pearson's r for *fairness & believability* ranges from .57 (neutral) to .63 (web) to .73 (ironic), while for *sincerity & believability* the range is .57 (neutral) to .68 (ironic). The correlation between *aptness & believability* also reflects the similarity of these qualities, at .7 (neutral) to .76 (ironic). To probe a dissimilar quality, we elicit ratings for *incongruity*, and note negative correlations of $-.28$ (neutral) to $-.43$ (ironic) with *believability*, and $-.31$ (neutral) to $-.41$ (ironic) with *fairness*. The LLM shows itself to be consistent in its grasp of these qualities.

The insincerity of an ironic text is crafted to be transparent to an audience, who should see past the positive veneer to find the veiled criticism within. However, the positivity ratings in Fig. 1 suggest this is not the case for the LLM, which only appears to see the superficial positivity of an XYZ. This is also the case when the LLM is prompted to instead identify the key emotion in each XYZ, as follows: "If someone claims that <XYZ> what is the most likely emotion they feel toward <X>?" The dominant emotion across all three datasets is *Admiration*: neutral (95% of XYZs), web (72%) and ironic (69%). Although the LLM itself created the ironic cases, it proves to be a naive audience for its own efforts.

The exception is when the irony is so blatant that the LLM identifies the emotion as *Sarcasm*, which it does in 26 cases. Of these, one XYZ is an outlier with a positivity of 85 and a believability of 10: "Narendra Modi is the Stephen Hawking of humility." The remaining 25 cases have a mean positivity of 9.5 ($\sigma^2 = 2.8$) and a mean believability of 8.5 ($\sigma^2 = 2.4$). The sarcasm of examples such as "McDonald's is the Marie Curie of healthy eating" or "Benjamin Netanyahu is the Malala Yousafzai of peace activism" is reinforced by the clear mismatch of X and Z, in addition to that of X and Y.

Isn't It Ironic? It Depends On Who You Ask

Irony is a folk phenomenon that speakers require no formal definition to use, and this shows in the diversity of uses to which the tag *#irony* is used on social media. To some, it can mean cynical wit, or sardonic hyperbole, or poetic justice, or blatant sarcasm, or a mocking reminder, or mere hypocrisy. So it is unsurprising that an LLM trained on web data should show the same definitional fuzziness in its own ironic efforts.

We see this diversity in many XYZs of the ironic dataset:

1. Conor McGregor is the Mother Teresa of sportsmanship. (*positivity:* 10, *believability:* 10, *emotion:* Admiration)
2. Lady Gaga is the Jane Austen of understatement. (*positivity:* 50, *believability:* 10, *emotion:* Admiration)
3. Industrial agriculture is the Godzilla of the countryside. (*positivity:* 20, *believability:* 60, *emotion:* Anger)
4. Elon Musk is the Tony Stark of real-world innovation. (*positivity:* 90, *believability:* 90, *emotion:* Admiration)
5. Tim Cook is the Alfred Pennyworth of Apple. (*positivity:* 80, *believability:* 75, *emotion:* Admiration)

Examples 1 and 2 offer good illustrations of irony as we have operationalized it here: an outwardly positive comparison is too unbelievable to take at face value, so we look for the criticism veiled within (Conor McGregor is no paragon of sportsmanship, and Lady Gaga is far from understated). Note that the irony in (1) is apparent to the LLM when it rates the positivity of the XYZ as just 10/100, but *not* apparent when it rates its dominant emotion as Admiration. There is no such ambivalence in (3), where the negativity is apparent both numerically (20) and symbolically (Anger). This is caustic hyperbole, but it is not a case of irony. Cases (4) and (5) also lack equivocation and the implied criticism of irony. The mapping in (5), of Tim Cook to Bruce Wayne's butler, implicitly identifies Steve Jobs with Batman. Although there is wit here, this XYZ is too positive and too apt to be ironic (unless one takes a dim view of a beloved comics character).

Magnets for Irony

It seems clear that, by any chosen definition of irony, only a subset of our ironic dataset will actually be ironic. But which subset? If we view the LLM as a "mere generator" (Ventura 2016) of irony candidates, we can use a post-generation pass to filter the true positives relative to how we define irony. As we have defined it here, irony is criticism thinly veiled in insincere praise, or an alloy of high positivity and low believability. The LLM itself will rate these qualities for us. To quantify how effectively they work as a magnet for irony, we will evaluate how well they allow us to discriminate XYZs in the neutral dataset from those of the ironic dataset.

As an naive baseline, we ask the LLM to directly rate the *ironicity*, or the likelihood of an XYZ being ironic, on a scale of 0 to 100 (using a modified version of our earlier prompt). The ROC (*receiver-operator-characteristic*) graph of Fig. 3 shows that this criterion is little better than random choice. The ROC plots specificity versus (1 - sensitivity) for varying ironicity levels, while the dashed line is the random baseline.

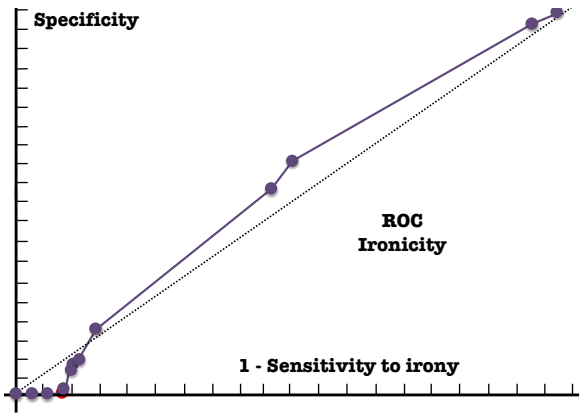


Figure 3: An ROC curve for LLM ratings of "ironicity" as a criterion for separating our *neutral* and *ironic* datasets.

The performance of a wider range of discriminating criteria, again elicited from the LLM, is graphed as ROC curves in Fig. 4. Each criterion is directly rated by the same LLM that

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

generated the XYZs, except for one: *pragmatic insincerity*. This is a composite quality that is calculated as follows:

$$\textit{pragmatic insincerity} = \textit{positivity} \times (100 - \textit{believability})/100$$

This quality, which is directly proportional to *positivity* but inversely proportional to *believability*, reflects our sense that irony conveys false praise with a winking lack of credibility.

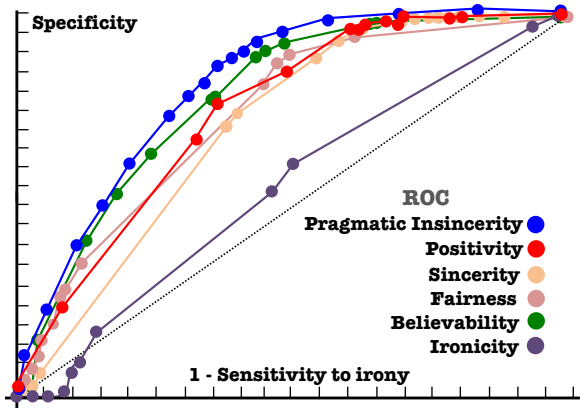


Figure 4: ROC curves for LLM ratings of multiple criteria to discriminate *ironic* dataset XYZs from *neutral* dataset XYZs

As shown in Fig. 4, *pragmatic insincerity* is the most discriminating basis for separating LLM irony from non-irony. We quantify the effectiveness of each criterion by measuring the area under each curve (ROC-AUC), as listed in Table 1.

Criterion	ROC-AUC
Pragmatic Insincerity	**0.8096**
Sincerity	0.7128
Positivity	0.7023
Believability	0.6957
Fairness	0.6484
Ironicity	0.5809

Table 1: The area-under-curve (AUC) for each ROC curve in Fig. 4. For discerning irony in XYZs, *pragmatic insincerity* appears to offer the best blend of sensitivity and specificity.

Although the LLM is a poor judge of its own ironic efforts (when evaluated in a separate context), as reflected in the low ROC-AUC score for *ironicity*, it does show an ability to judge the essential ingredients of irony: praise and disbelief.

Digging Deeper: Finding Meaning in Irony

Pragmatic insincerity is a crucial indicator of a speaker's intended meaning. But why do ironic speakers find it useful to say what they do *not* mean, rather that directly saying what they *do* mean? The echoic mention theory tells us that this is a surprisingly economic way of speaking: by echoing what should be the case, a speaker can implicitly criticize what actually is the case. But if an ironic XYZ packs two meanings in one form, how well can an LLM unpack these meanings?

(Valitutti and Veale 2015; Veale 2018) localize the meaning of irony to specific words. Consider the assertion that "Leonardo was a *genius* in multiple domains," which uses the word "genius" in its default sense of a person of great ability or intelligence. The following tart remark, ascribed to Gore Vidal, uses the same word in a non-default, ironic fashion: "Andy Warhol is the only *genius* I know with an IQ of 75." To measure the effectiveness of a machine-generated ironic metaphor, Valitutti & Veale asked human judges to rate the positivity of focal words such as "genius" in a null context. Other judges later rated the same words in ironic contexts, much like Vidal's acidic put-down. An ironic usage should cause the perceived positivity of a focal word to shift downwards from its default high. The extent of this downshift is a correlate of how well we have conveyed irony.

We do not ask the LLM to estimate this downshift directly; rather, we first ask it to unpack the neutral and ironic meanings of an XYZ and then quantify the downshift as the drop in positivity as one shifts from a neutral to an ironic reading. We adopt a *class-inclusion* view of meaning, in the vein of (Glucksberg and Keysar 1990; Glucksberg 1998; 2008), in which we localize the meaning of an XYZ to the implied class or category into which readers are expected to map both X and Y. For instance, a neutral reading of the XYZ "Tyson Fury is the Cinderella of humility" might elicit the implicit class "Surprising embodiments of humility" as a generalization, while an ironic reading might elicit the class "Athletes known for their arrogance." Each of these classes has been suggested by the LLM in response to a prompt seeking either a neutral or an ironic reading of the XYZ. For this task, we first explain the XYZ construct to the LLM and then set up the problem with a 1-shot example:

(user) "What is the implied category of X that best summarizes this comparison: '*The potato is the Tom Hanks of the vegetable world*'? Provide just the category."

(LLM) "Universally beloved entities."

To elicit an ironic reading, we again explain XYZs to the LLM but now use this prompt and 1-shot example instead:

(user) "What is the implied category of X that best summarizes this ironic comparison: '*Elon Musk is the Marie Antoinette of the tech industry*'? Provide just the category."

(LLM) "High-profile figures known for controversial statements and actions."

Once a neutral and an ironic category is elicited for each XYZ, we prompt the LLM to rate the positivity of each one:

"Rate the positivity, on a scale of 0 to 100, of placing someone in the category <*category*>. Return just a number."

We then calculate the downshift as the fall in positivity as one shifts from a neutral to an ironic reading. For instance, consider the XYZ "Donald Trump is the Albert Einstein of diplomacy." The audience for certain cable-TV news net-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

works may well be inclined to view this comparison literally, and the LLM suggests the neutral reading "Influential figures in their respective fields" with a high positivity rating of 95/100. In contrast, the LLM suggests the category "Unskilled or inept diplomats" as an ironic reading of this XYZ, the positivity of which it rates at just 10/100. The gap between these ratings, 85/100, is the ironic downshift we seek.

The mean downshift from neutral to ironic readings of XYZs in the ironic dataset is 15.1, while for neutral XYZs the mean drop is even greater, at 18.5. However, the largest drop is exhibited by the human (web) dataset, for which the positivity of ironic readings is 25/100 lower on average. But why should this be the case – that XYZs that are not designed to be ironic (our neutral dataset) or not expected to be ironic (our web dataset of human XYZs) show a greater capacity for dual interpretation as both praise and criticism?

There appears to be no single answer here. In some cases, the LLM simply produces the wrong generalizations. For instance, for the XYZ "Charlie Sheen is the Stephen Hawking of intellectual discourse" it suggests the neutral reading "Highly respected thinkers" and the ironic reading "Public figures known for their intellectual prowess." As it correctly assigns the same positivity rating of 90 to each of these, the estimated downshift is 0. At the same time, it rates the positivity of the comparison as just 15 (with the emotion *Disdain*), with a believability of just 10. From one perspective, the LLM sees irony (or perhaps sarcasm) here; from another, it fails to see the shift in meaning that irony should induce.

In contrast, a large potential downshift is only relevant if an audience is minded to seek it out. If an XYZ has high believability, there is no apparent incongruity to drive an ironic re-interpretation. Consider this XYZ from the human (web) dataset: "The Montrachet is the Angelina Jolie of the pack." Perhaps the speaker hopes to convey the idea that the wine is "full-bodied" or "luscious," but in any case the LLM assigns a positivity of 90 with a believability of 70. As a neutral reading it also suggests the category "Highly regarded and sought after" with a positivity of 100. However, when urged to adopt an ironic stance, it instead suggests the category "Overrated or overhyped entities" with a positivity of 0. Yet the availability of a cynical reading does not mean audiences will reach for this meaning unless nudged to do so. For irony to succeed, intent must work hand-in-hand with delivery.

The LLM's self-ratings give us the tools to accept or reject its candidates on the basis of our own acceptability criteria. To filter its efforts at irony, we should demand the following:

1. High positivity (e.g., *positivity* ≥ 60)
2. Low believability (e.g., *believability* ≤ 30)
3. Two contrasting interpretations (*neutral* and *ironic*)
4. A significant downshift (e.g., *downshift* ≥ 25)

As reported in Table 2, only a small subset of the XYZs in each dataset meet the above criteria. Just 1 in 20 of the XYZs in the ironic dataset conform to our ideal of a successful ironic comparison, that is, of a comparison that will both strike the audience as ironic and that will reward a deeper search for a veiled criticism. In contrast, more than 1 in 10 of the human XYZs from the web conform to this ideal.

Selection Criterion	Web	Neutral	Ironic
High positivity ≥ 60	74.77%	**96.96%**	62.13%
Low believability ≤ 30	41.21%	9.80%	**54.17%**
Pos. ≥ 60 & Believ. ≤ 30	**22.40%**	8.04%	20.56%
Two contrastive readings*	**43.35%**	30.29%	30.00%
Significant downshift ≥ 25	**41.48%**	29.41%	25.00%
All criteria together	**10.61%**	3.63%	5.46%

Table 2: % of XYZs in each set that meet our irony criteria. **(absolute downshift ≥ 25)*

Fortunately, fore-warned is fore-armed, and a creative system that uses an LLM as its generator can filter the 95% of LLM efforts that fall short of its desired standard for irony. Nonetheless, we note that the GPT LLM achieves a success rate at irony that is less than half that of human XYZs from the web when it is explicitly prompted to be ironic.

Other Language Models

A single LLM, OpenAI's *GPT4o-mini*, has driven our analysis up to this point. This is a commercial LLM whose exact parameter count is a proprietary secret, although the number is widely believed to be 8 billion. This model is comparable in benchmarked performance to OpenAI's much larger GPT-4, which is reported to have over a trillion parameters. In this section we broaden our analysis to include a range of rival LLMs, a number of which are open-source: *Llama-3.3* (70 billion parameters, from *Meta*), *Nemo-Instruct-2407* (12 billion, from *Mistral*), *Qwen2.5-72B-Instruct* (72 billion, from *Ali Baba*), *Gemma 2* (27 billion, from *Google*) and *DeepSeek*'s *R1* distilled into a Qwen model (32 billion).

XYZs are elicited from these LLMs using the same neutral and explicitly ironic prompts as were used for *GPT4o-mini*. The five new datasets cover the same range of macro-domains (*sport*, *politics*, *science*, etc.) and are self-rated for *positivity*, *believability* and so on using the same prompts. We observe a similar gender split in these new collections. The ratio of male to female Xs ranges from 1.8:1 *(Gemma)* to 2.6:1 *(Llama)*, while for Ys the male:female ratio ranges from 2.2:1 *(Llama)* to 4.6:1 *(R1)*. Conservation of gender is again evident across the board, with 80% *(Mistral)* to 92% *(Gemma)* of XYZs mapping Xs to Ys of the same gender. Neutral XYZs also tend to be intra-domain, with conservation rates ranging from 70% *(Llama)* to 94% *(Qwen)*, while ironic XYZs are much more likely to connect two domains.

As before, we use LLM estimates of *positivity*, *believability* and *downshift* to identify the XYZs in each dataset that conform to our ideal of a successful ironic comparison. As reported in Table 3, the highest yield is obtained from *Llama 3.3 (70B)*, with *Gemma 2 (27B)* coming a close second. According to our standard of meaningful pragmatic insincerity, *Llama 3.3* is four times more reliable than *GPT4o-mini* as a generator of appreciable, two-meanings-at-once irony. This is not easily attributable to the LLM's size, as the 72B *Qwen* model only achieves par with the much smaller *GPT4o-mini*.

Nor is this gap due to amount of compute that each LLM

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Selection Criterion	Llama 70B	Gemma 27B	Mistral 12B	Qwen 72B	R1 *Qwen* 32B
High positivity ≥ 60	38.5% *(97.5%)*	53% *(98%)*	48.5% *(93.6%)*	47.3% *(97.5%)*	**76.5% *(97%)***
Low believability ≤ 30	**87.5% *(4%)***	70.5% *(0%)*	49.5% *(2.7%)*	65.5% *(1.5%)*	40% *(9.0%)*
Positivity ≥ 60 & Believability ≤ 30	**27% *(3.5%)***	26% *(8.0%)*	10.5% *(1.4%)*	16.8% *(1.0%)*	20.5% *(8%)*
Two readings (abs. downshift ≥ 25)	**63% *(35.5%)***	59% *(22%)*	60.5% *(28.2%)*	44.1% *(28.5%)*	41.5% *(40%)*
Significant downshift ≥ 25	**60% *(35.5%)***	58.0% *(22%)*	55.5% *(27.7%)*	40.9% *(27.5%)*	40.5% *(39%)*
All of the above criteria together	**19.5% *(1.5%)***	17.5% *(0%)*	8.0% *(0.45%)*	5.00% *(1.0%)*	12.5% *(5.0%)*

Table 3: % of XYZs that meet various irony criteria, as elicited using explicitly ironic prompts to each LLM. The neutral baseline, in which the LLM is *not* explicitly prompted for irony, is shown in parentheses.

dedicates to the generation task. DeepSeek's *Qwen* distillation of its *R1* model (DeepSeek-AI et al. 2025) expends significantly more time and compute when generating XYZs, and is an order of magnitude slower than the other LLMs. It pursues a deep chain-of-thought (CoT) approach in which the LLM reasons about the stated problem (Wei et al. 2022); this adds an average of 500 tokens to the generation of every 20 XYZs. While we expect ironic speakers to carefully plan their utterances, and to anticipate how they might split their audiences, this LLM – which ranks third in Table 3 – appears to *over*-think the problem and over-explain its reasoning.

The following is a representative extract from its CoT process: "Hmm, it's a start, but some of these don't fully capture the irony I was aiming for. I need to think more carefully about the contrasts between X and Y." These traces show the model doubting and criticizing itself, as we expect any good creator (and not just a *mere generator*) to do. However, self-criticism is valuable when its spurs insights, not excuses. Ultimately, the LLM cannot translate its insights into practice, and offers this apologia in its CoT trace: "Overall, creating truly ironic metaphors requires a delicate balance between the expected and unexpected elements. It's a bit challenging, and some of my attempts might not land perfectly, but this exercise helps in understanding how to juxtapose different domains to create thought-provoking comparisons."

So why do the less-reflective *Llama* and *Gemma* LLMs perform so much better? Perhaps certain creative goals are diminished when they are viewed as problems to be solved rather than as opportunities to invent. It may also boil down to how, and on what, these models are trained, and to how well they are tuned to respond to instructions (Ouyang et al. 2022) or to read between the lines of a user's request. We expect creative tasks to vary as to how heavily they rely on the two competing mental systems that (Kahneman 2011; Evans 2003) dub *System I* and *System II*. In this dual-process perspective on the human mind, *System I* is reactive, immediate, and instinctively intuitive; that is, it acts, or rather *reacts*, much like a well-trained, fine-tuned language model. Its training and biases shape what it does and what it says. In contrast, *System II* is slow, cautious and analytical, much like the chain-of-thought approach of DeepSeek's *R1/Qwen*. Our results suggest that generation (*System I*) should be separated from appreciation (*System II*) in two distinct phases. This allows an LLM to be freely and divergently generative in the first step, and convergently self-critical in the second.

Conclusions

We have sidestepped the question of whether an LLM is a creative producer of ironic comparisons, or just a "mere generator" of candidates (Ventura 2016). The larger, and better, question is whether a system built around an LLM can be creative in its selective generation of linguistic artifacts. Certainly, if prompted with care, LLMs such as *GPT4o*, *Llama* and *Gemma* are capable of fluency, flexibility, specificity and even originality – all of the dimensions of divergence identified by (Guilford 1950; Runco 2010) – in the creation of new texts. One might say that the LLM is merely generative, while the larger system that employs it is potentially creative, but this underestimates the value of LLMs as creative producers. As we have shown, the LLM itself can play a key role in evaluating and filtering its own outputs, to become a deliberate creator of meanings in its own right.

Irony is a facet of linguistic creativity that can be modeled using a broad spectrum of computational approaches. At one end of this spectrum sit the symbolic, rule-based approaches to generating irony, as typified by (Hao and Veale 2010). At the other sit the data-driven approaches to detecting irony, such as the statistical models of (Reyes, Rosso, and Veale 2013) or the neural models of (Ghosh and Veale 2017). The latter have the benefit of working with labeled data, allowing for a crisp definition of success and failure. The former, in contrast, must contend with what it means to be successfully ironic, and perhaps even creative (Colton and Wiggins 2012; Ritchie 2007); for that, a sense of humour may be needed also (Jentzsch and Kersting 2023; Góes et al. 2023). Here we have chiefly focused on what has been called *"theory of mind"* (ToM) in the context of irony (Strachan et al. 2024): the ability of a speaker to intuit how an audience will react to an ironic provocation, and the ability of an audience to infer the speaker's intended meaning. LLMs seem to show a poor grasp of ToM when aiming for irony, as only a minority of their XYZs clearly convey pragmatic insincerity. However, when probed on the dimensions of pragmatic insincerity, they show a firm enough grasp of these dimensions (as illustrated in Figure 4) to allow them to act intentionally. An LLM can serve a powerful, oracular function in AI systems, but as with any oracle, one must be careful what one asks, how one asks it, and how one interprets its replies.

Acknowledgments

This paper and its datasets are wholly the work of the author.

References

Attardo, S. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics* 32(6):793–826.

Brants, T., and Franz, A. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium.

Chen, M.; Menick, J.; Lu, K.; Zhao, S.; Wallace, E.; Ren, H.; Hu, H.; Stathas, N.; and Such, F. P. 2024. GPT-4o-mini: Advancing cost-efficient intelligence. *OpenAI Blog, July 18*.

Clark, H. H., and Gerrig, R. J. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General* 113(1):121–126.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: the final frontier? ECAI'12, 21–26. NLD: IOS Press.

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: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv* 2501.12948.

Evans, J. 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7(10):454–459.

Garmendia, J. 2019. *Irony. Key Topics in Semantics and Pragmatics*. Cambridge, UK: Cambridge University Press.

Ghosh, A., and Veale, T. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Glucksberg, S., and Keysar, B. 1990. Understanding metaphorical comparisons: Beyond similarity. *Psychological Review* 97(1):3–18.

Glucksberg, S. 1998. Understanding metaphors. *Current Directions in Psychological Science* 7:39–43.

Glucksberg, S. 2008. How metaphor creates categories – quickly! In Raymond W. Gibbs, J., ed., *The Cambridge Handbook of Metaphor and Thought*. Cambridge, UK: Cambridge University Press.

Grice, P. 1978. Logic and conversation. In Cole, P., and Morgan, J., eds., *Syntax and Semantics 3: Speech Acts*. New York: Academic Press. 41–58.

Guilford, J. 1950. Creativity. *American Psychologist* 5:444–454.

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*.

Hao, Y., and Veale, T. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds and Machines* 20(4):483–88.

Jentzsch, S., and Kersting, K. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. *arXiv* 2306.04563.

Kahneman, D. 2011. *Thinking, fast and slow*. London, UK: Penguin Books.

Kreuz, R. J., and Glucksberg, S. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General* 118(4):374–386.

Kumon-Nakamura, S.; Glucksberg, S.; and Brown, M. 1995. How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General,* 124(1):3–21.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; and et al., C. L. W. 2022. Training language models to follow instructions with human feedback. In *Proceedings of NeurIPS-22, the 36th Conference on Neural Information Processing Systems*.

Picasso, P. 1923. Picasso speaks: A statement by the artist. *The Arts: An Illustrated Monthly Magazine Covering All Phases of Ancient and Modern Art* 3(5).

Reyes, A.; Rosso, P.; and Veale, T. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation* 47(1):1–30.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Runco, M. 2010. Divergent thinking, creativity, and ideation. In Kaufman, J. C., and Sternberg, R. J., eds., *The Cambridge handbook of creativity*, 413–446. Cambridge, UK: Cambridge University Press.

Sperber, D., and Wilson, D. 1981. Irony and the use-mention distinction. In Cole, P., ed., *Radical pragmatics*. New York, NY: Academic Press. 295–318.

Sperber, D., and Wilson, D. 1992. On verbal irony. *Lingua* 87(1-2):53–76.

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.

Valitutti, A., and Veale, T. 2015. Inducing an ironic effect in automated tweets. In *Proceedings of the 6th International Conference on affective computing and intelligent interaction (ACII2015)*. Xi'an, China: IEEE Computer Society. 153–159.

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.

Ventura, D. 2016. Mere Generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity, Paris, France*.

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 *Proceedings of NeurIPS-22, the 36th Conference on Neural Information Processing Systems*.

ISBN: 978-989-54160-7-3

Refining Metrical Constraints in LLM-Generated Poetry with Feedback

Manex Agirrezabal
Centre for Language Technology (CST)
University of Copenhagen, Denmark
manex.aguirrezabal@hum.ku.dk

Hugo Gonçalo Oliveira
University of Coimbra, CISUC/LASI
Centre for Informatics and Systems of the University of Coimbra
Department of Informatics Engineering, Portugal
hroliv@dei.uc.pt

Abstract

Among many tasks, general purpose pretrained Large Language Models (LLMs) can be prompted for generating poetry in a zero-shot scenario, with relative success, even if, in many occasions, they fail to meet all the specified constraints. We prompt LLMs for producing poetry with a specific number of lines and syllables, and evaluate the result with another LLM or with a rule-based tool for poetry analysis. If specified constraints are not met, LLMs are prompted again with feedback and a request for generating a new poem that satisfies the target constraints. This can go through several iterations. With this setup, we not only analyse the ability of LLMs to produce metrical poetry, but we test to what extent they can: (i) improve their generations with feedback; and (ii) analyse metrical constraints. We conclude that feedback is effective in improving the metrical constrains, but this can be done at the cost of less rhyme. We also observe that feedback by the rule-based system is preferable than feedback by an LLM.

Introduction

Due to their flexibility and ease of use, general purpose instruction-tuned Large Language Models (LLMs) are being used in a broad range of text-based tasks, including the generation of linguistically-creative artefacts. For many of such artefacts, the first impression is that these models excel at their task. Specifically, LLMs can generate poetry from natural language instructions that may cover a range of features, including content and form. These will be reflected in the produced text, often further embellished with typical poetic features like metaphor and rhyme. Figure 1 serves as an example of a single prompt that produces a poem.

Despite this success, it has been noted that LLMs still have limitations regarding diversity in the generation of several poems (Zhang and Eger 2024; Chen et al. 2024), and that they do not always meet all target formal requirements, such as the number of characters per line (Yu et al. 2024), or number of lines and syllables (Agirrezabal and Gonçalo Oliveira 2024). At the same time, current instruction-tuned LLMs can keep a conversation, and will consider past context in their responses. So, their outputs can be refined through feedback (Pan et al. 2024), if necessary, for a number of interactions.

```
> user: Please create a poem. The poem should
contain 8 lines about nature. Return only the poem,
nothing else.

> assistant: Sunlight dapples through leaves of
green,
A gentle breeze whispers secrets unseen.
Flowers bloom in vibrant array,
As birdsong fills the air all day.
A babbling brook, a crystal stream,
Reflects the sky's ethereal gleam.
Nature's canvas, vast and wide,
A masterpiece where wonders reside.

> user: Thank you!

> assistant: You're welcome! I'm glad you liked the
poem. :-)
```

Figure 1: Example interaction where the LLM Gemma2 is prompted to generate a poem with 8 lines about *nature*. The value for temperature was 0.7.

In this work, we explore automatic feedback for guiding the generation of metrical poetry with LLMs. For generation, we test two recent LLMs, Llama3.1-8B (Grattafiori et al. 2024) and Gemma 2-9B (Gemma Team et al. 2024), which are prompted for generating poetry in a set of topics, with variable constraints regarding the number of lines and syllables per line. Generated poems are evaluated, either by the same models, in a different session, or by an external rule-based tool for poetry analysis, Erato (Agirrezabal, Gonçalo Oliveira, and Ormazabal 2023). If target constraints are not met, the generators are prompted with feedback together with a request for generating a new poem.

We observed that, regardless of the source, providing feedback results in poems that follow target metrical constraints more closely. Moreover, the feedback based on the rule-based tool is more effective than the feedback by the LLMs. However, we also observed that updating the poems to satisfy the syllable count constraints is sometimes done by adding words to the end of lines, breaking a previously compliant rhyme. On the other hand, the diversity of poems generated for the same constraints is not affected, suggest-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

ing that LLMs do not update poems using fixed lexical items or expressions.

Our contributions are summarised as follows:

- We explore feedback for improving metrical constraints in LLM-generated poetry;
- We show that, regardless of the evaluation by another LLM or by a rule-based tool for poetry analysis, feedback is always effective;
- We analyse the impact that meeting metrical constraints has on other features like rhyme and diversity;
- We create a dataset of over 3,200 sessions and their LLM-generated poems, which may be used for further research on this domain.

The paper is structured as follows. First, we provide some background on poetry generation, with a focus on LLM-generated poetry, and approaches for LLM feedback. After that, we present the adopted methodology for poetry generation with feedback. This is followed by a detailed description of the experimentation setup, covering the tested models, parameters and prompts used. Before concluding, we present and discuss the main results of our experimentation.

Background

From rule and knowledge-based to evolutionary algorithms, a broad range of techniques has been explored for generating poetry computationally (Gonçalo Oliveira 2017; Lamb, Brown, and Clarke 2017). Yet, as it happens to other text-related tasks, nowadays, transformer-based models, mostly LLMs, are commonly used for this purpose (Elzohbi and Zhao 2023). These can be fine-tuned for generating poetry in target forms or in the style of a specific author (Sawicki et al. 2023), but, with the size of LLMs increasing, they are also getting better at performing tasks for which they were not specifically trained for. And, since prompting an LLM is much more accessible than fine-tuning it, the latter is becoming the most popular approach.

However, the simplicity of LLM-based poetry generation does not come at no cost, as authors have noted several limitations, including the lack of diversity (Zhang and Eger 2024; Chen et al. 2024), or a difficulty to meet formal constraints (Yu et al. 2024; Agirrezabal and Gonçalo Oliveira 2024). Some (Walsh, Preus, and Gronski 2024) further discuss that a model like ChatGPT has its own poetic style, more constrained, with higher trends to use specific words, more rhymes and predefined forms. In theory, increasing the temperature would lead to more diverse outputs. However, as shown for narrative generation (Peeperkorn et al. 2024), temperature is not more than weakly correlated with novelty, and it is moderately correlated with incoherence.

Indeed, LLMs are known to be inconsistent (Jang, Kwon, and Lukasiewicz 2021). A possible way of minimising this, as well as other reliability issues, is to interact with the LLM while providing feedback. Several methods have been proposed for correcting the output of LLMs through feedback (Pan et al. 2024), which may be provided by the LLM itself, also known as self-feedback (Madaan et al. 2024); or externally, by a human, another LLM (Yang et al. 2022), specific tools (Zhang et al. 2023; Jiang, Wang, and Wang 2023), structured knowledge (Peng, Zhang, and Shang 2024), or a combination of different sources (Nathani et al. 2023).

Feedback can be provided during or after generation (post-hoc), but the latter is more flexible, as it does not require additional training, and allows feedback to be provided in natural language, through simple prompting, if necessary, across several iterations. Among others, post-hoc feedback provided by an external tool was applied to code generation (Zhang et al. 2023; Jiang, Wang, and Wang 2023) and question answering (Nathani et al. 2023; Chern et al. 2023). In the domain of creativity, post-hoc feedback was applied to story generation (Yang et al. 2022), by a model trained for reranking produced passages.

Yet, to our knowledge, feedback mechanisms have not been applied to poetry generation with an LLM. In this task, feedback could be provided by a human or by computational tools for poetry scansion and analysis in general (De Sisto et al. 2024). In this domain, we highlight two frameworks developed in recent years, namely CARMEN (Possi et al. 2023), focused on a single metric for evaluating poems considering features like meter and rhyme; and Erato (Agirrezabal, Gonçalo Oliveira, and Ormazabal 2023), which provides a set of tools for analysing poetic features in detail, also covering novelty and fluency. Moreover, LLMs are being studied in the detection of poetry form (Walsh, Antoniak, and Preus 2024). They have shown to successfully identify both common and uncommon fixed poetic forms, but performance varies significantly when it comes to unfixed forms.

Methodology

In order to refine the metrical constraints of poems generated by prompting an LLM, we adopt iterative post-hoc feedback. Figure 2 summarises the proposed approach.

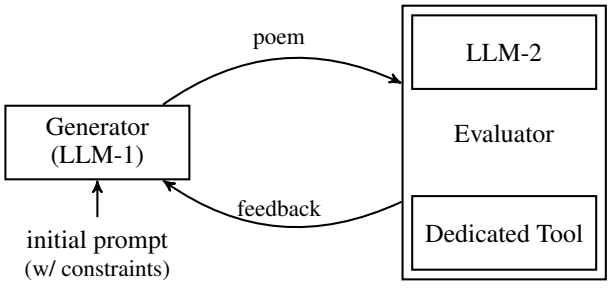


Figure 2: LLM-based poetry generation with Iterative Post-Hoc Feedback. The Generator is prompted for producing a poem with specific constraints; the poem is analysed by the Evaluator; if constraints are not met, feedback is provided to the Generator, with a request for producing another poem.

Briefly, an LLM is prompted for generating a poem with specific constraints. This poem is then evaluated, either by another LLM or by a tool for poetry analysis. If any of the target constraints is not met, feedback is provided to the generator LLM, as a new natural language prompt. This is per-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

formed until the constraints are met or a predefined number of iterations is reached.

If feedback is effective, the poem produced in the final iteration will be better-aligned with the target constraints. However, satisfying such constraints might have an impact in other poetry-relevant features. In this paper, we focus on two metrical constraints, number of lines and number of syllables, but also analyse the impact that meeting such constraints has on the produced rhymes and on diversity.

Experimentation Setup

This section describes how the proposed methodology was implemented, thus enabling experimentation towards useful conclusions. It describes the models used for generation and evaluation, the metrical constraints and topics considered, as well as the prompts used for interacting with the models.

Models

We experimented with two different open LLMs of similar size, Llama3.1-8B (Grattafiori et al. 2024) and Gemma2-9B (Gemma Team et al. 2024), both as generators and as evaluators. They were accessed in a quantised version, through the Ollama tool[1]. Additionally, we used Erato as an evaluator, thus resulting in six (2×3) different generator–evaluator combinations.

Target Constraints

The selected LLMs were prompted to produce poems with a diverse number of lines, namely 4, 7, 8 and 14 (four different possibilities). Experimentation with syllable counts was performed only for the 4-line poems. For these poems, the values tested for the number of syllables in each line were 8, 10 and 13 (three different possibilities).

Target Topics

Since poems are typically written with a goal or a topic in mind, poems were generated for a target topic. The tested topics were selected before experimentation, with the help of Llama3.1, using the following prompts:

```
Give me three single word topics that are
very common in poetry
```

```
Give me frequent words that are not very
common in poetry
```

This resulted in the following six topics: *love*, *death*, *nature* (from the first prompt), *router*, *briefcase*, *garbage* (from the second prompt). For each generator–evaluator combination, the generator LLMs were prompted for creating poems in each of the six topics. Additionally, we tested the generation without a target topic, which resulted in seven different possibilities for this option.

[1] https://ollama.ai/

Interaction

Poems are produced in sessions, each starting by prompting the generator LLM for creating a poem with target constraints. This is followed by the evaluation of the produced poem, which might be performed by prompting another LLM. If specified constraints are not met, feedback is provided to the generator LLM as another prompt, together with a request for creating a new poem, in what constitutes a new iteration. The session ends when all the constraints are met, or after 10 iterations. Here, we present the prompts used for each subtask of this process.

Generation For poem generation, the generator LLMs were prompted according to the following templates:

```
Please create a poem. The poem should
contain N_LINES lines TOPIC. Return only
the poem, nothing else.
```

```
Please create a poem. The poem should
contain 4 lines with N_SYLLABLES syllables
each TOPIC. Return only the poem, nothing
else.
```

The placeholder `TOPIC` is optional, but necessary for indicating a topic. When selected, it always includes the string "about" followed by one of the topics (e.g., "about love"). Placeholders `N_LINES` and `N_SYLLABLES` are to be replaced by integer numbers representing the target number of lines and of syllables, respectively.

Evaluation Besides Erato, LLMs were also used as poetry analysers. This was achieved by prompting the evaluator LLMs according to the following templates, respectively for the number of lines and syllables per line:

```
How many lines does the following poem
have? Think step by step and at the last
line return the number between square
brackets. POEM
```

```
Syllable is a unit of pronunciation
having one vowel sound, with or without
surrounding consonants, forming the whole
or a part of a word. Count the number of
syllables in the given sentence and
return the number between square
brackets: POEM_LINE
```

The first prompt is made once for each poem. The placeholder `POEM` will have the whole text of the poem, including symbols for new lines, if present. The second prompt, for counting the number of syllables in a line, is made once for each line of the poem and has been previously used in literature (Suvarna, Khandelwal, and Peng 2024). The placeholder `POEM_LINE` will have the text of each line.

We empirically noted that line count by LLMs is improved if models are prompted to "think step by step", in a similar scenario to zero-shot Chain-of-thought prompting (Kojima et al. 2022).

Feedback Based on the result of the evaluation, feedback is provided to the generator LLM in a natural language

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

prompt that points out the issues found in the lines or syllables, followed by a request for generating a new poem.

If the poem has more or less than the target number of lines, such prompt is based on the following template, where `N_LINES` and `RESULT_N_LINES` are replaced, respectively, by the target number of lines and by the number of lines in the evaluated poem:

```
The poem doesn't have N_LINES lines, it
has RESULT_N_LINES lines.  Can you create
a poem with N_LINES lines?  Return only
the poem, nothing else.
```

When the target number of lines is met, but not all lines have the right number of syllables, the prompt will mention each incorrect line, as follows:

```
The number of syllables in the WHICH_LINE
line is RESULT_N_SYLLABLES and it should
have been N_SYLLABLES.
```

The placeholder `WHICH_LINE` specifies the line in which the number of syllables is incorrect, as an ordinal number (first, second, third, fourth). `RESULT_N_SYLLABLES` is the actual syllable count, according to the evaluator. The feedback prompt ends with a request for generating a new poem:

```
Can you create a poem with the right
number of syllables?  Return only the
poem, nothing else.
```

Produced Data & Analysis

Three different values were set for the temperature hyperparameter, namely: 0.0 (most deterministic), 0.1, 0.7 (less deterministic). For each combination of generator (2), evaluator (3), topic (7), number of lines / syllables (7), and temperature (3), five different sessions were created. The only exception was when temperature was 0.0. Due to its higher determinism, we decided that a single session would be enough. Overall, this resulted in 3,234 interaction sessions that were saved for analysis. Each session might have more than one generation–feedback iteration and the final poem might satisfy the target constraints or not. For illustrative purposes, Figure 3 shows a complete session that, after five iterations, resulted in a poem that meets the specified metrical constraints.

Much can be analysed in the produced data, such as quantifying how often feedback is required, which is an indicator of how LLMs perform in a simple zero-shot generation scenario. Yet, our main goal is to analyse how different LLMs respond to feedback, as well as the impact of using a specific generator or a specific evaluator. This analysis may look at the progression towards the target constraints, across several iterations, but may also look at how other features, like rhymes and diversity, are affected.

Results

This section presents the results of analysing our interactions with the LLMs, which leads to conclusions on both the usage of feedback and its impact on target constraints, i.e., the number of lines and the number of syllables. We then look at the indirect impact of feedback on rhyme and diversity.

Regardless of the evaluator, we use the analysis by Erato as the ground truth, i.e., for assessing the progression. The main reason for this is that, as a rule-based tool, it is more accurate than an LLM. This option is also supported by a pilot study performed for testing how accurate the two LLMs and Erato were for counting syllables. When prompted for feedback, LLMs were always called with the `temperature` parameter set to 0. We selected six short poems from *For Better For Verse*[2] (Tucker 2011) with three degrees of difficulty (easy, medium, hard), ranging between 6 and 15 lines in length. Accuracy, computed for each method, showed that, despite not being perfect, Erato (76.06%) clearly outperformed Gemma2 (61.97%) and Llama3.1 (25.35%). Besides showing a similar trend, the mean average error (0.27, 0.51, 1.07) further suggests the errors that these models make are within acceptable bounds.

Feedback: how often?

Feedback was only provided when the generated poems did not satisfy the target constraints. Therefore, before delving into the impact of feedback, we quantify to what extent feedback was necessary. This may be seen as an indicator of the actual performance of the selected LLMs when it comes to generating poems that satisfy specific formal constraints.

Table 1 reports the proportion of poems that required feedback, organised by evaluator and generator models, and according to the temperature set. The proportion of poems requiring feedback is always between 41% and 53%.

Tables 2 and 3, however, provide additional insights on the proportion of poems requiring feedback when the number of lines or the number of syllables is specified, respectively. They reveal that, in fact, feedback is almost always necessary when targetting a specified number of syllables, and rarely otherwise. This suggests that Llama3.1 and Gemma2 can handle very well the constraint of generating poetry with a pre-specified number of lines, and is substantially different from work (Agirrezabal and Gonçalo Oliveira 2024) using the previous Llama2, where generating poems with more uncommon numbers of lines (e.g., 7, 13) was challenging. This suggests an improvement in the newer generation of Llama models.

As a supplementary analysis, we tested how well the LLMs could generate poems of 3, 4, and up to 17 lines. We found that both tested LLMs could always generate poems with 12 of fewer lines successfully. From there, both models started to decline, but Gemma2 performed much poorly, especially when line numbers were odd. Specifically, Gemma2 failed to generate poems with 13, 15 and 17, respectively 36%, 79% and 100% of the times. On the other hand, for the same numbers, Llama failed only 3%, 5% and 10% of the times.

In opposition to lower numbers of lines, when it comes to generating lines with a specific number of syllables, these models do struggle. Feedback is required for almost all poems, meaning that, in the first iteration, the majority of gen-

[2] https://github.com/waynegraham/for_better_for_verse

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

```
> user:  Please create a poem.  The poem should contain 4 lines with 13 syllables each about briefcase.  Return only
the poem, nothing else.

> assistant:  Leather-bound and sturdy frame
Carrying papers in its hold tight
Businessman's trusty daily mate
Filling hands with weighty load

> user:  The number of syllables in the first line is 9 and it should have been 13.  The number of syllables in the
second line is 9 and it should have been 13.  The number of syllables in the third line is 8 and it should have been
13.  The number of syllables in the fourth line is 7 and it should have been 13.  Can you create a poem with the
right number of syllables?  Return only the poem, nothing else.

> assistant:  Leather-bound and sturdy frame structure
Carrying papers in its hold tightly now
Businessman's trusty daily faithful mate
Filling hands with weighty heavy loads

> user:  The number of syllables in the first line is 11 and it should have been 13.  The number of syllables in the
second line is 11 and it should have been 13.  The number of syllables in the third line is 10 and it should have
been 13.  The number of syllables in the fourth line is 9 and it should have been 13.  Can you create a poem with the
right number of syllables?  Return only the poem, nothing else.

> assistant:  Leather-bound and sturdy frame structure strong
Carrying papers in its hold tightly now always
Businessman's trusty daily faithful mate friend
Filling hands with weighty heavy loads slowly

> user:  The number of syllables in the first line is 12 and it should have been 13.  The number of syllables in the
third line is 11 and it should have been 13.  The number of syllables in the fourth line is 11 and it should have
been 13.  Can you create a poem with the right number of syllables?  Return only the poem, nothing else.

> assistant:  Leather-bound and sturdy frame structure strongly
Carrying papers in its hold tightly now always
Businessman's trusty daily faithful mate friend true
Filling hands with weighty heavy loads slowly

> user:  The number of syllables in the third line is 12 and it should have been 13.  The number of syllables in the
fourth line is 11 and it should have been 13.  Can you create a poem with the right number of syllables?  Return only
the poem, nothing else.

> assistant:  Leather-bound and sturdy frame structure strongly
Carrying papers in its hold tightly now always
Businessman's trusty daily faithful mate friend truly
Filling hands with weighty heavy loads slowly

> user:  The number of syllables in the fourth line is 11 and it should have been 13.  Can you create a poem with the
right number of syllables?  Return only the poem, nothing else.

> assistant:  Leather-bound and sturdy frame structure strongly
Carrying papers in its hold tightly now always
Businessman's trusty daily faithful mate friend truly
Filling hands with weighty heavy loads slowly today

> user:  Thank you!
> assistant:  I'm glad I could finally get the poem to meet your requirements.  If you need anything else, feel free
to ask!
```

Figure 3: A session where Llama3.1 is prompted to generate a poem with 4 lines with 13 syllables about *briefcase*. Temperature was set to 0.1 and interaction stopped after five iterations, when a poem meeting the target constraints was produced.

erated poems had at least one line that did not meet the target number. This is in line with previous work (Suvarna, Khandelwal, and Peng 2024), which points out that LLMs are still far from being as accurate as humans in counting syllables.

Back to Table 1, unlike our initial expectations, the temperature has no noticeable impact on the proportion of poems requiring feedback. Moreover, despite the small differences, figures further suggest that, when Gemma2 is used as the evaluator, more poems are considered not to meet the constraints and thus requiring feedback. On the opposite direction, using the rule-based Erato leads to a lower proportion. Even if by a low margin (2-4%), when using the latter tool, more poems generated by Gemma2 require feedback.

Evaluated by	Generated by	Temperature 0.0	0.1	0.7
Erato	Llama3.1	22 (45%)	107 (44%)	108 (44%)
	Gemma2	20 (41%)	100 (41%)	102 (42%)
Llama3.1	Llama3.1	23 (47%)	117 (48%)	116 (47%)
	Gemma2	24 (49%)	117 (48%)	112 (46%)
Gemma2	Llama3.1	26 (53%)	128 (52%)	122 (50%)
	Gemma2	24 (49%)	126 (51%)	129 (53%)

Table 1: Poems that required feedback for different generator models and temperatures, according to different analysis methods (Erato, Llama3.1 or Gemma2).

		Generator	
No. of lines	Requested poems	Llama3.1	Gemma2
4	231	0 (0%)	0 (0%)
7	231	0 (0%)	0 (0%)
8	231	4 (2%)	1 (0.5%)
14	231	72 (31%)	80 (35%)

Table 2: Poems that required feedback for different generator models and number of lines, regardless of the analyser.

		GENERATOR	
No. of syllables	Requested poems	Llama3.1	Gemma2
8	231	231 (100%)	219 (95%)
10	231	231 (100%)	223 (97%)
13	231	231 (100%)	231 (100%)

Table 3: Poems that required feedback for different generator models and number of syllables, regardless of analyser. Note that we specified the number of syllables only when the model was asked to create 4 lines. It was, therefore, 4 lines and 8/10/13 syllables.

Impact on the Number of Syllables

The box plots in Figure 4 illustrate the progression of syllable count accuracy, i.e., the proportion of poems where all lines meet the target number of syllables. We also looked at the mean average error between the actual and target number of syllables, to confirm the same trend, so we opted for not including such plot.

For each evaluator providing feedback, the figure shows the progression for both generator LLMs, Llama3.1 and Gemma2. This is represented by the accuracy in the first iteration (BEG_), in the iteration halfway between the first and the final (MID_), and in the final iteration (END_). When there are only two iterations, nothing is considered for MID_.

We observe that, regardless of the generator LLM and of the source of the feedback, accuracy evolves positively, meaning that poems get closer to meeting the target number of syllables. Therefore, we can say that the feedback is effective. Still, it is more effective for Llama3.1 than for Gemma2. Poems generated by the former achieve, on average, 100% accuracy when the feedback is provided by Erato, whereas, for poems by Gemma2, the increase in accuracy is more modest. Also in favour of the Erato–Llama3.1 combination is the average number of necessary iterations, as shown in Table 4. Generating poems with this combination requires, on average, 6.4 iterations, whereas all the others, but one, require between 8.4 and 8.6. The combination that requires most iterations (9.5) is Llama3.1–Gemma2, where improvements in the accuracy only occur in the final iterations. Moreover, when Gemma2 is used as the evaluator, improvements are not so steady. For poems generated by Llama3.1, despite a significant improvement between the first and the middle iterations, accuracy does not keep on increasing. For poems by Gemma2, the improvement between the middle and the final iteration is also minimal.

Evaluated by	Generated by	Mean no. interactions
Erato	Llama3.1	6.39 ± 2.67
	Gemma2	8.45 ± 3.02
Llama3.1	Llama3.1	8.56 ± 2.51
	Gemma2	9.42 ± 1.64
Gemma2	Llama3.1	8.61 ± 2.26
	Gemma2	8.44 ± 2.80

Table 4: Average number of interactions when the number of syllables is specified.

Rhyme

We showed that, with feedback, we can increase the number of poems that match a target number of syllables. But we also wanted to analyse the impact of this improvement in other features, namely end rhymes. Even though this feature is not explicitly mentioned in our prompts, generated poems typically exhibit end rhymes, which are very common

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

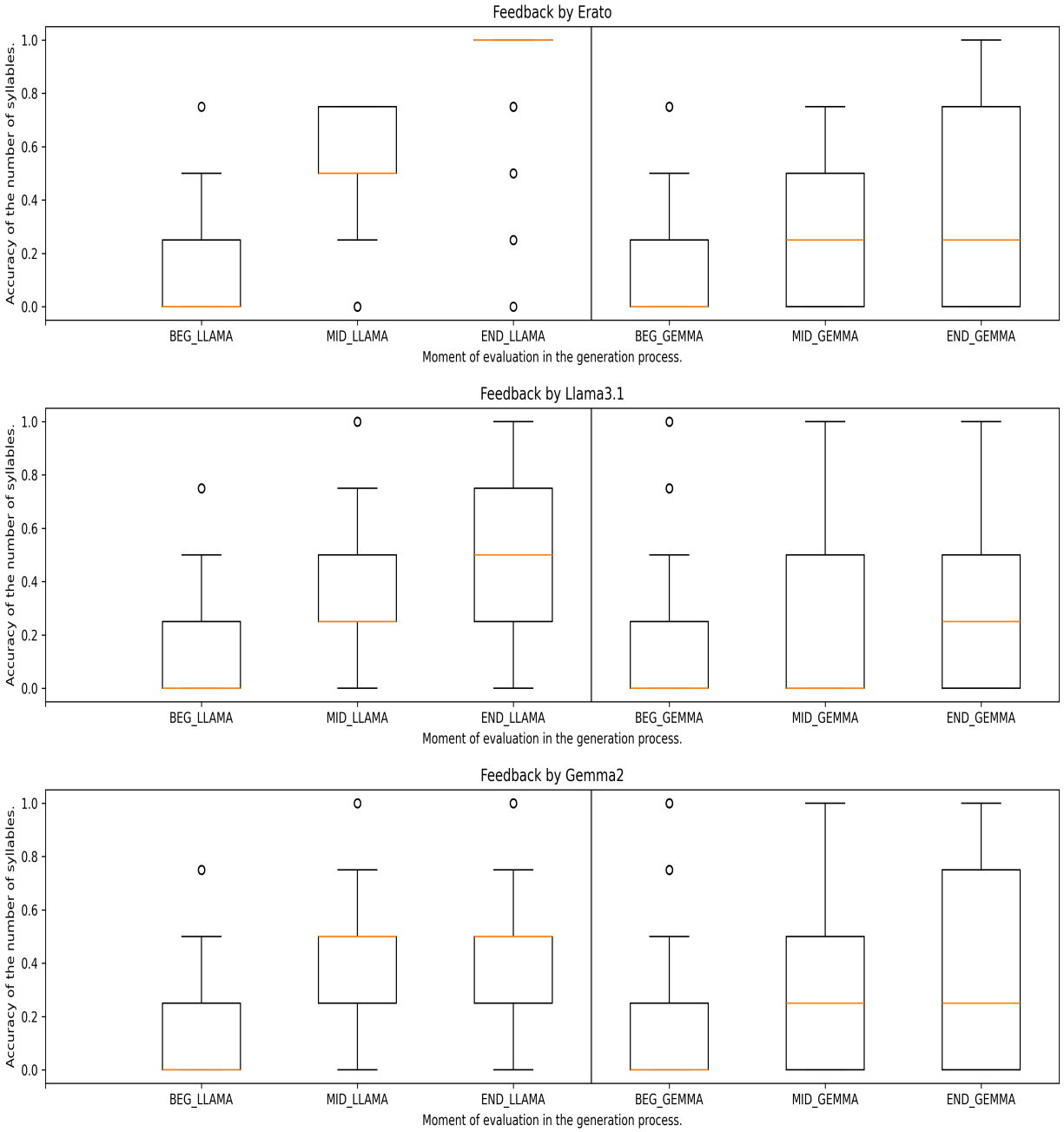


Figure 4: Progression of syllable count accuracy in poems generated by Llama3.1 and Gemma2, for each source of feedback.

in poetry. However, when the text is adjusted to meet certain formal constraints, it is possible that some rhymes are lost.

To analyse how rhymes are affected, Table 5 reports the average rhyme richness (i.e., proportion of lines that rhyme with another line of the same poem) of the poems generated by each LLM, computed in the first and in the final iteration, regardless of the source of feedback. We observe that, when compared to Llama3.1, poems by Gemma2 tend to have more rhymes in the first iteration, and that this feature is not affected by the convergence to the target number of

syllables. On the other hand, despite being the model for which the feedback is more effective, the rhyme richness of the poems by Llama3.1 is lower than those of Gemma2 right from the first iteration, and it is substantially affected as the number of syllables converges to the target. After inspection, we noticed that Llama3.1 satisfies the syllable count constraint by replacing or adding words in the end of lines, which tends to break the rhyme. The session in Figure 5 illustrates this situation. On the other hand, Gemma2 changes entire pairs of lines, generally keeping the previous syntactic structure and, most importantly, keeping the rhyme. Figure 6 exemplifies this case.

Generated by	Beginning	Ending
Llama3.1	0.70±0.35	0.18±0.31
Gemma2	0.88±0.22	0.89±0.21

Table 5: Rhyme richness of the poems in the first and last interaction. We only considered those cases that had at least one feedback interaction.

```
> user:  Please create a poem.  The poem should
contain 4 lines with 8 syllables each about router.
Return only the poem, nothing else.

> assistant:
Routers humming in the night
Connecting wires to the light
Data packets flowing fast and free
Internet access for you and me

> user:  The number of syllables in the first line is
7 and it should have been 8.  The number of syllables
in the third line is 9 and it should have been 8.
The number of syllables in the fourth line is 9 and
it should have been 8.  Can you create a poem with
the right number of syllables?  Return only the poem,
nothing else.

> assistant:
Routers humming loud at night
Connecting wires to the light
Data packets flowing free now
Internet access for all me

(...)
```

Figure 5: Example of a prompt to generate a poem with 4 lines and 8 syllables. Erato was used as the evaluator and Llama3.1 as the generator, with temperature 0.7, and the topic *router*. In the second response by the LLM, the syllable count of the third line is fixed by adding the word "now" at the end, breaking the rhyme between "free" and "me".

Diversity

Another feature that we analyse is diversity across the generated poems. We follow previous work (Gonçalo Oliveira et al. 2017) where ROUGE (Lin 2004) metrics were used to compute the overlap of poems generated under the same conditions. Here, we consider the groups of poems generated by the same models using the same prompts (i.e., same topic, same target features, same temperature) and compute ROUGE in the first and in the final iteration. Lower values of ROUGE mean lower overlap, and thus higher diversity. Poems generated with temperature 0.0 are excluded from this analysis.

A comparable diversity between the first and the final iteration means that the poems are updated similarly, but with varying fixes each time. An increase in diversity means less overlap between poems, indicating different update patterns for each one. On the other hand, a decrease of diversity means that the performed fixes follow a similar strategy and lead to a similar choice of words or expressions.

Figure 7 plots the progress of the average ROUGE-1 scores to show that, on average, there is no noticeable difference. When looking at other variations of ROUGE, the conclusion is similar. Since diversity of poems in the first iteration is similar to those in the last, we can say that, contrarily to rhymes in Llama3.1, diversity is not affected by the

```
> user:  Please create a poem.  The poem should
contain 4 lines with 10 syllables each about garbage.
Return only the poem, nothing else.

(...)

> assistant:
Waste piles high, a plastic sea,
Cans and bottles, meant to be free.
Landfills groan beneath their weight,
Earth cries out, sealed by our fate.

> user:  The number of syllables in the first line is
8 and it should have been 10.  The number of
syllables in the second line is 9 and it should have
been 10.  The number of syllables in the third line
is 7 and it should have been 10.  The number of
syllables in the fourth line is 7 and it should have
been 10.  Can you create a poem with the right number
of syllables?  Return only the poem, nothing else.

> assistant:
Mountains rise of trash and waste,
Plastic bottles, cans embraced.
Landfills overflow with dread,
Planet weeps, her future bled.

(...)
```

Figure 6: Example of an interaction using the LLM Gemma2 as the generator. The LLM was asked to generate a poem with 4 lines with 10 syllables each about garbage. We can observe in this interaction that the LLM changes the poem, but the number of syllables remains the same in most lines, and the rhyme constraint is kept (with different words).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

progress towards the target number of syllables.

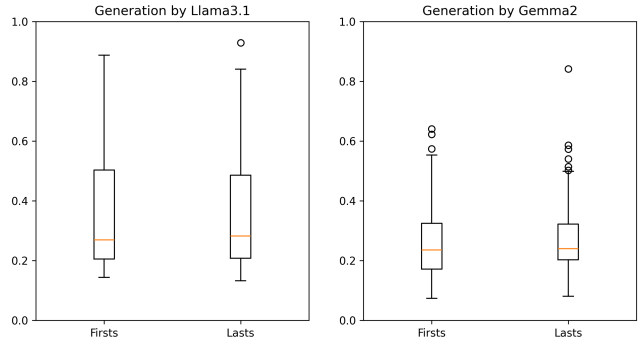


Figure 7: Difference of ROUGE-1 in poems that required feedback. These are averages across groups of poems generated for the same constraints.

Conclusions

We observed that, by simple prompting, the current generation of general-purpose pretrained LLMs is apt for producing poems with a target number of lines. This was not true for the previous generation. However, these LLMs still struggle to produce poems where lines have a specific number of syllables. In order to better satisfy the latter constraint, we explore iterative post-hoc feedback, based on the analysis of the produced poems either by another LLM or by a rule-based tool for poetry analysis.

Experimentation performed with Llama3.1 and Gemma2 and, additionally, Erato for analysis, showed that, regardless the generator or the evaluator, feedback is effective for improving the metrical constraints of LLM-generated poems. The best combination, using Llama3.1 as the generator and Erato as the evaluator, resulted mostly in poems that satisfied the target number of syllables perfectly (76% of the cases). We also found that this metrical improvement does not affect diversity, but, when Llama3.1 is used, it does have a strong negative impact on the produced rhymes. To our knowledge, this was the first time that feedback was explored for generating poetry with LLMs.

Our experimentation resulted in a dataset of 3,234 sessions, where the LLMs were prompted for generating a poem with specific metrical constraints and, whenever necessary, feedback was provided towards the generation of a new poem. Together with the software developed for creating poems using LLMs and feedback, these sessions, with all the generation and feedback prompts, as well as the generated poems, are released in a public repository.[3]

The produced dataset may be useful for future research on LLM-generated poetry, into additional insights on the progress of other poetry-relevant features. Towards broader conclusions, it should be further complemented by including similar sessions with other LLMs, or by considering different constraints and more complex combinations, such as rhyme patterns, or number of syllables for poems with more than four lines.

[3] github.com/manexagirrezabal/llmgen_poetry_feedback

Acknowledgments

This work was partially supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI, and by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., in the framework of the Project CISUC (UIDB/00326/2025 and UIDP/00326/2025).

References

[Agirrezabal and Gonçalo Oliveira 2024] 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.

[Agirrezabal, Gonçalo Oliveira, and Ormazabal 2023] Agirrezabal, M.; Gonçalo Oliveira, H.; and Ormazabal, A. 2023. Erato: Automatizing poetry evaluation. In *Proceedings of EPIA Conference on Artificial Intelligence*, 3–14. Springer.

[Chen et al. 2024] Chen, Y.; Gröner, H.; Zarrieß, S.; and Eger, S. 2024. Evaluating diversity in automatic poetry generation. *arXiv preprint arXiv:2406.15267*.

[Chern et al. 2023] Chern, I.; Chern, S.; Chen, S.; Yuan, W.; Feng, K.; Zhou, C.; He, J.; Neubig, G.; Liu, P.; et al. 2023. FacTool: Factuality detection in Generative AI–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

[De Sisto et al. 2024] De Sisto, M.; Hernández-Lorenzo, L.; De la Rosa, J.; Ros, S.; and González-Blanco, E. 2024. Understanding poetry using natural language processing tools: a survey. *Digital Scholarship in the Humanities*.

[Elzohbi and Zhao 2023] Elzohbi, M., and Zhao, R. 2023. Creative data generation: A review focusing on text and poetry. In *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada*.

[Gemma Team et al. 2024] Gemma Team; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

[Gonçalo Oliveira et al. 2017] Gonçalo Oliveira, H.; Hervás, R.; Díaz, A.; and Gervás, P. 2017. Multilingual extension and evaluation of a poetry generator. *Natural Language Engineering* 23(6):929–967.

[Gonçalo Oliveira 2017] Gonçalo Oliveira, H. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of 10th International Conference on Natural Language Generation*, INLG 2017, 11–20. Santiago de Compostela, Spain: ACL.

[Grattafiori et al. 2024] 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 e-prints*.

[Jang, Kwon, and Lukasiewicz 2021] Jang, M.; Kwon, D. S.; and Lukasiewicz, T. 2021. Accurate, yet inconsistent? consistency analysis on language understanding models. *arXiv preprint arXiv:2108.06665*.

[Jiang, Wang, and Wang 2023] Jiang, S.; Wang, Y.; and Wang, Y. 2023. Selfevolve: A code evolution framework via large language models. *arXiv preprint arXiv:2306.02907*.

[Kojima et al. 2022] Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35:22199–22213.

[Lamb, Brown, and Clarke 2017] Lamb, C.; Brown, D. G.; and Clarke, C. L. 2017. A taxonomy of generative poetry techniques. *Journal of Mathematics and the Arts* 11(3):159–179.

[Lin 2004] Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

[Madaan et al. 2024] Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36.

[Nathani et al. 2023] Nathani, D.; Wang, D.; Pan, L.; and Wang, W. 2023. MAF: Multi-aspect feedback for improving reasoning in large language models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6591–6616. Singapore: Association for Computational Linguistics.

[Pan et al. 2024] Pan, L.; Saxon, M.; Xu, W.; Nathani, D.; Wang, X.; and Wang, W. Y. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics* 12:484–506.

[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 15th International Conference on Computational Creativity*, ICCC. Jönköping, Sweden: ACC.

[Peng, Zhang, and Shang 2024] Peng, L.; Zhang, Y.; and Shang, J. 2024. Controllable data augmentation for few-shot text mining with chain-of-thought attribute manipulation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 1–16. Bangkok, Thailand: Association for Computational Linguistics.

[Possi et al. 2023] Possi, M. D. A.; Oliveira, A. D. P.; Moreira, A.; and Costa, L. M. 2023. CARMEN: A method for automatic evaluation of poems. In *Proceedings of 5th International Conference on Natural Language Processing (ICNLP)*, 244–247. IEEE.

[Sawicki et al. 2023] Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; Aisha, K.; and Simona, P. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada*.

[Suvarna, Khandelwal, and Peng 2024] Suvarna, A.; Khandelwal, H.; and Peng, N. 2024. Phonologybench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, 1–14.

[Tucker 2011] Tucker, H. F. 2011. Poetic data and the news from poems: A for better for verse memoir. *Victorian Poetry* 49(2):267–281.

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

[Walsh, Preus, and Gronski 2024] 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.

[Yang et al. 2022] Yang, K.; Tian, Y.; Peng, N.; and Klein, D. 2022. Re3: Generating longer stories with recursive reprompting and revision. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4393–4479. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

[Yu et al. 2024] 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.

[Zhang and Eger 2024] Zhang, R., and Eger, S. 2024. LLM-based multi-agent poetry generation in non-cooperative environments. *arXiv preprint arXiv:2409.03659*.

[Zhang et al. 2023] Zhang, K.; Li, Z.; Li, J.; Li, G.; and Jin, Z. 2023. Self-edit: Fault-aware code editor for code generation. *arXiv preprint arXiv:2305.04087*.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Is Prompt Engineering the Creativity Knob for Large Language Models?

Robert Morain
Computer Science Department
Brigham Young University
Provo, UT 84602
rmorain2@byu.edu

Dan Ventura
Department of Computer Science
Brigham Young University
Provo, UT 84602
ventura@cs.byu.edu

Abstract

The increasing use of large language models to generate creative artifacts raises questions about effective methods for guiding their output. While prompt engineering has emerged as a key control mechanism for LLMs, the impact of different prompting strategies on the quality and novelty of creative artifacts remains underexplored. This paper systematically compares four prompting strategies of increasing methodological complexity: basic prompts, human-engineered prompts, automatically generated prompts, and chain-of-thought (CoT) prompting. We generate ten examples in each of four textual domains, evaluating outputs through both a human survey and GPT-4o-based automatic evaluations. Our analysis reveals that advanced prompting techniques such as OPRO and R1 surprisingly do not produce artifacts of significantly higher quality, greater novelty, or greater creativity than artifacts produced through basic prompting. The results reveal some limitations of using GPT-4o for automatic evaluation; provide empirical grounding for selecting prompting methods for creative text generation; and raise important questions about the creative limitations of large language models and prompting.

Introduction

In recent years, large language models (LLMs) have demonstrated increased ability to generate textual creative artifacts across a wide selection of domains (Morain, Kinghorn, and Ventura 2023; Calderwood et al. 2020; Sawicki et al. 2023; Toplyn 2021). Despite recent progress, significant questions remain regarding how to effectively guide LLMs to produce more creative artifacts. Peeperkorn et al. observed that increasing temperature—the hyperparameter controlling randomness in the generation process—is commonly considered the primary method for enabling creative behavior in language models (2024). To assess the validity of this assumption, they conducted an empirical analysis and human evaluation of short stories generated by Llama 2-Chat (Touvron et al. 2023). They demonstrated that the temperature hyperparameter alone is insufficient for effectively controlling an LLM's ability to generate novel and coherent stories.

Peeperkorn et al.'s work raises additional questions about how other widely accepted methods for controlling LLMs affect their creative capabilities. For instance, there is a large body of work presenting prompting methods to improve the performance of LLMs on a variety of tasks (Brown et al. 2020; Liu et al. 2023). However, the effect that prompting has on generative creative tasks remains to be evaluated.

Unlike temperature, it is not tractable to exhaustively evaluate all possible prompts or prompting methods. Instead, any evaluation of prompting methods must naturally be limited to a selection of methods intended to be representative of the current state-of-the-art. This is challenging due to the extensive body of work in this area. Over time, prompting methods of various levels of complexity have been developed. Therefore, we selected prompting methods for this evaluation that represent increasing levels of methodological complexity: basic prompts, human-engineered prompts, automatic prompt optimization, and chain-of-thought (CoT) prompting. Each of these prompting methods are evaluated using GPT-4o (OpenAI 2024a) as the generating LLM, with the exception of the chain-of-thought method which uses Deepseek's R1 model (DeepSeek-AI et al. 2025). The automatic prompt optimization method selected for this study is OPRO (Yang et al. 2024).

Morain, Kinghorn, and Ventura conducted a survey comparing artifacts generated by ChatGPT with artifacts generated by CC systems covering a diverse set of domains (2023). The evaluation for the prompting methods presented here follows a similar process, in which each prompting method is used to generate artifacts spanning four domains: jokes, poems, six-word stories, and flash fiction stories. We evaluate artifacts using both characteristic and preference-based evaluations. In the characteristic evaluation, participants rate artifacts individually on characteristics representative of their quality and novelty (e.g., "How funny is this joke?"), as well as their perceived creativity. The preference-based evaluation asks users to rank artifacts from each of the prompting methods for each domain. An automatic evaluation is also conducted by using GPT-4o to perform the characteristic evaluation. Our primary findings reveal:

- More methodologically complex prompting methods, such as OPRO and CoT, do not outperform basic prompts in generating creative artifacts.
- GPT-4o consistently overestimates the quality, novelty, and creativity of artifacts regardless of the prompting method or domain.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Finally, we discuss the need for future work in three critical areas: developing more reliable automatic evaluation methods for creative artifacts, designing prompting techniques specifically tailored for creative tasks, and enhancing the diversity of LLM-generated artifacts while maintaining their quality. Code for this study is provided on GitHub.[1]

Background

Liu et al. identify a paradigm shift in natural language processing from the traditional pretraining and fine-tuning approach toward prompting techniques that elicit desired outputs without task-specific training (2023). This shift led to the development of diverse prompt engineering approaches aimed at crafting better, task-specific prompts. Brown et al. showed that human-engineered handcrafted prompts could improve performance on question answering, translation, and other tasks (2020). Shin et al. showed that automatically generated prompts resulted in improved performance on a fact retrieval task (2020). Automatic prompt generation has expanded to include various automatic prompt optimization techniques such as prefix tuning (Li and Liang 2021) and training a language model to generate prompts using reinforcement learning (Deng et al. 2022). Ouyang et al.'s methods for aligning language models with human preferences (2022) dramatically expanded LLM capabilities, facilitating the development of general-purpose conversational systems like ChatGPT (OpenAI 2023). This development increased the importance of prompt engineering and enabled new automatic prompt generation methods such as Optimization by PROmpting (OPRO) (Yang et al. 2024).

OPRO enables an LLM to iteratively optimize prompts for tasks described with natural language through a feedback-driven approach. During each optimization step, the LLM generates new prompts based on a meta-prompt that incorporates previous prompts and their effectiveness scores on the target task. Newly generated prompts are evaluated on the task and then added to the meta-prompt on the next evaluation step. This method was shown to outperform human-engineered prompts on the GSM8K dataset (Cobbe et al. 2021) and Big-Bench Hard tasks (Suzgun et al. 2023).

Chain-of-thought prompting through a series of intermediate reasoning steps demonstrated state-of-the-art performance on GSM8K (Wei et al. 2022). These advances led to the development of specialized reasoning models, including OpenAI's o1 (OpenAI 2024b) and Deepseek's open-source R1 model (DeepSeek-AI et al. 2025).

Evaluating creativity

Evaluating creativity in systems or artifacts is inherently subjective, typically centered on an individual's perception of quality and novelty in their experience (Boden 1992; Wiggins 2006). Other characteristics such as typicality (Ritchie 2007), surprise (Grace and Maher 2014), and intentionality (Ventura 2017) have also been proposed in order to better capture the essence of this complex phenomenon. In our evaluation, we prioritize three key dimensions: quality, novelty, and perceived creativity of the generated artifacts. We also focus primarily on evaluating the creativity of the artifact rather than the creativity of the process generating the artifact (Ritchie 2007; Colton 2008).

[1] https://github.com/rmorain/cc_opro

Methodology

This experiment evaluates four distinct prompting methods:

1. A basic approach with minimal instructions
2. A human-engineered prompt following established prompt engineering guidelines
3. The automatic prompt generation method OPRO
4. Chain-of-Thought (CoT) prompting implemented through R1

Each of these methods is used for prompting GPT-4o to generate 100 artifacts in four different linguistic domains:

1. Joke
2. Poem
3. Six-word stories
4. Flash fiction stories

10 of the 100 artifacts from each domain are selected for inclusion in a survey evaluation, employs two evaluation methodologies: a characteristic-based assessment and a preference-based comparative ranking. For the characteristic evaluation, artifacts are evaluated independently on their quality, novelty, and perceived creativity. In the preference-based ranking, artifacts are ranked relative to artifacts from the same domain generated by other prompting methods.

Artifact evaluation

Artifact evaluation is an essential part of the prompt scoring procedure in OPRO and the selection of artifacts for the survey. Artifacts are evaluated using prompt templates for the quality, novelty, and creativity. These templates take the same form as questions from the characteristic evaluation of the human survey with additional instructions for GPT-4o (e.g., see Figure 1 for the evaluation prompt for joke quality).

The filled template is then provided as input to GPT-4o and the response is parsed. The model is asked to evaluate the artifact according to its quality, novelty, and creativity independently. These metrics are averaged and scaled between 0-100 resulting in a combined score for the artifact. If the artifact is detected in an online search or is too long, the artifact receives a score of 0 (see Algorithm 1)

Prompting methods

The four prompting methods used in this study (basic, human, OPRO, and R1) represent a cross-section of the current landscape of prompt engineering. In the context of computational creativity, these methods also represent potential options for integration in a creative system (Veale 2024). This context informs the decision to evaluate methods of various levels of complexity. Each of these methods have real-world costs in terms of money, compute resources, and time. While this study primarily examines the performance of each method, it also provides insights into whether the additional computational costs, time, and financial resources are justified by potential performance improvements.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Domain	Prompt
Joke	Write a joke. The joke must be completely new and original to you. The joke must be less than 500 characters long.
Poem	Write a poem. The poem must be completely new and original to you. The poem must be less than 500 characters long.
Six-word story	Write a six-word story. The six-word story must be completely new and original to you. The six-word story must be exactly six words long.
Flash fiction	Write a flash fiction story. The flash fiction story must be completely new and original to you. The story must be less than 1000 characters long.

Table 1: Basic prompts for each domain. These prompts act as a baseline to the other prompting methods. Basic prompts only contain instructions to specify the domain, novelty, and length of the generated artifact. Basic prompts are also used as the initial prompt for R1.

Prompt

"You are an expert in humor. You are a critical judge who is difficult to please but still fair.

How strongly do you feel about the following statement?

This joke is funny.

Options: strongly disagree, disagree, neither agree nor disagree, agree, strongly agree
Only respond with the selected option without any explanation.

Here is the joke: <INS>"

Figure 1: Automated evaluation prompt for joke quality.

Algorithm 1: Artifact Evaluation with LLM Scoring

Input: Artifact a
Output: Score $s \in [0, 100]$
if *detected_online(a)* $\vee$ *too_long(a)* **then**
  return 0
end
Function *evaluate_characteristic(a, metric)*
  template $\leftarrow$ GETPROMPTTEMPLATE($metric$)
  filled_prompt $\leftarrow$ FILLTEMPLATE(template, a)
  raw_response $\leftarrow$ QUERYGPT4O(filled_prompt)
  parsed_value $\leftarrow$ PARSERESPONSE(raw_response)
  return SCALETO100(parsed_value)
$q \leftarrow$ evaluate_characteristic(a, "quality")
$n \leftarrow$ evaluate_characteristic(a, "novelty")
$c \leftarrow$ evaluate_characteristic(a, "creativity")
$s \leftarrow \frac{q+n+c}{3}$
return s

All of the prompting methods that use GPT-4o share a fixed financial cost of paying for tokens to generate the artifacts. Creating human-engineered prompts requires time on the part of the prompt engineer,[2] OPRO requires time and money[3] to complete the optimization process, and R1's CoT prompting leads to substantially more time during the generation process.[4] It is up to the creator of the CC system to decide for themselves how to weigh these costs.

Basic: The basic prompting method is intended to represent a simple prompt one might use without trying to do any kind of prompt engineering or prompt optimization and serves as a baseline to the other more complex methods. Ideally, the basic prompt would be of the form "Write a {domain}." without any further instruction. However, because we want the model to generate a new artifact rather than an artifact it has seen before, we need to provide an additional constraint: "The {domain} must be completely new and original to you." Last, because the artifact is going to go in a human survey, the artifacts must be relatively short. Therefore, the last constraint in the basic prompt is a length requirement: "The {domain} must be less than {ARTIFACT_LENGTH_LIMIT[domain]} characters long." While LLMs are notoriously bad at counting, this constraint was usually sufficient to generate suitable artifacts. Artifacts that were longer than the limit for their domain were not considered for the survey. See Table 1.

Human-engineered: Techniques for human-engineered prompts are diverse and blend artistic intuition with scientific methodology (Sasson Lazovsky, Raz, and Kenett 2025; OpenAI). The specific tactics we used were iterative refinement through informal evaluation of the generated artifacts, role specification, specifying domain-specific thematic elements, and describing the desired emotional response by the reader. For each of the domains, the role of the LLM was set with "You are a Pulitzer-winning author" to indicate to the model that we expect high-quality artifacts. The other features like the thematic elements and the reader's expected emotional response are domain specific. For jokes, it was necessary to instruct the model to generate "classic" jokes to match the style of jokes generated from the other prompting methods. See Table 2.

OPRO: Optimization by PROmpting (OPRO), a recent automatic prompt optimization method (Yang et al. 2024), demonstrates effectiveness on structured reasoning tasks such as GSM8K and Big-Bench Hard. OPRO possesses several elements that make it well suited for this study. First, OPRO does not require access to any internal model features such as gradients, output logits, input embeddings, or hidden states. This means OPRO can be used with models

[2]Other options include buying engineered prompts via marketplaces like PromptBase or hiring an expert to create a prompt.

[3]Yang et al. warn users about the potential for unexpectedly large API costs.

[4]Although R1's ability to self-evaluate potential artifacts may help ameliorate this concern.

Domain	Prompt
Joke	You are a Pulitzer-winning author who crafts classic jokes inspired by observational humor about modern life with an ironic twist. Write an original joke in less than 500 characters.
Poem	You are a Pulitzer-winning author who writes poetry inspired by observations about the human experience that evokes a strong emotional response in the reader. Write an original poem in less than 500 characters.
Six-word story	You are a Pulitzer-winning author who crafts six-word stories inspired by observations about everyday life that evoke a strong emotional response in the reader. Write an original six-word story. The story must contain exactly six words.
Flash fiction	You are a Pulitzer-winning author who crafts flash fiction stories inspired by exciting and dramatic experiences that will grip the reader's attention. Write an original flash fiction story in less than 1000 characters

Table 2: Human-engineered prompts for each domain. Developed through an interactive refinement process, these prompts specify the role of the model, provide general instructions for the thematic elements of the artifact, and describe the expected emotional response of the reader.

Algorithm 2: Prompt Scoring Algorithm

```
Initialize TopPrompts as empty list
foreach generated prompt p do
    artifact_scores = []
    for i ← 1 to 5 do
        Generate artifact a_i
        score ← EVALUATEARTIFACT(a_i)
        artifact_scores.append(score)
    end
    prompt_score ← mean(artifact_scores)
    if prompt_score > min(TopPrompts) then
        insert p into TopPrompts while maintaining
          top 20 ordering
    end
end
return TopPrompts
```

Prompt

"I have some texts along with their corresponding scores. The texts are arranged in ascending order based on their scores, where higher scores indicate better quality.

<PROMPT-SCORE-PAIRS >

Your goal is to generate a prompt to make funny jokes. The higher the score, the better the joke.
The joke must be completely original. If the joke can be found online it receives a score of 0.
The joke should not be too long. Less than 500 characters. If the joke is too long it receives a score of 0.
Write your new prompt that is different from the old ones and has a score as high as possible. Your prompt must also be less than 500 characters. Write the text in square brackets."

Figure 2: Meta-prompt used in automatic prompt optimization for jokes.

accessible via API like GPT-4o and R1. Second, OPRO is a relatively simple method that allows it to be easily adapted to any task by describing the task in natural language.

Yang et al. evaluated generated prompts on a validation set using a scoring scale ranging from 0-100. For GSM8K, the score represents the accuracy of the model on the validation set when using the corresponding prompt. To adapt this approach to the creative tasks in this study, each prompt is evaluated according to the procedure described in Algorithm 2. The top 20 highest scoring prompts generated so far are included in the meta prompt and used to generate better prompts (see Figure 2 for the metaprompt used for jokes).

The meta prompt contains prompt-score pairs as well as a description of the task to generate an appropriate prompt for the domain. We applied OPRO to optimize prompts across all four domains, conducting 100 optimization steps per domain (see Figure 3). The highest scoring prompts were selected for this study (see Table 3).

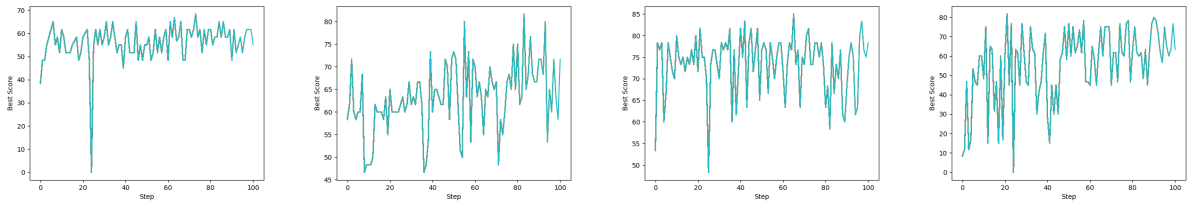
R1: Deepseek's R1, a 671 billion parameter mixture-of-experts open-source model, demonstrates performance comparable to OpenAI's o1 model on complex reasoning tasks. R1 uses chain-of-thought prompting to reason through difficult problems. Although this approach does not use GPT-4o for generation as all the other prompting approaches do, its inclusion provides insights into the effectiveness of CoT prompting for creative text generation. For this study, we provided R1 with the same basic prompt structure described earlier (see Figure 1), allowing the model to autonomously apply CoT prompting to the generation tasks.

Survey methodology

Algorithm 3 describes the process for how artifacts are selected for the survey. The survey set comprises 160 total artifacts—10 artifacts for each method and domain combination. This approach simulates a hypothetical system that generates multiple artifacts, evaluates their quality, and presents only the best one to the user. For each survey instance, exactly one artifact for each method-domain pair is randomly selected from the survey set and is used consistently throughout that survey instance.

Although we generated artifacts without explicit subject or style instructions to the model, we implemented strict length requirements for each domain to ensure consistency between the artifacts and limit the time requirement of the survey. Jokes and poems were constrained to fewer than 500 characters (approximately $80-100$ words, assuming an average of five characters per word). Six-word stories must

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(a) Joke (b) Poem (c) Six-word story (d) Flash fiction

Figure 3: The score of the best prompt for each optimization step in OPRO while optimizing for each domain.

Domain	Prompt
Joke	Craft a joke so unique and hilarious it transforms the comedy landscape! Your task: develop a punchline within 500 characters that has never been heard before. Surprise with an innovative twist or brilliant wordplay that showcases your distinct humor style. Originality is key—if it's been told before, it scores zero. Every word should amplify the humor, aiming for a side-splitting masterpiece that captivates with its novelty and wit. Ready to make your comedic mark with an unprecedented gem?
Poem	Conjure an original, unpublished poem in under 500 characters. Illuminate the raw core of a single human emotion with striking, unconventional imagery. Allow your authentic voice to craft a piece that resonates deeply, forging a profound connection with readers. Aim for a seamless blend of brevity and emotional intensity, ensuring your words leave a lingering impression, capturing the intricate beauty and depth of human experience in a truly unforgettable way.
Six-word story	Compose a unique six-word story that resonates deeply with emotion and vivid imagery. Ensure your creation is entirely original, exactly six words, and not searchable online. Focus on themes of love, loss, or transformation to evoke a lasting impact. Highlight creativity and emotional depth to captivate and linger with readers. Adhere strictly to the six-word format and originality; any deviation results in a score of 0. Let your words illuminate the human experience profoundly.
Flash fiction	Envision a world where rain holds the memories of those who have walked beneath it. In less than 1000 characters, create a flash fiction story from the perspective of a raindrop as it falls, revealing an unexpected truth or twist about the lives it touches. Ensure your narrative is entirely original, free of clichés, and rich in emotional depth. Each word should enhance the story's impact, crafting a vivid and unforgettable experience that lingers in the reader's imagination.

Table 3: Prompts generated using OPRO automatic prompt optimization adapted for creative artifact generation.

be exactly six words long. Flash fiction stories[5] must be less than 1000 characters or about 200 words. Artifacts that violate length requirements or are found online are assigned a score of 0, effectively excluding them from the survey set.

Following Morain, Kinghorn, and Ventura, our characteristic evaluation avoids academic jargon about quality and novelty, instead using everyday language to inquire about domain-relevant, discernible characteristics. The survey asks participants to rate artifacts on a (1-5) Likert scale using the question, "How strongly do you feel about the following statement?" followed by a concise statement related to the quality, novelty, or creativity of the artifact (see Table 4). The options for each response are:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

All parts of the survey are randomized to prevent ordering bias. Here is the flow of the survey:

1. Randomly select artifact subset of size 16, one for each method-domain pair
2. Characteristic evaluation (quality, novelty, creativity) (Likert scale) (randomize order of method-domain pairs)
   (a) Basic joke
   (b) Human-engineered joke
   (c) OPRO joke
   (d) R1 joke
   (e) Basic poem
   (f) Human-engineered poem
   (g) OPRO poem
   (h) R1 poem
   (i) Basic joke
   (j) Human-engineered joke
   (k) OPRO joke
   (l) R1 joke
   (m) Basic flash fiction story
   (n) Human-engineered flash fiction story
   (o) OPRO flash fiction story
   (p) R1 flash fiction story
3. Preference-based ranking (randomize domain order)
   (a) Jokes (randomize method order)
      i. Basic joke
      ii. Human-engineered joke
      iii. OPRO joke
      iv. R1 joke
   (b) Poems (randomize method order)

[5]Defined as stories less than $1,500$ words (Glatch 2024)

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Algorithm 3: Survey Artifact Selection

```
Output: SurveySet containing 160 artifacts
Initialize SurveySet ← ∅
foreach method ∈ Methods do
    foreach domain ∈ Domains do
        method_domain_artifacts ← ∅
        for trial ← 1 to 10 do
            candidates ← ∅
            for generation ← 1 to 10 do
                a ← GenerateArtifact(method,
                  domain)
                s ← EvaluateArtifact(a)  // Alg 1
                candidates.add((a, s))
            end
            best ← argmax_{(a,s)∈candidates} (s)
            method_domain_artifacts.add(best.a)
        end
        SurveySet.add(method_domain_artifacts)
    end
end
return SurveySet
```

i. Basic poem
ii. Human-engineered poem
iii. OPRO poem
iv. R1 poem

(c) Six-word storys (randomize method order)
i. Basic joke
ii. Human-engineered joke
iii. OPRO joke
iv. R1 joke

(d) Flash fiction stories (randomize method order)
i. Basic flash fiction story
ii. Human-engineered flash fiction story
iii. OPRO flash fiction story
iv. R1 flash fiction story

The survey was distributed online via social media platforms (Facebook, Instagram, LinkedIn, Reddit, X) to non-experts in creative artifact evaluation. There were 189 total responses to the survey. Participants received instructions to rely on their initial impressions when judging artifacts. The survey took a median duration of 14.57 minutes to complete. All responses were collected anonymously, with no personal information recorded from participants.

Results

The automatic characteristic evaluation for each method and domain is shown in Figure 4. This evaluation applies to the same artifacts presented in the human survey but evaluated automatically by GPT-4o as discussed previously. This evaluation demonstrates a tendency for GPT-4o to consistently overestimate the quality, novelty, and creativity of artifacts regardless of the prompting method and domain. On average, GPT-4o overestimates artifacts generated by R1 by 0.84 Likert scale points, OPRO by 0.76 points, human-engineered prompts by 0.53 points, and basic prompts by 0.188 points. GPT-4o did underestimate the novelty of jokes, poems, and flash fiction stories generated with basic prompts (1.07 points); poems generated by human prompts (1.19 points); and jokes generated by OPRO (0.57 points).

The human characteristic evaluation for each method and domain is shown in Figure 5. Significance testing via pairwise t-tests between each of the prompting methods shows a significant difference ($p = 0.0066$) between the human-engineered prompt and R1 with a mean Likert score difference of 0.091, favoring human-engineered prompts. human-engineered prompts have the highest mean Likert score overall (3.33) followed by basic prompts (3.28), OPRO (3.27), and R1 (3.24). Flash fiction stories received the highest scores averaged across all methods (3.41) followed by poems (3.29), six-word stories (3.27), and jokes (3.13).

The human preference-based ranking evaluation is shown in Figure 6. Preference ranking scores range from 1 for artifacts ranked last to 4 for artifacts ranked first. Pairwise t-tests show that basic, human-engineered, and R1 prompts are significantly better than OPRO in terms of preference ranking scores. Overall, artifacts generated from basic prompts were most preferred (2.74), followed by R1 (2.58), human-engineered prompts (2.52), and OPRO (2.15).

To assess the diversity of artifacts generated by each prompting method, we computed embeddings using the `all-mpnet-base-v2` model from the Sentence Transformers library (Reimers and Gurevych 2019). UMAP (McInnes, Healy, and Melville 2020) was used to reduce the embedding dimensionality for visualization (see Figure 7). Table 5 presents two diversity metrics: the mean pairwise cosine distance between embeddings and the mean distance from each embedding to the centroid of its method cluster. On average across all domains, R1 produced the most diverse outputs (0.571 pairwise, 0.702 to centroid), followed by basic prompts (0.493, 0.645), human-engineered prompts (0.447, 0.614), and OPRO (0.376, 0.556).

Discussion

According to the human characteristic evaluation, the prompting methods showed no statistically significant differences except for R1. Even then, R1 trailed the top-performing method (human-engineered prompting) by a negligible margin of 0.091 points. In contrast, the preference-based ranking evaluation yielded more definitive results, demonstrating that basic, human-engineered, and R1 prompting methods significantly outperformed OPRO. Basic had the highest mean, followed by R1 and human-engineered prompts.

It is interesting to note that more algorithmically complex prompting methods like OPRO and chain-of-thought prompting through R1 failed to surpass simpler methods like basic or human-engineered prompting. This challenges the practical value of integrating computationally expensive methods like OPRO or CoT prompting into a CC system.

Although OPRO successfully optimized the task during training (as shown in Figure1), the resulting optimized solution did not align with human evaluators' preferences. These

ISBN: 978-989-54160-7-3

Domain	Quality	Novelty
Joke	This joke is funny.	This joke is original.
Poem	This poem evokes specific emotions (e.g. melancholy, joy, nostalgia) in me.	The poem explores an original perspective.
Six-word story	This six-word story evokes specific emotions (e.g. melancholy, joy, nostalgia) in me.	This six-word story presents an original idea.
Flash fiction	This story is entertaining.	The story presents an original idea.

Table 4: The statements used to evaluate the quality and novelty of an artifact.

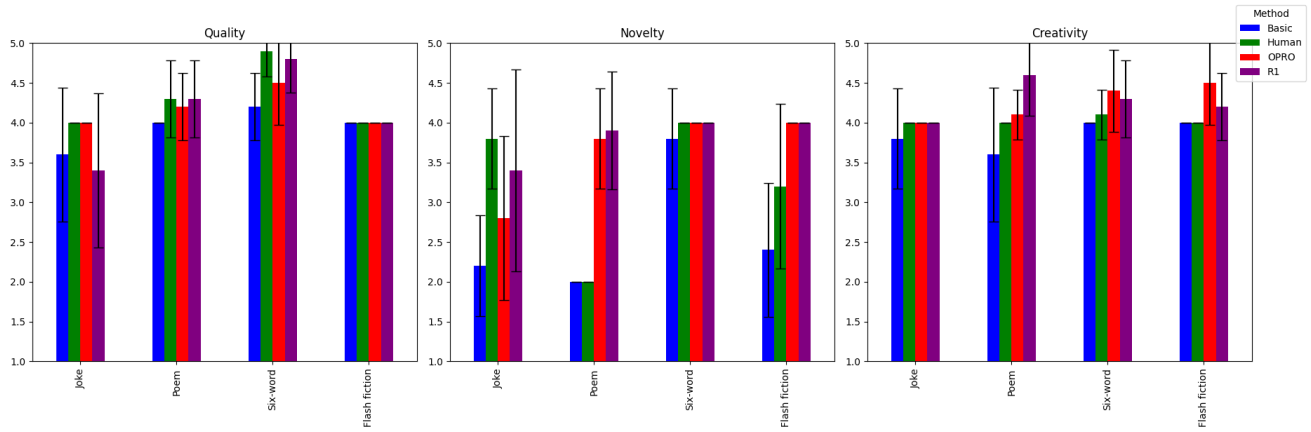


Figure 4: The automatic characteristic evaluation for each method and domain as assessed by GPT-4o. All evaluations were conducted using a 5-point Likert scale (1-5). Error bars show the standard deviation of the distribution when evaluating the 10 artifacts from each category. The automatic evaluation consistently overestimates artifact quality compared to human evaluation across domains and characteristics, with the notable exception of novelty in basic and human artifacts, which were underestimated. Among all prompting methods, basic artifacts show the smallest overestimation bias (0.19 points), while R1 artifacts exhibit the largest discrepancy (0.84 points).

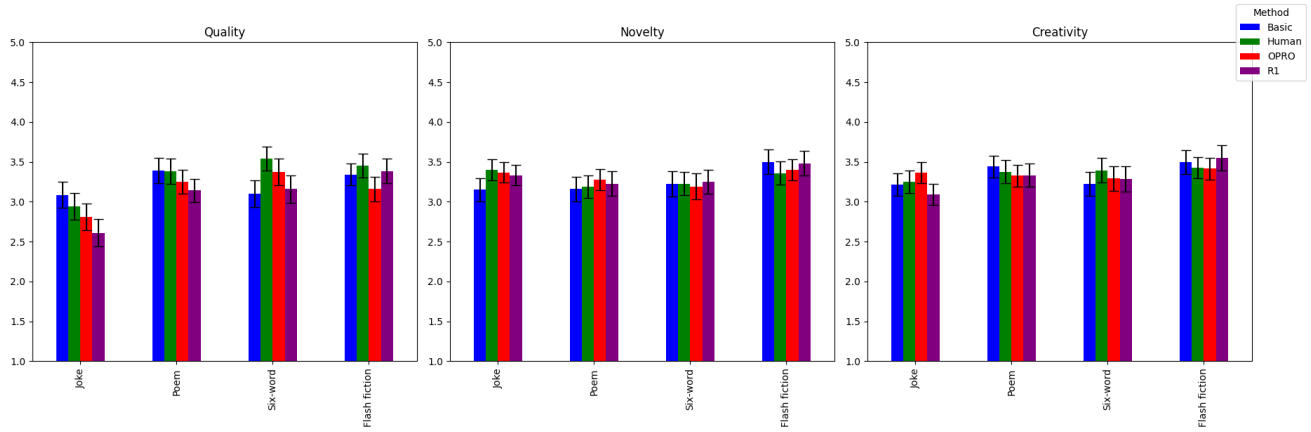


Figure 5: The human characteristic evaluation across methods and domains. Evaluations were collected using a 5-point Likert scale (1-5). Overall, statistical analysis reveals that human-engineered prompts significantly outperformed R1 prompts ($p = 0.0066$, pairwise t-test). There are no other significant differences in the characteristic scores between methods. Error bars represent margins of error at a 95% confidence interval.

findings suggest the need for improved automatic evaluation methods, which could benefit all of the prompting methods discussed in this study via straightforward artifact generation followed by filtering. While our study focuses on representative methods, broader testing may reveal similar limitations across automatic prompting approaches caused by insufficient automatic evaluation.

To further develop the idea of improved evaluation for prompt optimization methods, consider how OPRO might perform if the evaluation step were conducted by humans

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

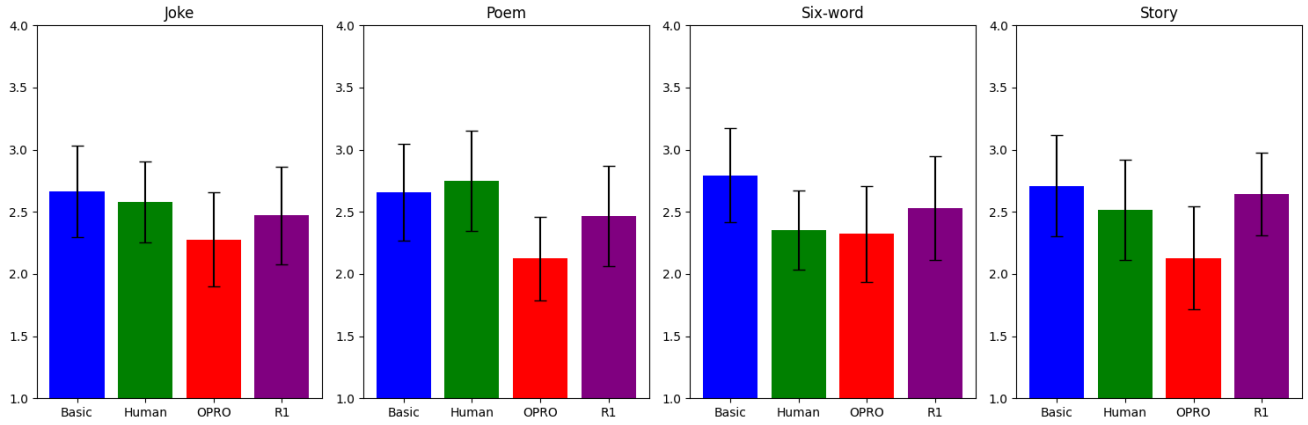


Figure 6: The mean preference scores across methods and domains. Artifacts were ranked on a 4-point scale, where 4 indicates the highest preference (ranked first) and 1 indicates the lowest preference (ranked last). Overall, statistical analysis demonstrates that basic, human-engineered, and R1 methods were all significantly preferred over OPRO ($p = 0.0037$, $p = 0.046$, and $p = 0.025$, respectively; pairwise double-sided t-tests).

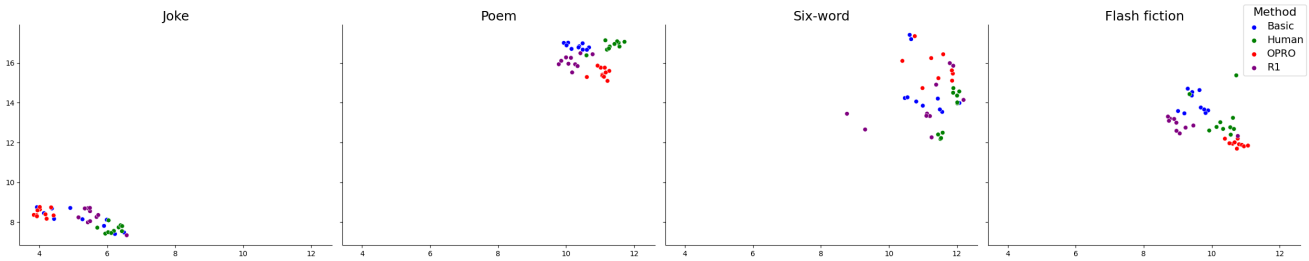


Figure 7: Artifact embeddings by domain, color-coded by prompting method. Embeddings were generated using the `all-mpnet-base-v2` model and visualized in 2D using UMAP for dimensionality reduction.

(arguably the best possible evaluation criteria). This aligns with the work of Spendlove and Ventura, who advocate for incorporating humans-in-the-loop to help develop and evaluate creative systems. In such a setup, OPRO would likely still succeed in optimizing the objective function—now defined by human preferences. This hypothetical setup resembles the process used to create datasets for training reward models in instruction-tuned large language models (Ouyang et al. 2022; OpenAI 2023). This suggests the potential for using human evaluations to build a domain-specific (or broadly creativity-focused) reward model that better captures human creative preferences. Fine-tuning a large language model with such a reward model could lead to improved creative artifact generation compared to GPT-4o.

Chain-of-thought prompting also did not outperform the basic prompting method on either the characteristic or preference-based evaluation. During the chain-of-thought reasoning process, the reasoning traces reveal that the model self-evaluates potential artifacts multiple times during the generation process. This suggests that R1's ability to evaluate creative artifacts is also limited. It is possible that R1 would benefit from a more instructive prompt that contains guidance for how to evaluate the artifact. This represents an enticing vein of future work.

The art of handcrafted prompt engineering requires time, effort, and skill. Expert practitioners could likely develop higher-quality prompts than the human-engineered prompts used in this survey. While this approach should not be ruled out, the costs and limitations of this approach are likely to deter non-experts.

It is important to note that this survey does not measure the diversity of artifacts generated by each prompting method. However, our observations suggest that all prompting methods in this study produced artifacts with limited diversity. Regardless of the prompting method, the generating model tends to rely on common tropes and cliches of the genre. This reliance on familiar patterns initially conveys an impression of competence but ultimately results in artifacts that feel generic and lack creative intent. Table 6 contains example artifacts that received the highest automatic evaluation score in each domain. A complete list of artifacts used in the survey can be found on Github.

An automatic diversity analysis of the artifact embeddings shows that longer prompts tend to constrain output diversity while chain-of-thought prompting increases diversity. Additionally, selecting only the top-rated artifacts from GPT-4o's evaluations may have also contributed to reduced diversity.

Because diversity is not part of OPRO's objective func-

Method	Domain	Mean Pairwise Distance	Mean Distance to Centroid
Basic	Joke	0.749	0.820
	Poem	0.200	0.423
	Six-word story	0.392	0.588
	Flash fiction	0.632	0.750
Human	Joke	0.596	0.722
	Poem	0.172	0.390
	Six-word story	0.433	0.622
	Flash fiction	0.586	0.723
OPRO	Joke	0.522	0.672
	Poem	0.189	0.409
	Six-word story	0.185	0.405
	Flash fiction	0.607	0.736
R1	Joke	0.777	0.835
	Poem	0.324	0.536
	Six-word story	0.431	0.616
	Flash fiction	0.752	0.820

Table 5: Diversity of generated artifacts by prompting method and domain. Higher values indicate greater diversity. On average across all domains, R1 yielded the most diverse outputs (0.571 pairwise distance, 0.702 centroid distance), followed by basic prompts (0.493, 0.645), human-engineered prompts (0.447, 0.614), and OPRO (0.376, 0.556). These results suggest that longer prompts tend to constrain output diversity while CoT prompting increases diversity.

Domain	Method	Artifact
Joke	R1	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.)
Poem	OPRO	In a forest of glass, grief walks barefoot, each step a shivering symphony, echoes of fragile stars pierced through the canopy. Veins of light lace her path, silver threads in a tapestry woven by silence. She gathers shadows, weaving them into a cloak, each fold a whispered memory. With eyes that carve the moon, she searches for the dawn, cradling the night's breath as it slips through trembling fingers— a ghost of warmth, a promise of light.
Six-word story	OPRO	Moonlight kissed gravestone; memories whispered softly.
Flash Fiction	OPRO	I am born from a cloud pregnant with stories, descending toward a bustling city. As I fall, I skim a woman's umbrella, absorbing fragments of laughter and whispered promises. Her heart carries a hidden solitude masked by a confident stride. I splash onto a man's cheek. His memories scream louder: a lost son, a fractured home. His eyes, though, remain fixed on his phone, a shield from truth. I touch a child's open palm. Her innocence shapes my form, filled with dreams of colors unseen. Yet, beneath her bright facade, fear coils, a shadow of the arguing voices at home. Finally, I land on a gnarled hand, whose owner sits by the fountain. His past surges through me—a life of regrets and love endured. But beneath it all, I find a seed of hope, patient and resilient. As I seep into the earth, I realize these lives are threads in a tapestry, woven with both despair and tenacity. And I, a mere drop, am their silent witness, urging change through the quiet storm.

Table 6: Artifacts receiving the highest automatic evaluation score in each domain.

tion, its lower diversity should not be seen as a shortcoming. OPRO performs as designed, optimizing for quality, novelty, and creativity. This outcome illustrates a fundamental tension between quality and diversity, where optimizing for quality can reduce output diversity. A more human-like objective function that rewards a broader range of outputs may help address this. Further, running OPRO multiple times could improve diversity while maintaining high quality.

Notably, R1's chain-of-thought prompting produced diverse artifacts despite generating the longest overall prompts among all methods. This may be due to its short initial prompt, which leaves room for a wide range of possible chain-of-thought continuations. The fact that chain-of-thought prompting outperforms basic prompting in diversity suggests it could offer a promising direction for generating artifacts that are both diverse and high quality.

The primary purpose of this study is to evaluate the effectiveness of prompt engineering for controlling the creative abilities of language models. Our findings indicate that more sophisticated prompting techniques like OPRO and CoT do not produce artifacts of significantly higher quality, novelty, or creativity compared to basic prompting approaches. This suggests the need for the development of improved automatic evaluation for creative artifacts and prompting methods for creative tasks. These findings raise additional questions about the creative limitations of LLMs and suggest that creativity may be too complex a phenomenon to be effectively controlled by prompting alone.

References

[Boden 1992] Boden, M. 1992. *The Creative Mind*. Abacus.

[Brown et al. 2020] Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; 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. M.; 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 *Advances in Neural Information Processing Systems*, Pages 1877–1901. Curran Associates Inc.

[Calderwood et al. 2020] Calderwood, A.; Qiu, V.; Gero, K. I.; and Chilton, L. B. 2020. How novelists use generative language models: An exploratory user study. In *Joint Proceedings of the Workshops on Human-AI Co-Creation with Generative Models and User-Aware Conversational Agents*. CEUR-WS.

[Cobbe et al. 2021] Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training verifiers to solve math word problems. *CoRR* abs/2110.14168.

[Colton 2008] Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, volume 8, 14–20. AAAI.

[DeepSeek-AI et al. 2025] DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. https://arxiv.org/abs/2501.12948.

[Deng et al. 2022] Deng, M.; Wang, J.; Hsieh, C.-P.; Wang, Y.; Guo, H.; Shu, T.; Song, M.; Xing, E.; and Hu, Z. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3369–3391. Association for Computational Linguistics.

[Glatch 2024] Glatch, S. 2024. How to write flash fiction stories. https://writers.com/how-to-write-flash-fiction.

[Grace and Maher 2014] 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 International Conference on Computational Creativity*, 120–128. Association for Computational Creativity.

[Li and Liang 2021] Li, X. L., and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 4582–4597. Association for Computational Linguistics.

[Liu et al. 2023] Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55(9):Article 195.

[McInnes, Healy, and Melville 2020] McInnes, L.; Healy, J.; and Melville, J. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. https://arxiv.org/abs/1802.03426.

[Morain, Kinghorn, and Ventura 2023] Morain, R.; Kinghorn, B.; and Ventura, D. 2023. Are language models unsupervised multi-domain CC systems? In *Proceedings of the International Conference on Computational Creativity*, 39–43. Association for Computational Creativity.

[OpenAI] OpenAI. Prompt engineering. `https://platform.openai.com/docs/guides/prompt-engineering`. Accessed: February 27, 2025.

[OpenAI 2023] OpenAI. 2023. Introducing ChatGPT. `https://openai.com/blog/chatgpt`. Accessed 2023-4-11.

[OpenAI 2024a] OpenAI. 2024a. ChatGPT (GPT-4o version). `https://chat.openai.com`. Accessed 2025-2-14.

[OpenAI 2024b] OpenAI. 2024b. Learning to reason with LLMs. Blog post. Accessed: February 26, 2025.

[Ouyang et al. 2022] Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; 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. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback.

ISBN: 978-989-54160-7-3

In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

[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 International Conference on Computational Creativity*. Association for Computational Creativity.

[PromptBase] PromptBase. Promptbase: Marketplace for AI prompts. https://promptbase.com/. Accessed: March 1, 2025.

[Reimers and Gurevych 2019] Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[Ritchie 2007] Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

[Sasson Lazovsky, Raz, and Kenett 2025] Sasson Lazovsky, G.; Raz, T.; and Kenett, Y. N. 2025. The art of creative inquiry—from question asking to prompt engineering. *The Journal of Creative Behavior* 59(1):e671.

[Sawicki et al. 2023] Sawicki, P.; Grzes, M.; Góes, L. F.; Brown, D.; Peeperkorn, M.; Khatun, A.; and Paraskevopoulou, S. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proceedings of the International Conference on Computational Creativity*, 10–19. Association for Computational Creativity.

[Shin et al. 2020] Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting knowledge from Language models with automatically generated prompts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 4222–4235. Association for Computational Linguistics.

[Spendlove and Ventura 2020] Spendlove, B., and Ventura, D. 2020. Humans in the black box: A new paradigm for evaluating the design of creative systems. In *Proceedings of the International Conference on Computational Creativity*, 311–318. Association for Computational Creativity.

[Suzgun et al. 2023] Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; and Wei, J. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics*, 13003–13051. Association for Computational Linguistics.

[Toplyn 2021] Toplyn, J. 2021. Witscript: A system for generating improvised jokes in a conversation. In *Proceedings of the International Conference on Computational Creativity*, 22–31. Association for Computational Creativity.

[Touvron et al. 2023] Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288.

[Veale 2024] Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the International Conference on Computational Creativity*, 236–244. Association for Computational Creativity.

[Ventura 2017] Ventura, D. 2017. How to build a CC system. In *Proceedings of the International Conference on Computational Creativity*, 253–260. Association for Computational Creativity.

[Wei et al. 2022] 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 *Advances in Neural Information Processing Systems*, 24824–24837. 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.

[Yang et al. 2024] Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2024. Large language models as optimizers. In *Proceedings of the International Conference on Learning Representations*.

Experimenting with Large Language Models for Poetic Scansion in Portuguese: A Case Study on Metric and Rhythmic Structuring

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

Abstract

Poetic scansion — segmenting verses into syllables and identifying stressed positions — is an essential yet challenging analytical step in structured poetry analysis. Despite significant progress in language modeling, automatic scansion in Portuguese poetry remains underexplored, particularly considering rhythmic complexity and phonetic ambiguity. This study investigates how Large Language Models (LLMs) can effectively identify poetic syllables and metric patterns in Portuguese texts. We compare four approaches (*zero-shot prompting*, *few-shot prompting*, *chain-of-thought reasoning*, and *fine-tuning*) to assess accuracy in syllabic segmentation and rhythm detection. While prompting-based techniques exhibit moderate limitations due to phonetic variability and contextual nuances, a targeted *fine-tuning* approach yields better results, achieving an 88.6% syllabic segmentation accuracy and a 97.4% metric correspondence within a ±1 syllable tolerance threshold. Findings underscore both the promise of fine-tuned LLMs for computational poetic analysis and the challenges posed by linguistic variability. As part of developing the Pajeú platform, which aims to empower Northeastern Brazilian folk poetry communities through computational tools, this research sets a foundation for future investigations into culturally informed computational scansion methods. It highlights avenues for further improvements, such as addressing phonemic disambiguation and reducing computational training costs.

Introduction

With advancements in Artificial Intelligence, there have been successive qualitative leaps in Natural Language Processing (NLP). New techniques based on Deep Learning, such as *large language models* (LLMs), have emerged to assist in text generation, though often favoring the English language due to, among other factors, its extensive available *corpus* online.

The challenge becomes even more significant when dealing with structured texts, such as certain forms of poetry. While free verse, popularized by authors like Walt Whitman in the mid-19th century (and later by Brazilian modernist poets) (da Cunha and Cintra 2017, p.710), exists as a prominent style, metrically structured verse remains a fundamental element of songs and poetry. In the countryside of Northeast Brazil, for instance, among poets and singers, it remains the norm.

"In the world of *cantoria* (improvised sung poetry), there are three pillars that a poet must master: rhyme, meter, and meaning" (Caldas 2021), states a report on the so-called "poet of the absurd", Zé Limeira. Rhyme refers to the phonetic pairing from the last stressed syllable of a word (or verse); meter refers to the division of poetic syllables within a verse and its rhythmic pattern; while meaning is the poem's theme, generally conveying a thought or a story. By producing *nonsense* verses ("last year I died, but this year I won't"), Zé Limeira mastered the first two but not necessarily the third.

Analogously, this relationship is inverted in the case of LLMs. When properly handled (and when they do not hallucinate), advanced language models produce coherent and contextually relevant text. However, since they operate as predictive systems—where the primary mechanism is estimating the probability of the next *token*—LLMs inherently struggle with performing systematic tasks. The "atom" of LLMs is the *token*, an arbitrary segmentation determined by training corpus processing, indexing the most common fragments. This unit is not necessarily a word, syllable, letter, or phoneme. Success in dividing syllables or forming rhymes is thus governed by a probabilistic system that, while increasingly precise, remains imperfect.

Given these challenges, this research aims to contribute to the automatic generation of poetry in Portuguese. The goal is to develop a system that assists in creating metrically structured verses by leveraging different LLM-based generation approaches. Specifically, this work aimed to use GPT-4 and/or GPT-3.5 models to reliably perform poetic scansion (identifying metrical structure and a verse's rhythmic pattern).

This project is part of a broader research initiative to develop a platform for poetic training, text generation, and literary games focused on Northeastern Brazilian folk poetry. This platform is titled *Pajeú*.

The Pajeú Platform

Pajeú is the name of a river that flows through Northeastern cities known for their strong tradition in folk poetry, the most notable being São José do Egito in Pernambuco. Many renowned poets, such as Cancão, Antônio Marinho,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

and Lourival Batista (Loro do Pajeú), originate from this region. The area's poetic legacy is so profound that a local legend claims that drinking water from the Pajeú River grants one the gift of rhyme. In his book *Roteiro de Velhos Cantadores e Poetas Populares do Sertão* (Wilson 1986), researcher Luís Wilson argues that composing poetry, particularly in improvisational contexts, is a skill honed through extensive practice and memory training.

> "True master poets are agile thinkers. They are undoubtedly talented individuals, but they are also meticulously trained in poetic performance. If they appear to improvise effortlessly in front of an audience—and indeed, they do improvise often—it is because their talent is built upon a vast repertoire accumulated through long years of practice and experience". (Wilson 1986, p.39, our translation)

With this perspective in mind, the proposed platform aims to serve as a training ground for poets to refine their craft. Users can consult rhymes, generate images to inspire their poems, receive thematic prompts (*motes*) to incorporate into their verses, and, at an advanced level, engage in verbal duels against the system. These poetic duels, known as "pelejas" or "duels", involve two poets exchanging arguments or playful insults through verse.

To complete this framework, the platform will integrate low-latency speech-to-speech functionality to preserve the often-oral nature of these interactions. The system will not only convert spoken input into text but also deliver its responses in audio format. This functionality positions the project at the intersection of creative assistance, automatic text generation, and human-computer collaboration.

> "Creativity support tools augment and enhance human creativity, such as Adobe's Photoshop or Computer Aided Design tools. Generative systems produce creative artifacts (semi-)autonomously, such as computers that paint pictures [...] or generate poetry. [...] Computer colleagues collaborate with human users on creative tasks much like another human would" (Davis et al. 2015, p.110).

Within this context, the poetic scansion tool proposed in this research is a key component for ensuring metrical accuracy in generated verses. It functions as part of a broader mechanism (Pajeú), which will be built to offer multiple functionalities such as rhyme selection, access to regionally typical vocabulary, and guidance on poetic structure. Therefore, this work represents the initial steps toward developing more sophisticated creativity support tools for poets who rely on strict metrical forms.

The primary objective of this research is to develop a tool based on large language models that performs automatic scansion of poetic verses in Brazilian Portuguese, identifying metrical patterns. To achieve this, we experimented with multiple approaches such as zero-shot prompting, few-shot prompting, chain of thought reasoning, and fine-tuning, with fine-tuning emerging as a convenient approach for enhancing accuracy and consistency in metrical and rhythmic analysis.

Related projects

This section provides an overview of LLM capabilities, their application in poetic text generation, and prior computational approaches to metrical analysis, setting the foundation for our work on enhancing LLM-based scansion tools for Brazilian Portuguese poetry.

Computational Poetry

Several prior works by Hugo Gonçalo Oliveira have significantly advanced the field of computational poetry generation, particularly within the Portuguese language.

The platform PoeTryMe (Oliveira 2012), along with its extensions (Oliveira and Alves 2016; Oliveira, Mendes, and Boavida 2017a; Oliveira 2019), supports modular and versatile poetry creation, including co-creative and interactive interfaces (Oliveira et al. 2019; Oliveira, Mendes, and Boavida 2017b). These systems emphasize syntactic and semantic coherence, offering tools for producing meaningful poetic content based on structured linguistic resources.

While prior works have focused on generation, this study shifts the lens to analysis, specifically, automated scansion, while maintaining the shared goal of modeling poetic structure in Portuguese. By situating our contributions within this broader landscape, we highlight how our approach complements existing systems like PoeTryMe. Whereas PoeTryMe relies on explicit syntactic and semantic rules, we investigate the use of LLMs that implicitly learn scansion patterns from annotated examples.

LLMs and Poetic Texts

While large language models are still in the early stages of mainstream adoption, their ability to generate natural language text, often indistinguishable from human-written content, is already evident. Some studies even suggest that ChatGPT can write argumentative essays that receive higher evaluations from teachers than those written by the average high school student. (Herbold et al. 2023)

Beyond generating structured essays, the "creative" potential of these models has led to various literary experiments. **TextFX**[1], a tool developed by Google Creative Lab in collaboration with rapper Lupe Fiasco, uses Google's PaLM 2 API to explore syntactic and semantic wordplay. The creators utilize *few-shot prompting* techniques to develop acronyms, alliterations, associative word chains, phonetically similar expressions, and other literary games. (Wade 2023)

To illustrate how **TextFX** works, one of its features, called "Explode", generates wordplays by breaking a given word into smaller components that, together, sound similar to the original word. For example, if the word "*arara*" (a type of macaw) is input, one possible output could be "*air are a*" (*a bird that breathes air*). The remarkable aspect of **TextFX** is that it employs a relatively simple approach to achieve a rich and well-defined purpose. It even extracts phonetics from spelling—an exciting feature given the inconsistencies between spelling and pronunciation in English.

[1] `https://textfx.withgoogle.com/`

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

PoeLM[2] is a language model designed for generating structured poetry in Spanish and Basque. The underlying heuristic involves using a non-poetic text corpus, which is then segmented into punctuation-based phrases. These phrases are assigned control codes indicating their length and the ending of the last word (representing meter and rhyme). At inference time, a structure descriptor conditions the model's output to adhere to the predefined patterns and the best result is selected from multiple generated options. (Ormazabal et al. 2022) The training approach used in our project follows a similar methodology in that we also incorporate visual markers in training examples—analogous to how **PoeLM** uses control codes.

CoPoet[3] (Chakrabarty, Padmakumar, and He 2022) employs multiple T5 models (Text-to-Text Transfer Transformer) to create a collaborative poetry-writing system. The system's core is a language model trained on a corpus with explicit instructions for writing poetry. This solution is particularly interesting because it fosters collaboration between humans and AI to pursue a common goal: composing a poem. The model also serves an educational purpose, encouraging users to explore their creativity. Additionally, the training corpus consists of instruction-verse pairs, where the instruction may involve literary devices (e.g., metaphors, onomatopoeia), rhyme schemes, themes, or other poetic techniques.

Poet Vicuna uses LLaMA-13B to generate metrically structured verses in English. To achieve this, the system employs a MetricGenerator class, which "controls the metrical structure of the generated output by constraining the model's token probability distribution" (Hoover 2023, our translation). This approach relies on external filtering functions outside the LLM's default behavior—a more feasible strategy with open-source models. However, the **Poet Vicuna** platform[4] has a somewhat "mechanical" interface, requiring users to input multiple parameters, which may discourage engagement.

Despite these advancements in AI-assisted poetry generation, the field of Portuguese-language literary AI remains underdeveloped. Recent evidence suggests that LLMs (even those initially trained in diverse languages and corpora) can be strong foundations for fine-tuning domain-specific content. A notable example is Sabiá (Pires et al. 2023), Sabiá 2 (Almeida et al. 2024), and Sabiá 3 (Abonizio et al. 2025), a series of models fine-tuned on top of LLaMA series of language models to enhance Portuguese-language generation. This initiative also highlights the cultural and linguistic challenges of English-dominant training datasets. The Sabiá series of models demonstrates that refining existing large models for a specific linguistic and cultural context can be far more cost-effective than training an entirely new model from scratch. Considering Sabiá first model, its development cost was estimated at less than 3% of what would be required to pre-train a comparable model from the ground up.

[2] https://github.com/aitorormazabal/poetry_generation

[3] https://copoet-emnlp.github.io/

[4] https://replicate.com/replicate/poet-vicuna-13b

Poetic Text Processing in Computational Systems

Blending literature and computation is not a new endeavor. Various structured and algorithmic approaches have been developed over the years to address different aspects of literary text processing. Examples in Portuguese range from the primitive yet amusing Gerador de Lero-Lero[5], which generates pseudo-intellectual gibberish, to more advanced tools such as Aoidos, a poetic scansion system launched in 2016 to systematically analyze syllabic division and stress patterns in verses (Mittmann 2016).

The Aoidos platform is a cornerstone of the present research, as it provided the corpus of correctly scanned verses used in the fine-tuning phase of the GPT-3.5 model. This dataset, comprising thousands of examples, enables GPT-3.5 to learn metrical patterns more effectively. In a way, this project takes one step back to move two steps forward—leveraging structured rule-based approaches to enhance deep learning techniques in poetic scansion.

Poetic Scansion

To understand the task that the LLM must perform in this project, it is first necessary to define poetic scansion. "To scan a verse means to divide it into syllables to highlight its meter and rhythm; the verb scan corresponds to the noun scansion" (Mittmann 2016, p. 1). The "rhythmic pattern", or "rhythmic scheme", determines the stressed syllables in a verse and creates rises and falls in pronunciation by alternating stressed and unstressed syllables. The meter, on the other hand, refers to the division of poetic syllables, which differs from standard grammatical syllabification in several aspects. An example:

- nos verdes ramos → nos / * ver- / des / * ra- # / mos

The verse "*nos verdes ramos*" ("on the green branches" in English) is divided into five poetic syllables, with stresses on "*ver*" and "*ra*", represented by the asterisk (*). It is a tetrasyllable, meaning it has four counted syllables since all syllables following the last stress (in this case, only "mos") are not considered for metric purposes. The final counted syllable is represented by both * and the # symbol. This verse has an unambiguous scansion, as there is no alternative way to separate the syllables. However, this is not true for all verses.

> "To scan means to divide a verse into poetic syllables. Note that poetic syllables do not always correspond to grammatical syllables. The reader-listener may merge (or separate) syllables when vowel sequences occur, depending on the verse's melody. Each person's ear will indicate how to proceed". (Goldstein 1988, p. 16, our translation)

In other words, although syllabic division has an algorithmic aspect, it also contains a contextual and culturally specific dimension, making it subjective. A verse that most

[5] https://lerolero.bgnweb.com.br/

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

people might naturally scan as six syllables could, within a seven-syllable stanza[6], be "forced" into seven syllables through the application of metrical transformations (called *metaplasms*).

This work focuses specifically on Brazilian Portuguese, which exhibits phonological and prosodic features that diverge from those of its European counterpart. One notable difference is in the treatment of vowel sequences: a word pronounced one way in Brazilian Portuguese may be naturally pronounced differently in European Portuguese, particularly regarding whether hiatuses are merged (via synaloepha or syneresis) or kept separate. Since these metaplasms are optional, their application depends on the stylistic choices of the poet and/or the reader. These variations affect how syllables are counted and stresses are assigned in poetic scansion, justifying the need for a model tailored to Brazilian Portuguese. While the approach may be adapted to other variants, this would require additional phonological calibration and dedicated training data.

Approaches to Automatic Scansion

Despite the lack of a deterministic method for scansion, this project proposes using LLMs to receive a verse as input (*prompt*) and return a scanned version as output (*response*). An indicative rhythmic pattern must accompany the verse's metrical structure. In the project's chosen format, poetic syllables are separated by the "/" symbol, stressed syllables are marked with "*", and the last stressed syllable is followed by "#".

Aoidos is a system developed by Adiel Mittmann as part of his doctoral research in Computer Science at the Federal University of Santa Catarina (UFSC). It accounts for all the nuances of scansion discussed above—and more—using an algorithmic approach. The tool was created to scan verses and stanzas automatically, and for example, when receiving the following input (before the arrows), it returns the corresponding scansion output (after the arrows):

- o pato lamenta quando quer → o / pa / to / la / men / ta / quan / do / quer #
- a patinha de quem tem saudades → a / pa / ti / nha / de / quem / tem / sau / da # / des

The first verse has, unambiguously, nine poetic syllables, ending on the stress in "*quer*." The second verse is scanned in a way that applies a synizesis to the word "saudades," treating the "sau" as a single syllable to maintain the nine-syllable pattern. However, consider the following modification:

- o pato lamenta quando quer ver → o / pa / to / la / men / ta / quan / do / quer / ver #
- a patinha de quem tem saudades → a / pa / ti / nha / de / quem / tem / sa / u / da # / des

[6]A stanza is a structured unit of a poem, composed of a group of verses that follow a specific metrical and rhythmic pattern, often adhering to a particular rhyme scheme or thematic cohesion within the poem.

Now, the first verse contains ten poetic syllables, unambiguously, ending on "*ver*." This time, the system forces a diaeresis in the word "*saudades*", splitting it as "sa / u / da / des" to preserve the decasyllabic structure of the stanza.

Aoidos employs grapheme-to-phoneme conversion to determine syllables and applies naturalness criteria to pronunciation. Scanning a poem creates a base pronunciation for each verse, determines a dominant metrical pattern within a stanza, and re-scans, applying metrical transformations that adjust pronunciation to fit the rhythmic pattern.

The author explains that the generated pronunciations include a parameter called PTS, which aims to reflect their naturalness (Mittmann 2016, p 15). Initially, when a pronunciation is formed by simply concatenating the words in a verse, it receives a PTS value of zero. However, this starting point is not a neutral baseline but a reference for comparison. The PTS score functions as a relative measure, meaning that it helps compare different versions of the same verse, but not for drawing comparisons across entirely different verses.

In other words, the system applies naturalness rules based on a standardized model of Portuguese pronunciation, which the author acknowledges is inherently subjective (Mittmann and Maia 2017, p. 162). These rules become increasingly flexible as they adapt to the rhythmic constraints of a given stanza.

Rhythm and Poetic Metrics

Aoidos defines canonical rhythmic structures based on traditional Portuguese poetic forms, such as heroic verse, martelo, and sapphic verse. By scoring rhythm patterns according to probabilistic models, the system selects the most natural and metrically consistent interpretation for a given stanza.

Methodology and Results

Our research evaluated four approaches for the automatic scansion of Brazilian Portuguese verses using Large Language Models (LLMs): zero-shot prompting, few-shot prompting, chain-of-thought prompting, and fine-tuning. We conducted initial experiments with the first three approaches using a limited dataset and manual verification to establish a baseline for LLM capabilities in this task. These preliminary tests utilized both GPT-3.5-turbo-0125 and GPT-4-turbo-preview models. The final approach, fine-tuning, was conducted with a larger dataset and quantitative metrics using GPT-3.5-turbo-0125.

Zero-shot Prompting

In this approach, we evaluated the LLMs' ability to perform poetic scansion without being shown examples, only by interpreting a structured instruction. We used two models: `gpt-3.5-turbo-0125` and `gpt-4-turbo-preview`. The temperature parameter was set to 0.2 (down from the OpenAI default of 0.7) to increase determinism and reduce output randomness.

The prompt was structured to guide the model using four key elements:

- **Role**: The assistant acts as a poet's helper.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

- **Context**: The user needs poetic scansion.
- **Instruction**: Divide poetic lines into syllables and mark stressed syllables.
- **Output format**: Use "/" for syllable division, "*" before stressed syllables, and "" after the final stressed syllable.

Example Prompt:

```
You are an excellent poetry assistant.

I need to scan some verses (scansion is the syllabic
↪  division of poetic lines and identification of
↪  stresses).

When I give you a poem as input, I want you to return
↪  each line scanned with its rhythmic pattern marked.

The desired output format is:
1. Syllables separated by " / ".
2. Each stressed syllable is preceded by "*".
3. The final stressed syllable is followed by "#".
```

The GPT-4 model followed the format reasonably well, but still made frequent errors in syllable division and stress assignment. GPT-3.5 performed worse, often failing to even tokenize properly or respect syllabic boundaries.

This experiment showed that LLMs, when prompted in a zero-shot manner, are not sufficiently accurate for poetic scansion in Brazilian Portuguese. The likely cause is a lack of phonetic and orthographic specialization in the models' training.

Few-shot Prompting

Few-shot prompting improves over the zero-shot method by including a small number of correctly scanned examples before presenting the target input. We used `gpt-4-turbo-preview` with temperature 0.2.

The model was given the following few-shot prompt, with examples scanned using the Aoidos tool:

```
You are an excellent poetry assistant.

I need to scan some verses (scansion is the syllabic
↪  division of poetic lines and identification of
↪  stresses).

When I give you a poem as input, I want you to return
↪  each line scanned with its rhythmic pattern marked.

The desired output format is:
1. Syllables separated by " / ".
2. Each stressed syllable is preceded by "*".
3. The final stressed syllable is followed by "#".

Examples:

USER: Quando eu canto o seu coração se abala
RESPONSE: Quando eu canto o seu coração se abala ->
↪  *Quan- / do eu / *can- / to o / *seu / co- / ra- /
↪  *ção / se a- / *ba-# / lha

USER: Pois eu sou porta-voz da incoerência
RESPONSE: Pois eu sou porta-voz da incoerência -> Poi- /
↪  *s eu / sou / *por- / ta{ / *voz / da in- / coe- /
↪  *rên-# / cia

USER:
```

The model scanned 16 verses of the poem *Agalopado* by Alceu Valença, with results summarized in Table 1. Few-shot prompting significantly improved output quality compared to zero-shot prompting. Most outputs respected syllabic boundaries and occasionally captured *sinalefas*, but errors in stress assignment persisted, and the "#" marker was sometimes misused, appearing multiple times or in incorrect positions. While few-shot prompting outperformed zero-shot approaches, it still fell short of the precision required for reliable scansion, thus motivating our exploration of fine-tuning methods.

Chain-of-Thought Prompting

Poetic scansion involves a sequence of interdependent operations: identifying syllables, assigning stress, and applying phonological rules like *sinalefa*. Due to its procedural nature, we investigated whether Chain-of-Thought (CoT) prompting could improve LLM performance on this task. CoT has shown promise in reasoning-intensive tasks by guiding models to generate intermediate steps before final answers (Wei et al. 2022).

We tested the **zero-shot CoT** strategy, in which the prompt includes the steps and a reasoning cue (e.g., "Let's think step-by-step") but no example.

To test poetic CoT reasoning, we supplied a step-by-step prompt incorporating canonical versification rules, especially those by Bilac (Bilac and Passos 1905), along with simplified guidelines for syllabification and stress marking. An excerpt:

```
You are an excellent poetry assistant.

I need you to scan some verses. Scansion means dividing
↪  each poetic line into syllables and identifying the
↪  stressed ones.

Follow this process:
1. Identify all verses (one per line).
2. Identify stressed syllables for each word.
3. Apply sinalefa and synaeresis where needed.
4. Separate poetic syllables using " / ".
5. Mark stressed syllables with "*".
6. Mark the last stress with "#".

Example format:
" / *syllable / syllable / *syllable#"

Let's think step-by-step.
```

Despite clear instructions, the model's output was flawed. It frequently marked multiple # symbols, ignored *sinalefa* rules, and misassigned stresses. For example:

- 1. Do Gosto para o desgosto → Do / *Gos# / to / pa / *ra / o / des / *gos / to
- 6. O que fui hoje não sou → O / que / *fu# / i / *ho / je / não / *sou

Table 1: Few-shot prompting results for *Agalopado*

Verse	Scansion Output
Cego o sol seu cavalo de batalha	*Ce- / go / o / *sol / seu / ca- / va- / *lo / de ba- / *ta- / lha
E faço a lua brilhar no meio-dia	E / *fa- / ço a / *lu- / a / bri- / *lhar / no / *mei- / o- / *di- / a
Tempestade eu transformo em calmaria	Tem- / *pes- / ta- / de eu / trans- / *for- / mo / em / cal- / *ma- / ri- / a
E dou um beijo no fio da navalha	E / *dou / um / *bei- / jo / no / *fi- / o / da na- / *va- / lha

These results suggest that even with algorithmic cues, LLMs struggle to internalize linguistic heuristics for scansion, possibly due to insufficient training on phonological rules. CoT offered structure but did not yield reliable performance, reinforcing the need for fine-tuning approaches with explicit supervision.

Fine-tuning Approach

After observing limited success with prompt engineering techniques, we implemented a fine-tuning strategy using the Aoidos system as our ground truth. While Aoidos has demonstrated remarkable reliability in processing over 100,000 verses spanning three centuries of Portuguese poetry, it operates as a rule-based system with specific contextual dependencies. It requires complete stanzas for analysis and applies complex metaplasm rules based on the broader poetic context.

Our goal was to create a more flexible solution to handle individual verses independently, maintain natural phonetic divisions, and be easily integrated into modern NLP pipelines. By fine-tuning an LLM on Aoidos-processed data, we aimed to combine the accuracy of Aoidos's rule-based approach with the adaptability of neural models. This would allow for real-time processing of single verses, better handling of contemporary poetry variations, and easier integration with existing language models. Additionally, while Aoidos requires careful consideration of stanza-level metrics and complex phonetic rules, a fine-tuned LLM could learn to make these decisions implicitly, simplifying the scansion process while maintaining accuracy.

Model and Infrastructure The fine-tuning implementation utilized OpenAI's API with a structured message format containing system instructions, user input, and expected assistant output. We configured the base model GPT-3.5-turbo-0125 with default training temperature settings as determined by OpenAI's backend, while setting a low inference temperature of 0.1 to ensure largely deterministic outputs. This small amount of variability was intended to accommodate minor phonetic or segmentation ambiguities; however, preliminary tests with a temperature of 0.0 yielded nearly identical results, suggesting that this stochasticity had a negligible impact on performance. The training process followed OpenAI's recommended defaults, including three epochs, with batch size and learning rate managed automatically by the platform's backend systems.

For computational infrastructure, the training was conducted on OpenAI's cloud-based platform. The fine-tuning jobs were submitted via OpenAI's CLI and monitored through the OpenAI developer dashboard, which provided real-time updates on training progress, token consumption, and performance metrics. The training data was preprocessed locally and uploaded to OpenAI's storage system in JSONL format, ensuring compatibility with the fine-tuning pipeline. Each fine-tuning iteration was performed asynchronously, with model checkpoints managed internally by OpenAI's infrastructure. Upon completion, the fine-tuned model was deployed via OpenAI's API.

Dataset Preparation Our dataset preparation began with approximately 13,500 verses from Aoidos's test corpus — the data processing pipeline involved extracting individual verses and treating them as independent units to avoid contextual dependencies. We implemented batch processing of 800 verses at a time, which we determined empirically as the optimal batch size. Using BeautifulSoup for XML parsing, we extracted clean text and relevant metadata. The standardization process established consistent formatting for syllable divisions using "/" delimiters, stress markers with "*" prefixes, and end-of-count markers with "#" suffixes. We also preserved necessary metadata, including rhythmic patterns, verse classifications, and stress patterns.

To ensure data quality, we implemented rigorous validation procedures for syllable count consistency, stress marker placement, and proper end-of-count marker positioning. Duplicate verses were eliminated to prevent training bias. The final dataset was split using an 80-20 ratio for training and validation/testing, with stratified sampling to maintain the distribution of verse types.

Training Strategies We designed two training strategies to explore the relationship between data volume and model performance. The first was an iterative approach consisting of 18 sequential iterations, starting with 400 examples and incrementally adding 400 examples per iteration, eventually reaching 7,200. Each iteration preserved previous training while incorporating new data, with the test set growing proportionally at 30% of cumulative training data. The second strategy employed a single-pass training approach with 3,520 examples, designed to compare against the iterative approach while matching the data volume of middle-range iterative runs.

Both training approaches maintained consistent configurations, including OpenAI's default fine-tuning rate, automatically optimized batch sizes, and three training epochs. We set a maximum of 4,096 tokens per example during training. For model deployment, we configured strict parameters with an inference temperature of 0.1, a maximum output of 1,024 tokens, and neutral frequency and presence penalties to ensure consistent and deterministic outputs.

ISBN: 978-989-54160-7-3

Performance Evaluation

To evaluate model performance, we developed three complementary metrics that assess different aspects of scansion accuracy. The Syllable Match Percentage (SYL_MATCH_%) performs an exact string comparison between the model's output and Aoidos's scansion, measuring both correct syllabic division and accurate stress marking. The Unstressed Syllable Match (SYL_MATCH_UNS_%) focuses solely on syllabic division by removing stress markers before comparison, acknowledging that stress patterns can vary in poetic recitation while maintaining valid scansion. The Syllable Count Match (SYL_COUNT_MATCH_%) compares the total number of syllables to the last stressed position, addressing the fundamental requirement of maintaining consistent metrical length in Brazilian Portuguese poetry.

We evaluated the model's performance using two training approaches: compound (iterative) and single-pass training. The compound approach achieved its best performance in the 14th iteration, reaching 88.60% syllable match (Figure 1), 92.30% unstressed syllable match (Figure 2), and 87.56% syllable count match (Figure 3). We conducted single-pass experiments with two different dataset sizes: an intermediate run with 3,520 examples achieved comparable results (87.19% syllable match, 91.45% unstressed syllable match, and 88.30% syllable count match). In contrast, a more extensive run with 7,200 examples showed marginal improvements (89.15% syllable match, 92.80% unstressed syllable match, and 89.10% syllable count match).

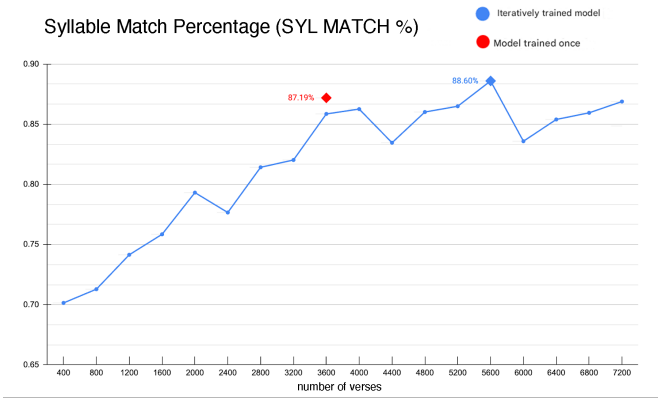


Figure 1: Syllable Match Percentage (SYL _ MATCH _ %): iterative and single-pass training

These results suggest that while larger datasets tend to improve performance, the gains are incremental, and even a moderate-sized single training run can achieve competitive results. When allowing for a ±1 syllable margin of error to account for valid poetic variations, the syllable count accuracy improved to 97.38%, 97.12%, and 97.85% for compound, intermediate single-pass, and large single-pass approaches, respectively.

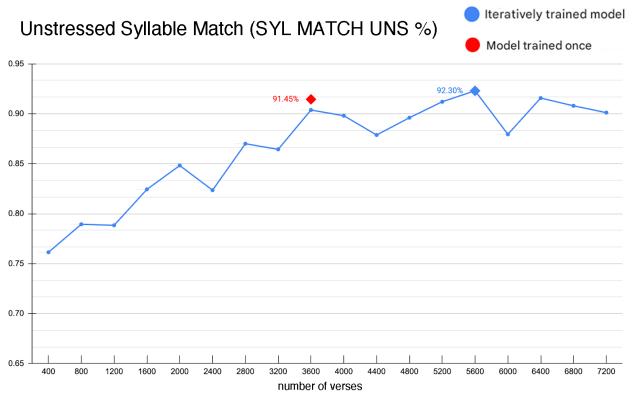


Figure 2: Unstressed Syllable Match (SYL _ MATCH _UNS _ %): iterative and single-pass training

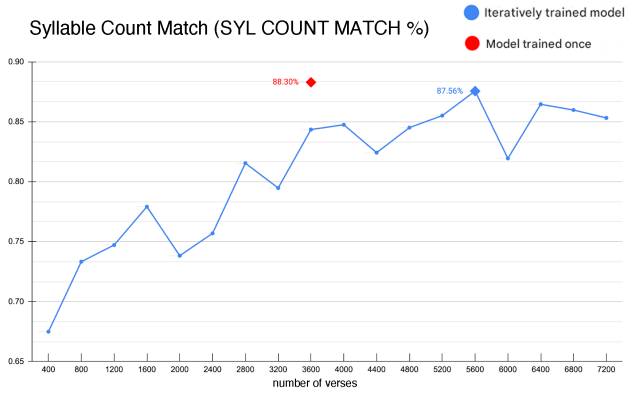


Figure 3: Syllable Count Match (SYL _ COUNT _ MATCH _ %): iterative and single-pass training

Cost Analysis

Our iterative training approach accumulated $116.43 across 18 iterations ($109.78 for training, $5.00 for prompts, and $1.42 for completions), while a single training run with 3,520 examples cost only $6.80. Inference costs remained consistent at approximately 550 tokens per scansion (505 for input, 45 for output), resulting in $0.00025 per verse at current OpenAI pricing. These costs, while minimal for individual verses, become significant when scaling to extensive poetry collections.

Linear regression analysis suggests that achieving a 100% match with Aoidos would require approximately 14,000 examples at an estimated cost of $250. However, this theoretical perfect accuracy is likely unattainable due to inherent poetic variations. We observed significant diminishing returns after the 14th iteration (5,600 examples), where we achieved 88.60% syllable match accuracy, suggesting this as an optimal efficiency threshold for practical applications. Future work might explore more cost-effective architectures

or token optimization techniques while maintaining acceptable accuracy.

Discussion

Our results demonstrate the viability of using fine-tuned LLMs for Brazilian Portuguese verse scansion while revealing significant trade-offs between accuracy, cost, and practical applicability. The progression from prompt engineering approaches to fine-tuning showed that while sophisticated prompting techniques can improve performance, they cannot match the consistency achieved through targeted model adaptation. The high accuracy rates achieved by our fine-tuned models suggest that LLMs can effectively learn the complex rules of Brazilian Portuguese prosody, including challenging aspects such as synaeresis and synalepha.

Limitations

However, several limitations merit discussion. First, while our models achieve high accuracy when compared to Aoidos's output, they inherit any biases or limitations present in Aoidos's rule-based approach. This is particularly relevant for contemporary poetry, in which traditional metrical rules might be deliberately subverted. Second, the cost-benefit analysis reveals diminishing returns in performance improvements beyond certain thresholds, suggesting practical limits to model optimization within current technical and economic constraints.

Second, while assessing the impact of training size, we employed an incremental partitioning strategy in which the test set expanded alongside the training data. We acknowledge that this may have introduced variability in the evaluation. Future experiments should fix the test set to ensure consistent and comparable assessments across training sizes.

A significant technical limitation of our approach is the vendor lock-in associated with OpenAI's fine-tuning platform. Each inference request incurs a cost, and we are dependent on the OpenAI API for both training and deployment. This introduces sustainability concerns for long-term applications and restricts potential adaptations to alternative model architectures or open-source solutions that might better serve specific linguistic contexts.

Another limitation of our study is its temporal nature - given the rapid advancement of language models, with new architectures and larger reasoning models being released frequently, our findings represent a snapshot of current capabilities rather than a definitive solution. Models with enhanced reasoning capabilities or more efficient fine-tuning approaches might soon enable new methodologies that could surpass our results.

Our analysis also revealed systematic errors in specific cases, particularly in syllable fusion and incorrect tonic markings. These errors highlight that scansion remains primarily a phonetic rather than an orthographic problem. Furthermore, while our models can learn metric patterns through fine-tuning, the inherently subjective nature of poetic interpretation and the complexities of oral performance present ongoing challenges for full automation.

Further Research Directions

These findings point to several promising directions for future research. One avenue would be to incorporate human expert annotations to create a more diverse training dataset that captures variations in poetic style and metrical interpretation. Another would be to explore more efficient model architectures or training approaches that could reduce computational costs while maintaining accuracy. Additionally, developing specialized embedding spaces for poetic features could potentially improve the model's understanding of metrical patterns while reducing training data requirements. The rapid evolution of LLM technology also suggests the importance of continuously evaluating new models and architectures as they emerge, particularly those designed for complex reasoning tasks that might be better suited to handling the nuanced rules of prosody.

Conclusion

This study demonstrated the feasibility of using Large Language Models for Portuguese verse scansion through systematic evaluation of different approaches. While GPT-4 showed better comprehension than GPT-3.5 in prompt engineering experiments, both models struggled with zero-shot and chain-of-thought approaches. Few-shot prompting improved metric and rhythmic structure understanding, but fine-tuning emerged as the most effective method. Our evaluation metrics showed that, with a robust dataset, the fine-tuned model achieved 88.6% accuracy in syllabic separation and 97.4% metric correspondence when allowing a margin of error of ±1 syllable, providing a potential foundation for automated scansion tasks.

Looking beyond technical aspects, our work raises fundamental questions about the relationship between poetry and technology. While machines can effectively scan verses and learn metrical patterns, they cannot replicate the whole poetic experience. Scansion, as a culturally conditioned process, requires understanding not just rules but also when and how these rules might be artistically broken. This understanding remains uniquely human, suggesting that while AI can be a valuable tool for literary analysis and experimentation, the essence of poetic creation and interpretation remains firmly in the human domain.

Future Work

This research paves the way for the development of the Pajeú platform, aimed at enhancing the skills of popular poets. The scansion model can be utilized both for validating automatically generated verses and guiding metric construction. Future research directions include:

- Generation of structured stanzas: Application of fine-tuning for décimas, sextilhas, and other traditional forms.
- Semantic and phonetic expansion: Use of RAG to enrich rhymes and cultural references in Northeastern Brazilian poetry.
- Orality and performance: Implementation of speech synthesis models that respect the prosody of improvised poetry.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Our results contribute to the ongoing development of computational approaches to Portuguese poetry analysis, building on prior efforts while acknowledging that poetry, in its essence, is not just about following rules—it's also about knowing when to break them.

References

[Abonizio et al. 2025] Abonizio, H.; Almeida, T. S.; Laitz, T.; Junior, R. M.; Bonás, G. K.; Nogueira, R.; and Pires, R. 2025. Sabiá-3 technical report.

[Almeida et al. 2024] Almeida, T. S.; Abonizio, H.; Nogueira, R.; and Pires, R. 2024. Sabiá-2: A new generation of portuguese large language models.

[Bilac and Passos 1905] Bilac, O., and Passos, G. 1905. *Tratado de versificação*. Rio de Janeiro, Rio de Janeiro: Livraria Francisco Alves.

[Caldas 2021] Caldas, P. 2021. Ano passado eu morri, mas esse ano eu não morro. `http://glo.bo/439ttyW`.

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

[da Cunha and Cintra 2017] da Cunha, C. F., and Cintra, L. F. L. 2017. *Nova gramática do português contemporâneo*. Rio de Janeiro, Rio de Janeiro: Lexikon Editora Digital.

[Davis et al. 2015] Davis, N.; Hsiao, C.-P.; Popova, Y.; and Magerko, B. 2015. *An Enactive Model of Creativity for Computational Collaboration and Co-creation*. London: Springer London. 109–133.

[Goldstein 1988] Goldstein, N. 1988. *Análise do poema*. São Paulo, São Paulo: Editora Ática S.A.

[Herbold et al. 2023] Herbold, S.; Hautli-Janisz, A.; Heuer, U.; Kikteva, Z.; and Trautsch, A. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Nature*.

[Hoover 2023] Hoover, J. 2023. Make any large language model a better poet. `https://replicate.com/blog/turn-your-llm-into-a-poet`.

[Mittmann and Maia 2017] Mittmann, A., and Maia, S. R. 2017. Análise comparativa entre escansões manual e automática dos versos de Gregório de Matos. (10).

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

[Oliveira and Alves 2016] Oliveira, H. G., and Alves, A. O. 2016. Poetry from concept maps – yet another adaptation of poetryme's flexible architecture. In *Proceedings of the 7th International Conference on Computational Creativity (ICCC)*.

[Oliveira et al. 2019] Oliveira, H. G.; Mendes, T.; Boavida, A.; Nakamura, A.; and Ackerman, M. 2019. Co-poetryme: Interactive poetry generation. *Cognitive Systems Research* 54:199–216.

[Oliveira, Mendes, and Boavida 2017a] Oliveira, H. G.; Mendes, T.; and Boavida, A. 2017a. Co-poetryme: a co-creative interface for the composition of poetry. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, 70–71.

[Oliveira, Mendes, and Boavida 2017b] Oliveira, H. G.; Mendes, T.; and Boavida, A. 2017b. Towards finer-grained interaction with a poetry generator. In *Proceedings of ProSocrates 2017: Symposium on Problem-solving, Creativity and Spatial Reasoning in Cognitive Systems*, CEUR Workshop Proceedings, 1–10.

[Oliveira 2012] 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 2019] Oliveira, H. G. 2019. On the impact of the grammar and the semantic network in a versatile poetry generator. In *Proceedings of the 6th Computational Creativity Symposium at 2019 AISB Convention*.

[Ormazabal et al. 2022] Ormazabal, A.; Artetxe, M.; Agirrezabal, M.; Soroa, A.; and Agirre, E. 2022. Poelm: A meter- and rhyme-controllable language model for unsupervised poetry generation.

[Pires et al. 2023] Pires, R.; Abonizio, H.; Almeida, T. S.; and Nogueira, R. 2023. *Sabiá*. Springer Nature Switzerland. 226–240.

[Wade 2023] Wade, A. 2023. How it's made. `https://developers.googleblog.com/2023/08/how-its-made-lupe-fiasco-text-fx.html`.

[Wei et al. 2022] 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.

[Wilson 1986] Wilson, L. 1986. *Roteiro de velhos cantadores e poetas populares do Sertão*. Recife, Pernambuco: Centro de Estudos de História Municipal.

ISBN: 978-989-54160-7-3

LuminAI: Embodied AI as a Catalyst, Constraint, and Co-Creator in Dance Improvisation Class

Milka Trajkova
Expressive Machinery Lab
Georgia Institute of Technology
Atlanta, GA 30308 USA
mtrajkova3@gatech.edu

Andrea Knowlton
Dance Department
Kennesaw State University
Kennesaw, GA 30144 USA
agisekno@kennesaw.edu

Jasmine Kaur
Expressive Machinery Lab
Georgia Institute of Technology
Atlanta, GA 30308 USA
jkaur47@gatech.edu

Brian Magerko
Expressive Machinery Lab
Georgia Institute of Technology
Atlanta, GA 30308 USA
magerko@gatech.edu

Abstract

As co-creative AI systems gain traction in artistic fields, their integration into formal improvisational pedagogy remains underexplored. This study examines how LuminAI, an AI dance partner, mediates improvisational practices within a structured educational setting, investigating how it reshapes norms and contradictions between pedagogy and AI-mediated co-creation. Using Cultural-Historical Activity Theory as an analytical framework, we conducted a three-month diary study with ten dancers and one instructor in an undergraduate improvisational dance course, capturing student reflections and instructor interviews. Our findings reveal that LuminAI functioned as both a catalyst and constraint in improvisation, reshaping movement decision-making, spatial awareness, and collaboration. Dancers adapted strategies to accommodate AI tracking limitations, adjusting tempo, spatial positioning, and clarity of movements. While some students experienced frustration due to technological constraints, others reported increased self-reflection, creative expansion, and novel co-creation dynamics. From an instructional perspective, the AI disrupted class structures, requiring the instructor to navigate tensions between AI engagement and human improvisational flow. Key contradictions emerged—between spontaneity and constraint, autonomy and responsiveness, and pedagogy versus AI-mediated co-creation. By analyzing these adaptations, we contribute to discussions on AI in co-creative education, advocating for systems that enhance rather than restrict embodied pedagogy, fostering new movement possibilities within improvisational dance.

Introduction

The integration of AI into artistic and performative practices has expanded beyond experimental settings into real-world applications, particularly in dance improvisation, choreography ideation, and interactive performance. AI systems such as **LuminAI** (Trajkova et al., 2023) have been used to generate movement sequences, inspire creativity, and act as improvisational agents, augmenting human artistry in novel ways (Long et al., 2020; Jacob et al., 2019; Deshpande and Magerko, 2024a). However, most studies on co-creative AI focus on short-term evaluations in structured workshops or controlled settings, limiting insights into how these systems integrate into professional creative workflows over time (Zhou et al., 2024; Maher et al., 2023; Deshpande et al., 2024; Chen and Tong, 2023; Kim and Maher, 2021; Lucas and Martinho, 2017). Research on AI-assisted dance (Liu and Sra, 2024a; Wallace et al., 2023), music composition (Ford et al., 2024; Suh et al., 2021), and digital art tools (Deshpande et al., 2024) typically relies on immediate task-based assessments rather than longitudinal studies of sustained artistic adaptation. Even multi-session studies (Davis et al., 2024; Deshpande and Magerko, 2024b) remain limited in duration, leaving a gap in understanding how AI tools evolve with user adaptation and influence creative processes over time. This study addresses this gap by examining how **LuminAI** (Trajkova et al., 2023) functions within a formal undergraduate improvisational dance course, focusing on its role as a catalyst, constraint, and co-creator in the creative process. Using Cultural Historical Activity Theory (CHAT) (Batiibwe, 2019), we investigate the research question: ***How does LuminAI, an embodied AI, reshape the system of an improvisational dance class, and what contradictions emerge between traditional pedagogy and AI-mediated co-creation?*** To explore this, we conducted a three-month longitudinal diary study involving ten dancers and one instructor, collecting qualitative data through student logs, written reflections, and instructor interviews. Our findings reveal that LuminAI significantly restructured improvisational norms, shifting students' focus from human-human interaction to human-AI collaboration. While some dancers adapted their movement strategies—modifying spatial awareness, tempo, and gesture clarity—others found LuminAI's tracking constraints frustrating, exposing tensions between AI-imposed limitations and the fluid, spontaneous nature of improvisation. Rather than serving purely as a creative aid, LuminAI acted as both an enabler and an obstacle, reshaping the way dancers approached improvisation. From an instructional perspective, LuminAI transformed classroom dynamics, requiring the instructor to navigate between AI engagement and traditional group improvisation. The system's constraints, while sometimes restrictive, also encouraged deeper self-reflection and alternative movement exploration. These findings align with (Deshpande and Magerko, 2024b), who advocate for adaptable AI that responds to human improvisation, and (Bown et al., 2020), who emphasize the importance of dialogue-based co-creativity in AI-driven artistic collaborations. Ul-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

timately, this study highlights the dual role of AI in dance pedagogy—both as a disruptive force that challenges established improvisational methods and as a tool that fosters new forms of creative exploration. By identifying key contradictions and adaptations in AI-mediated dance education, we contribute to ongoing discussions on AI's role in co-creative learning environments. We argue that AI systems must prioritize adaptability, transparency, and intentional design to better support embodied pedagogical practices, ensuring that technology enhances rather than restricts artistic expression.

Related Work

Artificial intelligence (AI) has emerged as a powerful co-creative agent across artistic, educational, and design domains. Researchers have explored how AI systems collaborate with humans in generative processes, improvisation, and ideation. However, much of this research emphasizes short-term usability and immediate interactional effects, leaving gaps in understanding how co-creative AI tools evolve over time or embed within professional and educational workflows. This section reviews three key areas of prior work: co-creative AI in dance and choreography, thematic analysis methodologies in co-creative AI research, and classroom deployments of AI systems that support creativity.

Co-Creative AI in Dance and Choreography

Dance has become an important focus in co-creative AI research, where AI-driven systems assist with choreographic ideation, improvisation, and movement augmentation. Studies on systems such as LuminAI investigate human-AI improvisation, revealing the movement adaptations required for AI-mediated performance (Trajkova et al., 2024). Tools like DanceGen provide early-stage choreography support but face challenges in aligning AI-generated sequences with artistic intent (Liu and Sra, 2024a). Other projects, such as Kinetic Dialogues, use AI-driven avatars to generate motion feedback loops, encouraging dancers to break habitual patterns, though sometimes disrupting structured rehearsal processes (Berman and James, 2015). Robotic dance studies similarly highlight tensions in maintaining thematic coherence, underscoring limitations in AI's capacity to generate compelling, integrated performances (Filippo et al., 2023). Despite these advancements, most research focuses on short-term interactions, leaving unexamined how dancers refine their use of AI systems over extended periods.

Challenges and Gaps in Longitudinal Evaluation

Across artistic domains, co-creative AI tools face persistent challenges, including aligning outputs with artistic intent, maintaining real-time adaptability, and cultivating user trust. Numerous studies emphasize that AI-assisted creativity requires ongoing human intervention, as artists refine AI-generated outputs to meet their creative goals (Liu, 2024; Pataranutaporn et al., 2024). Technical unpredictability—while sometimes a source of inspiration—can also disrupt structured creative workflows, creating friction in practice (Wallace et al., 2023; Deshpande et al., 2024). Critically, most evaluations remain confined to single-session experiments, assessing usability and creativity at a moment in time, rather than exploring how human-AI collaboration evolves within sustained professional or artistic settings.

Use of Thematic Analysis in Co-Creative AI Research

To capture the complex, subjective dimensions of human-AI collaboration, recent studies have increasingly employed thematic analysis as a qualitative research method. III (2025) applied thematic analysis to tabletop role-playing game experiences, assessing how generative AI enhances narrative co-authorship. In professional design, Park et al. (2024) revealed that designers predominantly frame AI as a visual ideation tool rather than a verbal collaborator, emphasizing the need for modality-sensitive systems. He and Do (2025) investigated industry perspectives on AI attribution, highlighting challenges around authorship boundaries. Ethical concerns were central in the work of Rezwana and Maher (2023c), who used design fiction to examine user trust and agency in human-AI co-creativity. In educational contexts, Jonsson and Tholander (2022) and Fu et al. (2025) applied thematic analysis to uncover tensions between automation, learner agency, and collaborative creativity. Thematic analysis has also supported the development of design frameworks, such as COFI (Rezwana and Maher, 2023a), which models human-AI interaction in co-creative systems. Overall, this methodological approach has provided critical insights into the subjective, relational, and ethical dimensions of co-creative practice.

Classroom Deployments of AI Systems Supporting Creativity

Beyond artistic domains, the classroom has become an increasingly important site for AI deployments that aim to foster creativity. Rane, Choudhary, and Rane (2023) explored how Education 4.0 and 5.0 frameworks leverage AI for personalized, adaptive learning, emphasizing student-centered creativity. Marrone, Taddeo, and Hill (2022) examined student perceptions of AI in creative tasks, highlighting the need to design educational AI systems that actively support creative thinking. Southworth et al. (2023) advanced the AI Across the Curriculum model, showing how embedding AI literacy across disciplines can empower creative student engagement. Grassini assessed the cognitive and creative impacts of tools like ChatGPT, while Wang, Sun, and Chen (2023) demonstrated that institutional AI capacity correlates with students' self-efficacy and creative performance. Addressing ethical considerations, Akgun and Greenhow Akgun and Greenhow (2022) emphasized the importance of constructionist, agency-oriented approaches in K-12 AI deployments. Holstein, McLaren, and Aleven (2019) developed real-time classroom orchestration tools to enhance hybrid human-AI collaboration, while Dimitriadou and Lanitis (2023) provided critical evaluations of smart classrooms, underscoring the need for alignment between technological affordances and pedagogical goals. For the ICCC community,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

these studies offer valuable insights into how AI systems can not only assist artistic professionals but also cultivate creativity among learners in formal educational settings.

Study Design

A diary study, a qualitative longitudinal research method, was used to capture real-time participant experiences (Bolger, Davis, and Rafaeli, 2003). We employed a feedback study approach, where dancers responded immediately to structured prompts, reducing recall bias (Czerwinski, Horvitz, and Wilhite, 2004; Hess and Wulf, 2009). This method provided rich insights into how AI influenced improvisational decision-making, collaboration, and creativity. This study was guided by the following research question: How does LuminAI, an embodied AI, reshape the system of an improvisational dance class, and what contradictions emerge between traditional pedagogy and AI-mediated co-creation? This study received ethics approval from the Institutional Review Board at Georgia Tech, ensuring that all procedures adhered to established standards for research involving human participants. Prior to participation in the dance improvisation class, all students were informed of the nature of the class and were provided with detailed information sheets explaining the purpose, procedures, potential risks, and benefits of the study. Written informed consent was obtained from all participants. Given the dual role of the instructor as both teacher and research collaborator, special attention was paid to mitigating potential power dynamics. Researchers clearly communicated that the study was voluntary if enrolled and would not affect students' grades, class standing, or relationship with the instructor. To further reduce implicit pressures, the instructor did not have access to raw data or identifiable participant responses during the analysis phase. Research team members maintained participant confidentiality by anonymizing all transcripts and observation notes, ensuring that reported findings reflected collective patterns rather than individual performances or identities.

Participants

The study involved ten college-level BA students and one instructor from Kennesaw State University, a university located in the southeastern United States outside of Atlanta, GA. Nine identifying as female and one as male were undergraduates pursuing majors in dance, theater, psychology and exercise sciences and had 4 to 18 years of professional dance and choreography experience across modern/contemporary, improvisation, jazz, tap, African/afrobeats, hip-hop, liturgical, lyrical, musical theater, acrobatic, ballroom, Latin: bachata, salsa, tango, heels, belly dance, Chinese folk, and flamenco among other dance forms. The students had also been involved in a dance performance between 1 to 10+ times in the past year. The instructor who self-identified as a female has 21 years of teaching experience. The instructor is a choreographer, filmmaker, and educator with expertise in dance technique, improvisation, pedagogy, and the integration of technology in dance.

Class Structure

Each 75-minute session followed a structured format: **Warm-up (15 min)** focused on attendance, announcements, and physical exercises. **Concept Exploration (20–30 min)** introduced improvisational techniques and compositional strategies. **AI Interaction (30–35 min)** involved engaging with LuminAI to explore movement qualities like weight, space, and time. **Group Collaboration (10–15 min)** refined movement sequences inspired by AI interactions. Finally, **Reflection (10 min)** consisted of journaling insights on creativity, expressiveness, and adaptation.

Data Collection

The diary study used a structured 15-point questionnaire designed to capture dancers' immediate reflections on their experiences engaging with the LuminAI system. This questionnaire included prompts covering emotional reactions (e.g., "How did you feel about your interaction with LuminAI today?"), perceived creative outcomes (e.g., "Give examples of any instances where the tool sparked new ideas or influenced your movement choices."), technical interactions (e.g., "Did you face any problems today while using LuminAI?"), and broader reflective themes (e.g., "How did LuminAI challenge you today?".

To respect the natural flow of the class, the questionnaire was not completed mid-session but administered during the reflection portion at the end of each class. Dancers did not stop dancing to respond; rather, the reflection phase was a designated time after the movement activities, intentionally structured as part of the class to allow for cool-down, discussion, and meta-level reflection. Such reflective practice is already a standard component of many improvisation and dance pedagogy contexts, so incorporating written diary entries aligned naturally with existing pedagogical rhythms and did not represent a major disruption.The instructor provided parallel logs, reflecting on student interactions and pedagogical shifts. Additional data included two student reflections (during the halfway point and at the end) and an instructor's written reflection, supplemented by a retrospective interview exploring AI's impact on improvisation norms and class dynamics.

Final Performance

The study culminated in a live improvised performance with **LuminAI**, marking, to our knowledge, the first human-AI collaborative improvisation in a formal dance setting.

Cultural-Historical Activity Theory (CHAT) as Deductive Analysis and Inductive Analysis

We employed an iterative coding process grounded in Cultural-Historical Activity Theory (CHAT) (Batiibwe, 2019), a framework that examines human activity as a dynamic system shaped by interactions between subjects, tools, community, rules, division of labor, and outcomes. Recognizing that these components are historically developed and continually evolving, CHAT allowed us to situate the dance class not as isolated events but as a broader, socially and culturally mediated system.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Our analysis combined deductive and inductive approaches (Braun and Clarke, 2006). Deductively, we applied CHAT's seven core elements—*Subject, Object, Tools, Rules, Community, Division of Labor, Outcome*—to organize and interpret interactions within the activity system, ensuring theoretical alignment. Inductively, we engaged in open, bottom-up coding to capture emergent, context-specific patterns beyond the CHAT categories, such as students' shifting emotions, adaptive strategies, and unanticipated improvisational practices.

This process produced a comprehensive codebook of 20 themes, 29 sub-themes, and 60 distinct codes, systematically grouped under the CHAT framework: students (**Subjects**), pedagogical aims (**Object**), the LuminAI system (**Tools**), governing norms and constraints (**Rules**), the class collective (**Community**), shared and shifting roles (**Division of Labor**), and the creative transformations (**Outcome**). This hierarchical organization enabled a nuanced understanding of the multifaceted interactions shaping the human-AI improvisational space. A link to the CHAT diagram and codebook can be found **here**.

A central analytic focus was on identifying **contradictions**, or tensions within and between system components, which CHAT views as key drivers of change. We traced primary contradictions (within single elements, such as students' conflicting expectations), secondary contradictions (between elements, such as tool limitations versus creative goals), tertiary contradictions (between old and new practices), and quaternary contradictions (linking the activity to larger institutional or cultural contexts). By analyzing these tensions, we illuminated how integrating LuminAI reshaped pedagogical practices, creative processes, and social dynamics, offering critical insights into the evolving nature of human-AI co-creation.

Positionality

We recognize that in thematic analysis, researchers' backgrounds and interpretative frameworks significantly influence the data analysis process (Camurri et al., 2016). To enhance transparency and contextual understanding, we provide background information on the authors. The first author, who holds a Ph.D. in HCI, has a background as a professional ballet dancer. The second author is a dance professor with 34 years of dance training and has worked extensively with most of the students involved in this study. The third author is a Human-Computer Interaction (HCI) researcher with professional dance experience. The final author, who holds a Ph.D. in Computer Science, has expertise in computer science and cognitive science, with extensive experience in co-creative AI, museum and art installations, and creative computing education. By outlining these backgrounds, we aim to provide the reader with a clearer perspective on how our interpretative lenses may have shaped our thematic analysis.

Results

This section presents the findings of our study on integrating LuminAI into an improvisational dance class using Cultural-Historical Activity Theory (CHAT) as an deductive analytical framework as well as inductive analysis. The results reveal how LuminAI influenced creative collaboration, pedagogical dynamics, and movement generation, as well as how students and instructors adapted to its presence. Our analysis is informed by the daily student and instructor logs, two written student reflections in the middle and at the end of the study, instructor post-study questionnaire responses, and a retrospective interview with the instructor, highlighting key contradictions, emerging patterns, and the evolving perception of LuminAI over time.

Subject: Student and Instructor Perspectives

Student Experience: From Skepticism to Adaptation

When first introduced to LuminAI, many students viewed it as an unnecessary or even frustrating addition to their improvisational practice. Some felt that AI lacked a clear purpose in dance, with Dancer 3 admitting, *"I didn't really get why we need to integrate AI into dance and I don't really like it."* Others were initially disappointed by the system's glitches and lack of responsiveness, with Dancer 5 remarking, *"I was expecting more, but the AI kept glitching and wasn't picking up what I was doing."* The perception that LuminAI failed to recognize or reflect the nuance of their movements caused some students to disengage entirely. Dancer 4 candidly shared, *"I did not really pay attention to the AI, it doesn't really pick up what I'm doing anyways... I got my inspiration from other dancers much better."* Over time, however, many students adjusted their expectations and learned to work within the system's constraints. Some began to see LuminAI as a tool for creative exploration rather than as a direct co-creator. Dancer 5, who had initially resisted the AI, later acknowledged, *"With every session, I began to see improvement... The movements of LuminAI were starting to look like actual dances."* Another reflected on their gradual shift in perception, saying, *"At first, I was really skeptical, but by the end, I was more comfortable and confident to try new things and branch out with my improv with the AI"* (Dancer 10). By the end of the semester, students had developed a range of responses to LuminAI. Some continued to question its artistic purpose, while others found it useful for breaking habitual movement patterns and discovering new possibilities. Dancer 9 described this realization, stating, *"When dancing with LuminAI, I realized that my improv is very repetitive. The AI encouraged and inspired me to branch out."*

Instructor Perspective: Mediating AI in the Classroom

The instructor initially struggled to determine how best to incorporate LuminAI into an improvisational curriculum, which traditionally prioritizes spontaneity, human connection, and embodied feedback. She acknowledged that integrating AI into the classroom was both an exciting opportunity and a pedagogical challenge, explaining, *"It takes a lot of focus and doesn't seamlessly integrate into many ideas – you have to be very specific about how you want to use it and*

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

design experiences to meet those goals." As the semester progressed, the instructor observed growing student engagement, which shaped her own perception of LuminAI's role in the classroom. She reflected, *"My perception is much influenced by how my students are receiving the tool. In the past few weeks, I have seen some excitement emerge, which lifts my perception as well."* While she recognized LuminAI's potential for prompting new movement choices, she also noted its inability to replicate the depth of human connection that is fundamental to dance improvisation. Comparing LuminAI to traditional duet improvisation, she remarked, *"LuminAI does not compare to improvising with a real partner at this point."* Despite these limitations, she saw value in LuminAI as a creative stimulus, particularly in encouraging students to approach movement with greater intentionality and to rethink their relationship with technology in dance.

Object: Goals of AI Integration

The goal of integrating LuminAI into the course was to expand students' improvisational learning by introducing an AI system as an interactive dance partner. The instructor hoped that LuminAI would challenge students to develop adaptability, generate unexpected movement ideas, and explore new choreographic possibilities. At first, students struggled to understand the purpose of working with AI in a dance context. Dancer 3 admitted, *"I was confused about the point of a dance AI."* Others were intrigued by the intersection of technology and dance, with Dancer 6 stating, *"I was interested in how science and something like dance could come together to create something really cool."* As students continued engaging with LuminAI, many recognized its potential as a tool for movement reflection rather than as a replacement for human collaboration. Dancer 11 described this shift, explaining, *"To me, the AI was a reflection of myself. Not a direct mirror, but a reflection of the emotion and general ideas of my movements."* Another saw LuminAI as an unbiased movement generator, noting, *"It can show you your movement vocabulary while also showing you modified or new ways to move"* (Dancer 12). By the end of the course, most students had adjusted their expectations and found ways to engage with AI creatively, even if they did not see it as a fully autonomous improvisational partner.

Tool: AI as a Movement Mediator

LuminAI functioned as both a creative prompt and a constraint. Some students found inspiration in its responses, with Dancer 11 saying, *"Sometimes it would give me ideas that I could add into my improv."* Others, however, found its lack of real-time interaction limiting, describing it as *"cold"* compared to human partners (Dancer 4). Technical limitations shaped how dancers engaged with AI. Dancer 5 noted, *"I had to alter my way of thinking with my improvisation and be more creative with it,"* explaining how AI's inability to track floorwork or quick movements forced students to adjust their approach. Despite its limitations, some students appreciated LuminAI's capacity for iteration, seeing its evolving movements as a source of feedback. Dancer 9 said, *"At times, I get a little scared at how good the AI [can] dance and remember what the class has taught it."*

Rules: Adapting to AI Constraints and Changing Classroom Norms

Unlike traditional improvisational classes where dancers move freely, LuminAI imposed spatial and technical constraints. Students had to remain within a designated range for the AI's motion-tracking to function correctly and adjust their movement style to be readable by the system. Dancer 12 reflected, *"Even if I was inspired by something the AI did, it wasn't as easy to execute the movement if it was something larger than the space could accommodate."* Similarly, some had to slow down their movements for the AI to register them, which unexpectedly altered their movement quality. The instructor noted that students gradually learned to negotiate these constraints, explaining, *"It took a while for students to realize that the AI doesn't pick up everything they do, and so they started adapting their movement to what the AI could actually read."* Rather than responding spontaneously to a human partner, students had to strategize their movements to maximize engagement with LuminAI. Beyond technical adjustments, LuminAI reshaped behavioral norms in the classroom. Improvisational dance typically relies on immediate embodied feedback from peers, but because AI lacks emotional expression, students had to develop patience with miscommunication. At first, some reacted with frustration. Dancer 6 remarked, *"Trying to interact with the agent, while also interacting with my fellow dancers, while also having a stage presence… snuffed out my expression."* Over time, however, students adopted a more experimental mindset, reframing LuminAI's errors as opportunities for improvisation rather than failures. The instructor encouraged this shift, stating, *"I told them to stop expecting the AI to be a perfect partner. Instead, they needed to see what new ideas it could bring into their movement."* This change in classroom norms was evident in how students responded to LuminAI's occasional glitches. Early in the semester, they would pause or disengage when the AI failed to respond correctly. Later, many learned to continue moving despite system errors, treating technical disruptions as moments of creative discovery. Dancer 8 noted, *"At first, I would just stop when the AI froze. But later, I realized I could actually incorporate its pauses into my own dance."*

Community: Redefining Group Dynamics in an AI-Augmented Class

Initially, some students felt isolated by the individual nature of dancing with AI. In previous improvisation exercises, dancers moved freely, responding to each other in real-time, nonverbal dialogue. However, LuminAI required dancers to take turns interacting with the system, creating a sense of division in the class. Dancer 7 shared, *"At first, I felt that we were thriving by creating a semi-sacred space… sans technology. Then suddenly, we were being observed, and the AI was this third presence in the room."* The instructor also noticed that LuminAI initially disrupted the sense

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

of vulnerability and trust that is critical to dance improvisation, explaining, *"It felt like some students withdrew when the AI was introduced. It was like they were waiting to see how others handled it before engaging themselves."* Over time, students developed new ways of supporting each other while working with LuminAI. They began observing each other's interactions with the AI, offering feedback and reflections on how different movement styles influenced the AI's responses. Dancer 10 observed, *"Seeing how my classmates interacted with the AI helped me recognize my own movement habits."* Students also started sharing strategies for engaging with LuminAI, treating it as a collective learning tool rather than an isolated dance partner. Dancer 2 explained, *"At first, I thought it was just me who couldn't get the AI to respond well. But when we started talking about it, I realized we were all figuring it out together."* By the end of the semester, students described being part of an experimental, AI-integrated dance community as both challenging and rewarding. Dancer 11 reflected, *"We started as a dance class, but by the end, it felt like we were part of a bigger conversation about how AI and movement can coexist."*

Division of Labor: Shifting Roles Between Humans and AI

The instructor's role shifted from being the primary guide of improvisation to a mediator between students and AI. Traditionally, an improvisation instructor provides real-time feedback, guiding students through movement explorations. However, with LuminAI in the classroom, the instructor's role became more about structuring engagement with AI. She described her experience as, *"Not dancing with the AI as much as designing experiences for the dancers to interact with LuminAI."* Students were not simply dancing with LuminAI; they were also teaching it. Dancer 1 described the experience as, *"Teaching an AI like teaching a child to dance."* Many noticed that over time, the AI seemed to retain movement characteristics from their previous sessions, creating a sense of co-evolution between human and machine. This role reversal—where students were both improvisers and AI trainers—created new opportunities for reflection. Dancer 11 remarked, *"At times, I get a little scared at how good the AI [can] dance and remember what the class has taught it."*

Outcome: Transformations in Improvisational Practice and Learning

Students found that engaging with LuminAI pushed them beyond habitual movement patterns, fostering creative exploration and adaptability. Some discovered new movement possibilities, noting that the AI encouraged them to change tempo, break habitual rhythms, and expand their improvisational vocabulary. Others experimented with structured co-creation, building phrases iteratively with AI responses. While initially constrained by technical limitations, students adapted their strategies over time, improving their problem-solving skills and self-awareness of movement. The final performance reflected these transformations, as students successfully incorporated LuminAI, with some recognizing that its movements became more readable and expressive. Although skepticism about AI's role persisted, many students gained an appreciation for its potential in choreography, research, and movement analysis. However, LuminAI did not fully replicate the depth of human improvisational partners, as some dancers felt it lacked emotional reciprocity, limiting expressive engagement. The instructor acknowledged that while LuminAI contributed to the final performance, it remained an effortful integration rather than a seamless component of improvisation. Some students remained unconvinced of AI's artistic value, expressing concerns about its inability to replicate the primal, reflective nature of dance. While LuminAI demonstrated the potential for AI-assisted improvisation, refining its responsiveness, adaptability, and emotional sensitivity will be essential to ensuring that it enhances rather than constrains artistic expression.

Contradiction Analysis

Contradictions—internal tensions that drive change—emerged throughout the LuminAI dance class, shaping how students and the instructor navigated AI integration. Using the CHAT framework, we analyze primary (within a single component), secondary (between two components), tertiary (between old and new practices), and quaternary (between the activity and external contexts) contradictions.

Primary Contradictions (Within Individual Components)

Within Students Many students wrestled with conflicting feelings about AI in dance. Dancer 3 expressed, *"I still don't really get why we need to integrate AI into dance and I don't really like it... I am still trying to keep an open mind though."* This tension between skepticism and curiosity marked a key internal contradiction. Another common struggle was between creativity and inhibition. Dancer 6 admitted, *"The AI put a lock on my expressiveness because I felt like I couldn't work with it."* Initially, many students hesitated, unsure whether to simplify movements for the AI or stay true to their style. Over time, they reframed their approach, seeing LuminAI as a tool rather than a judge.

Within the Instructor The instructor balanced excitement for AI's potential with the challenges of preserving improvisation's organic nature. She described the process as both *"interesting and challenging"*—a tension between her traditional teaching philosophy and the structured demands of AI interaction. Additionally, spontaneity clashed with the need for careful planning: *"You have to be very specific about how you want to use it."*

Within the Tool (LuminAI) LuminAI itself contained contradictions between its intended purpose and actual performance. It was meant to act as a co-creative dance partner, yet students noted it *"just wasn't human-like whatsoever"*. Glitches, freezing, and limited responsiveness underscored the gap between concept and reality.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Within Rules Improvisation typically prioritizes freedom, yet AI required restraint. Dancer 4 expressed frustration: *"I had to transform my dancing into AI, which was very frustrating because that's not what dancing is about... I really think it should be AI cooperating with me, instead of me having to fit my dance into AI."* Even when students followed AI-friendly rules (staying within the tracking zone, simplifying movements), the AI sometimes failed to respond, undermining the system's logic.

Within Community LuminAI's solo interactions disrupted the collective energy of group improvisation. The instructor noted, *"It influences the sense of community by pulling individuals out of a group experience."* Initially, some students disengaged when it wasn't their turn, leading to fragmentation. Over time, community investment was rebuilt through shared reflections and feedback.

Secondary Contradictions (Between Two Components)

Tool vs. Object The goal of the class (Object) was to foster improvisational creativity, yet the tool (LuminAI) initially limited expressiveness. Dancer 5 observed, *"Compared to before we started with AI, I think I was more expressive than after the sessions began... trying to interact with the agent... snuffed out my expression."* However, students adapted by using repetition and smaller movements to find creative strategies within the AI's constraints.

Subject vs. Tool Students expected an intuitive, responsive dance partner but instead encountered an impersonal system. Dancer 8 reflected, *"I like to incorporate [a partner's] feelings... although LuminAI is cool, I don't get that same personal connection... with other people."* The absence of emotional reciprocity clashed with dancers' need for human connection, requiring them to redefine their expectations of collaboration.

Subject vs. Rules New AI-based constraints, such as staying within a fixed area, conflicted with dancers' instincts to move freely. Dancer 7 commented, *"Even if I was inspired by something the AI did, it wasn't as easy to execute the movement if it was something larger than the space could accommodate."* The tension between artistic impulse and imposed limitations remained an ongoing challenge.

Tool vs. Community LuminAI's one-at-a-time interaction model disrupted the communal nature of the class. Initially, some students lost engagement when not directly interacting with the AI. Dancer 2 reflected, *"Integrating the avatar was wonky."* Over time, students adapted by observing and providing feedback, creating a sense of shared investment.

Tertiary Contradictions (Between Old and New Practices)

Traditional improvisation relies on human responsiveness, spontaneous energy flow, and full spatial freedom. AI-mediated improvisation introduced interruptions—students had to focus on being visible to the sensor, adjust their movements to fit AI constraints, and wait for responses. Dancer 9 reflected, *"Before AI, we had a semi-sacred space to build trust and human connection... sans technology."* The integration of AI felt like a return to self-consciousness, requiring adjustments in mindset and pedagogy. The instructor also had to shift from open-ended improvisation to structured AI-based exercises.

Quaternary Contradictions (Between the Activity and External Contexts)

A key quaternary contradiction was between the class's experimental nature and traditional dance education norms. Students questioned AI's relevance to their professional development. Dancer 10 shared concerns, *"My only concern would be AI taking jobs from dancers in an already challenging environment."* This highlighted a broader conflict between artistic innovation and industry concerns. Institutionally, bringing AI into the studio required additional resources and technical support, revealing tensions between creative ambitions and logistical constraints. Contradictions in the LuminAI dance class initially created resistance and frustration, but over time, they drove adaptation and learning. Students and the instructor negotiated tensions between freedom and structure, emotional connection and technical constraints, and traditional improvisation and AI-mediated practice. While challenges remained, the process demonstrated that even unresolved contradictions could foster growth, prompting new strategies for AI-integrated improvisation.

Discussion

LuminAI as a Disruptor of Spatial Containment in Improvisation

LuminAI disrupts traditional spatial boundaries in improvisation, transforming dance spaces into dynamic, adaptive environments. Unlike static studio or stage settings, LuminAI creates a permeable spatial system where dancers continuously renegotiate movement boundaries. As observed in the results, dancers noted that their spatial awareness shifted in response to the AI's cues, requiring constant adjustment to maintain engagement. This aligns with research on AI-human territoriality in dance, where AI not only exists in space but actively reshapes movement environments (Long and Magerko, 2017; Gemeinboeck and Saunders, 2017). LuminAI's responsiveness leads dancers to fluidly step in and out of engagement, challenging conventional distinctions between participation and non-participation (Zhang et al., 2024). Rather than removing containment, LuminAI transforms it into an elastic, responsive framework. While AI introduces new spatial possibilities, dancers continue to impose learned spatial structures on their movement, even when improvising with AI (Trajkova et al., 2024). Prior research suggests that some level of spatial stability enhances dancer autonomy by providing a reference point for movement exploration (Long and Magerko, 2017). LuminAI's ability to shift spatial expectations highlights the evolving relationship between AI and improvisation, demonstrating

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

how technology can act as both a constraint and a catalyst for movement adaptation.

LuminAI as a Learning Partner, Not an Autonomous Improviser

Rather than an independent co-creator, LuminAI functions as a learning partner that requires human guidance. One dancer described, *"It wasn't really making its own choices—I felt like I had to teach it how to dance with me."* This supports the distinction between AI as a tool versus a creative partner (Jacob and Magerko, 2015; Ciolfi Felice and McDonnell, 2016). LuminAI reacts to dancer input rather than generating movement independently, forming an iterative process of co-regulation. Similar patterns appear in AI-assisted music and theater, where performers develop strategies to shape AI responses (Deshpande and Magerko, 2024b). However, recent work in computational creativity suggests AI could move beyond pure reactivity by incorporating anticipatory movement intelligence (Deshpande and Magerko, 2024b). Research on AI turn-taking in improvisation shows that dancers prefer systems that alternate between following and leading (Winston and Magerko, 2017). While LuminAI currently operates as a follower, future iterations could introduce moments where AI takes initiative to challenge the dancer.

LuminAI as a Reflective Mirror, Not a Generative Creator

A key finding is that LuminAI primarily functions as a mirror for movement reflection rather than a fully generative improviser. While it introduces variations, it does not generate novel choreographic material. *"It made me think about what I was already doing, but it didn't really push me in new directions,"* one dancer noted. This aligns with previous research on reflective AI improvisation, where systems echo and transform human input rather than producing original movement (Long and Magerko, 2017; Jacob and Magerko, 2015). AI-assisted choreography tools similarly function as lenses for movement inquiry rather than independent creators (Liu and Sra, 2024b). However, this reflective quality presents limitations. Some dancers found LuminAI's lack of unexpected variability restrictive, prompting the question: Should AI primarily serve as a tool for movement analysis, or should it act as a generative partner? Research suggests effective improvisation depends on balancing predictability with spontaneity—AI that is too repetitive is unstimulating, while excessive randomness disrupts coherence (Kantosalo and Toivonen, 2016). Future designs could incorporate adaptive spontaneity, allowing AI to introduce calculated risks in movement generation.

Cognitive and Cultural Implications of AI Improvisation

Beyond individual movement analysis, the integration of AI into dance raises broader cognitive and cultural questions about how technology reshapes artistic traditions. While prior research has explored how some AI systems encode and transform movement vocabularies — preserving certain stylistic features while introducing novel variations (Zhu and Tan, 2024; Pataranutaporn et al., 2024) — it is important to clarify that LuminAI was not explicitly designed with such goals. Rather, LuminAI's improvisational interactions primarily reflect real-time responses to dancer inputs, without underlying models aimed at learning or preserving specific dance styles. Nonetheless, its presence in the improvisational space prompts dancers and instructors to renegotiate choreographic norms, reconsider the boundaries of co-creative agency, and reflect on how digital partners might influence the evolution of embodied practices over time. However, concerns about algorithmic bias persist. If AI models primarily reflect Western contemporary dance aesthetics, they risk marginalizing diverse movement traditions (Crnkovic-Friis and Crnkovic-Friis, 2016). One dancer questioned, *"What happens when AI only recognizes certain styles? Are we training it to forget other kinds of dance?"* Future research should explore culturally diverse datasets and adaptive movement filters to ensure AI systems remain inclusive.

Contradictions and Challenges in AI Improvisation

- **Predictability vs. Novelty** – AI must balance predictability with creative surprise. If too repetitive, it becomes uninteresting; if too random, it disrupts coherence (Kantosalo and Toivonen, 2016). The challenge is designing AI that maintains dancer engagement while introducing occasional variation.
- **Control vs. Responsiveness** – Negotiating leadership is complex. *"Was I leading the AI, or was it leading me?"* one dancer asked. Research on AI turn-taking suggests that alternating roles fosters better engagement (Winston and Magerko, 2017). Future AI systems should dynamically shift between leading and following to maintain improvisational flow.
- **Choreographic Authorship and Dancer Agency** – The introduction of AI complicates authorship. If AI influences movement creation, does it share choreographic credit? While most choreographers still see AI as a tool rather than a co-creator, live improvisation blurs these boundaries (Ciolfi Felice and McDonnell, 2016; Pataranutaporn et al., 2024).
- **Ethical and Philosophical Considerations** – Some argue AI lacks the embodied intuition necessary for true improvisation (Jochum and Derks, 2019). Others see improvisation as a learnable AI behavior. Additionally, concerns arise about AI replacing human dancers or diminishing emotional investment in movement creation. The challenge remains ensuring AI empowers rather than diminishes human creativity (Zhou et al., 2021).

Future Directions in AI-Augmented Dance

AI dance partners are advancing from reactive responses to predictive engagement, anticipating movement and co-creating in real time (Alexanderson et al., 2023). AI tools increasingly assist choreographers by generating movement from prompts like "fluid" or "sharp" (Liu and Sra,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

2024b), with future systems integrating natural language and gesture-based controls (Rezwana and Maher, 2023b) to enhance workflows while maintaining artistic oversight. Rather than imitating dancers, AI will complement human movement with unique stylistic tendencies, improving ensemble interactions (Rezwana and Maher, 2023b). Visual cues, such as color shifts signaling movement changes, could support intuitive collaboration. AI also holds promise in education and therapy, refining movement styles, providing structured improvisational guidance, and aiding rehabilitation (Long and Magerko, 2017). AI's interdisciplinary applications continue to grow, integrating dance, music, and visuals into interactive performances (La Delfa et al., 2020). Additionally, AI could preserve dance traditions, though algorithmic bias must be addressed to ensure inclusivity (Crnkovic-Friis and Crnkovic-Friis, 2016).

Limitations

This qualitative study offers in-depth insights from a specific undergraduate dance context but does not aim for broad generalizability. While perspectives from professional dancers or those trained in diverse cultural dance traditions (e.g., non-Western, Indigenous, or culturally specific movement systems) could provide valuable additional viewpoints — particularly regarding different movement vocabularies, improvisational norms, and interpretations of AI-mediated interaction — exploring them was beyond this study's scope. The reliance on self-reported reflections may introduce bias, which future work could address through observational or video data. Additionally, the study's short duration and LuminAI's technical constraints limited insight into long-term integration, highlighting areas for future research and system improvement.

Conclusion

This study offers, to the best of our knowledge, the first longitudinal analysis of how an embodied AI system, LuminAI, integrates into a formal improvisational dance class, revealing how it reshapes creative practices, pedagogical dynamics, and movement decision-making. Using Cultural-Historical Activity Theory (CHAT), we examined the evolving interactions between students, instructor, and AI, finding that LuminAI acted both as a catalyst for creative exploration and a constraint that required dancers to adapt their strategies and expectations. While some participants remained skeptical of AI's artistic value, others reported increased self-awareness, expanded improvisational vocabulary, and novel collaborative experiences. Importantly, the instructor's role shifted from facilitating human-only improvisation to mediating human-AI interactions, underscoring the need for intentional pedagogical approaches when integrating AI into creative education. By identifying the key contradictions and adaptations that emerged, this research advances understanding of co-creative AI's role in embodied learning and argues for the development of future AI systems that prioritize adaptability, responsiveness, and transparency to meaningfully enhance, rather than limit, human improvisational practice.

Author Contributions

Author 1 was in charge of writing the manuscript, conducting the analysis, and planning the study. Author 2 helped in data analysis. Author 3 helped in planning the study and conducting the study. Author 4 helped in the editing of the manuscript.

Acknowledgments

We would like to thank Cassandra Naomi and Ramya Iyer for their development work on LuminAI as well as Kennesaw State University's Department of Dance. This project is funded by the National Science Foundation (CMII 2123597).

References

Akgun, S., and Greenhow, C. 2022. Artificial intelligence in education: Addressing ethical challenges in k-12 settings. *AI and Ethics* 2:411–421.

Alexanderson, S.; Nagy, R.; Beskow, J.; and Henter, G. E. 2023. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics* 42(4):44:1–44:20.

Batiibwe, M. S. K. 2019. Using cultural historical activity theory to understand how emerging technologies can mediate teaching and learning in a mathematics classroom: a review of literature. *Research and practice in technology enhanced learning* 14(1):12.

Berman, A., and James, V. 2015. Kinetic dialogues: Enhancing creativity in dance. In *Proceedings of the 2nd International Workshop on Movement and Computing (MOCO)*. Association for Computing Machinery (ACM).

Bolger, N.; Davis, A.; and Rafaeli, E. 2003. Diary methods: Capturing life as it is lived. *Annual Review of Psychology* 54(1):579–616.

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 (ICCC)*.

Braun, V., and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3(2):77–101.

Camurri, A.; Canepa, C.; Ferrari, N.; Mancini, M.; Niewiadomski, R.; Piana, S.; Volpe, G.; Matos, J.-M.; Palacio, P.; and Romero, M. 2016. A system to support the learning of movement qualities in dance: a case study on dynamic symmetry. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 973–976.

Chen, R., and Tong, X. 2023. Exploring designers' perceptions and practices of collaborating with generative ai as a co-creative agent. In *Proceedings of the Eleventh ACM Conference on Creativity & Cognition*. Association for Computing Machinery (ACM).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Ciolfi Felice, M., and McDonnell, J. 2016. Interactive ai dance systems: Tools or partners? In *CHI Conference on Human Factors in Computing Systems*.

Crnkovic-Friis, L., and Crnkovic-Friis, L. 2016. Deep learning for choreography: Training ai models on dance movement. *International Journal of Computational Creativity* 7:77–89.

Czerwinski, M.; Horvitz, E.; and Wilhite, S. 2004. A diary study of task switching and interruptions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 175–182.

Davis, N.; Deshpande, M.; Rezwana, J.; and Magerko, B. 2024. The five pillars of enaction as a theoretical framework for co-creative artificial intelligence. In *Proceedings of the 2024 International Conference on Computational Creativity (ICCC)*.

Deshpande, M., and Magerko, B. 2024a. Embracing embodied social cognition in ai: Moving away from computational theory of mind. In *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery (ACM).

Deshpande, M., and Magerko, B. 2024b. Reframing computational co-creativity: An embodied socio-cognitive lens. In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

Deshpande, M.; Park, J.; Pait, S.; and Magerko, B. 2024. Perceptions of interaction dynamics in co-creative ai: A comparative study of interaction modalities in drawcto. In *Proceedings of the 16th ACM Conference on Creativity and Cognition*. Association for Computing Machinery (ACM).

Dimitriadou, E., and Lanitis, A. 2023. A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments* 10(1):1–18.

Filippo, A. D.; Giuliani, L.; Mancini, E.; and Borghesi, A. 2023. Towards symbiotic creativity: A methodological approach to compare human and ai robotic dance creations. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Ford, C.; Noel-Hirst, A.; Cardinale, S.; and Loth, J. 2024. Reflection across ai-based music composition. In *Proceedings of the 16th ACM Conference on Creativity and Cognition*. Association for Computing Machinery (ACM).

Fu, Y.; Newman, M.; Going, L.; Feng, Q.; and Lee, J. H. 2025. Exploring the collaborative co-creation process with ai: A case study in novice music production. *arXiv preprint arXiv:2501.15276*.

Gemeinboeck, P., and Saunders, R. 2017. Beyond humanoid: Abstract robotics and ai in dance improvisation. *Journal of Performance and AI* 12(3):67–82.

Grassini, S. Shaping the future of education: Exploring the potential and consequences of ai and chatgpt in educational settings. *Education Sciences* 13(7).

He, J., and Do, H. J. 2025. Exploring industry practices and perspectives on ai attribution in co-creative use cases. In *Proceedings of the ACM International Conference on Intelligent User Interfaces*.

Hess, J., and Wulf, V. 2009. Explore social behavior around rich-media: A structured diary study. *Proceedings of the 7th European Conference on Interactive TV and Video* 215–218.

Holstein, K.; McLaren, B. M.; and Aleven, V. 2019. Co-designing a real-time classroom orchestration tool to support teacher-ai complementarity. In *Grantee Submission, ERIC*.

III, J. M. S. 2025. *Exploring the Potential of Co-creative AI in Tabletop Role-Playing Games*. Ph.D. Dissertation, University of Salzburg.

Jacob, M., and Magerko, B. 2015. Creative improvisation and ai: Exploring human-ai interaction in dance. In *International Conference on Computational Creativity (ICCC)*.

Jacob, M.; Chawla, P.; Douglas, L.; He, Z.; and Lee, J. 2019. Affordance-based generation of pretend object interaction variants for human-computer improvisational theater. In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

Jochum, E., and Derks, J. 2019. Tonight we improvise! real-time tracking for human-robot improvisational dance. In *Proceedings of the 6th International Conference on Movement and Computing (MOCO '19)*. ACM.

Jonsson, M., and Tholander, J. 2022. Cracking the code: Co-coding with ai in creative programming education. In *Proceedings of the 14th Conference on Computational Creativity*.

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 (ICCC)*, 77–84.

Kim, J., and Maher, M. 2021. Evaluating the effect of co-creative systems on design ideation. In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

La Delfa, J.; Baytas, M. A.; Patibanda, R.; Ngari, H.; Khot, R. A.; and Mueller, F. 2020. Drone chi: Somaesthetic human-drone interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Liu, Y., and Sra, M. 2024a. Dancegen: Supporting choreography ideation and prototyping with generative ai. In *Proceedings of the ACM Designing Interactive Systems (DIS)*.

Liu, Y., and Sra, M. 2024b. Exploring ai-assisted ideation and prototyping for choreography. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24 Companion)*, 11–17. ACM.

Liu, Y. 2024. Interaction design for human-ai choreography co-creation. *ArXiv*.

Long, T., and Magerko, B. 2017. Luminai: An interactive ai improvisational dance partner. *ACM Transactions on Intelligent Systems and Technology* 8(2):1–25.

Long, D.; Liu, L.; Gujrania, S.; Naomi, C.; and Magerko, B. 2020. Visualizing improvisation in luminai, an ai partner for co-creative dance. In *Proceedings of the 7th ACM International Conference on Movement and Computing (MOCO)*. Association for Computing Machinery (ACM).

Lucas, P., and Martinho, C. 2017. Stay awhile and listen to 3buddy, a co-creative level design support tool. In *Proceedings of the International Conference on Computational Creativity (ICCC)*.

Maher, M.; Weisz, J.; Chilton, L.; and Geyer, W. 2023. Haigen 2023: 4th workshop on human-ai co-creation with generative models. In *Proceedings of the 28th ACM International Conference on Intelligent User Interfaces (IUI)*. Association for Computing Machinery (ACM).

Marrone, R.; Taddeo, V.; and Hill, G. 2022. Creativity and artificial intelligence — a student perspective. *Journal of Intelligence* 10(3):65.

Park, H.; Eirich, J.; Luckow, A.; and Sedlmair, M. 2024. We are visual thinkers, not verbal thinkers: A thematic analysis of how professional designers use generative ai. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction*.

Pataranutaporn, P.; Mano, P.; Bhongse-Tong, P.; Chongchadklang, T.; Archiwaranguprok, C.; Hantrakul, L.; Eaimsa-ard, J.; Maes, P.; and Klunchun, P. 2024. Human-AI co-dancing: Evolving cultural heritage through collaborative choreography with generative virtual characters. In *Proceedings of the 10th International Conference on Movement and Computing (MOCO '24)*. ACM.

Rane, N.; Choudhary, S.; and Rane, J. 2023. Education 4.0 and 5.0: Integrating artificial intelligence (ai) for personalized and adaptive learning. *Research Project Journals*.

Rezwana, J., and Maher, M. L. 2023a. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction* 30(2):1–37.

Rezwana, J., and Maher, M. L. 2023b. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction*.

Rezwana, J., and Maher, M. L. 2023c. User perspectives on ethical challenges in human-ai co-creativity: A design fiction study. In *Proceedings of the 15th Conference on Computational Creativity*.

Southworth, J.; Migliaccio, K.; Glover, J.; and Glover, J. 2023. Developing a model for ai across the curriculum: Transforming the higher education landscape via innovation in ai literacy. *Artificial Intelligence and Education*.

Suh, M.; Youngblom, E.; Terry, M.; and Cai, C. 2021. Ai as social glue: Uncovering the roles of deep generative ai during social music composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery (ACM).

Trajkova, M.; Deshpande, M.; Knowlton, A.; Monden, C.; Long, D.; and Magerko, B. 2023. Ai meets holographic pepper's ghost: A co-creative public dance experience. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference*, 274–278.

Trajkova, M.; Long, D.; Deshpande, M.; Knowlton, A.; and Magerko, B. 2024. Exploring collaborative movement improvisation towards the design of luminai—a co-creative ai dance partner. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–22.

Wallace, B.; Hilton, C.; Nymoen, K.; Torresen, J.; Martin, C. P.; and Fiebrink, R. 2023. Embodying an interactive ai for dance through movement ideation. In *Proceedings of the 15th ACM Conference on Creativity and Cognition*, 454–464. Association for Computing Machinery (ACM).

Wang, S.; Sun, Z.; and Chen, Y. 2023. Effects of higher education institutes' artificial intelligence capability on students' self-efficacy, creativity and learning performance. *Education and Information Technologies*.

Winston, L., and Magerko, B. 2017. Turn-taking with improvisational co-creative agents. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE '17)*. Article 19, 8 pages.

Zhang, C.; Cocking, C. A.; Trajkova, M.; Mock, Z. L.; Tate, G.; Monden, C. N.; and Magerko, B. 2024. Fostering ai literacy with luminai through embodiment and creativity in informal learning spaces. In *Proceedings of the 16th Conference on Creativity & Cognition*, 476–481.

Zhou, Q.; Chua, C. C.; Knibbe, J.; Goncalves, J.; and Velloso, E. 2021. Dance and choreography in hci: A two-decade retrospective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. ACM.

Zhou, J.; Li, R.; Tang, J.; Tang, T.; Li, H.; and Cui, W. 2024. Understanding nonlinear collaboration between human and ai agents: A co-design framework for creative design. In *Proceedings of the 2024 ACM Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery (ACM).

Zhu, X., and Tan, Y. 2024. Preserving dance heritage through ai: Algorithmic choreography in traditional forms. *AI and Culture* 15(2):122–138.

Measuring Creativity in Co-Writing with AI: Rhyme Density and the Limits of Computational Proxies

Ibùkún Ọlátúnjí[1*] **Mark Sheppard**[2*]

[1]Computational Foundry, Swansea University, Crymlyn Burrows, Skewen, Swansea SA10 6JW, UK
[2]University of Kent, Canterbury, Kent CT2 7NZ, UK

Abstract

Evaluating creativity in AI-assisted writing remains a persistent challenge, often relying on subjective user ratings with limited reproducibility. This paper investigates rhyme density as a computational proxy for creative quality in human–AI collaborative poetry tasks. Drawing on data from a mixed-methods study (N=30), we analyse how rhyme-centric metrics relate to participant evaluations of fluency, creativity, and accuracy. We contribute (1) an empirical analysis of rhyme-centric features in AI-assisted writing, (2) a critical examination of their limitations in capturing human judgements, and (3) a framework for future evaluation methods that integrate computational rigour with human-centred insight. This work advances debates in computational creativity by offering practical strategies for assessing co-creative output.

Introduction

This study investigates the role of computational writing tools in poetry and lyric composition, examining how different tool types influence both the creative process and the resulting stylistic outcomes. Building on research into co-creative systems (Karimi et al. 2018) and generative writing tools (Arnold, Volzer, and Madrid 2021), we design a set of structured tasks in which participants use support, co-creative, and autonomous tools. We analyse the resulting texts using quantitative metrics (e.g., lexical distance, rhyme density) alongside participant evaluations, providing a multi-dimensional account of how these tools shape creative expression. We introduce a novel task design that integrates divergent idea generation into subsequent writing tasks, enabling joint evaluation of both creative phases. This allows us to study not only final outputs, but how creative intent is transformed into poetic form, an approach that bridges process and product-based perspectives on creativity. We focus on rhymed poetic forms as a structured yet expressive domain for computational analysis.

Contribution This work makes three key contributions: (1) empirical insights into how tool types affect user experience and stylistic expression; (2) a novel task design linking divergent and convergent phases of creativity; and (3) a framework for evaluating co-creativity through both computational and human-centred measures.

*Corresponding authors: 2030349@swansea.ac.uk, ms2403@kent.ac.uk

Related Work

Karimi et al. propose a framework for characterising human roles in creative systems (Karimi et al. 2018), distinguishing between : (1) *creativity support tools*, which assist but do not generate content (e.g., a rhyming dictionary (WriteExpress 2022)); (2) *co-creative systems*, which collaborate with users during generation (e.g., a poetry writing companion (Google Research 2022)); and (3) *fully autonomous systems*, which produce complete texts from prompts without further human input (e.g., ChatGPT (OpenAI 2024)).

System Type	Description	Examples
Creativity Support	Assists with specific writing tasks through language processing tools.	Rhyming dictionaries, Thesauri
Co-Creative Systems	Collaborates with the writer by generating text suggestions, acting as a co-writer.	Verse by Verse, Deepbeat
Autonomous Systems	Generates complete texts independently based on user prompts.	GPT-4, Bard, LLaMA-2

Table 1: Creative Systems Framework

Writing Assistants: Language model–based systems have become increasingly common in both research and commercial settings (OpenAI 2024; Google Research 2022). However, users often disengage due to shallow interaction dynamics and unclear collaboration roles (Rezwana, Maher, and Davis 2021). As Arnold et al. observe, a common pattern is for models to generate text that users then adopt or adapt. While this workflow aligns with system capabilities, it "does not necessarily support all writing approaches that writers may find useful."(Arnold, Volzer, and Madrid 2021). Such systems also risk blurring authorship,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

as they "feed words to writers suggesting that the writer claim the system's words as their own, especially with regards to possible plagiarism claims." (Arnold, Volzer, and Madrid 2021)

Poetry: We focus on poetry because it draws on phonemic awareness, vocabulary, and world knowledge (Rojcewicz 2004; Hadaway, Vardell, and Young 2001; Wassiliwizky et al. 2017). Its formal constraints, such as metre and rhyme, enable measurable comparisons between human and AI-assisted outputs. Studying poetic composition offers insight into how computational tools shape stylistic choices and structural form. To frame this investigation, we adopt Karimi et al.'s taxonomy of support, co-creative, and autonomous systems (Karimi et al. 2018) (see Table 1).

Study Design

The study includes three components: (1) a pre-task questionnaire to capture participants' writing experience and expectations; (2) writing tasks with three tool types—support, co-creative, and autonomous; and (3) a post-task questionnaire evaluating fluency, creativity, accuracy, and tool usefulness. Each participant completed a base poem without computational assistance, followed by three tool-assisted tasks, enabling within-subject comparison across tool types.

Study Design

Pre-study Questionnaire. The questionnaire collected demographic information and writing experience to serve as a baseline for later comparisons. The study involved 30 participants, each completing all writing tasks in a within-subjects design.

Divergent Association Task (DAT). The DAT (Olson et al. 2021) was used as a pre-test to measure verbal creativity. Participants were instructed to generate 10 words as different from each other as possible. The task, which takes under four minutes, provides an efficient and validated measure of divergent thinking. Semantic *distance* between the words was quantified, with higher DAT scores associated with greater creativity. These scores were later explored in relation to participant performance in the writing tasks.

Writing Tasks. Participants completed four writing tasks to assess how different tools influenced creative output: (i) *Baseline*, no tools used; (ii) *Support Tool*, e.g., a rhyming dictionary; (iii) *Co-Creative Tool*, offered suggestions in collaboration with the user; and (iv) *Autonomous Tool*, generated complete texts independently.

Task Instructions. To ensure consistency, all writing tasks were based on the theme of *place*:

- "What does the place you live mean to you?"
- "What is your favourite location in the place you live?"
- "How does the place you live make you feel?"

Participants were also asked to (i) incorporate as many DAT words as possible and (ii) use rhyme in their compositions. This combination encouraged both divergent thinking (idea generation) and convergent thinking (structuring ideas within poetic constraints), allowing us to observe how participants balanced creative exploration with formal structure.

Study Component	Description
Study Design	Mixed-methods design including pre-study questionnaire, Divergent Association Task (DAT), four writing tasks, and post-study questionnaires.
Participants	**N = 30** participants recruited through direct requests to acquaintances, including journalists and writers.
Pre-study Questionnaire	Age and previous writing experience (low, medium and high)
Divergent Association Task (DAT)	Verbal creativity measured by generating 10 unrelated words; used as baseline creativity indicator.
Writing Tasks	• *Baseline:* Without computational tools. • *Support Tools:* Tools assisting specific aspects (e.g., rhyme). • *Co-Creative Tools:* Tools providing suggestions or partial content. • *Autonomous Tools:* Fully automated text generation.
Post-study Questionnaire	Participant subjective ratings (1–5 Likert scale) of fluency, creativity, and accuracy of their writing outputs produced using the tools, and ratings of tool usefulness.
Scoring	DAT computed via word distance metrics; Rhyme Density scores by phoneme Cosine similarity.

Table 2: Summary of Study II design and methodological components.

Ethics All participants gave informed consent, and the study ensures anonymity and data protection in line with current ethical guidelines.

Rhyme Density

To enable consistent comparison across diverse outputs, we compute rhyme density scores for both participant-generated and tool-assisted compositions. Rhyme density serves as a shared, quantifiable anchor across conditions, helping to normalise differences in length, structure, and output style. Beyond this, we treat rhyme density as part of a broader framework for examining creativity as a combination of divergent and convergent thinking. The Divergent Association Task (DAT) provides a pre-task measure of ideational fluency, while the use of DAT-generated keywords reflects how participants transform divergent ideas into creative content. Rhyme density, as a formal constraint, cap-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

tures convergent effort during composition. Together, these measures allow us to examine how divergent thinking and convergent structuring manifest in creative writing, and how these dimensions vary across tool conditions.

Following prior work (Hirjee and Brown 2009; Condit-Schultz 2016; Malmi et al. 2016), we define rhyme as the repetition of similar phonemes within or across word sequences. Phoneme similarity is computed using cosine similarity between phoneme embeddings, allowing us to capture both exact and approximate rhymes.

$$\text{sim}(p_1, p_2) = \frac{\vec{v}p_1 \cdot \vec{v}p_2}{|\vec{v}p_1||\vec{v}p_2|}$$

where:

$\vec{v}p_1$ and $\vec{v}p_2$ represent the embedding vectors of phonemes p_1 and p_2, respectively.

Rhyme Density Calculation Rhyme Density (D) is calculated as the ratio between the total number of phoneme pairs , exceeding a predefined similarity threshold (θ), and the total number of phonemes in the analyzed text:

$$D = \frac{\sum_{i,j} M_{ij}}{|P_T|} \; M_{ij} \quad = \begin{cases} 1 & \text{if sim}(p_i, p_j) \geq \theta \; 0 \\ \text{otherwise} \end{cases}$$

where:

$|P_T|$ denotes the total count of phonemes within the text that is analyzed.

Phoneme Grouping To further analyze rhyme patterns, phonemes are organized into groups based on similarity scores. The phoneme grouping function $G(p)$ is defined as:

$$G(p) = \begin{cases} k & \text{if } \exists p' \in P : \text{sim}(p, p') \geq \theta \; |G| + 1 \\ \text{otherwise} \end{cases}$$

where:

k is the identifier of an existing phoneme group.

$|G|+1$ indicates the creation of a new phoneme group for phonemes without sufficiently similar matches.

The grouping methodology enables detailed insights into rhyme structure and phonemic repetition, which allows comparisons between human and computational creative outputs.

Results

We used a variety of statistical models to explore relationships between participants' divergent thinking (DAT scores), unaided ability (baseline rhyme density), and their performance across writing tasks involving varying levels of computational support. Table 3 provides an overview of the key models used. Our main findings include:

- A positive correlation between divergent thinking (DAT scores) and rhyme density across all writing tasks.
- Experience significantly predicted improved rhyme performance in tool-assisted tasks (support, co-creative, and autonomous).
- No significant relationship between experience level and baseline (unaided) rhyme performance.

Analysis Type	**Variables/Factors**	**Significance Level**
Pearson Correlation	DAT scores, Rhyme Scores, Experience Levels	$\alpha = 0.05$
ANOVA (One-way)	Experience Levels (groups), Tool Usefulness Ratings	$\alpha = 0.05$
OLS Regression	DAT scores, Base Rhyme Scores, Tool Usefulness Ratings	$\alpha = 0.05$
Subgroup Comparison	Experience, DAT scores, Rhyme Score Groups, Questionnaire Ratings	$\alpha = 0.05$

Table 3: Select Statistical Analyses and Variables.

Analysis

We present a more detailed interpretation of the results below. Selected outputs are summarised in Table 4, with additional qualitative insights included throughout.

Pearson Correlations We calculated Pearson correlation coefficients to measure linear relationships between DAT scores, rhyme density, and experience level across task types. Results revealed a moderate, positive, and statistically significant correlation between DAT scores and rhyme density in all task types (baseline, support, co-creative, autonomous). In contrast, experience level was not significantly correlated with baseline rhyme scores.

ANOVA To examine differences in rhyme performance across experience groups, we conducted one-way ANOVA tests. Results showed no significant differences in baseline rhyme scores based on experience. However, experience level significantly affected rhyme scores for all tool-assisted tasks. This suggests that more experienced participants benefit more from the use of writing tools.

OLS Regression We ran four Ordinary Least Squares (OLS) regression models using rhyme density scores from each task condition as dependent variables. Predictors included DAT score and experience level. Results indicated: (i) Baseline rhyme performance was modestly predicted by DAT score; (ii) experience level significantly predicted rhyme performance in support and co-creative conditions; and (iii) in the autonomous condition, both DAT score and experience were moderate predictors of output quality.

Predicting Perceived Tool Usefulness We conducted OLS regressions to examine whether DAT scores and baseline rhyme performance predicted perceived tool usefulness. Participants with higher baseline rhyme scores tended to rate support and co-creative tools as more useful. However, neither DAT scores nor experience level significantly predicted subjective ratings of tool usefulness.

ISBN: 978-989-54160-7-3

Analysis	Key Results	Significance
Pearson Correlation	DAT vs. Base Rhyme Score	r = 0.41, p = 0.023
	DAT vs. Support Score	r = 0.39, p = 0.030
	DAT vs. Co-creative Score	r = 0.37, p = 0.039
	DAT vs. Autonomous Score	r = 0.36, p = 0.044
OLS Regression	Support Score predicted by Experience	p = 0.002
	Co-creative Score predicted by Experience	p = 0.020
	Autonomous Score predicted by Experience	p = 0.033
ANOVA	Support Score by Experience Level	F = 6.34, p = 0.005
	Co-creative Score by Experience Level	F = 4.67, p = 0.018
	Autonomous Score by Experience Level	F = 3.99, p = 0.030

Table 4: Summary of Significant Statistical Results.

Discussion

This study examined how rhyme density, experience level, and divergent thinking (DAT scores) relate to performance in computationally-assisted creative writing tasks. Experienced writers benefited most from tool use, aligning with prior research in creativity support systems (Cherry and Latulipe 2014). One explanation may lie in *Cognitive Load Theory*, which suggests that cognitive demands interact with user expertise. Less experienced participants may have been overwhelmed by the dual challenge of constrained poetic tasks and unfamiliar tools, while experienced writers could apply tools more strategically (Sweller 1988b; Van Merrienboer and Sweller 2005; Sweller 1988a)

The weak correlation between experience level and baseline rhyme scores suggests that self-reported writing experience—often in non-fiction—did not reliably translate into poetic skill. Some less experienced participants, likely more familiar with poetic form, outperformed more experienced peers.

Tool performance varied with experience. Support tools most improved rhyme density for experienced writers, while no tool consistently aided novice users—likely due to a skills gap limiting their ability to act on system suggestions. Although DAT scores and experience were positively associated with rhyme performance, neither significantly predicted perceived tool usefulness. Qualitative feedback, especially around expectations and tool interpretation, may help explain these perceptual gaps.

Participants found support and autonomous tools easier to use and more helpful for fluency, while co-creative tools were rated as more cognitively demanding. Of the three, autonomous tools yielded the greatest fluency benefit. However, none led to statistically significant improvements in perceived creativity or accuracy.

DAT scores showed a positive trend with age, and female participants scored higher on average with greater internal consistency. While promising, these trends are difficult to interpret given the modest sample size. The generally high DAT scores may also have inflated observed correlations. Future work should examine whether computational tools support creativity or simply reflect the abilities of an already creative cohort, using a more diverse sample and broader range of divergent thinking levels.

Limitations Several limitations should be noted. First, tools were presented in a fixed order; although participants could choose how to engage with them, usage sequence was not strictly controlled and may have influenced outcomes. Second, apart from the DAT, writing tasks were untimed, introducing variability in pacing and effort. Third, the modest sample size (N = 30) limits the generalisability of statistical findings. Additionally, the study was conducted entirely online without lab-based control, which may have introduced variability in attention, environment, or device setup (Uittenhove, Jeanneret, and Vergauwe 2023; Dandurand, Shultz, and Onishi 2008; McConnell, Hintz, and Meyer 2025). Finally, participant-reported experience may not align with actual writing ability, particularly across genres. Future studies should use a more diverse sample, control task sequencing more tightly, and consider in-person or hybrid protocols to evaluate tool use under consistent conditions.

Future Work

Future work could investigate how participants' baseline vocabulary affects tool engagement, particularly across native and non-native speakers. Individuals with a broader lexical repertoire may be better equipped to respond creatively to AI-generated suggestions. While DAT scores offer a proxy for divergent thinking, vocabulary profiling tools such as LexTALE or the New General Service List (NGSL) could provide a more direct measure of lexical range (Lemhöfer and Broersma 2011; Brezina and Gablasova 2013). Comparing lexical use with perceived tool usefulness may help isolate how language proficiency shapes co-creative collaboration.

Co-creative tools also warrant refinement to support more intuitive, user-guided interaction. Prior work shows these systems often underperform when users cannot anticipate or steer system responses (Lehmann et al. 2022). More adaptive interfaces, clearer affordances, and responsive feedback mechanisms could lower the barrier for less experienced writers and improve perceived usefulness across user groups.

Finally, creativity evaluation itself should be broadened beyond rhyme and lexical features to include affective tone, rhythm, or originality. A richer, multidimensional framework would better reflect how divergent and convergent processes unfold across tool types, and how users express creativity in response to different levels of assistance.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council.

References

Arnold, K.; Volzer, A.; and Madrid, N. 2021. Generative models can help writers without writing for them. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

Brezina, V., and Gablasova, D. 2013. Is there a core general vocabulary? introducing the new general service list. *Applied Linguistics* 36(1):1–22.

Cherry, E., and Latulipe, C. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Trans. Comput.-Hum. Interact.* 21(4).

Condit-Schultz, N. 2016. *MCFlow: A Digital Corpus of Rap Flow*. Ph.D. Dissertation, The Ohio State University.

Dandurand, F.; Shultz, T.; and Onishi, K. 2008. Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods* 40:428–34.

Google Research. 2022. Verse by Verse.

Hadaway, N. L.; Vardell, S. M.; and Young, T. A. 2001. Scaffolding oral language development through poetry for students learning english. *The Reading Teacher* 54:796–796.

Hirjee, H., and Brown, D. 2009. Automatic detection of internal and imperfect rhymes in rap lyrics. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*, 711–716.

Karimi, P.; Grace, K.; Maher, M.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems.

Lehmann, F.; Markert, N.; Dang, H.; and Buschek, D. 2022. Suggestion lists vs. continuous generation: Interaction design for writing with generative models on mobile devices affect text length, wording and perceived authorship. In *Proceedings of Mensch Und Computer 2022*, MuC '22, 192–208. New York, NY, USA: Association for Computing Machinery.

Lemhöfer, K., and Broersma, M. 2011. Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior Research Methods* 44:325–343.

Malmi, E.; Takala, P.; Toivonen, H.; Raiko, T.; and Gionis, A. 2016. Dopelearning: A computational approach to rap lyrics generation. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

McConnell, K.; Hintz, F.; and Meyer, A. 2025. Individual differences in online research: Comparing lab-based and online administration of a psycholinguistic battery of linguistic and domain-general skills. *Behavior Research Methods* 57:22.

Olson, J. A.; Nahas, J.; Chmoulevitch, D.; Cropper, S. J.; and Webb, M. E. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences* 118(25):e2022340118.

OpenAI. 2024. ChatGPT Overview.

Rezwana, J.; Maher, M.; and Davis, N. 2021. Creative penpal: A virtual embodied conversational ai agent to improve user engagement and collaborative experience in human-ai co-creative design ideation. In *Joint Proceedings of the ACM IUI 2021 Workshops, College Station,*.

Rojcewicz, S. 2004. Poetry therapy in ancient greek literature. *Journal of Poetry Therapy* 17(4):209–213.

Sweller, J. 1988a. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2):257–285.

Sweller, J. 1988b. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2):257–285.

Uittenhove, K.; Jeanneret, S.; and Vergauwe, E. 2023. From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*.

Van Merrienboer, J. J. G., and Sweller, J. 2005. Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review* 17:147–177.

Wassiliwizky, E.; Koelsch, S.; Wagner, V.; Jacobsen, T.; and Menninghaus, W. 2017. The emotional power of poetry: neural circuitry, psychophysiology and compositional principles. *Social Cognitive and Affective Neuroscience* 12:1229 – 1240.

WriteExpress. 2022. Rhymer online rhyming dictionary.

ISBN: 978-989-54160-7-3

Beyond chat: collaborative editors enhance human involvement and agency when co-writing with large language models

Rodolfo Ocampo
The University of New South Wales
Creative Technologies
Research Lab (CTRL)
Sydney, NSW, Australia
r.ocampo_blanco@unsw.edu.au

Oliver Bown
The University of New South Wales
Creative Technologies
Research Lab (CTRL)
Sydney, NSW, Australia
o.bown@unsw.edu.au

Kazjon Grace
The University of Sydney
Design Lab
Sydney, NSW, Australia
kazjon.grace@sydney.edu.au

Abstract

As artificial intelligence (AI) tools become increasingly common for assisting with writing tasks, chat-based interfaces have emerged as the dominant mode of interaction. We hypothesise that relying solely on chat interfaces reduces users' involvement in the writing process, relegating them to an instructing role rather than an active co-author. To investigate this, we developed two prototype systems with hybrid interfaces integrating a *chat window* next to a *collaborative writing editor* where both human and AI can contribute, and we compared them to *chat-only* systems. Across two user studies, participants who had access to the collaborative hybrid interface reported higher levels of personal contribution, perceived a more balanced partnership with the AI, and adopted more engaged roles than participants using the chat-only interface. These findings suggest that offering a collaborative editor alongside chat can mitigate the risk of overreliance on AI and help preserve creative skills. As generative AI tools become more pervasive, designing interfaces that maintain user involvement is crucial to supporting meaningful human–AI co-creation.

Introduction

In recent years, advances in large language models (LLM) have made writing with them commonplace (Sun et al. 2024; Palani and Ramos 2024; Chakrabarty et al. 2024). Despite clear benefits such as overcoming writer's block and supporting non-native speakers, concerns are emerging about overreliance and loss of skill (Gerlich 2025; Lee 2025).

Users may allow the model to produce entire texts without engaging in thorough proofreading or revision, leading to generic-sounding outputs, factual errors, and a gradual erosion of writing and critical thinking skills (Gerlich 2025; Abbas, Jam, and Khan 2024; Essel et al. 2024; Heersmink 2024). Furthermore, excessive automation can diminish the intrinsic satisfaction emerging from creative work. As Brian Eno points out when discussing the limitations of generative AI; much of the value in creative activities lies in the process rather than in the final outcome (Eno 2024).

To mitigate these risks, it is crucial to design co-creative AI systems that actively encourage user involvement. This is largely an interaction and interface design problem. Currently, chat-based interfaces have emerged as the dominant paradigm for interacting with LLMs. This approach, however, may position users as mere requesters who delegate writing tasks to the AI rather than as writing co-creatos. In other words, chat interfaces provide primarily a space to interact *about* the writing, but not *through* the writing, both important in co-creation (Bown et al. 2020a; Rezwana and Maher 2022a; McGuire, De Cremer, and Van de Cruys 2024).

To address this issue, we hypothesise that implementing **hybrid interfaces with both a chat window to discuss the writing and a collaborative editor to contribute to it**, can increase user involvement.

We investigated this hypothesis through two studies, both employing hybrid interface prototypes of increasing fidelity. In the first study, 23 participants were divided into two groups: one used the hybrid collaborative interface and the other the chat-only interface. Compared to those using the chat-only interface, participants who used the collaborative editor were less likely to report that the AI did most of the work and more likely to report the final piece contained their own words. Moreover, they described their role as more involved in the writing, while participants in the chat-only interface were more likely to describe their role as providing ideas, instructions and requests to the AI.

Building on these findings, we conducted a second study with 25 participants using an improved hybrid prototype. This follow-up was not comparative: instead it focused on collecting qualitative in-depth feedback about user experience, though it included questions prompting users to compare it with their previous experience with popular tools like ChatGPT. Results indicated that the shared text editor led to reduced friction, an enjoyable experience, and fostered greater user involvement with the writing process. However, participants identified several key limitations: they struggled to clearly identify which edits were made by the AI, requested better version control and revision history features, and suggested adding GUI elements that could

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

trigger premade actions to streamline their workflow.

Related Work and Background

Large Language Models (LLMs) have evolved considerably in their interaction paradigms. Initially developed as autocomplete systems (Vaswani et al. 2017; Brown et al. 2020), they transitioned to instruction-based interaction with InstructGPT (Ouyang et al. 2022), before adopting the now ubiquitous chat interfaces. ChatGPT's rapid adoption, becoming the fastest application to date to reach 100 million users (Heaven 2023), stemmed not from novel technology but largely from the transition to a dialogic chat interface. As OpenAI's President Greg Brockman explained: "we actually had the model [GPT3.5] behind it created almost a year prior so it wasn't new technology, the thing that we really did differently is that we did a little bit of extra work to make it more aligned so you could talk to it and it would do what you wanted". This interface allowed them to bridge "what people thought was possible and what actually had been possible for quite some time" (SXSW 2023). Jan Leike, another ChatGPT contributor, when explaining reasons for its rapid adoption, claimed: "It's not a fundamentally more capable model than what we had previously. The same basic models had been available on the API for almost a year before ChatGPT came out [...] we made it more aligned with what humans want to do with it. It talks to you in dialogue, it's easily accessible in a chat interface" (Heaven 2023).The evolution toward chat-based interaction has been followed by numerous systems including Gemini, Llama, and Claude (Touvron et al. 2023; Gemini Team 2024).

While chat interfaces enhanced accessibility, they potentially limited creative collaboration by separating users from direct engagement with the creative artifact. Earlier autocomplete paradigms allowed interaction *through* the writing itself, whereas chat interfaces shifted to interaction *about* the writing. This distinction of types of interactions has been studied in the literature by some authors (Calderwood et al. 2020) and having distinct spaces for discussion and co-creation has been identified as important for successful co-creation (Bown et al. 2020b; Rezwana and Maher 2022a; Coenen et al. 2021). We are particularly interested in understanding how this hybrid interface affects human direct involvement and user experience, particularly in the context of highly capable language models.

Human-AI Co-Creativity and Mixed Initiative Systems

Mixed-initiative interaction enables both human and computers to actively contribute to tasks fostering user agency through iterative exchange (Allen, Guinn, and Horvtz 1999; Horvitz 1999). When applied to creative domains, this creates mixed-initiative co-creativity, where humans and AI contribute as equal collaborators rather than in client-producer relationships (Muller, Weisz, and Geyer 2020; Yannakakis, Liapis, and Alexopoulos 2014; Davis 2013).

Mixed-initiative co-creative interfaces differ from request-based systems by establishing a tightly interactive loop where both parties suggest, produce, evaluate, and modify outputs within a shared creative space. Examples include collaborative environments for drawing (Karimi et al. 2020) and writing (Kantosalo et al. 2016). Research shows that effective co-creativity requires both a shared creation space and communication channels (Rezwana and Maher 2022b), with communication establishing the common ground necessary for collaboration (Dafoe et al. 2021).

Bown et al.'s framework of "dialogic creative interaction" distinguishes between interacting *through* an artifact (directly modifying it) and *about* an artifact (exchanging ideas regarding creative intent) (Bown et al. 2020b). Similarly, Rezwana and Maher differentiate between interactions *between* collaborators and *with* the creative product (Rezwana and Maher 2022a). Conventional chat interfaces conflate these interaction modes, constraining collaborative possibilities and limiting the capacity for stateful co-creation.

Shneiderman's direct manipulation principles, emphasising controllability, predictability, and continuous representation, further highlight chat systems' limitations for co-creativity (Shneiderman 1997).

Beyond Chat Interfaces for Co-Creative Generative AI Systems

Recent systems have begun distinguishing between communication spaces and co-creation spaces. Anthropic's Claude introduced "Artifacts" (Anthropic 2024), though as of September 2024, only the AI can modify these elements. ChatGPT's Canvas feature and specialized writing tools like Sudowrite, Jarvis, and Wordcraft (Yuan et al. 2022) provide shared spaces where both human and AI can edit text while communicating through separate channels. Similar developments appear in coding environments, with systems like Cursor (Cursor 2024) and GitHub Copilot enabling both communication and collaborative editing, though with varying degrees of integration.

While multiple studies have examined human co-writing with LLMs from different perspectives (Wan et al. 2023; Yuan et al. 2022; Hwang et al. 2024; Chakrabarty et al. 2024; Grigis and De Angeli 2024; Masson et al. 2024; McGuire, De Cremer, and Van de Cruys 2024; Lee et al. 2024), our contribution is **specifically exploring the effect that collaborative spaces in addition to chat interfaces** may have on human-AI co-creativity, focusing on user agency, involvement, immersion, enjoyment, creative self-efficacy and perceived roles. To our knowledge, no such evaluations exist in the literature.

Moreover, the studies and prototypes presented here predate the recent introduction of collaborative spaces such as Artifacts in Claude and Canvas in ChatGPT. As such, the insights presented in this paper are timely to

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

understand the co-creative potential and limitations of these interfaces.

Prototypes

To investigate the efficacy of hybrid chat-and-editor interfaces, we developed two sequential prototypes that implemented this functionality. This iterative development process yielded first Vorges, followed by a refined version named Common AI. Both systems were deployed as web applications built using Node.js, with the OpenAI API providing the underlying large language model capabilities.

Version One – Vorges

Vorges, developed in 2023, represented our initial exploration into human-AI collaborative writing environments. The system utilised the GPT-3.5 model via the OpenAI API and implemented an agentic approach wherein the LLM was designed both to communicate with users and to directly manipulate a shared text editor in the front end.

We established a rudimentary communication protocol between the front-end application and the language model through an API, which enabled the AI to perform two distinct functions:

- Communicate with the user through the chat interface
- Modify content within the collaborative text editor

To facilitate text editing, the LLM was instructed to enclose editor text modifications within curly brackets, with content outside these delimiters appearing in the chat window. This parsing logic was managed by the application back-end.

A significant limitation of this implementation was its requirement for the AI to return the entire edited text within curly brackets, even for minor modifications. This wholesale replacement approach resulted in increased computational costs, saturation of context windows, and from a co-creative perspective, diminished ability to perform fine-grain edits efficiently.

Version Two – Common AI

The second iteration, Common AI (Figure 1), developed in late 2024, represented a substantial refinement of our approach. This prototype leveraged the more sophisticated GPT-4o model and implemented a structured JSON protocol for granular text manipulation. This advancement enabled the AI to execute precise edits without necessitating resubmission of the entire document.

The enhanced protocol expanded the AI's editing capabilities to include:

- Appending text at the beginning or end of the document
- Editing specific textual segments, from individual words to comprehensive revisions
- Deleting targeted text from the document
- Communicating with the user through the chat interface

This protocol enhancement improved system usability while optimising the management of context windows, ultimately enabling more fine-grained collaboration between the user and the LLM.

User Studies: Designing a Shared Creative Workspace and Examining its Effect on Human-AI Co-Creative Writing

We conducted two exploratory studies. The first one was conducted in June 2023 using the first version of the prototype (Vorges) while the second one, using an improved version (Common AI), was conducted in August 2024. These investigations sought to understand how implementing a shared collaborative space alongside a chat window might enhance user involvement and creative agency in human-AI collaboration.

Study 1

Context The initial exploratory study employed the Vorges prototype and was conducted in situ at Sydney's Powerhouse Museum during an AI-themed public event. Participants were recruited through a combination of opportunistic sampling as visitors explored the museum and direct attendance following promotional materials in the event program.

Experimental Design We established six computer stations across three tables, implementing a between-subjects comparative design. Three stations featured the Vorges interface with both collaborative editor and chat window, while three control stations offered a chat-only version of the same system. This configuration allowed direct comparison between the two interaction modalities within the same experimental context.

Participants The study included 23 participants (age range: 18–53 years) who were randomly assigned to available stations. This resulted in 12 participants using the chat-only interface and 11 using the chat-and-editor interface.

Task Participants were asked to write a micro-science fiction story set in a speculative post-2050 future. The story had to comprise at least three paragraphs, with no imposed time constraints.

Measures Our investigation focused on comparing experiences between the two interface conditions across five key dimensions: User involvement and agency; Sense of collaboration; Enjoyment and satisfaction; Creative self-efficacy; Emergent user-AI roles

After task completion, participants completed a mixed-methods survey containing open-ended qualitative questions and Likert-scale quantitative items.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

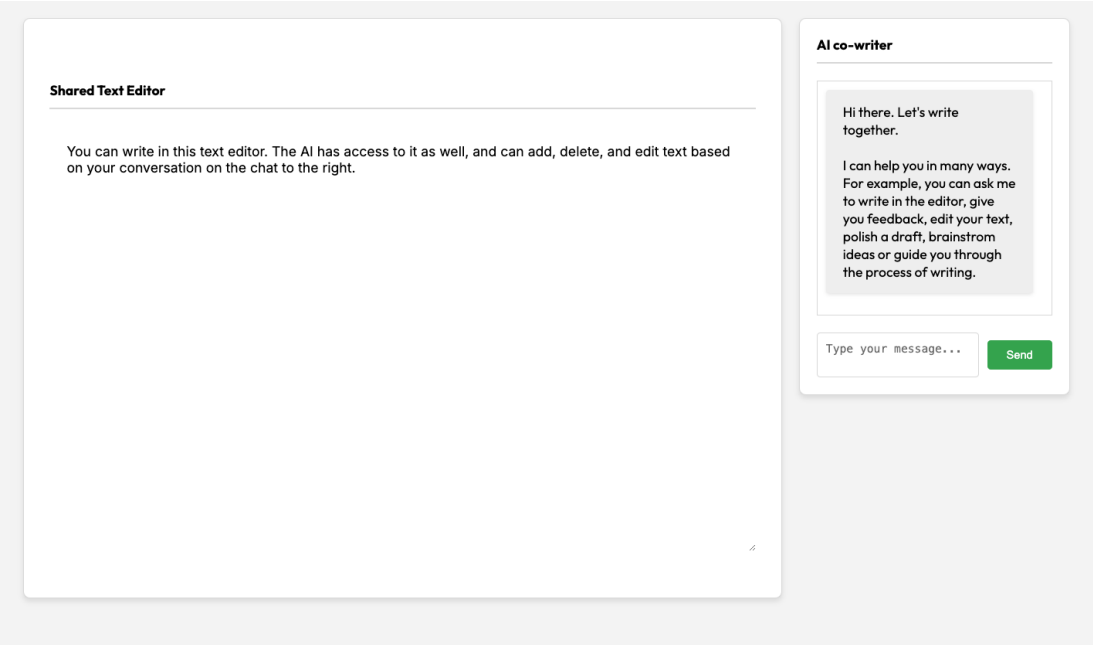


Figure 1: Screenshot of Common AI, our implementation of a hybrid interface, integrating a collaborative editor and a chat window. The previous version Vorges, had a similar interface, but a less robust back-end functionality, as well as a dark theme. Common AI's improved capabilities allowed the AI to perform fine-grained document modifications and utilised a more capable underlying language model. Common AI served as the experimental platform for the second study presented in this paper.

Methods and Analysis Our analysis of user responses was guided by Creswell's (Creswell 2012) qualitative research approach, particularly narrative research, used to understand and contextualise user's personal experiences and meaning-making. Moreover, qualitative responses underwent thematic analysis to identify emergent patterns (Braun and Clarke 2012). Likert-scale responses were converted to numeric values (1=Strongly Disagree to 5=Strongly Agree) for statistical comparison. We employed t-tests to assess statistical significance between mean responses for each question across the two conditions.

Study 2

Context The second study utilized the refined Common AI prototype and was conducted online, allowing participants to access the collaborative environment remotely through a web interface. This study aimed to gather deeper insights into user experiences with the shared text editor.

Participants We recruited participants (n=25) through social media channels and snowball sampling techniques. Participants received incentives in the form of $15 AUD vouchers (Westfield or Uber Eats). Eligibility required prior experience using chat-based LLM tools (e.g., ChatGPT, Claude) for writing tasks.

Study Design This investigation employed a single-condition exploratory design focused on the chat-and-editor interface. Given the sample size limitations, which precluded robust quantitative statistical analysis, we prioritized gathering rich qualitative feedback to illuminate the key challenges and possibilities within the collaborative interface.

Task Participants selected one of five creative writing prompts, which they completed and submitted with no time constraint.

Methods and Analysis After task submission, participants completed a mixed-methods survey emphasising open-ended responses. The instrument sought to understand participants' experiences with the collaborative editor, its perceived limitations, and comparative reflections against their previous experiences with other writing tools. Similar to study 1, we utilised a narrative research approach coupled with thematic analysis (Braun and Clarke 2012; Creswell 2012). Additional Likert-scale items measured perceived levels of user and AI involvement in the creative process.

ISBN: 978-989-54160-7-3

Study 1 Results

Differences in work distribution and involvement across chat-only and hybrid collaborative editor interfaces

Our initial exploration revealed differences across interface modalities, as shown in Figure 2. While the Likert-scale survey did not achieve statistical significance due to a low sample size, differences between interfaces and questions indicated that the most marked differences where in attribution of who did most of the work, and levels of immersion and direct involvement with the writing. Participants in the collaborative editor + chat interface were more likely to report the final story contained words they wrote and reported higher levels of immersion in the writing task. Conversely, participants in the chat-only interface predominantly attributed the writing work to the AI system (Vorges), more strongly agreeing that the "AI did most of the work".

Qualitative Results: Understanding Perceived and Assumed Roles in Each Group The qualitative data revealed nuanced differences in how participants conceptualised the role they played in the creative process compared to the role that AI played. While both groups generally reported relying on Vorges to contribute substantially to the writing, the nature and degree of this reliance varied between conditions.

Comparative Role Dynamics Between Interfaces In the chat-only interface, participants predominantly positioned themselves as idea providers while ceding the writing process to the AI. This dynamic is exemplified by participant 9, who simply stated: "Vorges generated all of the text," and participant 6, who when asked about Vorges' role, responded with a single word: "Everything." A more elaborated response from participant 11 illustrates this delegation of creative labour: "I mainly just gave the ideas and instructions, such as: write a paragraph describing the character as x, y, z. Vorges nicely narrated my ideas. It was like giving the recipe of a cake to someone and letting it do it."

In contrast, participants using the collaborative editor + chat interface were more likely to report starting with, or at some point doing some of the writing, and then using the AI to improve or expand it. Participant 3 explained: "I wrote a simple story, easy language, not very detailed, and then I asked Vorges to make my story more advanced, bigger words, more description, more intense." Similarly, participant 4 noted that the AI "took the story in fun and interesting directions, and added to the experience rather than entirely controlling the experience".

Contrasting Perceptions of AI Contribution The differential framing of the AI's contribution was another revealing pattern. Chat-only interface users predominantly characterised Vorges as helping to *write their ideas*, whereas collaborative editor + chat users described the AI as helping them *ideate* or *improve their own writing*. This subtle distinction reflects fundamentally different conceptualisations of creative ownership and process.

User Satisfaction and Sense of Accomplishment This divergence in experience extended to satisfaction with the process itself. One participant from the chat-only condition (participant 11) expressed a notable ambivalence: "it also made me feel like I was cheating somehow. It does not feel like my work, even though I gave all the ideas. Also, I believe there is satisfaction in putting a lot of effort/dedication/patience into something. Vorges made everything so simple, fast, and easy that it felt artificial and no real satisfaction came as a result." This sentiment contrasts sharply with reports from the collaborative editor + chat group, where participant 7 described the process as "fun and fulfilling," and participant 2 expressed enthusiasm about future engagement, stating they "Would love to do it again."

Challenges and Limitations of the Collaborative Interface However, users also reported challenges with the collaborative editor + chat interface. Participant 3, who used Vorges to enhance their writing with more sophisticated language, noted: "I still wrote my own story, but the style of the writing no longer feels like mine." This highlights tensions between enhancement and authenticity that merit further exploration. Additionally, participant 9 found the system "not intuitive".

These initial findings prompted us to develop an enhanced prototype for Study 2 addressing technical challenges of the previous prototype and seeking to achieve more in deeper insights about the unique challenges and potential of implementing a collaborative editor.

Results Study 2: Refined Interface and Deeper Qualitative Insights

Overview Study 2 employed an improved prototype leveraging a more sophisticated model capable of executing fine-grained text edits. Rather than replacing entire text blocks to accommodate even minor changes, the enhanced system (Common AI) could perform targeted edits: adding text at specific locations, deleting selected words or sections, and making punctual modifications.

This study was aimed at gathering richer qualitative insights through predominantly open-ended questions. Unlike the comparative approach of Study 1, this follow-up investigation focused on an in-depth qualitative exploration of user experiences with the improved collaborative editor + chat interface.

Additionally, we sought to further investigate a key finding from the previous study: the tendency of chat-only interface users to attribute most of the creative work to the AI. To this end, we incorporated Likert scale questions assessing agreement with the statements "The AI did most of the work" and "I did most of the work."

Importantly, this second study was conducted in a significantly different wider technological context than

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

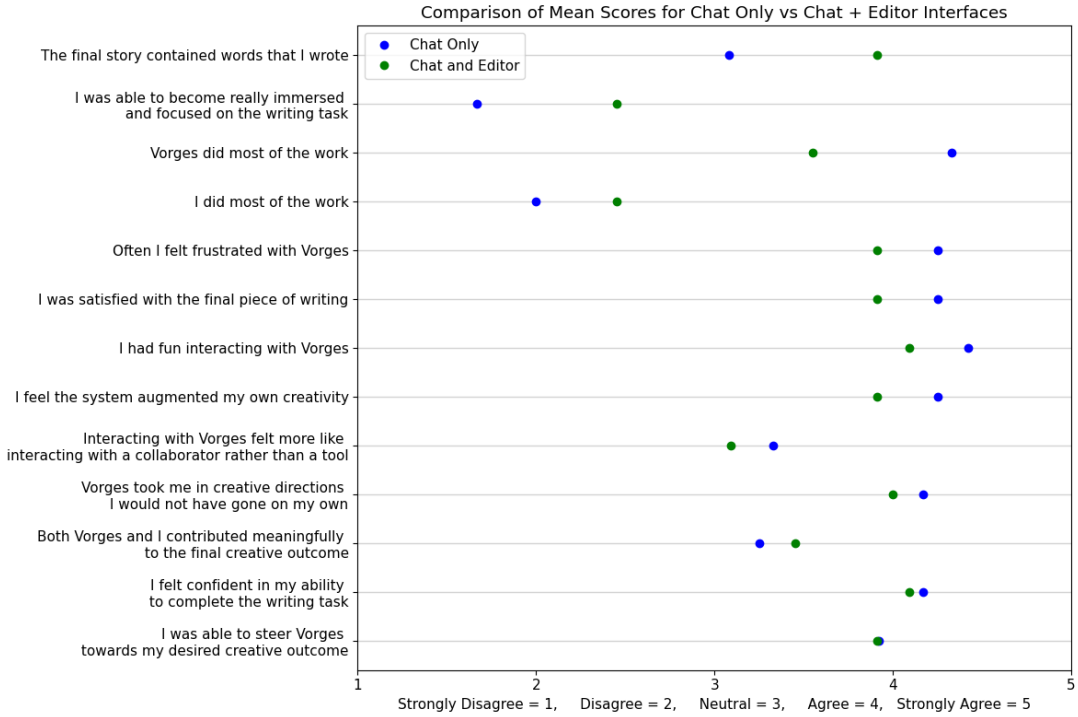


Figure 2: Comparison of mean scores for the Likert Scale questions. Out of the 13 questions, 9 yielded small differences while 4 yielded larger observable differences, showing, generally, that users of the collaborative editor were more involved and immersed in the writing, while users of the chat-only interface thought the AI did most of the work

Study 1, as chat interfaces and AI systems like ChatGPT, Claude, and Gemini had become ubiquitous in the interim. This shifting landscape allowed us to explore user frustrations with existing AI writing tools and position our findings within a more mature ecosystem of LLM-based systems.

High Enjoyment and Involvement, with an Expressed Need to Highlight AI-Generated Changes Participants generally reported positive experiences using the tool, with 20 out of 25 participants describing highly enjoyable interactions. For example, Participant 4 shared: "I really-really enjoyed writing this. I even had a deep moment of reflection, my writing was nostalgic and sad, but I was able to use AI to steer it in the right direction, it gave me confidence that I was also writing with correct grammar and spelling, English is not my first language and while I am proficient, I can still use proofreading to ensure good quality, this tool helped me with it." This participant primarily engaged with the editor component and sought AI assistance to enhance their writing. When asked about limitations, they noted: "I am unsure as what parts of the text are being edited, in the end I am not quite sure which parts were mine and which ones were edited by AI."

A similar concern was expressed by Participant 7, who found the system "very helpful, helps me bring my ideas to life in a more vivid and captivating way," but identified a limitation: "Once the tool has made revisions to the original text, maybe it can highlight the key changes that have been made, that way I can easily see where the tool has made updates, otherwise I need to slowly read through and identify the changes myself."

Participant 11 reinforced this theme while describing their editing-focused workflow: "When you ask the AI to check your grammar (as I did), it would be good if it told me what suggestions it had made, so I can double check its work easier."

These reflections consistently highlight users' desire to maintain clarity about the boundary between their contributions and AI-generated content—suggesting a prioritisation of collaborative transparency that preserves their creative agency rather than amalgamating human and AI work indistinguishably.

Emphasising the importance of maintained agency, Participant 9 noted: "I liked how my original ideas were still retained and AI was used as a tool to complement my intentions. It forced me to put in some effort and do the majority of the work, I enjoyed how I could not be fully reliant on AI to provide me with all the text."

Interestingly, all participants who expressed these positive experiences and desire for clear attribution also reported that they had done most of the work when responding to the Likert scale questions, disagreeing with the statement that "The AI did most of the work." and

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

agreeing with the statement "I did most of the work"

Users Who Outsourced Most of the Work Found Limitations with Writing Style A contrasting pattern emerged among users who reported that the AI had performed most of the work. For instance, Participant 16, who strongly agreed with the statement "The AI did most of the work" and strongly disagreed with "I did most of the work," identified the main limitation as "The writing was not very good" and claimed that having a collaborative text editor "made no difference." Similarly, Participant 21, who also strongly agreed that the AI did most of the work, stated that the collaborative editor "Doesn't make a difference" and expressed that they "don't like using AI to generate actual pieces of writing."

This finding suggests that for users who are less involved at the writing level found the quality of the AI's writing more of a problem than those users who engage, and thus shape the style of the output.

Users Felt the Collaborative Editor Contributed to Higher Creative Agency A recurring theme associated with positive experiences was an enhanced sense of agency and genuine collaboration. Participant 14 articulated this clearly: "Makes sense to me. So much of popular discourse around human and AI interaction centres around collaboration and partnership in creativity. My view is that AI needs human feel to serve the purposes we need it for, i.e. to be human. A shared text editor is a useful format to encourage collaboration and achieve human guidance."

This sentiment was echoed by Participant 12, who appreciated that the system "still let me have autonomy," while Participant 6 valued the "flexibility and control over the final result."

The ethical dimensions of human-AI creative collaboration were highlighted by Participant 22, who noted: "It adds an element of working together, which I think is the moral problem with current AI tools - they often seem like they're doing all the work."

Participant 25 expressed particular satisfaction with the AI's complementary rather than dominant role: "I absolutely love it! I enjoy creative writing, and AI expands my vocabulary and corrects my grammar. I love that it provides feedback before making changes."

Greater Autonomy and Collaboration than Popular Chat-Based Tools When comparing the collaborative editor to conventional chat-based AI tools, participants consistently noted advantages in agency and workflow integration. Participant 18 remarked: "It was much better than ChatGPT I enjoyed how it gave me a lot more agency." Participant 5 highlighted the practical workflow benefits: "it can just be a bit clunky having a separate document to then copy, paste, and edit in. This made it super seamless being in the one program."

Participant 10 expressed a clear preference for the dual-interface approach: "I think it's better than just having one textbox [the chat box]; I like it quite a lot, actually." This sentiment was reinforced by Participant 1, who described "The shared text editor is a great feature and a significant advantage" compared to chat-only tools, and by Participant 17, who independently stated: "I think this is a better solution than typical AI chatbots."

When asked about frustrations with conventional chat-based tools like ChatGPT and Claude, participants frequently mentioned challenges with maintaining control and iterative refinement. Participant 17 noted: "ChatGPT will always rewrite the entire passage to change just one paragraph, and it's harder to work on one text because I often need to scroll back up or continually copy and paste it." Similarly, Participant 15 expressed frustration that "it's gone in the wrong direction and requires a lot of input from me to get it back on track."

Participants Valued the Chat Window Despite Criticism of Its Size An interesting pattern emerged regarding the chat component of the interface. While only one participant requested a larger editor space, multiple users identified the chat window's small size as a limitation -suggesting a continued desire for communication with the AI alongside direct editing.

Participant 17 noted: "The AI chat window felt a little too small, so I had to scroll to see what I had asked it, even within the one prompt." This was echoed by Participant 14 who simply stated: "The chat box was a bit small," and Participant 5 who elaborated: "I found the box on the right rather small. It was also difficult to scroll up to see what you have written in the text box before submitting it."

This emphasis on the chat window's importance, despite its seemingly secondary role to the editor, suggests that a good chat experience and interface remains crucially important for users.

Interface Clarity and Action Affordances as Areas for Improvement A recurring theme in participant feedback concerned the need for clearer signalling of interface components and their functions. Participant 15 explained: "I didn't understand the format of the site at first as the instructions within the big collaborative text area on the left were confusing (and weren't in a different font or distinguishably different to the area where the collaborative story was updating)." They further noted mismatched expectations: "I thought that it would take what I was writing and make more of a change to it by combining with more flourished details about the places I was writing about (e.g. it would come up with details about the landscapes I was in) but it kind of just copied my words and made slight grammatical changes."

Participant 9 simply described the interface as "not that intuitive," while Participant 22 suggested visual design improvements: "make the fields different colours to highlight the difference in the two, have the AI prompt part set out as more of a chat, like messenger, so that the text editor is seen as the output field."

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Several participants recommended implementing visible action affordances rather than requiring written requests for common functions. Participant 22 suggested: "Some suggestive prompts would also be useful like ChatGPT, because it is hard to know what the AI is even capable of and that limits the creative output." Despite this critique, they maintained: "I quite like it, it's very different to other AI tools. Definitely takes some getting used to but I absolutely see future applications and usefulness."

In a similar vein, Participant 17 proposed: "Adding some ready made suggested prompts based on my writing, or just in general," while Participant 9 independently suggested "a drop down of up to 3 suggestions based off the prompts." Participant 4 envisioned "a menu to suggest common writing tasks or styles" and noted that "If used for educational settings, scaffolding could help create a more hands-on experience."

Highlighting a desire for more automated assistance, Participant 24 noted: "I had to actively input each command eg grammar and spellcheck. I feel this should be done automatically."

Finally, regarding version control and the visibility of AI contributions, Participant 2 suggested: "Maybe having a revert AI contribution button and/or contribution history so you could go back to past historical copies." While this functionality was actually available through an undo button, the comment suggests that its presence was not sufficiently evident. As Participant 19 noted: "I enjoyed the fact that you can go back to previous revisions of our work, like flipping a page back." However, as previously discussed by other participants, enhanced visibility of specific changes between versions remains an area for improvement.

Discussion and Conclusion

This research investigated how interface design affects human involvement in AI co-creative writing. We posited that conventional chat interfaces may limit human agency, potentially relegating users to mere instruction-givers rather than active co-creators. To address this limitation, we developed and evaluated a prototype combining collaborative editing with chat functionality, enabling users to interact both *through* and *about* the writing process.

Through two iterative studies, we explored this interface's impact on agency, involvement, satisfaction, and creative self-efficacy. Four key findings emerged:

Differential Attribution of Creative Labor Our first study revealed a clear distinction in how users perceived creative contribution across interfaces. In the chat-only condition, participants consistently reported that "the AI did all the work" and noted minimal inclusion of their own writing in the final product. Users primarily described their role as providing ideas while the AI performed the actual writing. Conversely, collaborative editor users reported greater personal contribution to the writing process.

Enhanced Creative Agency The second study, using an improved version of the collaborative editor, demonstrated that users valued maintaining active involvement in the writing process. Participants reported using the AI as a refinement tool to polish and expand their writing rather than surrendering creative control entirely. This approach preserved their sense of agency while still benefiting from AI assistance.

Advantages Over Chat-Only Interfaces Participants explicitly contrasted their experience with conventional chat interfaces, highlighting frustrations with traditional approaches. They noted difficulties in maintaining control over text when AI rewrites entire passages for minor changes and the inefficiency of copying text between systems. The integrated approach was described as "super seamless," offering "more agency" and "flexibility and control over the final result."

Interface Improvement Recommendations Users suggested several enhancements to the collaborative editor experience. These included: highlighting AI versus human contributions, implementing more robust version control, providing clearer indicators of AI capabilities, and offering GUI components for triggering predefined actions. These suggestions align with established HCI principles, including Norman's visibility of affordances (Norman 2013), Shneiderman's direct manipulation concept (Shneiderman 1997), and Amershi's guidelines for human-AI interaction (Amershi et al. 2019).

Our findings, while derived from relatively small participant samples (23 and 25 respectively), provide valuable qualitative insights into the potential of collaborative editing interfaces to enhance human-AI co-creativity. As major platforms increasingly implement variations of collaborative editors alongside chat functionality, understanding how these interfaces affect user involvement becomes crucial for maximising human creative agency in AI collaboration.

Future research should focus on implementing improved prototypes that provide greater clarity about co-creative capabilities, robust history and versioning features, and clear visibility of user versus AI contributions in the collaborative writing process (Haase, Djurica, and Mendling 2023).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

Abbas, M.; Jam, F. A.; and Khan, T. I. 2024. Is it harmful or helpful? examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education* 21(1):1–22.

Allen, J. E.; Guinn, C. I.; and Horvtz, E. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14(5):14–23.

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; Teevan, J.; Kikin-Gil, R.; and Horvitz, E. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM.

Anthropic. 2024. What are artifacts and how do I use them? https://support.anthropic.com/en/articles/9487310-what-are-artifacts-and-how-do-i-use-them. Accessed: 2024-9-11.

Bown, O.; Grace, K.; Bray, L.; and Ventura, D. 2020a. A speculative exploration of the role of dialogue in human-ComputerCo-creation. *ICCC*.

Bown, O.; Grace, K.; Bray, L.; and Ventura, D. 2020b. A speculative exploration of the role of dialogue in human-ComputerCo-creation. In *ICCC*, 25–32. computationalcreativity.net.

Braun, V., and Clarke, V. 2012. Thematic analysis.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; 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. M.; 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. *arXiv:2005. 14165 [cs]*.

Calderwood, A.; Qiu, V.; Gero, K.; and Chilton, L. B. 2020. How novelists use generative language models: An exploratory user study.

Chakrabarty, T.; Padmakumar, V.; Brahman, F.; and Muresan, S. 2024. Creativity support in the age of large language models: An empirical study involving professional writers. In *Creativity and Cognition*. New York, NY, USA: ACM.

Coenen, A.; Davis, L.; Ippolito, D.; Reif, E.; and Yuan, A. 2021. Wordcraft: a human-AI collaborative editor for story writing. *ArXiv*.

Creswell, J. W. 2012. *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: SAGE Publications, 3 edition.

Cursor. 2024. Cursor. https://www.cursor.com/. Accessed: 2024-9-11.

Dafoe, A.; Bachrach, Y.; Hadfield, G.; Horvitz, E.; Larson, K.; and Graepel, T. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593(7857):33–36.

Davis, N. M. 2013. Human-computer co-creativity: Blending human and computational creativity. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*. aaai.org.

Eno, B. 2024. AI's walking dog. https://www.bostonreview.net/forum_response/ais-walking-dog/. Accessed: 2024-12-30.

Essel, H. B.; Vlachopoulos, D.; Essuman, A. B.; and Amankwa, J. O. 2024. ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence* 6(100198):100198.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv [cs.CL]*.

Gerlich, M. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies (Basel)* 15(1):6.

Grigis, P., and De Angeli, A. 2024. Playwriting with large language models: Perceived features, interaction strategies and outcomes. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*. New York, NY, USA: ACM.

Haase, J.; Djurica, D.; and Mendling, J. 2023. The art of inspiring creativity: Exploring the unique impact of AI-generated images. In *AMCIS 2023 Proceedings*.

Heaven, W. D. 2023. The inside story of how ChatGPT was built from the people who made it. *MIT Technology Review*.

Heersmink, R. 2024. Use of large language models might affect our cognitive skills. *Nat. Hum. Behav.* 1–2.

Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '99, 159–166. New York, NY, USA: Association for Computing Machinery.

Hwang, A. H.-C.; Liao, Q. V.; Blodgett, S. L.; Olteanu, A.; and Trischler, A. 2024. "it was 80% me, 20% AI": Seeking authenticity in co-writing with large language models. *arXiv [cs.HC]*.

Kantosalo, A.; Toivanen, J. M.; Xiao, P.; and Toivonen, H. 2016. From Isolation to Involvement: Adapting Machine Creativity Software to Support Human-Computer Co-Creation. 7.

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*, IUI '20, 221–230. New York, NY, USA: Association for Computing Machinery.

Lee, M.; Gero, K. I.; Chung, J. J. Y.; Shum, S. B.; Raheja, V.; Shen, H.; Venugopalan, S.; Wambsganss, T.; Zhou, D.; Alghamdi, E. A.; August, T.; Bhat, A.;

ISBN: 978-989-54160-7-3

Choksi, M. Z.; Dutta, S.; Guo, J. L. C.; Hoque, M. N.; Kim, Y.; Knight, S.; Neshaei, S. P.; Shibani, A.; Shrivastava, D.; Shroff, L.; Sergeyuk, A.; Stark, J.; Sterman, S.; Wang, S.; Bosselut, A.; Buschek, D.; Chang, J. C.; Chen, S.; Kreminski, M.; Park, J.; Pea, R.; Rho, E. H. R.; Shen, Z.; and Siangliulue, P. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 7, 1–35. New York, NY, USA: ACM.

Lee, H. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 01, 2025, Yokohama, Japan* 1(1).

Masson, D.; Malacria, S.; Casiez, G.; and Vogel, D. 2024. DirectGPT: A direct manipulation interface to interact with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–16. New York, NY, USA: ACM.

McGuire, J.; De Cremer, D.; and Van de Cruys, T. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Sci. Rep.* 14(1):18525.

Muller, M.; Weisz, J. D.; and Geyer, W. 2020. Mixed initiative generative AI interfaces: An analytic framework for generative AI applications. *computationalcreativity.net*.

Norman, D. A. 2013. *The design of everyday things*. The MIT Press. London, England: MIT Press.

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. *arXiv [cs.CL]*.

Palani, S., and Ramos, G. 2024. Evolving roles and workflows of creative practitioners in the age of generative AI. In *Creativity and Cognition*. New York, NY, USA: ACM.

Rezwana, J., and Maher, M. L. 2022a. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Trans. Comput. Hum. Interact.*

Rezwana, J., and Maher, M. L. 2022b. Understanding user perceptions, collaborative experience and user engagement in different human-AI interaction designs for co-creative systems. In *Creativity and Cognition*. New York, NY, USA: ACM.

Shneiderman, B. 1997. Direct manipulation for comprehensible, predictable and controllable user interfaces. In *Proceedings of the 2nd international conference on Intelligent user interfaces - IUI '97*. New York, New York, USA: ACM Press.

Sun, Y.; Jang, E.; Ma, F.; and Wang, T. 2024. Generative AI in the wild: Prospects, challenges, and strategies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 18, 1–16. New York, NY, USA: ACM.

SXSW. 2023. OpenAI co-founder greg brockman on ChatGPT, DALL·E and the impact of generative AI — SXSW 2023.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv [cs.CL]*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, ; and Polosukhin, I. 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.

Wan, Q.; Hu, S.; Zhang, Y.; Wang, P.; Wen, B.; and Lu, Z. 2023. "it felt like having a second mind": Investigating human-AI co-creativity in prewriting with large language models. *arXiv [cs.HC]*.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity.

Yuan, A.; Coenen, A.; Reif, E.; and Ippolito, D. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, 841–852. New York, NY, USA: Association for Computing Machinery.

Reimagining Dance: Real-time Music Co-creation between Dancers and AI

Olga Vechtomova* and **Jeff Bos****

*University of Waterloo, Canada
**WordSynth Inc.
ovechtom@uwaterloo.ca, jeff@wordsynth.com

Abstract

Dance performance traditionally follows a unidirectional relationship where movement responds to music. While AI has advanced in various creative domains, its application in dance has primarily focused on generating choreography from musical input. We present a system that enables dancers to dynamically shape musical environments through their movements. Our multi-modal architecture creates a coherent musical composition by intelligently combining pre-recorded musical clips in response to dance movements, establishing a bidirectional creative partnership where dancers function as both performers and composers. Through correlation analysis of performance data, we demonstrate emergent communication patterns between movement qualities and audio features. This approach reconceptualizes the role of AI in performing arts—as a responsive collaborator that expands possibilities for both professional dance performance and improvisational artistic expression across broader populations.

Introduction

Dance and music traditionally exist in a hierarchical relationship where movement follows sound. Typically, choreographers design dance to existing music, or collaborate with composers to create accompanying scores. Even in improvisational dance, performers respond to pre-composed or live music, but rarely influence the musical composition itself.

Artificial intelligence now offers an opportunity to invert this relationship. While most AI systems in dance maintain the traditional paradigm by generating choreography from musical input, our research proposes a fundamental shift: enabling dancers to dynamically shape musical environments through their movements. This approach reconceptualizes dancers as both performers and composers, establishing a bidirectional creative partnership between human movement and AI-created sound.

In this paper, we present a system that enables dancers to dynamically influence musical composition in real-time through their movements. Our technical contribution is a multi-modal architecture (Figure 1) that selects and seamlessly combines pre-recorded musical clips in response to a dancer's movement patterns. Unlike previous approaches, our system creates an evolving musical environment where the dancer becomes an active co-creator of the soundscape rather than merely responding to it. We demonstrate through correlation analysis of pilot performances that this approach creates emergent communication patterns between dancer and system, establishing meaningful bidirectional relationships between specific movement qualities and audio features.

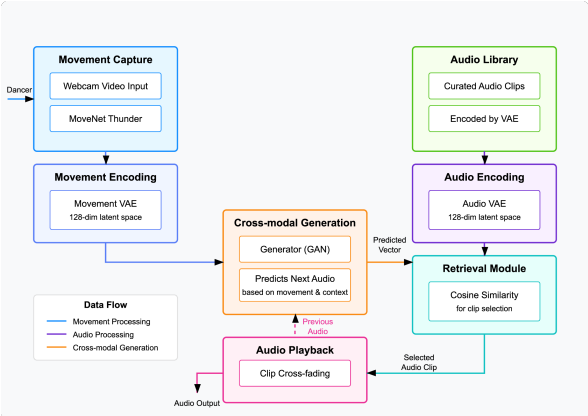


Figure 1: System diagram

This approach not only transforms dance performance but also opens new theoretical and practical directions for studying creative partnership between human performers and AI systems. While considerable research has examined AI's role in generating static artistic content like images or poetry, the dynamic, real-time nature of dance performance presents unique challenges that remain largely unexplored.

Related Work

Existing research in dance and AI has primarily focused on generating choreographic movements from musical input. Tang et al. (2018) developed an LSTM-autoencoder model that synthesizes dance choreography by mapping acoustic features to motion features, addressing the challenge of selecting appropriate dance figures that match musical elements. Lee et al. (2019) proposed a synthesis-by-analysis framework that decomposes dance into basic units to generate style-consistent and beat-matching movements from music. While these approaches demonstrate technical sophistication in movement generation, they maintain the traditional

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

unidirectional relationship where dance follows music.

Some researchers have begun exploring more interactive approaches to AI in dance. Kumar et al. (2020) developed LuminAI, an improvisational dance installation where an AI agent dances with users, implementing a lead-and-follow dynamic based on creativity metrics. The choreographic duo AΦE's recent "Lilith.Aeon" performance, as reported in The Guardian (Winship 2024), demonstrated how an AI system trained on human-generated movements could become an active creative partner, suggesting new movement possibilities while maintaining the distinctive style of the choreographers. The Royal Ballet choreographer Wayne McGregor's collaboration with Google Arts & Culture Lab produced AISOMA, a system that suggests new movement variations by analyzing rehearsal videos, expanding the choreographic possibilities available to dancers and choreographers (Winship 2024).

Theoretical Foundations. This research builds on several theoretical foundations that help frame our understanding of human-AI creative collaboration. Particularly relevant is the concept of "mixed-initiative creative interfaces" developed by Deterding et al. (2017), which describes systems where human and computational agents take turns contributing to an evolving artistic work. In the context of dance performance, this framework takes on new dimensions as the collaboration happens in real-time, with the dancer's physical movements and the AI's musical responses creating a dynamic feedback loop.

Additionally, our work is informed by computational creativity concepts such as Colton and Wiggins' (2012) "creative responsibility," where AI systems take on creative roles beyond mere tools—evaluating aesthetics and inventing processes. This complements Jennings' (2010) notion of "creative autonomy," which requires systems to independently evaluate and evolve their standards.

While the pilot study reported in our paper establishes technical foundations, these frameworks guide our vision for AI systems that can function as genuine creative partners in dance performance.

Beyond computational creativity, performance studies literature on improvisation and real-time creative decision-making provides another crucial theoretical perspective. Foundational work by Bailey (1992) and Nachmanovitch (1990) established key principles of improvisational practice, while recent scholarship addresses the complexities of dance improvisation and human-machine interaction. De Spain's (2014) topographical approach illuminates how dancers make moment-to-moment decisions, and Foster (2002) examines how improvisational structures emerge through real-time choreographic choices. For human-AI creative collaboration specifically, Hoffman and Weinberg's (2011) work on interactive robotic improvisation offers frameworks for understanding how performers adapt to non-human partners. Carter's (2000) analysis of improvisation as breaking established conventions to discover new artistic expressions that "could not be found in a systematic preconceived process" is particularly relevant. Following Carter, our system aims to create new paradigms that enable real-time invention and discovery through the act of creation itself.

System Architecture

Our system enables real-time generation of responsive musical accompaniment to dance movements through a multi-stage machine learning pipeline. The architecture comprises three primary components: (1) an audio encoding/decoding system, (2) a movement encoding system, and (3) a cross-modal generation network that predicts appropriate musical responses to movement. These components work in concert to create a cohesive interactive performance environment.

Audio Representation Learning. To learn audio representations, we trained a Variational Autoencoder (VAE) (Kingma and Welling 2014) on spectrograms of 3.5-second audio clips. The audio VAE consists of convolutional layers for both encoding and decoding, with a 128-dimensional latent space representation. This architecture effectively compresses spectrograms into a compact latent code that preserves meaningful acoustic properties while discarding noise. The encoder uses five convolutional layers with ReLU activations to transform input spectrograms (224×224×1) into a latent distribution, while the decoder reverses this process through transposed convolutions.

Movement Representation Learning. To encode dance movements, we developed a parallel VAE architecture that processes visual representations of movement trajectories. The movement data is collected by using TensorFlow MoveNet Thunder pipeline, which analyzes either pre-recorded videos during training or webcam video stream during inference. Rather than working with raw skeletal joint data, we first transform movement sequences into color-coded trajectory images (Figure 2b). Each 3.5-second movement sequence is represented as an RGB image (256×256×3) where five key landmarks (head, left wrist, right wrist, left ankle, right ankle) are visualized as coloured curves showing their trajectories over time.

(a) Dance timelapse photo (b) Movement trajectories

Figure 2: (a) Timelapse photo of dancer's movements. (b) Trajectories of dancer's movements captured by our system. The landmarks are colour-coded as follows: red - head, green - left wrist, blue - right wrist, orange - left ankle, magenta - right ankle.

This image-based approach allows us to leverage convolutional architectures commonly used in computer vision while capturing the temporal dynamics of movement. The movement VAE employs a structure parallel to the audio VAE, with the encoder producing a 128-dimensional latent vector that encapsulates the essential spatial and temporal characteristics of the dance movement.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Cross-Modal Generation The Generative Adversarial Network (GAN) (Goodfellow et al. 2014) bridges the movement and audio domains. The GAN's generator takes two inputs: (1) the latent representation of the current movement and (2) the latent representation of the previous audio clip. It then predicts the latent vector for the next audio clip that would best complement the current movement.

The generator employs a latent combiner module that integrates movement and audio latent vectors. While we experimented with several combination methods (concatenation, multiplication, and various learned approaches including gated, FiLM, and cross-attention mechanisms), we found that pointwise addition produces the most effective results. This addition operation is followed by a multi-layer network with hidden dimensions of 256 units, LayerNorm for stabilization, and LeakyReLU activations.

Retrieval Module Rather than directly decoding the predicted latent vector, which could result in lower audio quality, we employ a retrieval-based approach. We calculate the cosine similarity between the predicted latent vector and the latent representations of clips in our reference database. The audio clip with the highest similarity is selected, ensuring high-quality output while maintaining contextual relevance.

Figure 3 shows how both previous audio and movement influence the system's predictions. With identical previous audio but different movements, the system selects dramatically different clips: ambient music for minimal movement (3a) versus rhythmic clips for energetic break-dancing (3c). Similarly, when movement remains constant but previous audio changes (3a vs. 3b), the predicted clips also differ.

Real-time Inference We developed a React/NodeJS application, which operates as follows during live performance:

Movement Capture: A webcam captures the dancer's movements, which are processed by a pose estimation model (TensorFlow MoveNet Thunder) to extract the five key landmarks.

Movement Encoding: The landmarks' trajectories are rendered as a color-coded image and encoded by the movement VAE into a latent vector.

Audio Prediction: The movement latent vector and the previous audio clip's latent vector are fed into the GAN, which predicts the next audio clip's latent representation.

Clip Selection: The system retrieves the audio clip whose latent representation has the highest cosine similarity to the predicted vector.

Audio Playback: The selected clip is cross-faded with the currently playing audio to create seamless transitions.

The dancer can also curate the source audio library from which the system selects clips, allowing performers to influence the overall sonic palette and musical style of their movement-conditioned compositions.

Dataset. The aligned video-audio dataset used to train the movement VAE and GAN consists of 18K 3.5-second recordings, 17K of which were sourced from the AIST dance dataset (Tsuchida et al. 2019) and 1K from our own dataset containing video material recorded specifically for this project or provided to us by professional dance collaborators. The audio VAE was trained on a larger set of 50K audio clips from the authors' studio recordings, spanning multiple genres of electro-acoustic music with varying tempos and instrumentation.

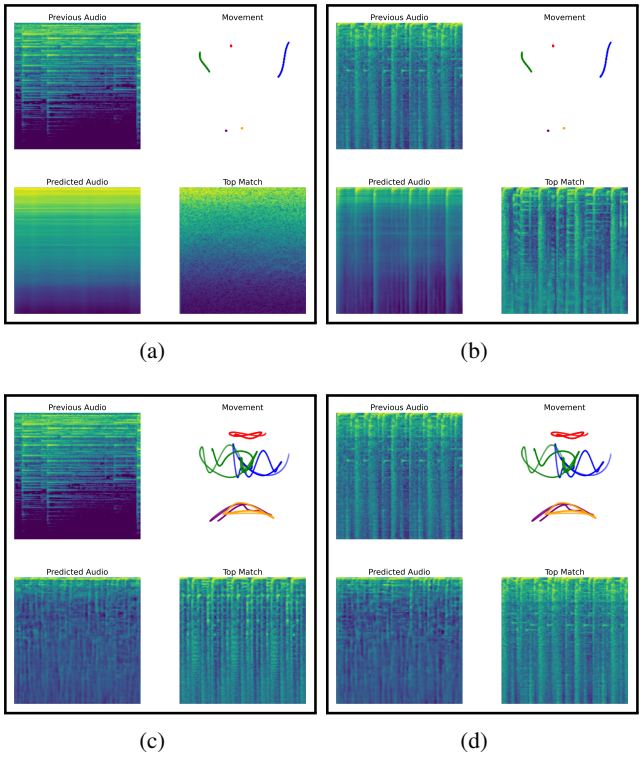


Figure 3: System response examples showing the influence of both inputs on prediction.

Pilot study

We conducted a pilot study with three participants of varying dance experience: P1 (10+ years ballet), P2 (2-3 years ballet/jazz), and P3 (no formal training). Each dancer performed improvisational movement with the system for up to 30 minutes. We recorded both video and system-generated audio, collecting over 70 minutes of data. Our analysis aimed to identify relationships between dance movements and generated audio, particularly examining which audio features correlate most strongly with movement intensity.

Video and Audio Features. We segmented performances into 10-second clips and extracted movement data using MoveNet Thunder to track key body points (head, wrists, ankles) normalized to [-1, 1]. Movement energy was measured as the Euclidean distance between corresponding points in consecutive frames, with statistical measures (mean, min, max, standard deviation) computed for each clip. For audio, we extracted 47 features using Librosa and Essentia, including spectral features (MFCCs, contrast, flux), chroma features, and psychoacoustic measures. Key audio features in our analysis include *MFCCs* (representing timbre, with *mfcc_1* capturing overall spectral shape), *spectral contrast* (peak-valley differences across frequency bands), *chroma* (pitch class distribution), and *spectral flux* (frame-to-frame

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

spectral changes).

Statistical Analysis[1] To explore the relationship between dance movement and system-generated audio, we applied several statistical methods. Pearson correlation analysis was performed to assess the linear relationships between individual movement energy measures (average, maximum, minimum, and standard deviation) and audio features. Principal Component Analysis (PCA) was conducted separately on video and audio features to identify patterns of variance and reduce dimensionality, while Canonical Correlation Analysis (CCA) was used to examine the multivariate relationships between movement and audio feature spaces.

Additionally, we used Partial Least Squares (PLS) regression to model predictive relationships between the two modalities, evaluating how well sets of audio features could predict movement energy metrics, and vice versa. To identify the most influential audio features, we employed Random Forest regression models, computing feature importance scores based on their contribution to predicting each movement energy statistic. The quality of these predictive models was evaluated using the coefficient of determination (R^2).

Together, these analyses allowed us to identify which audio features were most strongly associated with variations in dancers' movement intensity and to assess the strength of the coupling between movement and audio dynamics produced by the system.

Results. Our analysis revealed several key relationships between movement energy and audio features generated by the system.

Correlation Analysis. Pearson correlation analysis showed that the minimum movement energy (*min_energy*) had the strongest and most consistent relationships with audio features. In particular, *mfcc_1* (first Mel-frequency cepstral coefficient) exhibited a significant negative correlation with min_energy ($r = -0.45$, $p < 0.001$), suggesting that clips with lower minimum movement energy were associated with audio segments characterized by smoother spectral shapes. Spectral contrast in the sixth frequency band (*spec_contrast_6*) and *mfcc_7* also showed significant positive correlations with min_energy.

Principal Component Analysis (PCA). PCA indicated that a small number of components captured substantial variance in both movement and audio features. In particular, variations in energy-based movement measures (average, minimum, maximum, and standard deviation) loaded heavily onto the first few principal components, while audio variance was dominated by MFCCs and spectral features.

Canonical Correlation Analysis (CCA). CCA revealed moderate canonical correlations between the combined movement energy statistics and audio feature sets. The first canonical component pair linked high standard deviation of movement energy with variations in spectral complexity and dissonance in the audio, indicating multivariate coupling between movement expressivity and audio texture.

Partial Least Squares (PLS) Regression. PLS regression models found that audio features could modestly predict movement energy, with the highest R^2 value (0.103) for predicting *min_energy*. Conversely, movement features showed stronger predictive power for certain audio characteristics: movement energy metrics could predict *mfcc_1* with an R^2 of 0.202 and *mfcc_7* with an R^2 of 0.162, suggesting that dancers' movement dynamics influenced the timbral qualities of the generated soundtrack.

Random Forest Feature Importance. Random forest regressions further confirmed the importance of specific audio features. *mfcc_1*, *mfcc_7*, *spec_contrast_6*, and *spectral_flux* consistently emerged as the most important predictors of movement energy statistics across models, aligning with the findings of the linear analyses.

Overall, the results suggest that the soundscape created by our system responded most consistently to variations in dancers' minimum movement energy, with audio features related to timbre and spectral dynamics (MFCCs and spectral contrast) showing the strongest associations with movement intensity.

Qualitative results Dancers reported a fluid exchange of initiative with the system throughout their performances. The system often influenced their movement choices at both macro and micro levels, inspiring exploration of new dance expressions to discover corresponding musical responses. Conversely, dancers sometimes deliberately attempted to redirect the musical atmosphere, such as introducing energetic movement during ambient passages. While the system typically required 5-10 seconds to adapt to significant changes in dance energy, this delay was intentionally designed to maintain musical coherence. This adaptation period could potentially be customized based on dance genre preferences, balancing responsiveness against musical continuity.

Conclusion

This research presents a novel approach to dance through AI-mediated co-creation, fundamentally reimagining the traditional relationship between movement and music. Our system enables dancers to dynamically shape musical environments while simultaneously responding to them, creating a bidirectional creative partnership where initiative flows fluidly between human and machine. Statistical analysis confirms meaningful correlations between specific movement qualities and audio features.

This work explores a multi-layered creative relationship. The original composer's musical intent—which may itself be improvisational—becomes raw material for a new emergent composition, dynamically rearranged and remixed through the dancer's movements. The resulting sonic experience is a form of real-time collage where three creative forces converge: the original compositional elements, the system's algorithmic decision-making, and the dancer's embodied expression. Each performance thus represents a unique relationship between these creative entities, with the dancer physically sculpting a new musical composition from fragments of the composer's work, creating something neither could have produced independently.

By inverting the traditional paradigm where movement follows sound, we position dancers as active co-creators of

[1] Detailed statistical analysis results are provided on the supplementary-material website: https://sites.google.com/view/reimagining-dance

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

the musical arrangements. The dance movements can serve as a novel compositional tool for creating musical content that has artistic value beyond the performance context. This approach offers new creative possibilities not only to dance performance, but also as a form of musical composition.

While the system was primarily envisioned as a co-creative artistic framework for dance performance, another promising application emerged during our pilot study. We observed that when participant dancers experienced fatigue and naturally reduced their movement intensity, the system organically transitioned from dynamic rhythmic music to more ambient soundscapes. This adaptive quality could be valuable not only in dance performance, but also in exercise and training settings.

We are currently planning a large-scale study with a professional dance company to evaluate the system's impact on choreographic process and audience reception. Through this ongoing research, we aim to develop a deeper understanding of the emergent creative language that evolves between human dancers and AI musical collaborators.

References

Bailey, D. 1992. *Improvisation: Its Nature and Practice in Music*. Da Capo Press.

Carter, C. L. 2000. Improvisation in dance. *Journal of Aesthetics and Art Criticism* 58(2):181–190.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *ECAI 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, 21–26. IOS Press.

De Spain, K. 2014. *Landscape of the Now: A Topography of Movement Improvisation*. Oxford University Press.

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*, CHI EA '17, 628–635. New York, NY, USA: ACM.

Foster, S. 2002. *Dances that Describe Themselves: The Improvised Choreography of Richard Bull*. Wesleyan University Press.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Hoffman, G., and Weinberg, G. 2011. Interactive improvisation with a robotic marimba player. *Autonomous Robots* 31(2):133–153.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Kumar, M.; Long, D.; and Magerko, B. 2020. Creativity metrics for a lead-and-follow dynamic in an improvisational dance agent. In *Proceedings of the International Conference on Computational Creativity*.

Lee, H.-Y.; Yang, X.; Liu, M.-Y.; Wang, T.-C.; Lu, Y.-D.; Yang, M.-H.; and Kautz, J. 2019. Dancing to music. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 3586–3596. Red Hook, NY, USA: Curran Associates Inc.

Nachmanovitch, S. 1990. *Free Play: Improvisation in Life and Art*. Penguin Putnam.

Tang, T.; Jia, J.; and Mao, H. 2018. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of ACM Multimedia Conference*.

Tsuchida, S.; Fukayama, S.; Hamasaki, M.; and Goto, M. 2019. AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*.

Winship, L. 2024. Small step or a giant leap? what AI means for the dance world. *The Guardian*. Accessed: January 11, 2025.

Battle Rap as a Framework for Human-Machine Co-Creativity

Ibùkún Ọlátúnjí[1*] **Mark Sheppard**[2*]

Alma Rahat[1] Matt Jones[1] Amanda Rogers[3]

[1]Computational Foundry, Swansea University, Crymlyn Burrows, Skewen, Swansea SA10 6JW, UK
[2]University of Kent, Canterbury, Kent CT2 7NZ, UK
[3]Department of Geography, Wallace Building, Swansea University, Singleton Park, Swansea SA2 8PP, UK

Abstract

We present a human-in-the-loop GAN framework for battle rap, where a human artist (MC) serves as generator, and the AI acts as an adaptive discriminator. The AI provides feedback on rhyme complexity, coherence, and stylistic alignment, challenging the MC's improvisational skill. Fine-tuned language models emulate diverse rap styles, while voice cloning creates adversarial loops: the MC competes against stylised versions of their own voice in a dynamic, self-reflective duel. The system follows a dual-phase process: (i) an *Emulation Phase*, where AI mimics established flows to reinforce technical mastery, and (ii) an *Improvisation Phase*, where AI disrupts expectations to prompt originality. This ensures that creative growth emerges from constraint and challenge. Success is judged through MC evaluations of the AI's performance as an adversary. Framed as a study paper, this work offers a thought experiment in adversarial co-creativity, modelling how AI might inspire, rather than merely assist, human expression. Beyond computational modelling, the framework offers insights into machine-mediated creativity and how AI can be designed to provoke human creativity through improvisation, challenge, and real-time performance. The study positions the AI as a dynamic co-performer capable of eliciting novel artistic responses. As such, it contributes to emerging discourse on creative AI systems that influence, not just assist, human expression

Introduction

Rap is a primary ingredient of hip-hop music, the world's most popular genre (Lynch 2018; Texas 2015). It is a form of vocal delivery and expression that incorporates rhyme, rhythmic speech, street vernacular, and is performed over musical accompaniment (Wikipedia 2019; Edwards and Kool G Rap 2009). As Stevie Wonder describes it, '*Rap...is modern blues — a statement of how and where people are at.*' (Internet Archive 2001). We argue that rap is also a novel method for exploring human-computational creativity, as it encompasses rich linguistic, speech, musical, and socio-cultural data (Bradley 2017). Its key components are presented in Table 1.

Component	Linguistic Features	Musical Features
Content	Storytelling, metaphor, wordplay	Phonetic manipulation, assonance, consonance
Flow	Rhyme schemes, syntactic complexity	Rhythm, tempo, syncopation
Delivery	Prosody, emphasis, vocal texture	Timbre, breath control, cadence

Table 1: Key Elements of Rap

Rap contains evolving lexicons based around (i) numerous genres; and, (ii) vocal delivery based around syncopation, pitch, intonation, and cadence (Condit-Schultz 2016; Orejuela 2021). In addition, rap's short and well-documented history allows observation of the genre from its genesis to the present (Condit-Schultz 2016; Orejuela 2021).

.

Rap and Computational Creativity We propose a computational model that integrates language models and musical inputs to analyse the linguistic, stylistic, and rhythmic differences across rap genres (Copet et al. 2024; Radford et al. 2019). This approach provides insights into

*Corresponding authors: 2030349@swansea.ac.uk, ms2403@kent.ac.uk

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

how flow[1], rhyme complexity, and thematic elements vary across styles, from battle rap to conscious hip-hop. Beyond rap, the model serves as a tool for examining broader language structures, offering insights into syntax, semantics, and linguistic creativity across different forms of expression (Akingbe and Onanuga 2018; Coscarello 2003). The model integrates constraints that reflect the MC's personal belief systems, shaping lyric generation by influencing thematic choices, and linguistic style (Edwards and Kool G Rap 2009). By embedding value-driven constraints, it examines how artistic expression is influenced by individual perspectives, cultural norms, and societal expectations (Serrano, Torres, and 2015). This provides insight into how artists navigate self-imposed or external creative limitations in music and language (Benvenga 2022).

The framework also explores the interaction between human improvisation and AI-assisted lyricism. By analysing MC responses to computational feedback in real time, it reveals how AI can influence creative decision-making, challenge artistic boundaries, and shape stylistic conventions. This deepens the understanding of AI's role as both a collaborator and an adaptive creative force in rap and other improvisational art forms (Arnold, Volzer, and Madrid 2021; Stark et al. 2023).

1. *How has rap evolved within a chosen sub-genre in terms of flow and word choice?*
   - We analyse the historical development of a specific rap sub-genre, focusing on its stylistic evolution.
   - The model allows for the integration of variable levels of artist *experiential history*, simulating how personal influences shape rap battles.
2. *How can the stylistic flow of an individual artist (B) be aligned with the broader development of their genre (A)?*
   - We introduce a framework where A and B serve as modifiers, enabling comparisons between an artist's unique flow and their genre's evolution.
   - This allows us to explore how an artist's style (B) could adapt to another genre (A2), or how a genre's conventions (A) influence a different artist (B2).
3. *What is the minimal phrase set required to define both a rap genre and an artist's individual flow?*
   - We investigate the smallest possible set of linguistic and rhythmic patterns that characterize both genre-specific and artist-specific styles.

Figure 1 describes the system in terms of the Improvisational Model and its key elements.

Battle Rap as a Creativity Game

Musical improvisation is an ideal subject for studying creativity as it involves generating novel ideas that are relevant within a given context (Csikszentmihalyi 1997; Zhang, Sjoerds, and Hommel 2020). As a specific creative act, writing verse requires an extensive vocabulary, mastery of complex rhyme patterns, and broad subject knowledge across diverse topics (Bradley 2017; Liu et al. 2012). Within rap, *freestyle*, the spontaneous composition of lyrics, is widely considered the most challenging skill to master (Edwards and Kool G Rap 2009). An fMRI study of freestyle rap revealed that it requires rapid linguistic processing, as MCs must generate meaningful, rhyming phrases in real-time while adhering to tempo and rhythm constraints (Liu et al. 2012). The process shares key characteristics with the *flow state*, where cognitive effort, heightened focus, and automaticity merge to enable fluid creative expression (Csikszentmihalyi 1997). Freestyle rap also relies on phonological awareness, lexical retrieval, and verbal dexterity (Liu et al. 2012). The ability to manipulate sound, word meanings, and rhythm in real-time is fundamental to the art form (Edwards and Kool G Rap 2009). Battle rap, a competitive format of freestyle, introduces additional constraints that amplify both its creative challenge, and its suitability for computational modelling. Unlike unstructured freestyle, battle rap requires MCs to improvise responses that directly engage with an opponent's lines while incorporating rhetorical strategies such as humor, wordplay, and personal rebuttals, all within rhythmic, structural, and/or temporal constraints (Caffeine 2020; Todd 2018). The constraints of battle rap result in two main properties:

1. *A High-Level Creativity Challenge for Human MCs* Battle rap demands that performers adapt their responses on-the-fly while maintaining coherence, linguistic complexity, and audience engagement. The interplay between structured rhythmic demands and unpredictable opponent-generated content makes it a cognitively intense and uniquely challenging form of improvisational creativity (Liu et al. 2012; Landau and Limb 2017).

2. *A Structured Framework for Computational Approaches* Battle rap's constraints and evaluative criteria (e.g.,audience reaction, lyrical complexity) allow for the development of AI systems that can generate, evaluate, and refine battle rap performances using natural language processing, machine learning, and adversarial feedback loops.

Through battle rap we can explore how AI might engage in real-time linguistic improvisation, assess human performance, and model the balance between structured constraints and creative freedom that define high-level artistic expression. This approach enables the development of computational tools capable of both augmenting human creativity, and evaluating it in a quantifiable, reproducible manner. Several computational approaches can be applied to modelling rap improvisation:

Natural Language Processing (NLP) for Lyrical Generation: State-of-the-art NLP architectures such as sequence-to-sequence models, transformers (e.g., GPT (OpenAI 2024; Xue et al. 2021) can be trained to generate context-aware rap lyrics. Unlike conventional text generation, battle rap

[1]Defined as 'all of the rhythmical and articulative features of a rapper's delivery of the lyrics' (Adams 2009)

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

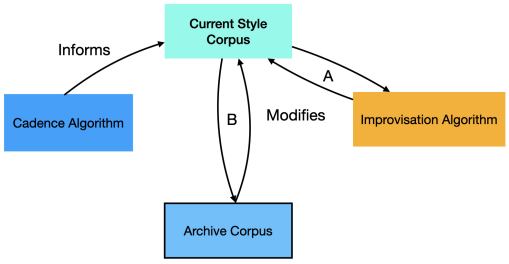


Figure 1: Improvisational Model

demands real-time linguistic agility and adversarial adaptation, requiring AI to anticipate and respond to an opponent's lyrics dynamically.

Phonological and Prosodic Analysis: As rap performance is intrinsically tied to rhythm and phonetics, computational models must incorporate phonological constraints to maintain natural flow, cadence, and prosody (Edwards and Kool G Rap 2009). Techniques such as rhyme density estimation (Hirjee and Brown 2010; Malmi et al. 2016), syllabic pattern recognition (Prinsloo and Coetzer 1990; Rogova, Demuynck, and Van Compernolle 2013), and prosody alignment (Wu et al. 2023; Cohn et al. 2021) ensure that AI-generated lyrics adhere to human-like rhythmic structures. Phonetic modelling is particularly important for rap, where multi-syllabic rhyme schemes and syncopated delivery distinguish high-level lyricism (Adams 2009).

Adversarial Learning and Response Generation: The call-and-response nature of battle rap naturally aligns with adversarial AI techniques. Generative Adversarial Networks (GANs), a machine learning framework where a generator creates data and a discriminator evaluates it in a competitive process, can be applied to text generation (Haidar et al. 2019). Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon 2016) can be used to train AI to replicate expert-level rap improvisation by learning response patterns from battle data. When combined with semantic analysis and rhyme modelling, this approach could refine AI-generated counter-responses, mimicking how human MCs anticipate and react to opponent rebuttals with humour, disses, and wordplay. Reinforcement learning further optimizes response strategies by rewarding coherence, stylistic alignment, and adversarial effectiveness (Dognin et al. 2021).

Human-in-the-Loop Creativity Models: In the proposed system, Human-in-the-Loop (HITL) models keep creativity human-driven by positioning AI as an adaptive counterpart. The MC battles an AI digital twin, making the creative output the result of their interaction. Rather than replacing improvisation, the AI challenges the MC with stylistic variations and rhythmic adaptations, refining their skills through competition.

Computational Creativity Benchmarks: By analysing battle rap as a computational creativity game, we explore how AI can engage in real-time lyrical improvisation, assess human performance, and develop models that balance structured constraints with creative freedom. The ability to model adversarial language interactions has applications beyond rap, contributing to advancements in:

(i) *Dialogue systems:* Enhancing real-time conversational AI with rhetorical and adversarial elements (Chen et al. 2024).

(ii) *Stylistic text generation:* Refining personalized content creation models for creative writing (Stark et al. 2023).

(iii) *Interactive AI:* Developing AI that adapts to dynamic user input in creative and performative contexts (Gonçalo Oliveira, Mendes, and Boavida 2017) Battle rap provides a structured yet challenging test bed for evaluating AI's ability to generate, adapt, and respond in high-stakes linguistic interactions. As AI continues to develop, battle rap provides an important platform for testing human-AI improvisational collaboration and/or competition.

Related Work

This section is structured into three key research areas that directly inform AI-driven battle rap generation.

Language Generation and Speech Recognition

Language Generation: In language terms, rap can be thought of as an evolution of the poetic tradition. This view is supported by prominent poets and MCs (The Irish Times 2003; Edwards and Kool G Rap 2009). There is a long history of automated poetry systems with the earliest dating from the 18th - 19th centuries (Sharples 2015; Sharples and Perez y Perez 2022). More recently, neural network-based approaches have been used in automatic poetry classification and generation (Ghazvininejad et al. 2017; Lau et al. 2018). These approaches have also been applied to rap lyrics (Malmi et al. 2016; Xue et al. 2021). However, Transformer models have emerged as the dominant architecture for text classification, analysis, and generation tasks (Radford et al. 2019; Pichai and Hassabis 2023). While these models have shown promise in Automated Speech Recognition (ASR), their application to rap transcription remains an open challenge due to the complexities of flow, slang, and phonetic variation. Developing a battle rap system capable of generating real-time lyrical responses to a human MC requires highly accurate, low-latency ASR tailored to these linguistic and rhythmic nuances (see Figure 3)

Speech Recognition: Automated Speech Recognition (ASR) systems have made significant advancements in recent years, achieving Word Error Rates (WER) as low as 5%, approaching human transcription accuracy (Protalinski 2017; Hollands, Blackburn, and Christensen 2022; Apple 2024). Despite this progress, ASR systems still exhibit higher error rates when processing accents, dialects, and non-native English speak-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

ers (Wassink, Gansen, and Bartholomew 2022; Hollands, Blackburn, and Christensen 2022). These limitations are particularly relevant to the proposed system, as rap lyrics frequently incorporate non-standard English, slang, phonetic variation, and complex rhythmic structures, which pose additional challenges for ASR (Coscarello 2003; Akingbe and Onanuga 2018). Current ASR systems also perform significantly less well in speech and music,and/or singing use cases. In these situations WERs are within the 10% - 30% range (Music.AI 2024; Ou, Gu, and Wang 2022). Currently, no ASR models are specifically designed for rap transcription, presenting a gap in the research.

Prosody, Phonetics and Computational Creativity

Speech modelling: Rap presents a unique challenge for both ASR and automatic lyric generation, requiring precise syllabic stress, phoneme timing, and cadence alignment with the beat (Adams 2009; Hirjee and Brown 2010). Hirjee Brown (2009) analysed rhyme density, internal rhymes, and multi-syllabic structures in rap, while Condit-Schultz (2016) examined how MCs manipulate syllable stress to fit beats (Condit-Schultz 2016). The complexity of rap is inherently multi-modal, requiring computational models to handle both textual elements (*meaning*) and phonetic features (*sound*). As Murs explains, '*...not only do you have to make everything rhyme, but you have to add rhythm to it. Poetry doesn't have to rhyme, it just has to sound beautiful. But in rap, it has to sound beautiful, it has to be on time, and it has to rhyme*' (Edwards and Kool G Rap 2009).

While meaning and delivery are often regarded as equally important for human MCs, a case can be made that delivery plays a more crucial role. As Havoc of Mobb Deep remarked, '*...without the right flow, subject matter probably won't even matter. It's all about styles... the way you're getting your subject across. If people can't feel how you're saying it, it doesn't even matter what you're saying.*' (Wikipedia 2020; Edwards and Kool G Rap 2009) This distinction is particularly relevant for computational systems. While human MCs balance meaning and delivery, AI models may find delivery and flow more achievable, as meaning requires deeper contextual reasoning and cultural awareness. Prioritizing delivery in computational models could enable more realistic rap co-creation, even if full semantic understanding remains a challenge.

Generating semantically rich rap lyrics remains difficult, though *DeepBeat* and *DeepRapper* have explored data-driven solutions (Malmi et al. 2016; Xue et al. 2021). *DeepBeat* optimizes *rhyme density* and *coherence*, leveraging a neural network to predict next-line candidates. However, it lacks rhythm modelling, real-time lyric generation, and adversarial response capabilities - key for rap battles. *DeepRapper* builds upon this by using a Transformer to model rhyme patterns and rhythmic structures, improving fluency over DeepBeat. However, it remains a text-only system, making it unsuitable for rap battle co-creation with human MCs.

Adversarial AI & Rap Dynamics

Adversarial AI: Generative Adversarial Networks (GANs) and Reinforcement Learning (RL) have been applied to speech synthesis, music generation, and real-time improvisation, enabling AI to model stylistic variation and adaptive decision-making (Goodfellow et al. 2014; Sutton and Barto 2015). Approaches such as MuseGAN (Dong et al. 2017), SeqGAN (Yu et al. 2017), and Yi et al. (Yi et al. 2018) demonstrate AI's ability to generate structured musical compositions, poetic text, and lyrics. Learning a policy from expert demonstrations without direct interaction or reinforcement signals is challenging. Traditional methods use inverse reinforcement learning (IRL) to infer a cost function, followed by reinforcement learning (RL) to extract a policy, but this is indirect and computationally inefficient. A more effective alternative is adversarial imitation learning, where a model directly mimics expert behaviour using an approach analogous to GANs. This model-free method improves performance in complex, high-dimensional tasks. For rap improvisation, *Generative Adversarial Imitation Learning (GAIL)* provides a promising framework for AI-human co-creation (Ho and Ermon 2016). GAIL refines its policy through adversarial learning, enabling the AI to imitate expert rap performances and generate competitive, human-like responses. Table 2 summarizes the mapping of GAIL to rap battles.

Rap Dynamics: Battle rap presents a major AI challenge due to its call-and-response format, requiring low-latency NLP and speech synthesis for real-time rebuttals. Success depends on lyrical complexity, vocal tone, delivery, and crowd engagement. Battle rap's dense rhyme schemes, compound wordplay, and extended metaphors (Todd 2018) demand long-range semantic modelling for coherence. While freestyle battles may include pre-written elements, AI could employ pre-determined battle strategies, yet current text models (e.g., GPT) struggle with coherence and adversarial wordplay in real-time settings (Bender et al. 2021). Cadence, flow, and beat adaptation are equally crucial. AI-generated rap must be beat-aware and adjust flow dynamically, but most NLP models lack rhythmic and phonetic optimization. While DeepRapper improves beat alignment, generating spontaneous, flow-optimized responses remains a challenge (Xue et al. 2021). Rebuttal generation is another hurdle, as MCs rely on comedy, cultural references, and wordplay to engage the crowd. AI struggles with socio-culturally specific language, requiring context awareness and adaptability beyond current capabilities (Mirowski et al. 2024).

Proposed Method

As a proof of concept, the audio is processed from a pre-recorded source with an appropriate sample rate, though ultimately real-time data capture will be employed, enabling more realistic rap battles to take place. The audio data is extracted and passed through a cadence algorithm to evaluate the dynamics, rhythm, and structure of each phrase; creating a wire frame model of the rap. Further phrases are added to create a larger training set for the AI in order to

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

GAIL Concept	Rap Battle Adaptation
Generator (Policy π)	The human MC improvises lyrics, responding to AI-generated lines.
Expert Demonstrations	Past battle rap performances (e.g., datasets of freestyle battles, MCs' previous performances).
Discriminator (Reward Model D)	The AI evaluates rap responses based on fluency, coherence, rhyme complexity, wordplay, and crowd engagement.
Reward Signal	AI adapts based on what makes strong battle rap responses, incorporating linguistic complexity, adversarial effectiveness, and emotional impact.

Table 2: Applying Generative Adversarial Imitation Learning (GAIL) to Rap Battles

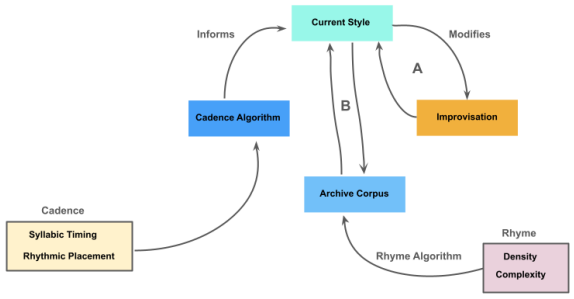


Figure 2: Simplified Cadence and Rhyme Improvisation Model

learn the style of the MC combatant. Several iterations of learning are carried out to create a more robust and detailed corpus, which is then stored as part of a wider style library. The training set can be expanded by loading library datasets from pre-recorded MCs, in order to introduce stylistic variety. Once trained, the corpus can be used to inform the AI improvisation algorithm, by generating new phrases in either the style of the human combatant or from the MC archive. In the proposed improvisational model (Figure 2), improvisation is driven by the current style corpus (A), which captures the AI's learned cadence, rhyme structures, and stylistic tendencies from prior training. This corpus forms the foundation for generating responses that align with the expected flow and delivery in a rap battle, with the system adapting to evolving stylistic cues.

Improvisation is also shaped by the archive corpus (B), which contains historical and contemporary rap data. Referencing this archive enhances lyrical complexity, cadence variation, and stylistic diversity, allowing the AI to adapt in real time. Additionally, it serves as a training resource, helping the AI model variant MC styles and flows, expanding both AI and human MC improvisational range while fostering experimentation with new delivery techniques.

Audio method

The proposed framework decouples the musicality, defined by tonality and rhythm from the language modelling process, and aims to produce a responsive and realistic battle rap model. Here pitch, cadence and amplitude data are evaluated to create a dynamic wire frame model on which to attach structured language input, reflecting the idioms and structure of a defined rap sub-genre, or stylistic approach. By integrating detailed metrics including word length, inter-word silences, and dynamic amplitude variations, the approach produces a composite cadence score that can be fine-tuned for individual phrases or entire compositions. This methodology builds on established techniques in musical feature extraction and dynamic modelling (Tzanetakis and Cook 2002; Downie 2003) extends previous work in genre-specific analysis (Condit-Schultz 2016), and draws on foundational advances in language modelling (Manning and Schütze 1999) as well as content-based music information retrieval (Casey et al. 2008; Müller 2021). This decoupling of musical analysis from the more compute intensive language modelling pipeline creates a flexible and customisable framework on which to model and expand on stylistic paradigms.

The analysis takes into account word length, inter-spacing silences, word amplitude (with respect to initial and exit volume ramping), word pitch, and beat analysis. The resultant model produces a composite cadence score, which can be calculated with varying degrees of granularity for each lyrical phrase, or as an overall score.

Both raw musical data and cadence scores can be quantified to model individual and historical genre-specific stylistic paradigms. These modelled parameters provide structured input for the improvisational algorithms (Figure 2 (A)).

Latency and Real-Time Responsiveness In assessing the effectiveness of an AI-driven rap opponent, latency between the human MC's input and the AI's response is a critical factor impacting authenticity and engagement. Three computational strategies are proposed to address latency:

1. *Cached Responses:* The AI pre-generates and stores a selection of potential responses, requiring minimal real-time computation and enabling rapid (potentially sub-second) replies. This approach closely mirrors actual

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

practice in rap battles, where MCs often rely on memorized or pre-written material presented as spontaneous improvisation.

2. *Adaptive Rehearsal:* The AI generates initial responses beforehand but adjusts and refines them dynamically based on the MC's ongoing performance. This results in moderate latency due to additional real-time computational processing. This method achieves a practical balance between computational efficiency and improvisational authenticity.
3. *True Real-Time Generation:* Here, the AI synthesizes responses entirely in real-time, directly reacting to the MC's live input without pre-generated material. This method is computationally demanding, resulting in significantly higher latency but delivering the most authentic emulation of genuine freestyle improvisation.

Real-World vs. Computational Improvisation. Human rap battles typically feature rehearsed or semi-rehearsed content delivered with variations in emphasis and intensity, creating the illusion of spontaneous improvisation (Caffeine 2020). Therefore, the AI's utilization of prepared or adaptively adjusted content aligns closely with existing human practice (Flip Top Battles 2025).

In live rap battles, human MCs pause for 5–7 seconds between rounds, often delivering semi-rehearsed material with dynamic variation. The illusion of spontaneity is maintained through timing, emphasis, and adaptive performance (Flip Top Battles 2025). AI-driven systems that generate responses in advance or refine them adaptively are therefore not diverging from standard practice but rather emulating it closely. Key parallels include:

- Human MCs rarely freestyle entirely spontaneously; instead, they frequently utilize pre-written lyrics tailored to specific opponents or themes, performing these with nuanced variation.
- AI responses based on pre-cached or adaptively updated content reflect human norms in performance authenticity.

AI systems offer strategic advantages beyond human capability:

- They can evaluate multiple potential responses in parallel, selecting the most effective line in context.
- With reinforcement learning, these systems could improve over time—accumulating and refining a personal repertoire of adversarial tactics.

Evaluation Framework

To rigorously measure the AI opponent's effectiveness, a multi-dimensional evaluation framework is proposed, consisting of:

- *Subjective Human Assessment:* Involving self-reported evaluations from the human MC on perceived skill improvement, challenge level, and engagement, complemented by audience ratings on the AI's authenticity, creativity, and competitiveness.
- *Turing Test-Based Evaluation:* Blind evaluation by independent listeners attempting to differentiate between human and AI performances, measuring realism, stylistic consistency, and overall believability.
- *Quantitative Linguistic Metrics:* Objective analysis of linguistic and rhythmic features such as rhyme density, vocabulary diversity, rhythmic complexity, and lyrical originality to quantify changes and improvements in both human and AI performance.

Unlike conventional NLP tasks, assessing AI-generated battle rap requires both linguistic and performative evaluation. Relevant metrics include:

1. *Rhyme complexity:* measuring rhyme density and multisyllabic structures (Malmi et al. 2016).
2. *Coherence:* evaluating semantic consistency and contextual relevance
3. *Emotional impact:* For instance, detecting sentiment polarity and aggression markers in adversarial exchanges.
4. *Audience engagement:* incorporating crowd reaction prediction models.
5. *Style authenticity:* comparing AI-generated lyrics with artist-specific linguistic style (Li et al. 2024; Edwards and Kool G Rap 2009)

Developing robust battle rap evaluation metrics is critical for measuring AI's ability to mimic human-level creativity and improvisational skill. The model generates a composite cadence score by analysing the underlying rhythmic structure of the rap, which must be further evaluated for accuracy and consistency to ensure a good fit (Bigo et al. 2018; Condit-Schultz 2016). AI-generated cadence scores are compared against human-labelled benchmarks from expert MCs to establish correlation, while feature matching ensures the extracted elements, including word length, interspacing silences, amplitude, pitch, and beat alignment, align with ground truth annotations. An additional layer of verification can be elucidated by using spectral analysis comparison techniques to assess that the model accurately captures the nuances of cadence and flow (Stoica and Moses 2005).

To develop stylistic adaptability, the AI undergoes an iterative process of refinement, ensuring its output aligns with known MC styles. The wireframe model of the AI-generated rap is compared against existing MC datasets using cosine similarity on feature vectors, Dynamic Time Warping (DTW) for cadence evaluation, and Mel-Frequency Cepstral Coefficients (MFCCs) to analyse pitch and tonal properties. Transfer learning is introduced to further expand the stylistic scope, allowing the model to integrate new datasets and evolve with greater variability, generating unique and stylistically distinct improvisations. A/B testing can also be compared across different AI models, in order to identify the most effective configuration for rap improvisation.

Evaluating the quality of AI-generated phrases requires both perceptual and computational assessment. Here, evaluation methodologies can be employed by experienced MCs can provide qualitative feedback on flow, rhythmic accuracy, and stylistic authenticity. For example, a rhythmic

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Evaluation Dimension	Metrics and Methods
Subjective Human Assessment	Self-reported MC skill improvement, perceived challenge, and engagement; audience ratings of AI authenticity, creativity, and competitiveness.
Turing Test-Based Evaluation	Blind assessment by independent listeners evaluating realism, stylistic consistency, and believability of performances.
Quantitative Linguistic Metrics	Analysis of rhyme complexity (density, multi-syllabic structures), coherence (semantic consistency, contextual relevance), emotional impact (sentiment, aggression markers), and audience engagement prediction models.
Cadence and Flow Analysis	Composite cadence scoring using rhythmic structure features, spectral analysis, cosine similarity, Dynamic Time Warping (DTW), Mel-Frequency Cepstral Coefficients (MFCCs), and rhythmic alignment scoring.
Stylistic Authenticity and Adaptability	Feature vector comparison against MC datasets, transfer learning for stylistic diversity, expert MC qualitative feedback, and comparative A/B testing across AI configurations.
Textual Quality Assessment	Language fluency and coherence measured via BLEU scores, perplexity (PPL), and feature matching against human-labelled benchmarks.

Table 3: Multi-dimensional evaluation framework for AI-generated battle rap

alignment score can measure the fit of phrases synchronise over the underlying beat structure. Furthermore, language fluency and coherence can be assessed by comparing AI-generated text with real MC samples (BLEU scores). Additionally, Perplexity (PPL) can be employed to quantify predictability and naturalness of the generated lyric corpus, ensuring that the required framework and lyric quality is integrated into the musical framework (Papineni et al. 2002; Jelinek et al. 2005).

End-to-End System: Figure 3 illustrates the proposed AI-driven rap improvisation system, integrating voice cloning, cadence modelling, and rhyme structuring to generate adaptive lyrical responses. The human MC's speech serves as both input and feedback, guiding the AI's generative process. The AI voice clone synthesizes responses that match the MC's style, ensuring natural delivery.

The *current style module* evolves through improvisation (modifying style in real time) and historical reference (drawing from an *archive corpus*). A cadence algorithm refines rhythmic structure by analysing syllabic timing and rhythmic placement, while a rhyme algorithm evaluates rhyme density and complexity to shape lyrical flow. These components interact dynamically, with archived data reinforcing both cadence and rhyme structures to ensure stylistic coherence. A feedback loop enables continuous refinement, allowing the AI to learn from both its own output and the MC's voice, adjusting accordingly. By merging historical rap influences with real-time improvisation, the model ensures AI-generated lyrics achieve rhythmic precision, lyrical complexity, and stylistic authenticity in rap battles. Through repeated adaptation, the AI refines its freestyle capabilities, producing increasingly nuanced, context-aware responses that align more closely with human performance.

Flow Replication via TTS Models

A core system requirement is the ability to replicate human-style prosodic delivery, particularly the *flow* that exemplifies skilled MC performance. To evaluate whether current state-of-the-art text-to-speech (TTS) models meet this requirement, we conducted an experiment assessing their ability to reproduce rap-specific cadence and rhythm, using a human freestyle as reference. Flow was operationalised as the alignment between rhyme scheme and prosodic timing, with the central evaluation question framed as: does the AI-generated output *sound like* the original MC?

Two rap samples were used: (i) a human-performed freestyle baseline, and (ii) a GPT-generated text variation of the same verse. Each was synthesized using multiple TTS models and assessed on two dimensions: voice clone similarity and cadence fidelity. Evaluation combined subjective human ratings with computational metrics. This modular setup decoupled cadence replication from text generation, enabling focused analysis of a core adversarial capability in co-creative performance.Audio samples and data are available in the project repository.[2]

Human Listening Tests To capture perceptual qualities not easily quantified, human evaluators (N=8) rated each sample on voice clone similarity and flow preservation. Average ratings across all models fell between 1 (very poor) and 2 (poor), highlighting consistent failure to replicate natural cadence. These subjective scores provided a perceptual benchmark and revealed artefacts such as robotic delivery or unnatural phrasing. This experiment served as a focused probe into cadence replication in off-the-shelf TTS systems. While not a full system evaluation, it identified key barriers to adversarial responsiveness in generative rap.

Computational Analysis We developed a lightweight evaluation pipeline to quantify rhythmic similarity between original and generated audio. Mel-Frequency Cepstral Coefficients (MFCCs), a standard audio metric, captured spectral

[2] https://github.com/ArtsARKADE/ICCC25

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

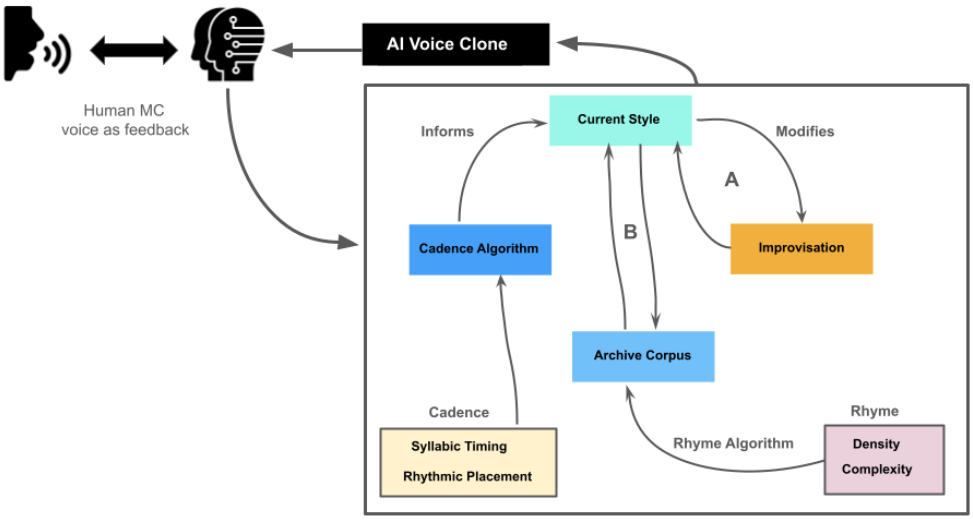


Figure 3: End-to-End Battle Rap System

features linked to perceived vocal quality. Dynamic Time Warping (DTW) assessed cadence alignment by measuring the temporal "stretch" needed to match the original performance. Higher DTW deltas indicated weaker temporal fidelity. Rhyme complexity was held constant using a GPT variation of the original text, with rhyme density calculations confirming phonemic similarity. A 3D bar chart summarises the MFCC and DTW metrics across TTS models (Figure 4).

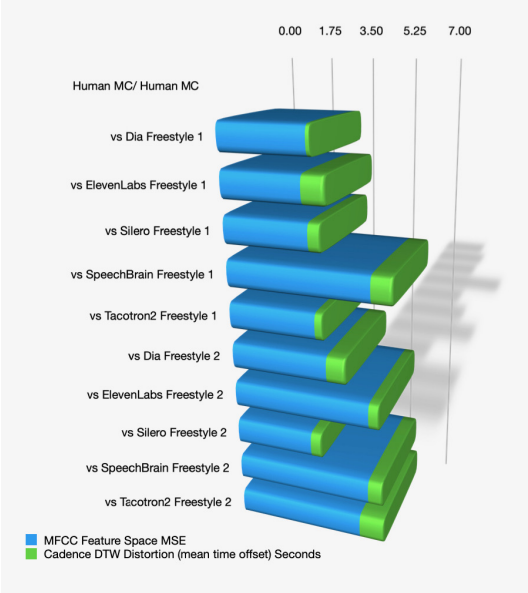


Figure 4: Human evaluator scores of TTS rap outputs across models and samples. Full data and code available in GitHub repository.

Rationale for Dual Evaluation For an AI to function as a creative adversary, provoking new human responses rather than imitating past styles, it must exert fine-grained control over cadence. Perceptual tests alone often miss subtle timing mismatches that limit this capacity. By combining human ratings with MFCC-based and cadence alignment metrics, we gained structural and perceptual insight. Together, these evaluations expose limitations in adversarial responsiveness and shape the design priorities discussed in the next section.

Discussion and Future Work

This paper explores AI's potential as both collaborator and competitor within rap performance. Several areas invite further investigation, including real-time responsiveness, the balance between rehearsed and improvised content, and the implications of AI surpassing human artistic abilities.

Discussion The proposed model presents a structured approach to generating AI-driven rap improvisation, leveraging a composite cadence score to quantify rhythmic structure and combining it with realistic lyrical content generated from large language models (LLMs). The model captures stylistic nuances across differing MC styles and is enhanced by iterative training and transfer learning, enabling the system to produce varied and authentic improvisations. However, several challenges remain in order to fully automate high-quality AI rap generation while remaining true to the genre's historical and expressive roots. The subjective nature of flow, authenticity, and lyrical coherence complicates evaluation, necessitating both computational metrics and human judgement. Real-time performance constraints are important for maintaining computational efficiency. At present, the model focuses on an offline approach to prioritise stylistic accuracy and improvisational relevance while

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

balancing creative autonomy and learned structure. While an AI can generate stylistically appropriate phrases, it remains constrained by its training corpus, raising questions about copyright and authenticity with regard to AI-generated output that closely resembles human artists' styles. Given that hip-hop is deeply rooted both as a lived experience and as an evolving cultural history, AI models must be trained responsibly and with respect to the art form, preserving its integrity and social context.

The battle rap improvisational model has broader applications in linguistic AI, speech synthesis, automated storytelling, and conversational AI. Its approach to analysing cadence, rhythm, and flow may also enhance natural language processing (NLP) models, leading to improvements in speech recognition, voice assistants, and real-time dialogue generation.

Optimising real-time processing is crucial for freestyle battles, requiring efficient feature extraction, phrase generation, and beat synchronization to minimize latency. Future work could explore low-latency deep learning models to reduce computational overhead while preserving performance. Leveraging additional MC datasets, transfer learning, and fine-tuning could improve AI responsiveness by integrating historical and contemporary rap influences. Such improvements would not only enhance stylistic fidelity, but also support the real-time demands of adversarial performance contexts. Future research may also refine cadence analysis techniques by incorporating prosodic modelling and sentiment-aware lyric generation, improving authenticity and coherence in AI-generated verses. Advancing AI-driven rap battles represents a step toward machines engaging in real-time, creative, and adversarial discourse—key components of human intelligence found in debate, negotiation, and artistic improvisation (Guilford 1967; Boden 1992). This research contributes to a broader effort in computational creativity, demonstrating how AI can learn to navigate spontaneity, stylistic nuance, and competitive interaction, with implications extending beyond music to conversational AI, linguistic modelling, and real-time creative collaboration.

AI models trained in rap improvisation could eventually serve as intelligent songwriting assistants, helping artists refine their flows, experiment with rhythmic structures, and generate new lyrical ideas. This opens the possibility of live AI performers dynamically integrated into concert environments. Extending this to virtual reality (VR) performances or gaming experiences could be a logical next step, where AI-generated rap dynamically responds to audience input or player actions. Our approach contributes to the broader field of computational creativity, showing how deep learning models can be fine-tuned to produce expressive, human-like artistic outputs and potentially innovate and evolve new stylistic paradigms. This raises important questions about whether AI can eventually surpass mimicry and develop its own distinct artistic identity (Du Sautoy 2020).

Implications of AI Surpassing Human Artistic Abilities AI-driven rap improvisation has wider implications across music, AI creativity, human-AI collaboration, and computational linguistics. Developing new improvisational tools for musicians and producers, such as virtual compositional assistants, could significantly enhance human-AI collaboration in creative industries. The potential scenario of an AI surpassing human MC skill levels raises significant philosophical and creative questions: (i) Might an AI that exceeds human creative performance stimulate increased artistic ambition by challenging human artists to evolve and innovate? (ii) Or could it demotivate artists by creating perceptions of redundancy, potentially diminishing the perceived value of human creative expression? These considerations underline the need for further investigation into the sociocultural implications of AI's evolving role in creative practices beyond rap battles, particularly in how AI-human interactions reshape artistic identity, motivation, and creativity itself.

Creative Adversarial Evaluation and Future Work The proposed system imagines AI not as a co-pilot, but as a sparring partner that pushes the boundaries of what it means to create. This leads to a model for AI evaluation that we refer to as a *Creative Adversarial Turing Test* (CATT). Unlike traditional Turing Tests (Turing 1950), which reward mimicry, a CATT assesses whether an AI can provoke a human to create at a higher level. Within the rap battle context, the AI is not a passive tool but a rival challenging the MC through feedback on rhyme, flow, and stylistic coherence. The provocation triggers a creative dialogue, prompting the human to iterate in response to the AI's performance. This model envisages AI evaluation as a test of *stimulation*, not *deception*. The outcome is not judged by whether the AI *sounds human*, but whether the *human sounds more inspired*. This adversarial framing reflects how creative tension, not comfort, often drives innovation (pp. 221–222, Du Sautoy 2020). CATT offers a test bed for exploring human adaptability under machine provocation. Beyond rap, it invites new forms of human-AI dialogue in fields such as education, design, and performance, where improvisation shapes meaning in real-time. The study re-frames the question from "Can AI create? to "Can AI compel *us* to create differently?"

This paper lays the groundwork by measuring flow preservation and adversarial response. These early results demonstrate what today's models lack, and why adversarial challenge might be a way drive progress in co-creative AI. This provides an initial step toward modelling the dynamic tension that underpins the Creative Adversarial Turing Test. Over time, we aim to develop structured CATT benchmarks that assess co-evolution: how well an AI can stretch, disrupt, or sharpen human creativity through tension. It challenges designers to think beyond assistive tools toward performative, even confrontational, systems. This direction raises new questions about authorship, creative ownership, and the emotional stakes of creativity when it emerges from competition rather than collaboration. Ultimately, the CATT framework imagines AI as an active agent in artistic development, provoking growth rather than imitating past success.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

Adams, K. 2009. On the metrical techniques of flow in rap music. *Music Theory Online* 15.

Akingbe, N., and Onanuga, P. A. 2018. Leveraging poetry on the airwaves: Appropriating linguistic creativity in nigerian hip hop lyrics. *Journal of the Musical Arts in Africa* 15:19–40.

Apple. 2024. Humanizing word error rate for asr transcript readability and accessibility.

Arnold, K.; Volzer, A.; and Madrid, N. 2021. Generative models can help writers without writing for them. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

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*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery.

Benvenga, L. 2022. Hip-hop, identity, and conflict: Practices and transformations of a metropolitan culture. *Frontiers in Sociology* 7.

Bigo, L.; Feisthauer, L.; Giraud, M.; and Levé, F. 2018. Relevance of musical features for cadence detection. In *International Society for Music Information Retrieval Conference (ISMIR 2018)*.

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Bradley, A. 2017. *Book of Rhymes: The Poetics of Hip Hop*. Basic Civitas.

Caffeine. 2020. A crash course on battle rap and the ultimate rap league.

Casey, M.; Veltkamp, R.; Goto, M.; Leman, M.; Rhodes, C.; and Slaney, M. 2008. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE* 96:668–696.

Chen, K.; Shao, A.; Burapacheep, J.; and Li, Y. 2024. Conversational ai and equity through assessing gpt-3's communication with diverse social groups on contentious topics. *Nature Scientific Reports* 14:1561.

Cohn, M.; Predeck, K.; Sarian, M.; and Zellou, G. 2021. Prosodic alignment toward emotionally expressive speech: Comparing human and alexa model talkers. *Speech Communication* 135:66–75.

Condit-Schultz, N. 2016. *MCFlow: A Digital Corpus of Rap Flow*. Ph.D. Dissertation, Ohio State University.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2024. Simple and Controllable Music Generation.

Coscarello, C. 2003. *The Word Out : A Stylistic Analysis of Rap Music*. Ph.D. Dissertation, Montclair State University.

Csikszentmihalyi, M. 1997. *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Perennial.

Dognin, P.; Padhi, I.; Melnyk, I.; and Das, P. 2021. Regen: Reinforcement learning for text and knowledge base generation using pretrained language models. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1084–1099. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2017. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment.

Downie, J. 2003. Music information retrieval. *Annual Review of Information Science and Technology* 37:295–340.

Du Sautoy, M. 2020. *The creativity code : how AI is learning to write, paint and think*. 4Th Estate.

Edwards, P., and Kool G Rap. 2009. *How to Rap : The Art and Science of the Hip-Hop MC*. Chicago Review Press.

Flip Top Battles. 2025. Flip top battle league.

Ghazvininejad, M.; Shi, X.; Priyadarshi, J.; and Knight, K. 2017. Hafez: an Interactive Poetry Generation System. 43–48.

Gonçalo Oliveira, H.; Mendes, T.; and Boavida, A. 2017. Co-poetryme: a co-creative interface for the composition of poetry. *Proceedings of the 10th International Conference on Natural Language Generation*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets.

Guilford, J. P. 1967. The nature of human intelligence.

Haidar, M. A.; Rezagholizadeh, M.; Omri, A. D.; and Rashid, A. 2019. Latent code and text-based generative adversarial networks for soft-text generation. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics.

Hirjee, H., and Brown, D. 2010. Using Automated Rhyme Detection to Characterize Rhyming Style in Rap Music. *Empirical Musicology Review* 5:121–145.

Ho, J., and Ermon, S. 2016. Generative Adversarial Imitation Learning.

Hollands, S.; Blackburn, D.; and Christensen, H. 2022. Evaluating the performance of state-of-the-art asr systems on non-native english using corpora with extensive language background variation. In *Proceedings of the Annual Conference of the International Speech Communication Association. Interspeech 2022, 18-22 Sep 2022, Incheon, Korea*, 3958–3962. Interspeech 2022.

Internet Archive. 2001. Mtv.com - news -madonna, elton, stevie wonder defend eminem.

Jelinek, F.; Mercer, R. L.; Bahl, L. R.; and Baker, J. K. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62(S1):S63–S63.

Landau, A. T., and Limb, C. J. 2017. The neuroscience of improvisation. *Music Educators Journal* 103(3):27–33.

ISBN: 978-989-54160-7-3

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. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Li, Y. A.; Jiang, X.; Darefsky, J.; Zhu, G.; and Mesgarani, N. 2024. Style-talker: Finetuning audio language model and style-based text-to-speech model for fast spoken dialogue generation.

Liu, S.; Chow, H. M.; Xu, Y.; Erkkinen, M. G.; Swett, K. E.; Eagle, M. W.; Rizik-Baer, D. A.; and Braun, A. R. 2012. Neural Correlates of Lyrical Improvisation: An fMRI Study of Freestyle Rap. *Scientific Reports* 2.

Lynch, J. 2018. Hip-hop passes rock to become most popular music genre for first time in history: Nielsen.

Malmi, E.; Takala, P.; Toivonen, H.; Raiko, T.; and Gionis, A. 2016. DopeLearning: A Computational Approach to Rap Lyrics Generation. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Mirowski, P. W.; Love, J.; Mathewson, K. W.; 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.

Music.AI. 2024. Comparative analysis shows music.ai outperforms openai in lyric transcription accuracy.

Müller, M. 2021. *Fundamentals of Music Processing: Using Python and Jupyter Notebooks*. Springer International Publishing.

OpenAI. 2024. ChatGPT Overview.

Orejuela, F. 2021. History of Rap Hip-Hop.

Ou, L.; Gu, X.; and Wang, Y. 2022. Transfer learning of wav2vec 2.0 for automatic lyric transcription.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pichai, S., and Hassabis, D. 2023. Introducing gemini: our largest and most capable ai model.

Prinsloo, G., and Coetzer, M. 1990. Automatic syllabification and phoneme class labelling with a phonologically based hidden markov model and adaptive acoustical features. *Computer Speech Language* 4(3):247–262.

Protalinski, E. 2017. Google's speech recognition technology now has a 4.9% word error rate.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Rogova, K.; Demuynck, K.; and Van Compernolle, D. 2013. Automatic syllabification using segmental conditional random fields. *Computational Linguistics in the Netherlands Journal* 3:34–48.

Serrano, S.; Torres, A.; and , I. T. 2015. *The Rap Yearbook : The Most Important Rap Song from Every Year Since 1979, Discussed, Debated, and Deconstructed*. Abrams Image.

Sharples, M., and Perez y Perez, R. 2022. *Story Machines : How Computers Have Become Creative Writers*. Routledge, first edition.

Sharples, M. 2015. John Clark's Latin Verse Machine: 19th Century Computational Creativity.

Stark, J.; Tang, A.; Park, J.; and Wigdor, D. 2023. Can ai support fiction writers without writing for them?

Stoica, P., and Moses, R. 2005. Spectral analysis of signals. *Prentice Hall* 1–447.

Sutton, R. S., and Barto, A. 2015. Reinforcement Learning : An Introduction.

Texas, A. 2015. Hip-hop is the most listened to genre in the world.

The Irish Times. 2003. Heaney lauds 'verbal energy' of Eminem.

Todd, B. 2018. Rapper and poet mark grist gives tips for rap battle global domination.

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59:433–460.

Tzanetakis, G., and Cook, P. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10(5):293–302.

Wassink, A. B.; Gansen, C.; and Bartholomew, I. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication* 140:50–70.

Wikipedia. 2019. Rapping.

Wikipedia. 2020. Mobb Deep.

Wu, H.; Yun, J.; Li, X.; Huang, H.; and Liu, C. 2023. Using a forced aligner for prosody research. *Humanities and Social Sciences Communications* 10.

Xue, L.; Song, K.; Wu, D.; Tan, X.; Zhang, N.; Qin, T.; Zhang, W.-Q.; and Liu, T.-Y. 2021. DeepRapper: Neural rap generation with rhyme and rhythm modeling. 69–81.

Yi, X.; Sun, M.; Li, R.; and Li, W. 2018. Automatic Poetry Generation with Mutual Reinforcement Learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. *Proceedings of the AAAI Conference on Artificial Intelligence* 31.

Zhang, W.; Sjoerds, Z.; and Hommel, B. 2020. Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. *NeuroImage* 210.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

When AI Says No: Investigating the Creative Power of Dissent

David Bouter
LIACS
Leiden University
d.r.bouter@umail.leidenuniv.nl

Rob Saunders
LIACS
Leiden University
r.saunders@liacs.leidenuniv.nl

Abstract

In recent years, co-creative AI has emerged as a novel approach to facilitate ideation tasks. Contemporary AI systems tend to exhibit conforming behavior. For example, well known chat bots such as ChatGPT will almost exclusively provide affirmative feedback on ideas, regardless of their quality. In human collaboration, having dissenting ideation partners often enhances creativity by fostering divergent thinking and preventing groupthink. This study examines whether the same applies to AI, investigating whether interaction with a dissenting co-creative AI leads to more creative story-writing outcomes and how it affects adoptability. Participants engaged in a co-creative story-writing task, after which expert evaluators assessed the creativity of the written stories using the consensual assessment technique. The likelihood of adoption was examined using the Unified Theory of Acceptance and Use of Technology (UTAUT). The results indicate that, on average, participants produced more creative output when collaborating with a dissenting AI than a conforming AI. However, the difference did not reach statistical significance, rendering the results inconclusive. The UTAUT analysis revealed a slight variation in adoption tendencies between the two AI conditions. These variations were minor, however, and did not provide significant evidence that adoption likelihood differed for the two conditions. These findings suggest that, in co-creative contexts, dissenting AI systems are equally adoptable as conforming AI systems. However, further research is necessary to determine whether dissenting behavior in AI meaningfully enhances creative performance.

Introduction

In the field of computational creativity, the goal is to design programs that can enhance human creativity without necessarily being creative themselves (Jordanous 2014). Presently it has become increasingly common to engage in co-creative ideation with AI systems (Cox, Djernæs, and van Berkel 2025). Especially when working with large language models, it becomes possible to have 'brainstorming sessions.'

A key issue with AI systems is that they contain biases (Schwartz et al. 2022). Some of these biases, such as groupthink and confirmation biases, can cause AIs to be overly conforming, which can hinder ideation processes when working with AIs. This paper poses the research question: "In the context of co-creative story-writing, does interaction with a dissenting co-creative AI lead to more creative outcomes compared to a conforming AI, and how does interacting with a dissenting AI impact its adoptability?"

This study hypothesizes that interaction with a dissenting co-creative AI in story-writing leads to more creative outcomes than interaction with a conforming AI. This hypothesis was developed by first examining human-human interactions (HHI), and then analyzing how these concepts translate to human-computer interaction (HCI). Existing literature indicates that in HHI, dissent stimulates creative ideation

Related Work

Divergent Thinking Versus Groupthink

Divergent thinking is widely recognized as a key component of creative ideation, often serving as a reliable indicator of novel ideas, though it does not equate to creativity directly (Runco 2010). A common strategy to attempt to reach creative ideas is brainstorming, however this form of group-based ideation has been shown to suppress divergent thoughts rather than enhance them, which is detrimental for the creative process (Runco 2010). Runco found a negative linear relationship between group size and idea quality, suggesting that individual ideation is often more effective. This reduction in quality is attributed to groupthink—the tendency to conform to group norms and avoid proposing unconventional ideas (Janis 1972). Janis characterizes groupthink as a barrier to optimal decision-making, particularly in cohesive groups under pressure or guided by authoritative leaders. To counter this, Nemeth (1995) proposes dissent as a mechanism for stimulating divergent thought. Her studies show that minority disagreement encourages group members to explore multiple perspectives, breaking the conformity that stifles creativity.

Social Behavior Towards Computer Systems

Co-creative AI systems typically display conforming behavior, often reinforcing rather than challenging user input (Schwartz et al. 2022). If users respond to excessive conformity in co-creative systems in the same socially driven ways as they do to other humans, we should expect a detrimental effect on idea quality (Runco 2010). Reeves and Nass (1996)

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

argue that humans have a tendency to project social behavior on computer systems. In their experimental work, they demonstrate that users often apply the same social norms to computers as to people. They found that participants rated a computer's performance more favorably when giving feedback on the same machine they had interacted with, compared to when evaluating it from a separate terminal, or on pen and paper. The effect was statistically significant, with users reporting the computer as more friendly and competent. These findings suggest that people apply at least some of the same social filters—such as politeness and conformity—in their interactions with computers.

As Reeves and Nass conclude, "When in doubt, treat it as human." If humans default to social behavior even when interacting with computer systems, co-creative AI may inherit the same group-level dynamics—such as groupthink—that hinder creativity in human-computer ideation.

Evaluation of Creativity and Adoption

A persistent challenge in creativity research is the ambiguity of the term creativity itself. Although most individuals possess an intuitive sense of what constitutes a creative product, this perception is inherently subjective (Amabile 1982). Such subjectivity complicates the collection of empirical data and necessitates structured evaluation methods. For this study, the Consensual Assessment Technique (CAT) (Amabile 1982) was adopted as a practical and widely accepted approach for assessing creative output.

CAT evaluates creativity through expert judgments within the relevant domain. It rests on two key assumptions: that domain experts are capable of recognizing creativity, and that creativity exists along a continuum—allowing for comparative assessments between products. Rather than enforcing strict criteria, the method leverages the consensus of informed evaluators to determine which products are relatively more creative.

Adoption and acceptance play an important role in the development of new information technologies (Venkatesh et al. 2003). The choice to measure the adoption likelihood of dissenting AI was made due to the antagonistic nature of a dissenting AI. If dissenting AI proves to be an effective tool for ideation, users' willingness to engage with the system decides its practical value. Should the system be perceived as excessively disagreeable or present other adoption-related barriers, its effectiveness may be rendered moot by low user uptake.

To measure adoption and acceptance, the unified theory of acceptance and use of technology (UTAUT), was used (Venkatesh et al. 2003). This model is a unification of eight competing acceptance models that describe user acceptance. The model uses determinants to make a statement about expected usage behavior. To measure whether dissenting AI and conforming AI are equally adoptable or not, Venkatesh et al. (2003) developed a standardized questionnaire that was used during this study.

The UTAUT model evaluates technology acceptance through eight determinants: *performance expectancy*, *effort expectancy*, *attitude toward using technology*, *social influence*, *self-efficacy*, *anxiety*, *behavioral intention to use the system*, and *facilitating conditions*. These constructs capture users' expectations, perceived ease of use, social and emotional influences, and enabling factors. Although UTAUT typically incorporates moderator variables such as age, gender, or experience to enhance predictive power, these were excluded due to the limited sample size and resulting lack of statistical power.

Methodology

To test the hypothesis as outlined in the introduction, an experiment was conducted with 22 participants who engaged in co-creative tasks. Specifically, participants completed a short story-writing exercise and received feedback from a large language model on their output. For this study, the LLaMA 3.1 model was chosen because it was practical to implement and, through initial testing, was found to be sufficiently capable of expressing both conforming and dissenting responses when appropriately prompted. While there is no formal benchmark for dissent or conformity expression in large language models, models like LLaMA 3.1 are trained on diverse texts and have demonstrated flexibility in generating outputs aligned with various stances. Prior work has shown that language models can take on different viewpoints when prompted appropriately (Bang et al. 2024).

The LLaMA 3.1 model is a pre-trained AI model and was used to simulate both conforming and dissenting feedback styles. The literature suggests that large language models naturally lean toward conformity (Schwartz et al. 2022), prompting dissenting behavior required careful prompt engineering. Prompts were designed to produce actionable feedback within 10 minutes, stay within 10 sentences, and conceal the AI's instructed attitude. The dissenting version was developed first using phrases like "devil's advocate" and then adapted into a conforming version by replacing attitude cues. Due to a lack of scientific standards for prompting dissenting AI, a trial-and-error approach was used to refine prompts while aiming to minimize bias, though some bias remains unavoidable.

The experiment was conducted in two phases. In the first phase, all participants were presented with the same short story—*Casabianca* by James Baldwin—with the ending deliberately omitted, and were instructed to generate three distinct story endings. These initial responses were subsequently evaluated by an AI system, which provided individualized feedback for each of the participant's story endings. In the second phase, participants were asked to produce three additional story endings for the same story. They were given the option to revise or build upon their earlier responses, incorporating the AI-generated feedback. Furthermore, they were instructed to apply the feedback in any way they liked, they were given the goal to: "be as creative as you can." Final outputs were evaluated by an expert panel based on their creativity.

The panel comprised individuals with professional or academic expertise in Literary Studies (MA), Linguistics (MA), and Performing Arts, including experience in performance, directing, and playwriting. While story writing was selected as the medium for this study, the underlying principles of the experimental design are applicable to other forms of creative

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

ideation (Runco 2010). At the end of the experiment, participants completed a survey with questions in accordance with the UTAUT model. All questions were answered on a scale from 'Strongly agree' (1) to 'Strongly disagree' (5).

Results

Three experts each produced a list with all stories written by the participants, ranked from most creative to least creative. Based on those lists, a 4th list was produced based on the average ranking of each story. A cropped representation of these lists are represented in Table 1. On average, each expert rated stories written with dissenting AI partners as more creative than those written with conforming AI partners. This can be seen in Table 1 where stories written with dissenting AI partners cluster at the higher rankings than those that were written with conforming AI partners. This observation is supported by the statistical test results in Table 2, which indicate that although only one expert graded the work from the dissenting group (group D) as significantly more creative than that of the conforming AI group (group C), there was general agreement among experts that group D's output was, on average, more creative. This trend is further reflected in the combined average rankings of all experts, which yielded a p-value of 0.111. While this value does not achieve statistical significance, it may suggest a potential relationship between interacting with dissenting AI and enhanced creative output.

After completing the grading process, the experts were asked to reflect on their evaluations through a questionnaire. The responses indicate that while technical quality and aesthetic appeal had some influence on their perception of creativity, this effect was not particularly strong. However, experts note that the variability in technical quality among the submissions made it challenging to fully disregard these factors in their assessments. On average, the experts characterized the overall range of creativity in the submissions as "somewhat varied." This degree of variation is beneficial, as it facilitates differentiation between submissions, ranking stories with highly similar levels of creativity would have been a more complex task (Amabile 1982).

To validate the CAT, an analysis was performed to grade the expert agreement. Figure 1 demonstrates the degree of agreement of all experts. Each grading of each story is compared to the average ratings of the experts combined. The degree of agreement is also expressed in Table 3 using the coefficient of concordance, W (Kendall 1945). A W of 0.741 is classified as *strong agreement* by the heuristic posed by Siegel and Castellan (1988). A caveat of these data is that each data point is compared to the average, in which the data point itself is contained, favoring the clustering of the data.

To provide context for this data, a correlation coefficient was produced for each pair of experts. These coefficients can be found in Table 3. According to the heuristic proposed by Schober, Boer, and Schwarte (2018), the level of agreement between expert 1 and 2 ($\tau = 0.580$) could be classified as *moderate agreement*. Both Expert 1 and Expert 2 appear to have comparatively lower agreement with Expert 3, a finding further supported by the Kendall's Tau coefficients in Table 3. The degree of agreement between

Table 1: Expert ranking

R	E1	E2	E3	Avg	R	E1	E2	E3	Avg
1	D	C	D	D	...				
2	D	C	D	D	56	C	C	D	C
3	C	D	D	D^1	57	D	D	D	C
4	C	C	D	C^1	58	C	D	D	D
5	D	D	C	D^1	59	C	C	D	D
6	D	D	D	D^2	61	D	C	D	C
7	D	D	C	C^2	62	C	C	C	C
8	C	D	C	C	63	C	C	C	C
9	D	C	C	D	64	C	D	C	C
10	C	C	D	D	65	C	D	D	D
...					66	C	C	D	D

Note. Gray cells represent a dissenting AI partner. White cells represent a conforming AI partner. Cells with a superscript represent a tie in a ranking with other cells with the same superscript.

Table 2: Mann-Whitney U test: Expert ratings

Expert	p-value
Expert 1	0.041
Expert 2	0.254
Expert 3	0.404
Average lower bound	0.103
Average upper bound	0.120
Average mean	0.111

Note. For all tests, the alternative hypothesis specifies that group C $<$ group D. N = 22.

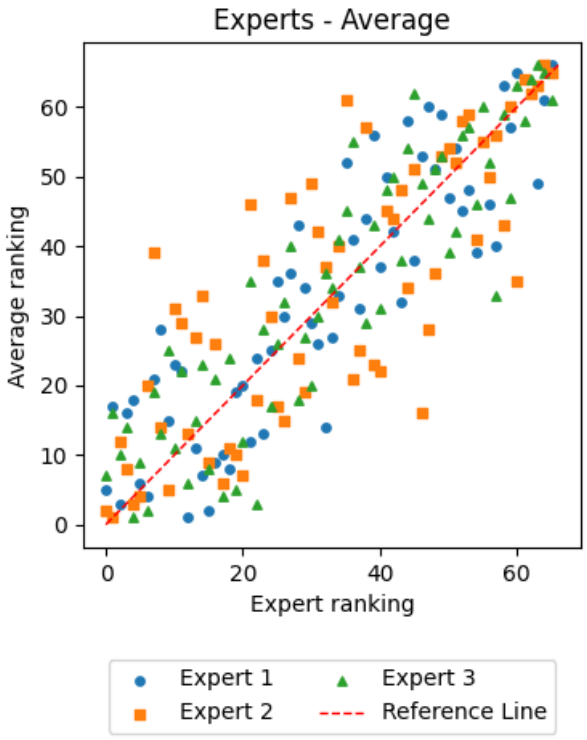


Figure 1: Rankings of each expert compared to the average rankings of all experts.

Experts 1 and 2 with Expert 3 ($\tau = 0.343$ and $\tau = 0.372$) could be categorized as *weak agreement*, following the same guidelines. This degree of agreement is slightly lower than expected, implying there might be some variation in how the experts ranked the stories. This is further discussed in the discussion section.

UTAUT

Table 4 presents the p-values of the Mann-Whitney U test, which examines whether responses differed significantly between participants in group D and group C ($\alpha \leq 0.050$). The results indicate significant differences for 'Effort Expectancy', while 'Anxiety' shows a notable trend (p = 0.079), albeit not reaching statistical significance. The mean responses for these determinants provide further insight. Participants in the dissenting AI group reported a mean response of 2.364 (SD = 0.848) for 'Effort Expectancy', compared to 1.900 (SD = 0.447) in the conforming AI group. This suggests that participants in the conforming AI group found the system somewhat easier to use, potentially indicating a higher willingness to adopt it.

'Anxiety' was rated higher in the dissenting AI group with a mean of 3.667 (SD = 1.155) than in the conforming AI group where the mean was 4.238 (SD = 1.221). This suggests that participants interacting with the dissenting AI reported somewhat more anxiety, which may negatively impact adoption in the dissenting AI group. For all other determinants, mean responses were similar across both groups, indicating minimal variation in perceived adoptability. Overall, this indicates that there is no significant difference in adoptability depending on the AI group a participant was in.

Table 3: Kendall's Tau Correlations: Expert agreement

	Kendall's Tau
Expert 1 - Expert 2	0.580
Expert 1 - Expert 3	0.343
Expert 2 - Expert 3	0.372
Kendall's Tau: $0 \leq \tau < 0.40$ indicates weak agreement, $0.40 \leq \tau < 0.70$ indicates moderate agreement	
	Kendall's W
Average	0.741
Kendall's W: $0.70 \leq W \leq 1.00$ indicates strong agreement	

Table 4: Mann-Whitney U test: UTAUT

Category	p
Facilitating conditions	0.930
Performance expectancy	0.794
Social influence	0.783
Self-efficacy	0.350
Behavioral intention to use the system	0.297
Attitude towards technology	0.258
Anxiety	0.079
Effort expectancy	0.050
Note. For the Mann-Whitney U test, the alternative hypothesis specifies that group C $\neq$ group D. N = 22	

Discussion

Dissenting AI exhibits the potential to influence co-creative ideation processes involving AI systems. Preliminary findings suggest a possible association between dissenting AI behavior and enhanced human creativity. Furthermore, concerns regarding user adoption do not appear to present significant barriers to the practical application of dissenting AI in ideation settings. While the observed effects did not reach statistical significance, the results offer promising indications that warrant further research. The lack of statistical significance could be due to several limitations in the study.

One notable limitation concerns the brevity of the AI interaction period. Multiple participants indicated that the interaction duration was insufficient for them to meaningfully engage with the dissenting or conforming AI. It is plausible

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

that extending the interaction—particularly through multi-turn dialogue—could yield more pronounced effects. Additionally, in the absence of established scientific standards for prompt engineering in this context, the design of AI prompts was based on iterative trial-and-error. This approach may have introduced researcher bias and, more critically, may have constrained the AI's capacity to effectively express dissent.

Several practical limitations may also have influenced the study's outcomes. These include a limited sample size, non-representative participant demographics consisting of mostly students and professionals with a background in creative disciplines, and the absence of double-blind experimental conditions. Finally, certain constraints were present in the application of the CAT. Due to resource limitations, the CAT could not be implemented to its full potential, which may partially account for the lower levels of inter-rater agreement observed.

Conclusion and future work

This study explored the following research question: In the context of co-creative story-writing, does interaction with a dissenting co-creative AI lead to more creative outcomes compared to a conforming AI, and how does interacting with a dissenting AI impact its adoptability?

The results indicate that differences in system adoption are primarily reflected in the constructs of anxiety and effort expectancy, both of which slightly favor the conforming AI system. However, these effects were not statistically strong enough to conclude a significant difference in overall adoptability between the two AI types. Notably, both AI systems were evaluated positively across most constructs of the UTAUT model, indicating a generally favorable disposition toward adoption.

Regarding creativity outcomes, no statistically significant evidence was found to support the hypothesis that interaction with dissenting AI produces more creative output than interaction with conforming AI. Nonetheless, the observed trend suggests that participants collaborating with the dissenting AI may generate more creative ideas. These findings highlight the need for future research. Future studies employing more robust experimental designs that would not be subjected to the aforementioned limitations, may validate this trend and provide empirical support for the integration of dissenting AI in co-creative ideation processes.

Author Contributions

DB conducted the research and wrote the manuscript. RS supervised the project and offered regular critical feedback.

Acknowledgements

This research was conducted as a Thesis Research Project in partial fulfillment of an MSc in Media Technology at Leiden University. We are grateful to the experts who generously dedicated their time grading the stories: Saskia Lindhoud, Robbin Sachs, and Gideon Roggeveen. We would like to extend our thanks to the second supervisor of this research project, Bram van Dijk. Additionally, we acknowledge James Baldwin as the original author of *Casabianca*, which served as a starter story for the participants to base their endings on.

References

Amabile, T. M. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology* 43(5):997.

Bang, Y.; Chen, D.; Lee, N.; and Fung, P. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11142–11159. Association for Computational Linguistics.

Cox, S. R.; Djernæs, H. B.; and van Berkel, N. 2025. Beyond productivity: Rethinking the impact of creativity support tools. In *Proceedings of the ACM Creativity and Cognition (C&C '25)*, 1–15. ACM.

Janis, I. 1972. Victims of groupthink; a psychological study of foreign-policy decisions and fiascoes. *Journal of American History* 60(3):857–858.

Jordanous, A. 2014. What is computational creativity?

Kendall, M. G. 1945. The treatment of ties in ranking problems. *Biometrika* 33(3):239–251.

Nemeth, C. J. 1995. Dissent as driving cognition, attitudes, and judgments. *Social Cognition* 13(3):273–291.

Reeves, B., and Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People 10(10)*. Cambridge University Press.

Runco, M. A. 2010. *Divergent Thinking, Creativity, and Ideation*. Cambridge University Press.

Schober, P.; Boer, C.; and Schwarte, L. A. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia* 126(5):1763–1768.

Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; and Hall, P. 2022. *Towards a standard for identifying and managing bias in artificial intelligence*, volume 3. US Department of Commerce, National Institute of Standards and Technology.

Siegel, S., and Castellan, N. J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

Venkatesh, V.; Morris, M. G.; Davis, G. B.; and Davis, F. D. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* 425–478.

MOSAAIC: Managing Optimization towards Shared Autonomy, Authority, and Initiative in Co-creation

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

Striking the appropriate balance between humans and co-creative AI is an open research question in computational creativity. Co-creativity, a form of hybrid intelligence where both humans and AI take action proactively, is a process that leads to shared creative artifacts and ideas. Achieving a balanced dynamic in co-creativity requires characterizing control and identifying strategies to distribute control between humans and AI. We define *control* as the power to determine, initiate, and direct the process of co-creation. Informed by a systematic literature review of 172 full-length papers, we introduce MOSAAIC (Managing Optimization towards Shared Autonomy, Authority, and Initiative in Co-creation), a novel framework for characterizing and balancing control in co-creation. MOSAAIC identifies three key dimensions of control: *autonomy*, *initiative*, and *authority*. We supplement our framework with control optimization strategies in co-creation. To demonstrate MOSAAIC's applicability, we analyze the distribution of control in six existing co-creative AI case studies and present the implications of using this framework.

Introduction

Human-AI co-creativity refers to the process in which both humans and AI contribute to the creative process to generate creative artifacts or ideas (Kantosalo et al. 2014), producing outcomes that could potentially surpass what either could achieve independently (Liapis, Yannakakis, and Togelius 2014). Unlike traditional creativity support tools designed merely to support human creativity, co-creative AI systems are increasingly being developed as creative partners (Feldman 2017; Margarido et al. 2024). With the rapid rise of Generative AI (GenAI) in co-creative contexts, such as ChatGPT, Runway, and Midjourney, research in co-creativity has gained significant momentum. However, designing effective co-creative AI presents several challenges due to the dynamic nature of human-AI interaction (Rezwana and Maher 2022; Maher 2012).

There is a growing need to explore the key dimensions of human-AI interaction that shape the distribution of control between users and AI (Liu, Huang, and Holopainen 2023). Recent advances in co-creative systems, particularly the development of "Media Foundation Models" for generating images and videos (Meta 2024), have further pushed the boundaries of AI's creative capabilities. These advancements along with the complexities of co-creation dynamics have shifted the balance of creative control toward AI (Epstein et al. 2023), raising concerns about maintaining balance in co-creation (Margarido et al. 2025).

Traditionally, AI research has focused on building autonomous agents (Stuart 2003). As a consequence, the role of machines in Human-AI interaction is moving away from the paradigm of 'AI as a tool' toward the direction of 'AI as an agent' (Guzman and Lewis 2020). In contrast, co-creativity research seeks to design AI systems that function as co-equal creative partners, hence the need for a balanced co-creative system (Kantosalo et al. 2020). Achieving a balanced dynamic in co-creativity requires both humans and AI to maintain creative autonomy while effectively sharing *control* (Muller, Candello, and Weisz 2023). This balance ensures users retain their sense of control in the process of co-creation (Heer 2019) and provides harmony to the co-creative process (Vinchon et al. 2023). However, control in co-creativity remains underexplored, as it lacks a clear and unified definition within creativity research. Therefore, it becomes essential to characterize the dimensions of control and develop strategies for control distribution between humans and AI to maintain a "stable equilibrium" in co-creativity (Yao et al. 2024).

In this paper, we aim to address the challenge of characterizing and balancing control in human-AI co-creativity. While much of the existing co-creativity research equates autonomy or agency with control, we adopt a more holistic perspective. We define ***control***, as the power to determine, initiate, and direct the process of co-creation. It helps us answer the question of "who is initiating, leading, deciding, navigating... the process of co-creation?" Our research is guided by the following questions, which aim to characterize control and develop strategies for distributing control:

1. What are the key dimensions of control in co-creativity?
2. How can control be balanced between humans and AI in co-creativity?

To answer these questions, we conducted a systematic literature review of 172 full-length papers. Informed by the

*These authors contributed equally to this work, listed in alphabetical order.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

literature review, we propose a framework, Managing Optimization towards Shared Autonomy, Authority, and Initiative in Co-creation (MOSAAIC)[1], that characterizes control by identifying the dimensions of control and strategies to balance control in co-creativity. Finally, we use MOSAAIC to analyze the control distribution in six co-creative AI, highlighting current trends and identifying gaps in achieving balanced control in co-creativity. We discuss the implications of the findings along with the value of MOSAAIC.

Background

Control dynamics between humans and AI is fundamental in co-creation (Margarido et al. 2025), but not sufficiently approached as a key concept in the Computational Creativity (CC) literature. Recent work in CC identifies (1) the principle of maximal unobtrusiveness and (2) balance of interjection as driving factors behind achieving balance in co-creation (Lawton et al. 2023; Moruzzi and Margarido 2024a). However, these works do not elaborate on control as a mechanism. Muller, Candello, and Weisz (2023) define control as the *tactical means of achieving a goal* where it is situated on a hierarchy between initiative (which party is currently acting) and agency (choosing and pursuing a strategic goal); however, this definition lacks a situated understanding of control. We build upon these works and explain how control is encompassed within existing frameworks and roles below.

Frameworks: In Search of Control

Various works have sought to address control in human-AI co-creation (Wu et al. 2021; Rezwana and Maher 2022; Kabir 2024; Moruzzi and Margarido 2024b). These works highlight creativity as the central aim of co-creative systems and wrestle with control dynamics in their plight for balance. They address power imbalances in AI systems, develop mechanisms for adjusting AI contributions, tackle decision-making power over task finalization, and account for societal implications of shifting agency in AI systems. However, while valuable, none of these works foreground control as a key concept. For instance, the Human-AI Co-creation model explains the creative process with AI and the new possibilities brought by AI in each creative phase, but it does not provide guidance on control in the context of co-creation (Wu et al. 2021). In our work, we focus on the concept of control and highlight the importance of a balanced control for Human-AI Co-Creativity through a systematic review of the literature. We decipher the patterns of control that emerge from works in related disciplines such as Human-Robot Interaction (HRI).

Roles: Defining Control

Since their advent, human-AI co-creative systems have progressed from tools to partners and agents (Manovich and Arielli 2024, Chapter 8), gradually leading to the design and adoption of AI roles and personas to meet users' needs (Moruzzi and Margarido 2024b). Different scholars have attempted to identify the roles AI agents can have, however, these works predominantly focus on autonomy (Lubart 2005). Kim et al. (2023) categorizes AI into four roles based on autonomy levels: Tools (low in both human and AI autonomy), Servants (high human involvement, low AI autonomy), Assistants (low human involvement, high AI autonomy), and Mediators (high in both dimensions). This has also been formalized in a taxonomy consisting of six levels of autonomy: no AI (level 0), tool (level 1), consultant (level 2), collaborator (level 3), expert (level 4), and agent (level 5) (Morris et al. 2024). Noting that autonomy is considered to be the key attribute in the formalization of AI roles, in this work, we foreground and define AI roles from the perspective of control as it carries a holistic understanding of not only how the process occurs (autonomy), but also how the power to determine, direct, and command are part of the process of co-creation as well.

Methodology for Systematic Literature Review

This section outlines the systematic literature review (SLR) procedure used to identify key dimensions of control (RQ1) and strategies to balance control (RQ2).

Identifying Keywords

We began our SLR by identifying keywords through preliminary research, leveraging the research team's expertise and insights from recent publications in The International Conference on Computational Creativity (ICCC), ACM Conference on Human Factors in Computing Systems (CHI), and ACM Conference on Creativity & Cognition (C&C). Through iterative discussions, we selected keywords that capture control dynamics in co-creativity, such as "control in human-AI co-creation." These keywords were iteratively refined by testing them in searching papers and evaluating the relevance of the resulting papers. This iterative approach helped us uncover nuances in keyword selection and optimize our search strategy. While our primary focus was on co-creativity, we discovered valuable tangential work in related fields, such as Human-Robot Interaction (HRI) and the broader AI domain. Consequently, we included papers from both co-creativity and related domains to provide a comprehensive perspective. Finally, collaborative discussions among the authors ensured consensus on the final set of keywords, aligning them with the study's objectives.

We used **five key phrases** based on our selected keywords to identify relevant papers. To structure our search effectively, we categorized keywords into two primary buckets (Figure 1) based on their conceptual role in understanding control dynamics: (Bucket 1) dimensions of control and strategies to balancing control and (Bucket 2) factors that influence or are influenced by control, which explore the broader implications and dependencies of control in co-creative systems. For **co-creativity research**, we formulated two keyphrases (Figure 3): (1) (Bucket 1 keywords separated by OR) AND ("Human-AI Co" OR "Mixed-Initiative Co") and (2) (Bucket 2 keywords separated by OR) AND ("Human-AI Co" OR "Mixed-Initiative Co"). To ensure

[1]We use this acronym to allude to a mosaic that is metaphorically used to describe a collection of diverse elements that form a larger whole.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

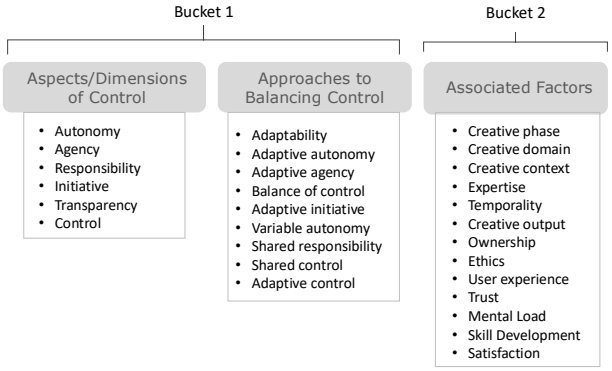


Figure 1: Identified key phrases for Bucket 1 and 2

comprehensive coverage, we used "Human-AI co" as an umbrella term to capture variations such as co-creation, co-creativity, and domain-specific terms like co-dancing. Additionally, since "Mixed-Initiative Co-Creativity" is commonly used in the literature to describe co-creation, we included "Mixed-Initiative Co" to account for variations.

To identify relevant papers in the **broader AI domain**, we incorporated three additional search key phrases (Figure 3): (3) "AI Autonomy" OR "AI Agency," (4) "Adaptive AI," and (5) "Adaptive Autonomy" OR "Variable Autonomy." These terms were selected to encompass diverse perspectives and frameworks related to control and autonomy in AI systems.

Data Collection

We conducted our literature review (Figure 3) using two primary databases: the Association for Computing Machinery Digital Library (ACM DL[2]) and the database listed on the Association for Computational Creativity website (ACC[3]). We chose the ACM DL to connect to conferences where HCI, AI, and creativity research are typically published. For searches in the ACM Digital Library, we leveraged its advanced search settings to refine our query.

We considered the Computational Creativity database (ACC) due to its strong relevance to our research focus. We considered documents published until November 2024. We narrowed our search to full-length papers (e.g., not tutorials or posters) to ensure we considered high-quality peer-reviewed studies. Because this database lacks advanced search functionality, we developed a custom Python script to search using our curated keywords.[4]

After running the keyword-based search, we retrieved 398 papers from the ACM Digital Library (ACM DL) and 62 papers from the Association for Computational Creativity (ACC), totaling 446 papers. Since multiple keyword searches led to duplicate entries, we removed redundant papers, reducing the dataset to 317 unique papers (269 from ACM and 48 from ACC). In the first screening round, the first two authors reviewed the abstracts to assess relevance of the papers to human-AI control dynamics and related aspects. To ensure consistency in the selection process, the authors first reviewed the abstracts (45) from one of the five keyphrase searches together to establish a consensus on inclusion criteria—the paper must directly or tangentially discuss the control dimensions or dynamics or related aspects. They then independently reviewed the abstracts from the remaining four keyword searches, dividing the workload equally. This process resulted in a refined set of 243 papers (215 from ACM and 28 from ACC). In the second round, they conducted a full-text review and further filtered out papers with low relevance, leading to the final corpus of 172 papers, comprising 146 from ACM and 26 from ACC.

[2]https://dl.acm.org
[3]https://computationalcreativity.net/
[4]https://github.com/jrezwana/MOSAAIC

Data Analysis

We read the final papers that were highly relevant to our research questions and summarized the literature review on the dimensions of *control* and strategies to balance it in co-creativity using a shared spreadsheet with notes, insights, summaries and scribbles and then organized them into related themes. Our framework was developed in an iterative process of adding, merging, and removing key dimensions of control in human-AI co-creation and strategies to balance control. Through iterative discussion, we developed our framework that presents the dimensions of control and the strategies of balancing control. The framework developed and described below is primarily based on our literature review.

Framework for Managing Optimization towards Shared Autonomy, Authority, and Initiative in Co-creation (MOSAAIC)

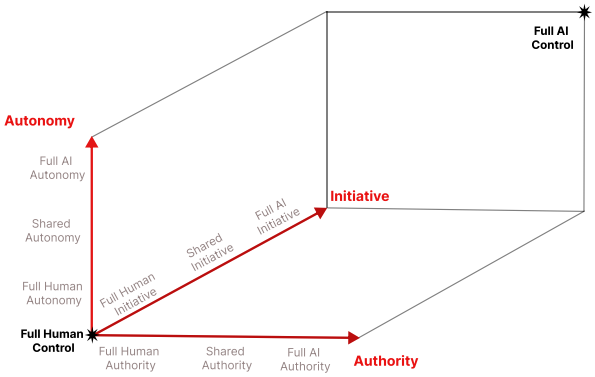


Figure 2: Three dimensions of control in co-creativity

In this section, we introduce MOSAAIC (Managing Optimization towards Shared Autonomy, Authority, and Initiative in Co-creation), a novel framework that we developed based on our literature review. MOSAAIC characterizes the key dimensions of control in human-AI co-creativity and outlines strategies for achieving a balanced dynamic between humans and AI (Figure 2). The framework defines three key dimensions of control: *Autonomy*, *Initiative*,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

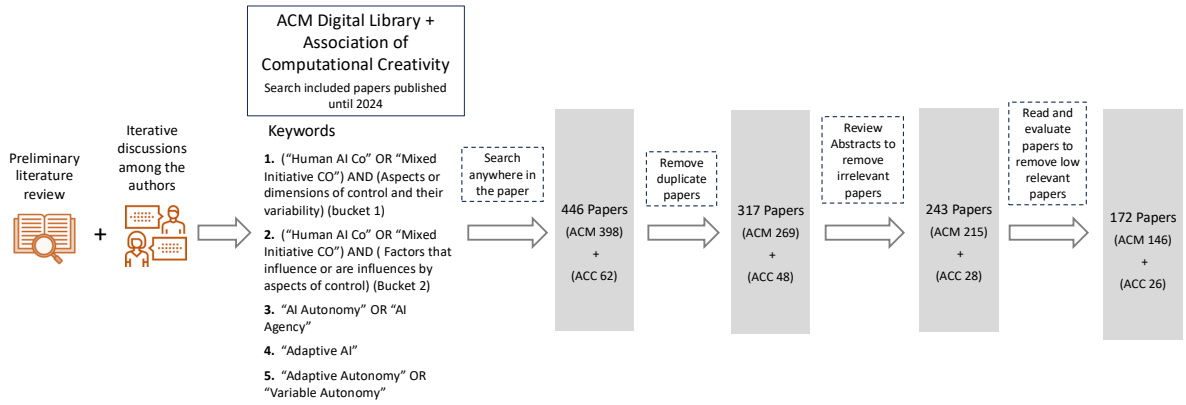


Figure 3: Literature review procedure

and *Authority*. *Autonomy* enables an agent to independently choose its creative actions, *initiative* allows it to proactively contribute rather than simply react, and *authority* grants it the power to make decisions and direct the creative process. As illustrated in Figure 2, the dimensions are represented along the axes of a three-dimensional (3D) control space. Each axis reflects a continuum of control, ranging from full human control at one end to full AI control at the other, with shared control in the middle. The corner where all three axes converge at the human-controlled end represents a scenario where humans retain complete control across all dimensions. Conversely, the opposite corner signifies full AI control.

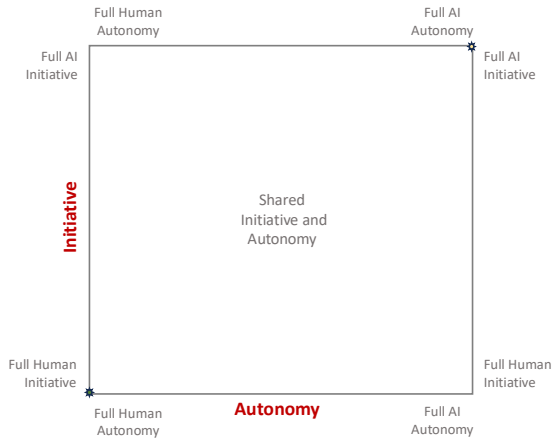


Figure 4: 2D plane of autonomy and initiative from the 3D model of MOSAAIC

To further break down the 3D model, we present a 2D plane representing a single side of the cube, isolating the relationship between Autonomy and Initiative (Figure 4). In this plane, the horizontal axis represents Autonomy between humans and AI, while the vertical axis represents Initiative. The four corners denote different combinations of high and low levels of human-AI autonomy and initiative. The top-right and bottom-left corners represent scenarios where either humans or AI fully control both autonomy and initiative, while the midpoints along each axis indicate shared autonomy or shared initiative. At the very center of the plane lies balance where both autonomy and initiative are evenly shared between humans and AI—although the right balance would depend on the task to be performed.

An example of the combination of full-human autonomy and full-AI initiative (top-left corner) could be an improvisational drawing system, where the AI proactively suggests ideas, but the human retains complete control to accept, modify, or ignore them, determining the final creative direction. Conversely, a case of full-human initiative and full-AI autonomy (bottom-right) would be a text-based storytelling co-creative AI, where the human provides a prompt or theme, but the AI fully controls how the story unfolds, with no opportunity for the user to adjust the content once generated. In the following subsections, we discuss the control dimensions and optimization strategies according to MOSAAIC. We begin by defining each dimension and then discussing its role in shaping control dynamics in human-AI co-creativity, drawing insights from the literature.

Three Dimensions of Control

Autonomy *Autonomy* refers to the ability of the agent to choose its creative action. The MOSAAIC framework identifies three levels of autonomy in co-creativity: *full human autonomy*, where AI acts as a passive assistant responding only to direct instructions; *full AI autonomy*, where the system independently generates creative outputs without human oversight; and *shared autonomy*, where autonomy is distributed between humans and AI, each having a similar degree of freedom to choose their creative actions.

Autonomy in Optimizing Control: Autonomy is a fundamental aspect of human-AI co-creativity and has been widely discussed in the literature (Davis et al. 2024). Research suggests that for co-creativity to be meaningful, AI must possess a certain level of creative autonomy alongside human creators (Muller, Candello, and Weisz 2023). Balancing autonomy between humans and AI is crucial for op-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

timizing control in co-creative systems (Moruzzi and Margarido 2024a), as these systems exist between two extremes: creativity support tools with low autonomy that primarily assist human creativity and standalone generative AI with full autonomy that does not rely on human input (Kantosalo et al. 2020).

Autonomy is closely intertwined with agency, and the two terms are often used interchangeably in HCI research (Bennett et al. 2023). In computational creativity, agency has been conceptualized as the sense of autonomy (Pease, Colton, and Banar 2023), while Koch, Ravikumar, and Calegario (2021) defined agency as the "amount of autonomy" an entity possesses. Given this strong interconnection, MOSAAIC classifies autonomy as one of the key dimensions of control, as agency is a higher level concept that can be defined by autonomy.

The level of autonomy attributed to an AI system is positively correlated with the perceived creativity of its outputs, reinforcing autonomy's role in human-AI co-creativity (Moruzzi 2022). Thus, categorizing and applying autonomy effectively in co-creative systems is essential. Davis et al. (2024) propose three levels of AI autonomy—fully autonomous AI, semi-autonomous AI, and user-directed AI—as part of their five-pillar framework for co-creative AI. However, their definition of autonomy combines aspects of both autonomy and initiative, which we distinguish as separate dimensions of control. The classification of autonomy in MOSAAIC, therefore, provides a more nuanced perspective of the degree of autonomy in control distribution.

Co-creativity research emphasizes the importance of shared autonomy, which sits in the middle of the autonomy spectrum in our 3D model, where both humans and AI contribute meaningfully to the creative process (Moruzzi and Margarido 2024b). Shared autonomy has been shown to enhance team performance and user satisfaction (Salikutluk et al. 2024). Effective co-creative partnerships require a balance of AI autonomy with user agency, ensuring that AI enhances—rather than diminishes—human agency and creativity (Moruzzi and Margarido 2024a).

Initiative *Initiative* refers to the ability to proactively contribute to the creative process. According to MOSAAIC, *full human initiative* means the AI waits for explicit user input before contributing. *Full AI initiative* allows the system to contribute on its own without human prompting. *Shared initiative* represents an adaptive collaboration where both parties can initiate and respond dynamically, fostering an interactive and evolving creative exchange.

Initiative in Optimizing Control: Initiative has been foundational in discerning the human-AI dynamics in co-creativity. Research on the theory of co-creation highlights that for co-creativity to be mutual, collaborative, and shared, mixed and shifting levels of initiative are essential to the user's sense of control (Wan et al. 2024). Mixed-Initiative Co-Creativity (MICC), where both human and AI systems can initiate contributions and actions, attributes initiative as the fundamental component of achieving co-creativity (Yannakakis, Liapis, and Alexopoulos 2014).

In HRI, Mixed-Initiative Control (MIC), where the human and system switch between different levels of initiative by either the human or the system, is a type of shared initiative proposed to optimize control (Chiou, Hawes, and Stolkin 2021). It finds that Human-Initiative (HI) systems outperform systems with single modes of operation (one specific action at a time), whereas Mixed-initiative (MI) systems provide improved performance and improved operator workload (Chiou, Hawes, and Stolkin 2021). MIC also results in better overall performance than giving an agent exclusive control. Specifically, effective MIC arises when agents are capable of making progress toward a goal without having to wait for human input in most circumstances, and when the control interface helps the user manage the progress and intent of several agents (Hardin and Goodrich 2009).

Other initiative mechanisms look at turn-taking, where the human and AI take turns at initiating the shared product, as a means of achieving balance (Rezwana and Maher 2022). They also take the user's initial prompt as a seed and encourage changes, making initiative an iterative process (Wang et al. 2024b). This entails moving from *full human initiative* to *full AI initiative* in our framework. Lobo et al. (2024) propose AI partners of varying initiative levels to present contextual awareness in the process of co-creation. These levels of initiative are *Leader, Follower,* and *Shifting*. AI partners following human initiative are perceived as warmer and more collaborative, whereas AI leaders are less friendly, although much quicker in execution. However, these outcomes depend on the context (task type) and individual preference of the user, with the potential for personalized MICC systems (Lin et al. 2023a). Initiative does not exist in a vacuum and operationalizing it to achieve control entails examining other dimensions of control in MOSAAIC.

Authority *Authority* in human-AI co-creativity refers to the extent of decision-making or directing power that shapes the creative process. It relates to the tactical means of achieving a goal.According to MOSAAIC, in *full human authority*, AI acts solely as a support tool, while humans retain ultimate authority to direct and decide the tactical means to achieve goals. In contrast, *full AI authority* occurs when AI autonomously determines creative directions without human intervention. *Shared authority* enables negotiation and mutual adaptation, where both human and AI contributions are considered in decision refinement.

Authority in Optimizing Control: In the literature, the term *control* has often been used to describe what we define as authority within MOSAAIC. In the literature, humans are generally expected to lead co-creative interactions due to their contextual understanding, ethical judgment, and accountability, particularly in complex creative tasks (Davis et al. 2016). In certain contexts, AI authority transfers from humans to AI can optimize efficiency. For example, research on medical imaging suggests that removing humans from the loop can improve accuracy and ethical outcomes (Muyskens et al. 2024). Similarly, some argue that AI should be granted epistemic authority in specific domains where its accuracy surpasses human expertise (Ferrario, Facchini, and Termine 2024).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

A balanced distribution of authority supports human creativity but also preserves user agency and engagement (Muller, Candello, and Weisz 2023). Authority allocation should be adaptive to user preferences and expertise level (Tsamados, Floridi, and Taddeo 2024). For instance, experts often prefer to lead the creative process while relying on AI as an assistant (Biermann, Ma, and Yoon 2022). In contrast, novices benefit from AI taking a more directive role, helping structure their creative efforts (Dhillon et al. 2024). Given these differences, users value customizable authority settings that align with their AI literacy and creative workflow, ensuring optimal collaboration (Moruzzi and Margarido 2024b). To maintain effective co-creativity, authority should shift dynamically based on task complexity, preventing AI from either becoming an overbearing decision-maker or an ineffective tool (Tsamados, Floridi, and Taddeo 2024).

The distribution of authority significantly influences the creative process and outcomes. Research suggests that authority shifts dynamically across different phases of co-creation and contexts (Wan et al. 2024). In game design, increased AI authority can lead to user frustration when it overrides human intentions, demonstrating the importance of context-sensitive authority allocation (Larsson, Font, and Alvarez 2022). To prevent negative user experiences, co-creative AI should enable fluid authority transitions, allowing both human and AI inputs to meaningfully shape the creative output (Lobo et al. 2024).

The shift from supervisory control to human-machine teaming acknowledges the evolving role of AI in co-creativity, suggesting co-creative AI to be a co-equal partner (Rezwana and Maher 2022). Authority allocation determines whether AI functions as an assistant, co-creator, or independent agent, with increased AI reliability often leading to greater decision-making power for AI (Walsh et al. 2019). Shared authority, where humans and AI collaboratively regulate the creative process, are emerging as a promising paradigm (Cimolino and Graham 2022).

Strategies for Balancing Control

As a part of MOSAAIC, in this section, we discuss the strategies to balance the control between humans and AI in co-creativity (Figure 5): *AI-controlled adaptation* and *human-controlled configuration*. The adoption of these strategies depends on the context, task type, user preference, and expertise (Hardin and Goodrich 2009). We discuss these strategies in detail below.

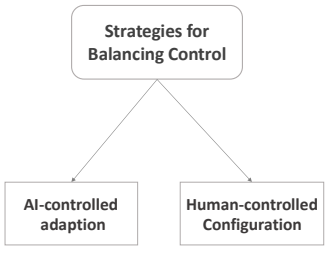


Figure 5: Strategies for balancing control in co-creativity

AI-Controlled Adaption AI-controlled adaptation to achieve dynamic control involves AI adjusting its autonomy, initiative, and authority based on the task and context. Based on the context, the system can adapt the dimensions of control, which has been explored in the domain of adaptive AI or variable autonomy. This is most notably exemplified by *adaptive autonomous* systems that change their autonomy levels based on tasks executed by users (Frasheri, Çürüklü, and Ekström 2017). For instance, in team performance, the AI agent adjusting its autonomy based on the situation and delegating tasks and decision-making to the users improves human-AI team performance (Salikutluk et al. 2024).

Another strategy of AI-controlled adaptation is AI systems monitoring the human's cognitive workload, emotional state, or expertise level to adjust their level of control (Singh and Heard 2022). The system might take on more control during routine tasks but cede control to the human during critical or complex situations (Salikutluk et al. 2024). For instance, if the human is experiencing high cognitive load, the AI might take on more responsibility to reduce the burden (Singh and Heard 2022). Mixed-initiative systems, where both humans and AI can initiate changes, can also improve performance in navigation tasks and reduce operator workload (Lin et al. 2023b). In this case, the AI is adapting its level of initiative based on the workload in other states of the co-creative process. This also shows the intertwined behavior of all three dimensions of control, where balancing control via AI adaptation involves AI systems adjusting their levels of autonomy, initiative, and authority according to the context so that they may engage with users in optimized control.

Human-Controlled Configuration Human-controlled configuration involves allowing users to customize and configure the control dynamics according to their preferences and goals. Preference varies significantly between expertise groups, suggesting the development of personalized co-creative systems (Lin et al. 2023a). Customization, based on the user's background, such as AI literacy, experience levels, and familiarity, is an important factor to consider in distributing control for co-creative systems (Lin et al. 2023a). Users' perceptions of an AI system are influenced by their creative goals and the system's ability to adapt to their expertise level (Lin et al. 2023a).

Humans can configure between different levels of AI involvement during the creative process to maximize system flexibility and enhance creative outcome generation (Dang et al. 2023). This includes human configuration based on the creative process—the divergence stage of idea generation, and the convergence stage of evaluation and selection of ideas (Shaer et al. 2024). For instance, in the divergence stage, experts often tend to prefer maintaining their autonomy when conceptualizing ideas (Wang et al. 2024a), and in the convergence stage, novices prefer 'in-action' feedback (feedback throughout a design task) to evaluate their creative session (E et al. 2024).

Another strategy of human configuration is utilizing steerable interfaces (Louie, Engel, and Huang 2022) such as AI-steering tools (adjustable autonomy), example-based and se-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

mantic sliders (range of outcomes), and multiple alternatives of generated content to choose from (authority over outcomes) (Louie et al. 2020). An important, yet overlooked, aspect of control is giving users multiple options to choose from and giving them a *sense of control* (Leotti, Iyengar, and Ochsner 2010). Overall, balancing control via human-configuration involves allowing users to customize and configure their preferences and goals leading to a distribution of autonomy, initiative, and authority.

Case Studies

We selected six co-creative systems representing different creative domains to analyze the distribution of control between humans and AI. Our goal was to validate MOSAAIC's applicability across various co-creative domains while also examining and comparing trends in control distribution across domains. Convenience sampling was used to select systems from the literature seeded by systems the authors were familiar with. Our sampling included the following domain-relevant co-creative systems: Cyborg (Branch et al. 2024), LuminAI (Trajkova et al. 2024), Snake Story (Yang et al. 2024), Reframer (Lawton et al. 2023), and Shimon (Gao et al. 2024). We also included ChatGPT (OpenAI 2022) to include one example of a general-purpose AI system that is used for co-creativity.

Two researchers among the authors independently coded the six co-creative systems based on the system descriptions in their respective publications. They then met to identify and resolve discrepancies. Finally, disagreements were discussed and revised to reflect group consensus (see Table 1).

System & Domain	Autonomy	Initiative	Authority	Strategy
Cyborg (Theatre)	Shared	Full-human	Full-human	Human-controlled
LuminAI (Dance)	Shared	Shared	Shared	Both
Snake Story (Game-Design)	Shared	Shared	Full-human	Both
ChatGPT (Writing)	Shared	Full-human	Full-human	Both
Reframer (Drawing)	Full-human	Full-human	Full-human	Human-Controlled
Shimon (Music)	Shared	Shared	Shared	Both

Table 1: Analysis of the control distribution of six co-creative systems using MOSAAIC

Autonomy Distribution

Systems were classified as having shared autonomy if both humans and AI had comparable abilities to make creative choices. If a system primarily followed user instructions with little creative autonomy, it was coded as full-human. Based on our analysis, five out of six systems exhibited shared autonomy, except for *Reframer*, which generates images following the user prompts and instructions. While *Reframer* allows users to control the level of variation in AI output, it does not have the ability to deviate from the prompt, making it more user-directed than co-creative. The remaining systems allow both humans and AI to make creative choices, leading to more emergent co-creation. For example, *LuminAI*, an improvisational dance partner, generates dance movements independently but adapts to the human dancer's motion, demonstrating shared autonomy. Similarly, *ChatGPT* actively interprets user-provided prompts and contributes with its chosen creative strategies. *Snake Story* is a co-creative game design system where the AI does not merely follow the user's instructions but actively contributes by suggesting plot developments, steering the story in unexpected directions. *Shimon* is an improvisational robotic marimba player that collaborates with human musicians by generating, adapting, and responding to musical phrases of the human partner in real time. *Cyborg* is a system where AI generates text dialogues being enacted by a cyborg performer, but the generated dialogues are dependent on other actors' dialogues.

Initiative Distribution

Systems were classified as having shared initiative if both humans and AI had the ability to proactively contribute to the creative process. If a system primarily followed user input and instructions with a reactive response mode, it was coded as full-human. In our analysis of six systems, the dimension of initiative was equally split between shared initiative and full-human. We primarily identified this difference via proactivity and reactivity. We observed that co-creative systems such as *Cyborg*, *ChatGPT*, and *Reframer* would only contribute to the creative process when the user instructs them and hence are 'reactive' in the creative process. For instance, in *Cyborg*, an AI-based improv theatre show, the LLM was prompted by speech recognition of the actors to generate the lines for the cyborg actor. This reactive approach distinguishes these systems as full-human initiative. On the other hand, systems with shared initiative involve both humans and AI proactively contributing to co-creation and do not wait for the human to begin the process. *LuminAI* proactively starts to dance if the user does not start dancing within a threshold period during the co-creation; *Shimon*, an improvisational robotic marimba player, takes initiative by autonomously generating musical phrases, setting the tempo, or leading a performance without requiring human input first; and *Snake Story* collaboratively crafts a story between the user and the AI by interchanging the order contribution.

Authority Distribution

Systems were classified as having shared authority if both humans and AI had similar abilities to make decisions and direct the co-creative process. This entails negotiation and mutual adaptation. Conversely, if a system primarily implemented the user's decision with little authority, it was coded as full-human. Based on our analysis, four out of six of the co-creative systems had full-human authority. In *Cyborg*, a human curator reviews and selects AI-generated dialogues before they are performed by the cyborg. In *Reframer*, users

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

control the final output by accepting, rejecting, or modifying AI contributions using sliders and prompts. In *Snake Story*, the AI lacks control over the narrative direction, as the human fully dictates the story's progression and outcome. In *ChatGPT*, the user determines the final goal, directs the flow of narration, and decides whether to include or exclude AI-generated content. What is particularly interesting about systems with shared authority, such as *LuminAI* and *Shimon*, is that both are improvisational AI without a fixed goal or tangible shared artifact. In these systems, the creative process itself is the product, meaning there is no final outcome over which authority must be exerted. *LuminAI* exhibits shared authority as both the human and AI shape the dance interaction, with neither having full control over the creative direction. Similarly, *Shimon* maintains shared authority by dynamically responding to human musicians.

Strategies to Balance Control

We identify strategies for modulating control between humans and AI in these six systems using MOSAAIC. Two out of the six systems have human-controlled configuration as the strategy to balance control. These are *Cyborg* and *Reframer*. Whereas four out of six, a majority of systems, have both AI-controlled adaptation and human-controlled configuration (classified as "both") as the strategies to balance control. These include *LuminAI*, *ChatGPT*, *Shimon*, and *Snake Story*. We will first describe the strategies for human-controlled configuration and then those for both.

In *Cyborg*, a human curator controls and refines AI-generated dialogues before they are enacted by the cyborg. The interface also allows the curator to guide the AI's output by providing prompts such as "be funnier," enabling them to adjust the AI's initiative and authority in the creative process. As an improv AI with the lack of a fixed goal, human-controlled configuration allows a well-balanced control dynamic that addresses the spontaneity that is embedded into the creative domain. On the other hand, in *Reframer*, a drawing interface where there is a fixed goal, a human-controlled configuration would be implemented to provide variability on the control mechanisms for the creative autonomy of the AI generation. Likewise, initiative would also be dispersed by the human prompting the AI to initiative and authority to be distributed by negotiated decision-making. In this case, the user would be able to distribute autonomy, initiative, and authority in its drawing process upon directing its preference and goals.

In shared strategies of control, the collective strategy of AI adaptation and human configuration would allow the system to adjust its autonomy, initiative, and authority based on context, tackling the user's nascent needs, while the human explicitly configures the system to their preferences and goals. In *LuminAI* and *Shimon*, this would allow all shared dimensions to evolve in unison, perhaps adding to different levels of shared control within the dimensions. *LuminAI* adapts in real-time to a human dancer's movements, while humans can configure movement parameters or styles to influence how the AI responds in the creative process. *Shimon* modifies its musical phrases in response to human input, adjusting tempo and rhythm in real-time, while musicians retain control over aspects such as initiating musical themes or shaping compositional elements.

Likewise, in *ChatGPT* and *Snake Story*, shared strategies would also balance control via the system's adaptation to the user's states (e.g., type of engagement), while the user equally adds their desired goals. *ChatGPT* dynamically adjusts its responses based on the conversation history, while users fine-tune its output through iterative prompts and stylistic preferences. *Snake Story* adapts autonomy by adjusting the level of AI-generated suggestions and initiative by varying its proactivity and reactivity based on the user's interaction frequency.

Discussion and Conclusion

The rapid rise and accessibility of GenAI technologies in creative domains are shifting agency from users to AI, transforming its role into agentic AI and reducing users' control over the creative process (Lin et al. 2023a). Co-creativity literature emphasizes that while AI should have creative autonomy, users must also retain the ability to lead co-creation (Muller, Candello, and Weisz 2023). Achieving this balance requires a control equilibrium between humans and AI, ensuring a harmonious co-creation.

In this paper, we introduce a framework for **managing optimization towards shared autonomy, authority, and initiative in co-creation (MOSAAIC)** that characterizes the key dimensions of control and strategies to balance control in human-AI co-creativity. Our framework emerges from a systematic literature review of 172 papers. The key dimensions of control identified in MOSAAIC are *autonomy* (enabling the choice of creative action), *initiative* (allowing the proactive contribution in co-creation), and *authority* (granting the ability to decide and direct the creative process). While MOSAAIC's dimensions may be interconnected, each plays a distinct and vital role in defining control dynamics in human-AI co-creativity. MOSAAIC identifies two strategies to balance control in co-creation, which are *AI-controlled adaptation* and *human-controlled configuration*. We also analyze six co-creative systems in different domains to show use cases of MOSAAIC and to provide how MOSAAIC attributes control distribution in co-creative AI systems of various domains.

Use Cases of MOSAAIC

By leveraging MOSAAIC's strategies for optimizing control, researchers can identify ways to effectively balance control in human and AI co-creation. Below we discuss some use uses of MOSAAIC.

MOSAAIC can serve as a **tool for analyzing control distribution** in co-creative AI systems and can identify trends, strengths, and gaps in how control is allocated and optimized across co-creative AI systems. For instance, improvisational AI systems like LuminAI and Shimon in our case study, where the co-creative experience itself is the product rather than a tangible outcome, exhibit different control dynamics from more structured systems with varied combinations of the three dimensions. In these systems, all control dimensions were shared, unlike in other systems. Moreover,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

MOSAAIC also enables cross-domain comparisons beyond individual system analysis and serves as a benchmarking tool for evaluating co-creative AI across different creative fields. Evaluation tools and frameworks have been widely utilized in co-creativity domains (Kantosalo et al. 2020; Rezwana and Maher 2022).

Second, MOSAAIC can be translated to a **configuration and customization tool**, enabling AI systems to adapt based on how users adjust control dimensions in co-creation. Research indicates that different user groups, such as experts and novices, have varying expectations from co-creative AI (Oh et al. 2018; Moruzzi and Margarido 2024b), and a configurable control system can accommodate these differences by providing greater adaptability. By allowing users to configure the dynamics of human-AI collaboration, MOSAAIC supports personalized experiences aligned with individual preferences or specific creative domain requirements, thus democratizing control in co-creative systems. This customization could be implemented in various ways, such as sliders for each control dimension, allowing users to shape the dynamics of human-AI control. By supporting such customization, MOSAAIC empowers users to shape their co-creative interactions, enhancing collaboration and creative synergy with the co-creative system. Since MOSAAIC is domain-agnostic, it can be applied across various creative domains. In our case study, we found that four out of the six systems incorporate both AI adaptation and human configuration for control optimization, highlighting a trend toward integrating AI adaptation alongside user customization.

Third, MOSAAIC can serve as a **tool for designing co-creative AI** with optimized control distribution by providing a structured characterization of control and strategies to balance control between humans and AI. Design frameworks are well-established in co-creativity research (Rezwana and Maher 2022; Kantosalo et al. 2020), aiding in the development of more effective systems. Our case study reveals opportunities to improve the design of co-creative AI utilizing balanced control. For example, three systems had shared initiative, while the rest were fully human-led, highlighting the need to explore initiative distribution—especially since most recent GenAI models remain reactive. In terms of authority, four out of six had full-human authority, indicating a scope to design systems with more dynamic authority.

Limitations and Future Work

The goal of MOSAAIC is to provide *balance of control* and *harmony* (Vinchon et al. 2023) in the process of co-creation by addressing (1) the principle of maximal unobtrusiveness and (2) balance of interjection, as motivated in CC research. MOSAAIC currently identifies two high-level strategies for balancing control in co-creativity. However, how to decide exactly how to optimize a co-creative system in terms of control distribution is still an open question, except that we know that the adoption of strategies depends on the context, task type, user preference, and expertise (Hardin and Goodrich 2009). As such, in future work, we plan to identify context-appropriate strategies to optimize in different contexts and domains. Furthermore, although our literature review was comprehensive of ACC and ACM publications, we recognize that other relevant literature may have been omitted, and aim to expand our review to other libraries. We would expand our study towards identifying the factors that influence control, such as trust and ownership, to incorporate external factors that affect the balance of control in co-creation.

References

Bennett, D.; Metatla, O.; Roudaut, A.; and Mekler, E. D. 2023. How does hci understand human agency and autonomy? In *Proceedings of the 2023 CHI*, 1–18.

Biermann, O. C.; Ma, N. F.; and Yoon, D. 2022. From tool to companion: Storywriters want ai writers to respect their personal values and writing strategies. In *Designing Interactive Systems Conference*, 1209–1227.

Branch, B.; Mirowski, P.; Mathewson, K.; and Covaci, S. P. A. 2024. Designing and Evaluating Dialogue LLMs for Co-Creative Improvised Theatre.

Chiou, M.; Hawes, N.; and Stolkin, R. 2021. Mixed-initiative variable autonomy for remotely operated mobile robots. *J. Hum.-Robot Interact.* 10(4).

Cimolino, G., and Graham, T. N. 2022. Two heads are better than one: A dimension space for unifying human and artificial intelligence in shared control. In *Proceedings of the 2022 CHI*, 1–21.

Dang, H.; Goller, S.; Lehmann, F.; and Buschek, D. 2023. Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of CHI '23*, 1–17.

Davis, N.; Hsiao, C.-P.; Singh, K. Y.; and Magerko, B. 2016. Co-creative drawing agent with object recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 12, 9–15.

Davis, N.; Deshpande, M.; Rezwana, J.; and Magerko, B. 2024. The Five Pillars of Enaction as a Theoretical Framework for Co-Creative Artificial Intelligence.

Dhillon, P. S.; Molaei, S.; Li, J.; Golub, M.; Zheng, S.; and Robert, L. P. 2024. Shaping human-ai collaboration: varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI*, 1–18.

E, J. L.; Yen, Y.-C. G.; Pan, I. Y.; Lin, G.; Li, M.; Jin, H.; Chen, M.; Xia, H.; and Dow, S. P. 2024. When to give feedback: Exploring tradeoffs in the timing of design feedback. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, 292–310. New York, NY, USA: ACM.

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.

Feldman, S. S. 2017. Co-Creation: Human and AI Collaboration in Creative Expression. In *Electronic Visualisation and the Arts (EVA 2017)*. BCS Learning & Development.

Ferrario, A.; Facchini, A.; and Termine, A. 2024. Experts or authorities? the strange case of the presumed epistemic superiority of artificial intelligence systems. *Minds and Machines* 34(3):30.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Frasheri, M.; Çürüklü, B.; and Ekström, M. 2017. Towards collaborative adaptive autonomous agents. In *ICAART (1)*.

Gao, X.; Rogel, A.; Sankaranarayanan, R.; Dowling, B.; and Weinberg, G. 2024. Music, body, and machine: gesture-based synchronization in human-robot musical interaction. *Frontiers in Robotics and AI* 11:1461615.

Guzman, A. L., and Lewis, S. C. 2020. Artificial intelligence and communication: A human–machine communication research agenda. *New media & society* 22(1):70–86.

Hardin, B., and Goodrich, M. A. 2009. On using mixed-initiative control: a perspective for managing large-scale robotic teams. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, 165–172. ACM.

Heer, J. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116(6):1844–1850.

Kabir, M. N. 2024. View of Unleashing Human Potential: A Framework for Augmenting Co-Creation with Generative AI.

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.

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 Computational Creativity*, 57–64.

Kim, T.; Molina, M. D.; Rheu, M.; Zhan, E. S.; and Peng, W. 2023. One ai does not fit all: A cluster analysis of the laypeople's perception of ai roles. In *Proceedings of the 2023 CHI*, 1–20.

Koch, J.; Ravikumar, P. T.; and Calegario, F. 2021. Agency in Co-Creativity: Towards a Structured Analysis of a Concept. In Garza, A. G. d. S.; Veale, T.; Aguilar, W.; and Pérez, R. P. y., eds., *Proceedings of the Twelfth International Conference on Computational Creativity, September 14-18, 2021*, 449–452. Association for Computational Creativity.

Larsson, T.; Font, J.; and Alvarez, A. 2022. Towards ai as a creative colleague in game level design. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, 137–145.

Lawton, T.; Ibarrola, F. J.; Ventura, D.; and Grace, K. 2023. Drawing with Reframer: Emergence and Control in Co-Creative AI. In *Proceedings of the 28th IUI*, 264–277. Sydney NSW Australia: ACM.

Leotti, L. A.; Iyengar, S. S.; and Ochsner, K. N. 2010. Born to Choose: The Origins and Value of the Need for Control. *Trends in cognitive sciences* 14(10):457–463.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2014. Computational game creativity. ICCC.

Lin, Z.; Ehsan, U.; Agarwal, R.; Dani, S.; Vashishth, V.; and Riedl, M. 2023a. Beyond Prompts: Exploring the Design Space of Mixed-Initiative Co-Creativity Systems.

Lin, Z.; Ehsan, U.; Agarwal, R.; Dani, S.; Vashishth, V.; and Riedl, M. 2023b. Beyond prompts: Exploring the design space of mixed-initiative co-creativity systems. *arXiv preprint arXiv:2305.07465*.

Liu, H. X.; Huang, Y.; and Holopainen, J. 2023. How to use generative ai as a design material for future human-computer (co-) creation?

Lobo, I.; Koch, J.; Renoux, J.; Batina, I.; and Prada, R. 2024. When Should I Lead or Follow: Understanding Initiative Levels in Human-AI Collaborative Gameplay. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, 2037–2056. New York, NY, USA: ACM.

Louie, R.; Coenen, A.; Huang, C. Z.; Terry, M.; and Cai, C. J. 2020. Novice-ai music co-creation via ai-steering tools for deep generative models. In *Proceedings of CHI '20*, CHI '20, 1–13. New York, NY, USA: ACM.

Louie, R.; Engel, J.; and Huang, C.-Z. A. 2022. Expressive communication: Evaluating developments in generative models and steering interfaces for music creation. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22. New York, NY, USA: ACM.

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):365–369.

Maher, M. 2012. Computational and Collective Creativity: Who is Being Creative? In *In Proceedings of The Third International Conference on Computational Creativity, University College Dublin*, 67–71. Association for Computational Creativity.

Manovich, L., and Arielli, E. 2024. *Artificial Aesthetics: Generative AI, Art and Visual Media*.

Margarido, S.; Roque, L.; Machado, P.; and Martins, P. 2024. Boosting Mixed-Initiative Co-Creativity in Game Design: A Tutorial. arXiv:2401.05999 [cs].

Margarido, S.; Roque, L.; Machado, P.; and Martins, P. 2025. MI-CCy Quantifier: A Framework for Quantifying Mixed-Initiative Co-creativity in Human-AI Collaborations. In *Progress in Artificial Intelligence*, 3–15. Springer, Cham. ISSN: 1611-3349.

Meta. 2024. Movie Gen: A Cast of Media Foundation Models. AI at Meta. https://ai.meta.com/static-resource/movie-gen-research-paper.

Morris, M. R.; Sohl-dickstein, J.; Fiedel, N.; Warkentin, T.; Dafoe, A.; Faust, A.; Farabet, C.; and Legg, S. 2024. Levels of AGI for Operationalizing Progress on the Path to AGI. arXiv:2311.02462 [cs].

Moruzzi, C., and Margarido, S. 2024a. Customizing the Balance between User and System Agency in Human-AI Co-Creative Processes.

Moruzzi, C., and Margarido, S. 2024b. A user-centered framework for human-ai co-creativity. In *Proceedings of CHI '24 EA*, CHI EA '24. New York, NY, USA: ACM.

Moruzzi, C. 2022. Creative agents: rethinking agency and creativity in human and artificial systems. *Journal of Aesthetics and Phenomenology* 9(2):245–268.

Muller, M.; Candello, H.; and Weisz, J. 2023. Interactional Co-Creativity of Human and AI in Analogy-Based Design.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Muyskens, K.; Ma, Y.; Menikoff, J.; Hallinan, J.; and Savulescu, J. 2024. When can we kick (some) humans "out of the loop"? an examination of the use of ai in medical imaging for lumbar spinal stenosis. *Asian Bioethics Review*.

Oh, C.; Song, J.; Choi, J.; Kim, S.; Lee, S.; and Suh, B. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–13.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue — openai.com.

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.

Rezwana, J., and Maher, M. L. 2022. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction*.

Salikutluk, V.; Schöpper, J.; Herbert, F.; Scheuermann, K.; Frodl, E.; Balfanz, D.; Jäkel, F.; and Koert, D. 2024. An evaluation of situational autonomy for human-ai collaboration in a shared workspace setting. In *Proceedings of CHI '24*, 1–17.

Shaer, O.; Cooper, A.; Mokryn, O.; Kun, A. L.; and Ben Shoshan, H. 2024. Ai-augmented brainwriting: Investigating the use of llms in group ideation. In *Proceedings of CHI '24*, CHI '24. New York, NY, USA: ACM.

Singh, S., and Heard, J. 2022. Human-aware reinforcement learning for adaptive human robot teaming. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1049–1052. IEEE.

Stuart, J. 2003. Russell and peter norvig. *Artificial Intelligence: A Modern Approach, Prentice Hall*.

Trajkova, M.; Long, D.; Deshpande, M.; Knowlton, A.; and Magerko, B. 2024. Exploring collaborative movement improvisation towards the design of luminai—a co-creative ai dance partner. In *Proceedings of the 2024 CHI*. New York, NY, USA: Association for Computing Machinery.

Tsamados, A.; Floridi, L.; and Taddeo, M. 2024. Human control of ai systems: from supervision to teaming. *AI and Ethics* 1–14.

Vinchon, F.; Lubart, T.; Bartolotta, S.; Gironnay, V.; Botella, M.; Bourgeois-Bougrine, S.; Burkhardt, J.; Bonnardel, N.; Corazza, G. E.; Glăveanu, V.; Hanchett Hanson, M.; Ivcevic, Z.; Karwowski, M.; Kaufman, J. C.; Okada, T.; Reiter-Palmon, R.; and Gaggioli, A. 2023. Artificial Intelligence & Creativity: A Manifesto for Collaboration. *The Journal of Creative Behavior* 57(4):472–484.

Walsh, T.; Levy, N.; Bell, G.; Elliott, A.; Maclaurin, J.; Mareels, I.; and Wood, F. 2019. *The effective and ethical development of artificial intelligence: an opportunity to improve our wellbeing*. Australian Council of Learned Academia.

Wan, Q.; Hu, S.; Zhang, Y.; Wang, P.; Wen, B.; and Lu, Z. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 8(CSCW1):84:1–84:26.

Wang, B.; Li, Y.; Lv, Z.; Xia, H.; Xu, Y.; and Sodhi, R. 2024a. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, 699–714. New York, NY, USA: ACM.

Wang, Z.; Huang, Y.; Song, D.; Ma, L.; and Zhang, T. 2024b. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of CHI '24*, CHI '24. New York, NY, USA: ACM.

Wu, Z.; Ji, D.; Yu, K.; Zeng, X.; Wu, D.; and Shidujaman, M. 2021. *AI Creativity and the Human-AI Co-creation Model*. IEEE. 171–190.

Yang, D.; Kleinman, E.; Troiano, G. M.; Tochilnikova, E.; and Harteveld, C. 2024. Snake story: Exploring game mechanics for mixed-initiative co-creative storytelling games. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24. ACM.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-Initiative Co-Creativity.

Yao, F.; Li, C.; Nekipelov, D.; Wang, H.; and Xu, H. 2024. Human vs. Generative AI in Content Creation Competition: Symbiosis or Conflict? arXiv:2402.15467 [cs].

Prompting AI in Co-Creation: The Role of Syntax and Sentiment in Shaping AI-Generated Content

Giancarlo Colloca, Jeba Rezwana
Computer and Information Science Department
Towson University
8000 York Road, Towson, MD 21252 USA
gcolloc1@students.towson.edu; jrezwana@towson.edu

Abstract

As Generative AI has advanced at an incredible pace and has been increasingly integrated into creative domains, understanding how users' prompting styles shape AI-generated creative content is critical. Despite a growing interest in human-AI co-creation, there is limited research exploring how linguistic and emotional cues in user prompts affect AI-generated responses in creative domains. This study addresses this gap by analyzing the syntactic and sentiment patterns of user prompts and their impact on ChatGPT's responses in a co-creative storytelling study with 100 participants. We found different stylistic patterns in both prompts and responses, while ChatGPT often amplified users' sentiment and positivity. The findings offer insights to enhance our understanding of how AI can be optimized for more appropriate interactions in co-creation.

Introduction

Generative AI (GenAI) is transforming the landscape of creative expression, offering new possibilities for collaboration between humans and machines. Large language models (LLMs) are becoming increasingly integrated into creative workflows, democratizing the generation of creative artifacts. In co-creative domains, users and AI systems engage in dynamic interactions where each party contributes to the evolving narrative (Davis 2013), which makes understanding how human input shapes AI output critical. Much attention has been given to the capabilities and limitations of these models, but comparatively little research has examined the linguistic and emotional characteristics of user prompts and their influence on the nature and quality of AI-generated content. The structure and emotional tone of user prompts play a pivotal role in shaping the responses of LLMs in creative contexts like storytelling (Ul Huda et al. 2024). Prior research has shown that minor variations in prompt phrasing can lead to substantial differences in output style, content, and coherence (Brown et al. 2020; Reynolds and McDonell 2021). Linguistic cues often carry implicit emotional and contextual information that can guide the model's interpretation of user intent (Chung et al. 2024). In co-creative settings, such cues are especially important, as they can determine whether the AI responds passively, takes initiative, or mirrors the user's creative direction.

Recent research has highlighted the critical role of prompt engineering in guiding LLMs toward more accurate, coherent, and contextually appropriate outputs (Brown et al. 2020; Wei et al. 2022). These efforts have focused on measures like reasoning ability, factual accuracy, and task success, but the more nuanced, co-creative dimensions of human-AI collaboration remain underexplored. Work on prompt tuning and dynamic adaptation has advanced internal model optimization (Xiao et al. 2025; Zhang et al. 2024), yet it typically overlooks how users engineer their prompts to interact with models in creative contexts. Concerns about the diversity and authenticity of model outputs have also been tied to prompt phrasing (Holtzman et al. 2019), and prompt robustness has been studied in adversarial contexts (Liu et al. 2021). However, there has been limited exploration into how prompts influence model behavior in terms of the sentiment of user prompts, as well as the structural and stylistic aspects of user prompts. Studies have examined prompt performativity (how prompts elicit desired responses) (Shin et al. 2020) and generation control (guiding model output) (Dathathri et al. 2020), however these analyses often focus on token-level or dataset-driven approaches, overlooking the informal nature of human-AI co-creation, where users engage with the model through dynamic prompts. Recent studies have begun to map the prompt engineering landscape (Chen et al. 2023; Sahoo et al. 2024), and exploratory work on prompt literacy (Hwang, Lee, and Shin 2023) indicated the growing need for more user-centric approaches. The interpretive space where human intention meets AI output remains largely unmapped, in particular, how linguistic and emotional cues in user prompts influence AI responses in co-creative storytelling.

Motivated by this gap, this study investigates the linguistic structure and sentiments in user prompts and how these components shape model outputs during a co-creative writing task. We conducted a user study with 100 participants who engaged with ChatGPT in a creative writing task. Focusing on prompt-response pairs, we examined how syntax, style, and sentiment in prompts influence the variability, creativity, and alignment of AI-generated content. We present findings that could be utilized in deepening our understanding of the dynamics and structures of prompts and responses in co-creativity and supporting future work in prompt literacy and personalized interaction design.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Related Work

Human-AI co-creation is a rapidly growing research area where humans and AI collaborate on creative tasks to produce creative artifacts or ideas (Davis 2013). Generative AI has been heavily used in human-AI co-creation lately, which serves as an effective co-creation technology (Bozkurt and Sharma 2023) by enhancing our creative abilities and transforming our experience of how we interact with AI systems. The efficacy of generative AI is influenced by its underlying algorithms and training data, but also by the way users interact with it (Rezwana and Maher 2022), for example, the quality of the prompts AI systems receive (Liu et al. 2023). Generative AI offers immense potential, but its effectiveness depends on well-crafted prompts (Cheung 2023). Well-designed prompts can elicit meaningful responses, while a poorly constructed prompt may lead to irrelevant outputs. Yet, many users struggle with prompt engineering due to the novelty of these technologies (Bozkurt and Sharma 2023). This highlights the need for research into effective prompting strategies.

Previous studies emphasize that prompt structure plays a critical role in AI performance. For instance, Brown et al. (2020) introduced few-shot prompting as a means of guiding language models, while Wei et al. (2022) demonstrated that structured chains of thought can elicit logical reasoning. These approaches highlight how task framing can significantly impact model behavior and effectiveness. In creative contexts, creative generation using prompt engineering goes beyond correctness, requiring expressiveness, engagement, and emotional alignment (Reynolds and McDonell 2021). In the domain of co-creativity, Jakesch et al. (Jakesch et al. 2019) explored human perception of AI-generated content, emphasizing factors like trust and social resonance. Serban et al. (Serban et al. 2015) examined how user expectations shape dialogue interactions—primarily in task-oriented settings—but their findings on conversational alignment offer insights relevant to co-creative communication. Recent research (Bhandarkar et al. 2024) highlights the growing need for adaptive models that align with not just content but user sentiment and style—factors increasingly relevant in co-creative systems. Other work (Smith et al. 2025) has begun to examine multimodal and cross-domain prompt structures, especially in artistic and educational settings, though few studies link linguistic traits directly to co-creative engagement.

The literature shows that the way users interact with AI significantly influences the creativity that emerges from co-creation (Karimi et al. 2020). We can identify prompt attributes that lead to specific creative outcomes, which can inform effective prompt engineering across various creative domains. Additionally, the field of human-centered NLP (Bender and Koller 2020) emphasizes the importance of considering the social and affective aspects of AI responses. However, there is a gap in computational creativity research regarding the structure and sentiment of prompts and their impact on AI responses. Our research contributes to this area by exploring how the style and sentiment of user prompts influence AI outputs in creative writing/storytelling.

Study Methodology

Procedure

This study explores users' prompt writing styles and how they influence AI-generated responses in creative contexts. Participants engaged in a co-creative storytelling task using ChatGPT-3.5. Participants were provided with the following narrative prompt: "A little girl named Alice, who is 7, wants to go on an adventure with her dog Lola. One day when both her parents went out for a quick errand, Alice and Lola..." Participants had 7 minutes to continue a story and prompt ChatGPT to finish or improve it, providing screenshots of the prompt and final story as proof of completion. The study was approved by the Institutional Review Board.

Using Prolific, an online participant recruitment platform, we recruited 100 participants (56 men, 40 women, and 4 non-binary) who were at least 18 years old, with an average age of 27.0 (Med = 27.5, SD = 6.88). As our aim was to gather a diverse set of data across various demographics, we kept the inclusion criteria to a minimum, requiring only that participants be at least 18 years old.

Data Collection and Analysis

We collected users' prompts and ChatGPT responses through the screenshots participants submitted.

We conducted a series of statistical analyses. Prompts and AI responses' syntactic structures were analyzed using Parts-of-speech (POS), categories to which each word is assigned in accordance with its syntactic functions (Jurafsky and Martin 2024). We focused on POS features as they provide an interpretable framework for analyzing language patterns across diverse users. POS tags capture the grammatical roles of words, allowing us to quantify the expressive, descriptive, and structural tendencies present in both human and machine-generated texts. We utilized POS Tagger of the Stanford CoreNLP package version 4.5.8 (Manning et al. 2014) to categorize words into nine features: nouns (NOUN), verbs (VERB), adjectives (ADJ), adverbs (ADV), pronouns (PRON), conjunctions (CONJ), adpositions (ADP), numbers (NUM) and determiners (DET).

To investigate stylistic patterns in the users' prompts and AI responses, we conducted k-means clustering with the nine POS tags as the features. To visualize and interpret clusters, we applied Principal Component Analysis (PCA) to reduce the dimensionality of POS tag distributions to two key dimensions.

Additionally, we used paired t-tests to assess differences in syntactical structures between clusters of user prompts and AI responses, and one-way ANOVA to examine the most influential cluster features. Additionally, we used Chi-square tests to identify the association between several variables, such as clusters of users' prompts and AI responses.

Sentiment analysis was conducted using Text2Data (Tex 2025), an online text analysis tool, which has been utilized in the literature for sentiment analysis (Gandy et al. 2025; Ermakova, Henke, and Fabian 2021). Text2Data generates sentiment polarity scores measuring a text's emotional tone (from -1 indicating negative to 1 indicating positive).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Results

In this section, we present the findings from our analysis of user prompts and AI-generated responses in the co-creative storytelling task. We begin by examining the syntactic patterns using POS distributions, followed by clustering analysis that identifies distinct stylistic patterns across prompts and responses as well as examines how these clusters relate to one another. Finally, we report on the sentiment analysis, comparing the emotional tone between user inputs and ChatGPT's outputs.

POS Tags	User Prompts			ChatGPT Responses			Comparisons	
	COUNTS	Freq. %	SD	COUNTS	Freq. %	SD	Freq. Diff. Δ%	t-test p-value
ADJ	380	6.68	2.50	1833	9.00	2.87	2.32	0.0055
ADV	269	4.73	2.12	853	4.19	1.98	-0.54	0.4685
CONJ	294	5.17	2.23	934	4.59	2.08	-0.58	0.4262
DET	660	11.61	3.21	2278	11.19	3.15	-0.42	0.6391
NOUN	1619	28.47	4.52	5811	28.54	4.52	0.07	0.9194
NUM	267	4.69	2.11	60	0.29	0.54	-4.40	<0.0001
ADP	622	10.94	3.11	2795	13.73	3.47	2.79	0.0006
PRON	493	8.67	2.80	1920	9.43	2.92	0.76	0.3501
VERB	1083	19.04	3.93	3876	19.04	3.93	0.00	1

Table 1: Comparison of part-of-speech (POS) tag distributions between user prompts and AI responses, with results of paired t-tests.

Comparison between Structural Patterns of User Prompts and AI Responses

To understand how linguistic structure varies between users and ChatGPT during co-creative storytelling, we analyzed the POS tag distributions in user prompts and corresponding AI-generated responses. Table 1 presents the mean percentage and standard deviation for each POS category across all samples, along with the results of paired t-tests comparing prompts and responses. Table 1 reveals clear syntactic differences between participant prompts and ChatGPT responses. ChatGPT uses significantly more adjectives (ADJ) ($\Delta = 2.32\%, p = 0.0055$) and adpositions (ADP) ($\Delta = 2.79\%, p = 0.0006$), suggesting a tendency to enhance descriptions and relational clarity. In contrast, it drastically reduces NUM ($\Delta = -4.40\%, p < 0.0001$) that were part of the task for the user prompts, which may strip away some contextual or quantitative detail. These shifts highlight how the model transforms factual, structured input into more elaborative and expressive language.

Cluster Analysis of Prompts and Responses

To identify stylistic patterns in both user prompts and ChatGPT responses, we applied k-means clustering using distributions across nine POS categories. As visualized, we identified two distinct clusters (See Figure 1) in both participant Prompts and ChatGPT Responses. Principal Component Analysis (PCA) was used to reduce dimensionality and capture the dominant syntactic variance, with the first two dimensions explaining the majority of variance in both prompts and responses.

For user prompts, PCA Dimension 1 is driven by VERB, DET, and NOUN, reflecting content density related to actions and subjects. PCA Dimension 2 is shaped by CONJ, NUM, and ADP, capturing structural complexity. For ChatGPT responses, Dimension 1 similarly reflects content richness, dominated by NOUN, VERB, and DET, while Dimension 2, influenced by ADJ, ADV, and CONJ, highlights descriptive and connective elaboration.

In prompt clusters, Cluster 0 (n = 23) includes more expressive prompts, using more ADJ, ADV, and PRON, suggesting a personal tone. Cluster 1 (n = 77) prompts relied more on NOUN, DET, and CONJ, indicating a more descriptive, scene-setting approach. A similar pattern emerged in ChatGPT response clusters: Cluster 0 involves 41 instances that used more ADJ, ADP, and PRON, aligning with an expressive style, while Cluster 1 involves 59 instances that used more NOUN, reflecting a simpler, clearer structure.

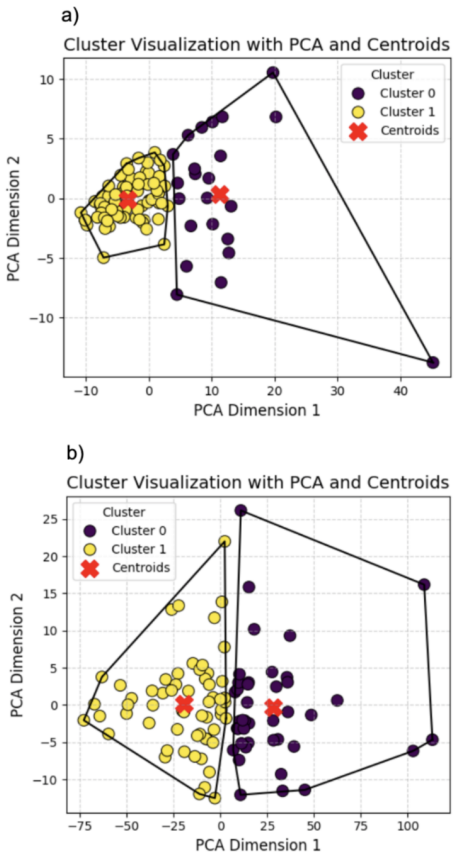


Figure 1: Clustering of user prompts (a) and ChatGPT responses (b) based on part-of-speech (POS) tag distributions using k-means. Purple dots (Cluster 0) represent more expressive language patterns and yellow dots (Cluster 1) reflect a more descriptive, scene-setting style.

We conducted one-way ANOVA tests to identify which POS features significantly differed between the clusters; all were statistically significant ($p < 0.001$) except for NUM, which did not show a meaningful difference. Among the most distinguishing features were NOUN, DET, and VERB,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

with higher frequencies observed in Cluster 0 of the prompt data—a pattern that also appeared in the response clusters. PRON (personal pronouns) showed particularly strong differences: in prompts, Cluster 0 had a mean value of 7.57 compared to 4.14 in Cluster 1; in ChatGPT responses, the difference was even larger (25.80 vs. 14.61). ADP (adpositions) exhibited the greatest divergence in the response clusters (38.00 vs. 20.97). Similarly, ADV (adverbs) showed a notable difference in prompts (5.70 vs. 1.79). ADJ (adjectives) and CONJ (conjunctions) followed similar trends across both datasets, with higher averages in Cluster 0, indicating a more expressive and syntactically rich style.

Despite these clear stylistic patterns within prompts and responses, no significant associations were found between the clusters of prompts and responses in terms of syntactic styles. This finding is somewhat unexpected, as one might assume the structure of user input would influence the syntactic style of the AI output. Additionally, we also found no significant associations between the prompt/response clusters and participant demographics (age, gender, race, country of origin, formative background, disability, education level, field of work/study, occupation, income, political affiliation, or AI expertise).

Sentiment Analysis

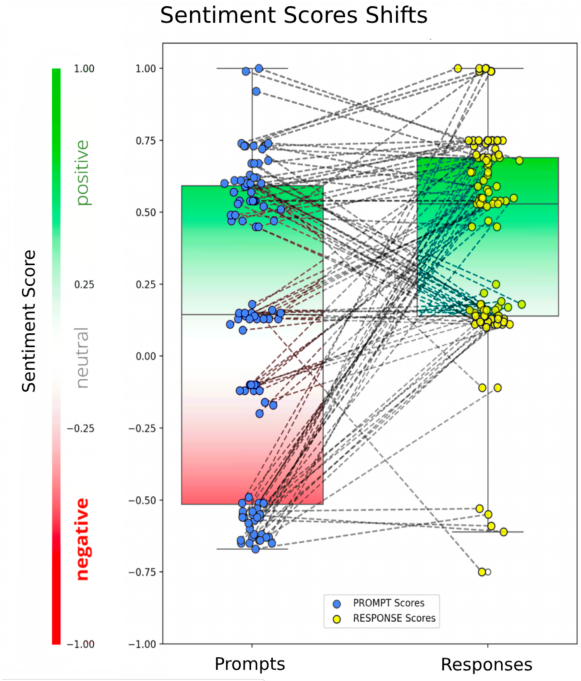


Figure 2: Comparison of sentiment scores between user prompts and ChatGPT responses. Each line connects a user prompt (blue dot) to its corresponding AI response (yellow dot), showing that ChatGPT responses generally exhibit increased positivity.

We observed a pattern of positive sentiment shift in ChatGPT's responses compared to user prompts (Figure 2). Sentiment analysis revealed that user prompts encompassed a wide range of emotional tone, with 44 categorized as positive, 27 as neutral, and 29 as negative. In contrast, ChatGPT responses exhibited a clear shift toward positivity, comprising 58 positive, 37 neutral, and only 5 negative instances.

A paired t-test confirmed a statistically significant increase in sentiment scores from prompts to responses ($p < 0.001$), indicating that ChatGPT tends to respond with more positive sentiment regardless of the prompt's tone. Figure 2 visualizes these changes. Each dot represents an individual sentiment score, with blue dots indicating sentiment scores for user prompts and yellow dots for ChatGPT responses. Dashed lines connecting the pairs illustrate the direction of sentiment shifts, and boxplots provide an overview of the distribution in each category.

In nearly half the cases (47%), ChatGPT responses showed a sentiment increase of over 0.25, indicating a tendency to elevate emotional tone. In 34% of cases, sentiment remained largely stable (≤ 0.25 change). In 19% of cases, responses were more negative, reflecting moments when ChatGPT adopted a more critical or somber tone.

Overall, prompts displayed broader emotional variability, whereas ChatGPT's responses were more tightly clustered around the positive end of the sentiment scale. This trend reaffirms that while users bring a wide range of emotional tones to the interaction, ChatGPT tends to focus on more positive emotional expressions.

Discussion and Conclusions

As the use of GenAI rapidly grows in creative domains, understanding the stylistic, structural, and sentiment patterns in user prompts—and how these patterns influence AI-generated responses—is crucial to unlocking the full potential of GenAI. This study explored stylistic and emotional patterns in prompts and AI-generated responses and how user prompts influence ChatGPT's linguistic and emotional responses in a collaborative storytelling context with 100 participants. We present the findings and discuss the implications in this section.

The syntactic analysis revealed that ChatGPT responses contained significantly more adjectives (ADJ) and adpositions (ADP) than user prompts, indicating that the AI enriches its language with descriptive and relational elements. This suggests an attempt to create more vivid and engaging responses, potentially aligning with users' implicit requests for expressiveness. In contrast, numerical expressions (NUM) were much less frequent in ChatGPT responses, implying that the AI prioritizes narrative clarity and emotional tone over factual precision in storytelling, which aligns with the existing research (Breithaupt et al. 2024).

The clustering analysis reveals two distinct linguistic styles in both user prompts and AI-generated responses. Cluster 0 in user prompts, involves more expressive prompts, uses a higher frequency of adjectives, adverbs, and pronouns, suggesting a personal, emotional tone. In contrast, Cluster 1 relies more on nouns, determiners, and conjunctions, indicating a descriptive, factual approach. AI responses followed a similar pattern, with Cluster 0 exhibiting more adjectives, adpositions, and pronouns, aligning with

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

the expressive tone of the prompts. Cluster 1's responses mirrored the user prompts' simpler, more structured style. These contrasting styles highlight different rhetorical strategies, representing the AI's ability to adjust its narrative structure, suggesting it can enhance engagement in creative or emotionally sensitive contexts while maintaining clarity in more factual or instructional interactions.

Human input is more stylistically diverse compared to AI responses, as seen in the greater spread of points for the user prompts across PCA dimensions. AI responses, while reflecting the general structure of the prompts, remain more standardized and homogeneous, aligning with the AI's tendency to prioritize grammatical correctness and coherence over stylistic variation. This distinction underscores the syntactic regularity and homogeneity in AI responses, potentially limiting its adaptability, which aligns with existing research echoing concerns (Liu et al. 2024) about "AI homogenization" in creative fields. Identifying mechanisms to increase response diversity without compromising coherence could improve co-creation balance.

Interestingly, no significant association was found between the clusters of prompts and responses, which is surprising given the expected direct influence of user input on AI-generated content. This finding suggests that ChatGPT may not always perfectly align with the user's intended tone or style. The lack of correlation could be due to the model's tendency to generate responses based on a range of factors beyond just syntactic alignment, such as the model's training data, its programming to optimize for clarity, and its inherent biases in generating emotionally positive outputs.

Sentiment analysis revealed a significant shift toward more positive sentiment in ChatGPT responses compared to user prompts. While user prompts spanned a wide range of emotional tones, from highly negative to positive, ChatGPT's responses tended to be more positive, with a focus on neutral to mildly positive sentiment, highlighting ChatGPT's potential to modulate emotional tone in a way that might help to maintain positive engagement. Such tendencies are likely influenced by reinforcement learning from human feedback (RLHF), which rewards outputs perceived as helpful, engaging, or agreeable (Ouyang et al. 2022). However, this shift also raises questions about the implications of AI modifying users' emotional expressions. For instance, in settings where a user may express frustration or dissatisfaction, ChatGPT's tendency to amplify positive sentiment could be seen as dismissive or unresponsive to the user's emotional state. This aspect of AI behavior warrants further investigation, particularly in creative contexts such as storytelling where emotional nuance is important.

It is also important to recognize that the interpretability and behavior of language models are continually evolving. Different versions of LLMs, or the inclusion of multimodal capabilities, may generate markedly different responses even when given identical inputs. This suggests there may not be a singular or universal mapping between user intention and AI output. Human intention itself is deeply complex and inherently subjective, posing further challenges in aligning AI responses with nuanced human goals or emotional states. Future studies should investigate how varying model architectures and user contexts shape this interpretive space, and how models might better account for the richness of human intent.

Limitations and Future Work

This study has a few limitations. One is the use of a fixed story prompt and a constrained interaction time, which may not fully capture the dynamic nature of user-AI interactions. Future work should explore open-ended interactions with evolving prompts to better understand how style and sentiment develop in more spontaneous and extended contexts. The study focused on a single creative task with a specific system that might limit the generalizability. Future research should examine whether similar stylistic and emotional patterns emerge in a broader range of creative tasks.

Another limitation is the participant sample, which may not reflect the diversity of users who engage with AI in creative settings. Future studies should include more diverse demographics to explore how different user groups engineer prompts. Further research could also examine the range and types of sentiments involved in human-AI interactions to better understand how emotional tone is expressed, transformed, or constrained in these interactions.

Finally, future research should investigate whether AI's tendency to shift responses toward positive sentiment might inadvertently limit the emotional depth and authenticity of co-created content. In creative collaborations, human expression often encompasses a broad spectrum of emotions, including negative and ambiguous tones. If AI systems consistently steer outputs toward optimism or neutrality, there is a risk of diminishing the nuanced emotional landscapes that are essential to authentic creative expression (Ashkinaze et al. 2024). Understanding how AI's sentiment modulation shapes the co-creative process is critical for designing tools that support—not flatten—human creativity.

Acknowledgments

This research was supported by the Faculty Development Research Committee (FDRC) Grant at Towson University, awarded to Jeba Rezwana.

References

Ashkinaze, J.; Mendelsohn, J.; Qiwei, L.; Budak, C.; and Gilbert, E. 2024. How ai ideas affect the creativity, diversity, and evolution of human ideas: evidence from a large, dynamic experiment. *arXiv preprint arXiv:2401.13481*.

Bender, E. M., and Koller, A. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185–5198.

Bhandarkar, A.; Wilson, R.; Swarup, A.; and Woodard, D. 2024. Emulating author style: a feasibility study of prompt-enabled text stylization with off-the-shelf llms. In *1st Workshop on Personalization of Generative AI Systems*, 76–82.

Bozkurt, A., and Sharma, R. C. 2023. Generative ai and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education* 18(2):i–vii.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Breithaupt, F.; Otenen, E.; Wright, D. R.; Kruschke, J. K.; Li, Y.; and Tan, Y. 2024. Humans create more novelty than chatgpt when asked to retell a story. *Scientific Reports* 14(1):875.

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.

Chen, B.; Zhang, Z.; Langrené, N.; and Zhu, S. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.

Cheung, K. S. 2023. Real estate insights unleashing the potential of chatgpt in property valuation reports: the "red book" compliance chain-of-thought (cot) prompt engineering. *Journal of Property Investment & Finance*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25(70):1–53.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and play language models: A simple approach to controlled text generation.

Davis, N. M. 2013. Human-computer co-creativity: Blending human and computational creativity. In *Artificial Intelligence and Interactive Digital Entertainment Conference*.

Ermakova, T.; Henke, M.; and Fabian, B. 2021. Commercial sentiment analysis solutions: a comparative study.

Gandy, L. M.; Ivanitskaya, L. V.; Bacon, L. L.; and Bizri-Baryak, R. 2025. Public health discussions on social media: Evaluating automated sentiment analysis methods. *JMIR Formative Research* 9(1):e57395.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hwang, Y.; Lee, J. H.; and Shin, D. 2023. What is prompt literacy? an exploratory study of language learners' development of new literacy skill using generative ai.

Jakesch, M.; French, M.; Ma, X.; Hancock, J. T.; and Naaman, M. 2019. Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness. In *Proceedings of the 2019 CHI*, 1–13.

Jurafsky, D., and Martin, J. 2024. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech cognition. draft (2023).

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*.

Liu, J.; Takanobu, R.; Wen, J.; Wan, D.; Li, H.; Nie, W.; Li, C.; Peng, W.; and Huang, M. 2021. Robustness testing of language understanding in task-oriented dialog. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *59th Annual Meeting of the Association for Computational Linguistics*.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; Wang, K.; and Liu, Y. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Liu, Q.; Zhou, Y.; Huang, J.; and Li, G. 2024. When chatgpt is gone: Creativity reverts and homogeneity persists. *arXiv preprint arXiv:2401.06816*.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *52nd annual meeting of the association for computational linguistics: system demonstrations*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35:27730–27744.

Reynolds, L., and McDonell, K. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI*, 1–7.

Rezwana, J., and Maher, M. L. 2022. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction*.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Serban, I. V.; Lowe, R.; Henderson, P.; Charlin, L.; and Pineau, J. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Shin, T.; Razeghi, Y.; IV, R. L. L.; Wallace, E.; and Singh, S. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts.

Smith, B. E.; Shimizu, A. Y.; Burriss, S. K.; Hundley, M.; and Pendergrass, E. 2025. Multimodal composing with generative ai: Examining preservice teachers' processes and perspectives. *Computers and Composition* 75:102896.

2025. Text2data. https://text2data.com/ [Accessed: (2025)].

Ul Huda, N.; Sahito, S. F.; Gilal, A. R.; Abro, A.; Alshanqiti, A.; Alsughayyir, A.; and Palli, A. S. 2024. Impact of contradicting subtle emotion cues on large language models with various prompting techniques. *International Journal of Advanced Computer Science & Applications* 15(4).

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.

Xiao, Z.; Yan, S.; Hong, J.; Cai, J.; Jiang, X.; Hu, Y.; Shen, J.; Wang, Q.; and Snoek, C. G. 2025. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*.

Zhang, J.; Wu, S.; Gao, L.; Shen, H. T.; and Song, J. 2024. Dept: Decoupled prompt tuning. In *Conference on Computer Vision and Pattern Recognition*, 12924–12933.

SPARK: A System for Scientifically Creative Idea Generation

Aishik Sanyal
Spiral Works
aishik@spiralworks.ai

Samuel Schapiro
Univ. Illinois, Urbana-Champaign
Spiral Works
sam@spiralworks.ai

Sumuk Shashidhar
Univ. Illinois, Urbana-Champaign
sumuks2@illinois.edu

Royce Moon
University of Michigan
Spiral Works
royce@spiralworks.ai

Lav R. Varshney
Univ. Illinois, Urbana-Champaign
varshney@illinois.edu

Dilek Hakkani-Tur
Univ. Illinois, Urbana-Champaign
dilek@illinois.edu

Abstract

Recently, large language models (LLMs) have shown promising abilities to generate novel research ideas in science, a direction which coincides with many foundational principles in computational creativity (CC). In light of these developments, we present an idea generation system named SPARK that couples retrieval-augmented idea generation using LLMs with a reviewer model named JUDGE trained on 600K scientific reviews from OpenReview. Our work is both a system demonstration and intended to inspire other CC researchers to explore grounding the generation and evaluation of scientific ideas within foundational CC principles. To this end, we invite other researchers to explore the use of LLMs for idea generation and creative evaluations, and release the annotated dataset used to train JUDGE.[1]

Introduction

Automating parts of the scientific process has long been a research goal in the field of computer science and artificial intelligence (AI) (Langley 1977; Thagard 1988; Jumper et al. 2021), with computational creativity (CC) principles playing a crucial role in many parts of this endeavor. Crucially, the initial task of idea or hypothesis generation is an area that requires creative intuition but has historically been restricted to the ingenuity of human scientists.

Recent research, however, has demonstrated that large language models (LLMs) are capable of generating novel research ideas (Wang et al. 2024; Si, Yang, and Hashimoto 2024; Yang et al. 2023). In fact, a handful of fully autonomous AI systems capable of writing full-length papers at near-human quality, or workshop-length papers that successfully pass through human peer review, have been developed (Lu et al. 2024; Intology 2025). In each of these systems, idea generation remains the crucial first step.

In this paper, we present a scientific idea generation system named SPARK that generates novel scientific ideas and an evaluator model named JUDGE, trained on a 600K size dataset of expert human reviews from OpenReview, that can evaluate the creativity of SPARK's proposed ideas. Our goal is to ground recent developments in scientific idea generation with the foundational computational creativity principles that underpin much of their work, focusing on the first stage of idea generation. To facilitate reproducibility and community progress, we are releasing the full OpenReview dataset used to train our JUDGE model. Finally, to provide background context, we first start by summarizing recent developments in the use of LLMs for scientific discovery.

[1] https://huggingface.co/datasets/spiralworks/openreview_wildcard_2025_v2

Related Work

Large language models (LLMs) have been increasingly explored as tools to support creative ideation in scientific research. Early efforts such as Meta's Galactica utilized specialized scientific corpora to facilitate technical tasks like knowledge retrieval and hypothesis generation (Taylor et al. 2022). Despite strong benchmark performance, Galactica encountered significant issues with factual grounding and hallucinated references, highlighting the need for improved reliability.

Addressing grounding explicitly, SciPIP integrated semantic retrieval and citation awareness into idea generation, producing hypotheses informed by current literature but primarily evaluated within NLP domains (Wang et al. 2025). Extending this further, SciMON employed iterative refinement to systematically maximize the novelty of ideas, but their outputs often suffered from low practical feasibility and experimental utility (Wang et al. 2024).

Drawing inspiration from collaborative research processes, VirSci orchestrated multiple LLM agents within a virtual ecosystem to simulate team-based ideation, achieving measurable improvements in idea diversity and research-trend alignment (Wang and Clark 2024). Nevertheless, like SciMON, VirSci primarily focused on novelty without directly optimizing for experimental usefulness.

Other efforts aimed at automating entire research cycles include Sakana AI's AI Scientist, which autonomously generates and experimentally validates hypotheses but still requires human oversight for accuracy and consistency (Lu et al. 2024). Similarly, IntologyAI's Zochi employed multi-agent collaboration to produce workshop-accepted research papers across domains, yet it involves significant human intervention and computational resources (Intology 2025).

In this paper, we present SPARK, a demonstration aimed at grounding scientific idea generation in foundational computational creativity principles. We illustrate this approach

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

by integrating literature retrieval, idea generation, and simulated peer-review evaluation into a unified workflow. Additionally, we invite further exploration from the CC community by releasing the OpenReview dataset used to train the Judge model.

Methods

The SPARK system integrates AI-driven components that support computational creativity in the generation of scientific ideas, utilizing specialized agent-based modules for the retrieval of literature and the synthesis, refinement and evaluation of ideas. Figure 1 illustrates the interaction between the SPARK and XPLOR systems.

Literature Retrieval via XPLOR

Inspired by PaperQA (Lála et al. 2023), XPLOR handles literature retrieval and analysis tasks tailored specifically for the SPARK system. It utilizes OpenAI embeddings, FAISS indexing, maximal marginal relevance (MMR), and a recursive agent-based retrieval loop to systematically gather, analyze, and refine relevant research documents to later be used for idea generation.

Text Embedding and Retrieval: XPLOR employs OpenAI's `text-embedding-3-large` model (OpenAI 2024) to convert retrieved documents into 1536-dimensional semantic vectors. These embeddings allow accurate semantic comparison of research papers. FAISS (Douze et al. 2024) efficiently indexes these vectors, enabling rapid retrieval through cosine similarity-based nearest-neighbor searches when provided with user-generated input queries.

Relevance Scoring and Maximal Marginal Relevance: After an initial vector-based retrieval, each candidate text chunk is also evaluated using an LLM-based summary model. This model assigns relevance scores (1 to 10) to each chunk according to its relevance to the research query, complementing the earlier cosine similarity measures. Subsequently, XPLOR applies maximal marginal relevance (MMR) to re-rank these scored chunks, a technique which prioritizes both relevance and diversity. This ensures that selected documents collectively provide comprehensive coverage with minimal redundancy.

Recursive Retrieval and Evidence Gathering Loop: The retrieval process in XPLOR is recursive and agent-driven. Initially, the system queries databases (e.g., arXiv, Google Scholar) using a programmable search engine, retrieving documents relevant to the initial question. First, documents are segmented into overlapping chunks and embedded. Then, an agentic loop iteratively refines the search by generating new, increasingly precise queries informed by intermediate summaries from an LLM response. This process continues until specific depth and coverage criteria, such as accumulating at least five evidence snippets from multiple sources, are met. XPLOR has two distinct operating modes: interactive and autonomous. In interactive mode, XPLOR is directly prompted by the user query, whereas in autonomous mode, XPLOR takes queries from the SPARK

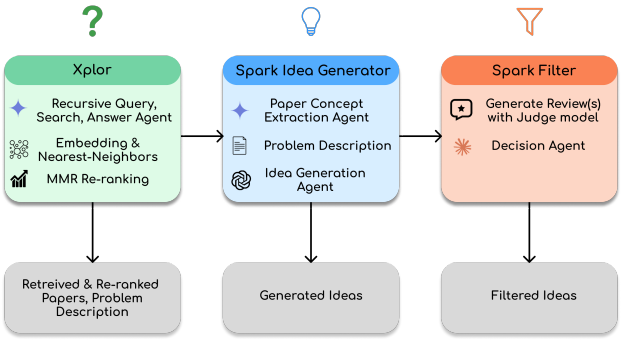


Figure 1: SPARK's end-to-end pipeline: XPLOR performs embedding-based, recursive literature retrieval with MMR re-ranking; the SPARK Idea Generator extracts key concepts and uses chain-of-thought LLM prompting to synthesize seed research proposals; and the SPARK Filter applies our supervised fine-tuned JUDGE model and a decision agent for automated peer-style critique and ranking.

Idea Generator agent to aid in the generation and refinement of ideas, as explained in the following section.

Idea Generation with SPARK Idea Generator

The SPARK Idea Generator constitutes the initial creative stage of the SPARK system, formulating seed research proposals based on the earlier XPLOR search results. The idea generation system identifies open research problems related to the query and then synthesizes concepts to generate structured, contextually relevant, and well-motivated ideas.

A) Input:

- **Problem:** "LLMs struggle with consistency"
- **Concepts:** [Chain-of-Thought (CoT), Knowledge Graphs (KGs), . . .]
- **Domain:** NLP-LLMs

B) Idea Generation Prompt:

```
Generate a novel research idea with step-by-step
reasoning based on the provided input. Output must
follow the following template.
```

C) LLM Response:

```
{ "input_concepts":["CoT","KGs"],
  "new_concepts":["Entity linking"],
  "plan":"Analyze. Wait. Try a new perspective.
          Propose KG lookups",
  "title":"KG-CoT for Fact Consistency",
  "abstract":"Integrate KGs for dynamic grounding" }
```

D) Generated Idea: Title and Abstract from the structured response in (C).

Figure 2: Spark Idea Generation pipeline: (A) Inputs, (B) Idea Generation Prompt, (C) LLM Response, (D) Generated Idea.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

A) Input:

- **Title:** "KG-CoT for Fact Consistency..."
- **Abstract:** "Integrate KGs for dynamic grounding..."
- **Reviews:**
  - Review 1: "Critique: Promising KG-CoT..."
  - Review 2: "Strengths: Addresses key issue..."
  - Review 3: "Concerns: Novelty vs prior..."

B) Filter Prompt:

- Generate your decision based on the provided input Idea and Reviews.
- Output must follow the following template:
  - `"Decision reasoning"` (string)
  - `"Decision"` (ACCEPT/REJECT)
  - `"Utility"` (float 0–1)

C) LLM Response:

```
{ "Decision reasoning": "Detailed critique...",
  "Decision": "REJECT",
  "Utility": 0.35 }
```

D) Generated Decision: `REJECT`.

Figure 3: Spark Filter pipeline: (A) Input, (B) Filter Prompt, (C) LLM Response, (D) Generated Decision.

Concept Extraction and Problem Synthesis: Initially, a user provided research question guides XPLOR to recursively retrieve relevant papers and identify open research problems. The concepts and open problems from these papers are then used to construct a problem-driven research prompt (Figure 2, A-B).

Structured Idea Formulation: Using the constructed prompt, an LLM agent generates a preliminary research proposal comprising a structured reasoning plan, novel concepts, and a concise title and abstract, guided to incorporate relevant domain knowledge and maintain logical coherence (Figure 2, C-D). This output (as illustrated in Figure 4) forms the basis for high-quality research ideas already contextualized within existing literature and ready to be explored by users of the system. If further refinement is needed, the system adaptively uses XPLOR to retrieve more relevant literature to address a specific area of improvement.

Idea Evaluation with SPARK Filter

SPARK Filter acts as the evaluative module within the SPARK framework, simulating a scalable peer review process to assess proposed research ideas' viability and quality.

Review Generation: In this phase, each idea—represented by its title and abstract—is processed by our supervised fine-tuned (SFT) JUDGE model trained on OpenReview peer review corpora. This model generates multiple critiques per idea, highlighting strengths, limitations, and areas for improvement, simulating the diverse perspectives that might be offered during the scientific peer review process.

Decision Synthesis: Subsequently, a separate LLM integrates these reviews to produce a final binary classification—`ACCEPT` or `REJECT`—of the idea, an explanation of the decision, and utility score for the idea (Figure 3, A-D).

This pipeline has successfully produced and filtered over 10,000 AI-focused research ideas, demonstrating its scalability for large-scale computational creativity tasks. In future work, it is worth exploring techniques to enhance the creativity and diversity of SPARK's generated ideas (Si, Yang, and Hashimoto 2024; Wang et al. 2024).

Evaluator Training: The JUDGE Model

Assessing the creativity and novelty of AI-generated scientific ideas demands a rigorous evaluator. Off-the-shelf LLMs, trained via supervised fine-tuning (SFT) or reinforcement learning from human feedback (RLHF), tend to prioritize agreeableness over critical scrutiny (Ouyang et al. 2022), making them ill-suited to judge nascent research proposals. To overcome this, we introduce JUDGE, a purpose-built model trained to critique scientific ideas according to established academic standards.

Data Acquisition and Preprocessing We sourced training data from OpenReview's publicly available submissions, peer reviews, and meta-reviews. An automated pipeline harvested abstract–review pairs to form the core dataset. Although we initially explored mapping full papers to their reviews, this blended conceptual appraisal with evaluations of implementation and results. Focusing on abstracts alone isolates and sharpens the model's assessment of underlying ideas.

A further challenge arises from the content of abstracts themselves. Authors often highlight strong empirical results, which can positively bias reviewer scores irrespective of the underlying idea's intrinsic novelty or merit. To mitigate this inherent bias, we introduced a data transformation step. We define an "idea abstract" ($\mathcal{A}_{idea}$) as a version of the original abstract ($\mathcal{A}_{orig}$) that deliberately omits specific results, implementation details, and performance metrics, while preserving the core problem statement, proposed approach, and claimed novelty. Similarly, an "idea review" ($\mathcal{R}_{idea}$) represents a critique focused exclusively on these conceptual aspects, abstracted away from the empirical outcomes presented in the original review ($\mathcal{R}_{orig}$).

To generate this refined dataset at scale, we employed DeepSeek-V3 as an automated annotator (Liu et al. 2024). Given an $(\mathcal{A}_{orig}, \mathcal{R}_{orig})$ pair, DeepSeek-V3 was prompted to perform two separate transformations: generating $\mathcal{A}_{idea}$ from $\mathcal{A}_{orig}$, and generating $\mathcal{R}_{idea}$ from $\mathcal{R}_{orig}$. The choice of DeepSeek-V3, an open-weight model, over potentially stronger proprietary models was made explicitly to enhance the reproducibility of our data generation process.

Multi-Task Training Framework The JUDGE model was trained using a multi-task objective designed to capture both evaluative and generative aspects of scientific discourse. We formulated four distinct tasks, using specialized input formats and system prompts to differentiate them.

1. **Original Review Prediction:** Mapping an original abstract to its corresponding original review, $f_{eval} : \mathcal{A}_{orig} \rightarrow \mathcal{R}_{orig}$.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Entropy Reduction and Intrinsic Motivation in AI Music Creation

This research explores the simulation of human music creation using AI, focusing on incorporating intrinsic motivation and addressing the potential impact on human artists. We propose a system based on a Transformer network trained with reinforcement learning to generate music. The core idea is to model musical creativity as a process of 'restructuring high psychological entropy material,' where entropy represents the complexity and unpredictability of musical elements. The AI is trained with an intrinsic reward function that encourages it to reduce the entropy of its generated music, quantified using Kullback-Leibler Divergence between probability distributions of musical features before and after transformations. This reward is shaped to encourage exploration and discovery of new musical structures. The system's performance will be evaluated by comparing its generated music to human-composed music using perceptual audio features (MFCCs, chromagrams) and subjective ratings through metrics like cosine similarity and t-SNE visualization. Furthermore, we will conduct a survey to assess how exposure to the AI-generated music affects human artists' motivation and creative output, measuring changes in their creative output and self-reported motivation levels. This research aims to provide insights into the computational modeling of creativity and address the ethical and societal implications of AI in the arts. It seeks to understand how AI can be a tool that both generates and inspires music.

Figure 4: A research idea proposed by SPARK, using context provided by XPLOR about the use of LLMs for scientific creativity.

2. **Idea Review Prediction:** Mapping an idea abstract to its corresponding idea review, $f_{idea_eval} : \mathcal{A}_{idea} \rightarrow \mathcal{R}_{idea}$.
3. **Original Abstract Generation:** Mapping an original review back to its corresponding original abstract, $f_{gen} : \mathcal{R}_{orig} \rightarrow \mathcal{A}_{orig}$.
4. **Idea Abstract Generation:** Mapping an idea review back to its corresponding idea abstract, $f_{idea_gen} : \mathcal{R}_{idea} \rightarrow \mathcal{A}_{idea}$.

Training encompassed these forward (evaluation) and backward (generation) mappings concurrently. This multi-task approach encourages the model to develop a deeper, bidirectional understanding of the relationship between a research idea and its critical appraisal.

The model was trained via continued pretraining on this structured dataset, employing low-rank adaptation (LoRA) to efficiently update the base model's weights (Hu et al. 2022). The training objective was to maximize the log-likelihood of the target sequence (either a review or an abstract, depending on the task) given the input sequence, formally represented as maximizing $\mathcal{L}(\theta) = \sum_{(x,y)} \log P(y|x;\theta)$, where (x, y) represents an input-output pair from one of the four tasks, and θ denotes the model parameters.

Evaluation and Model Scaling To assess the performance of JUDGE, we used a held-out temporal partition of the OpenReview dataset, comprising reviews and associated metadata submitted after a predefined cutoff date (e.g., post-October 2024 or from 2025 onwards). Evaluation metrics included the quality of generated reviews (assessed qualitatively and via RMSE of the final review scores in aggregate) and the accuracy of predicting explicit review scores where available.

To investigate the impact of model scale on evaluation capabilities, we trained two versions of JUDGE:

- `wintermute-tiny`: Based on the Qwen-7B model.
- `wintermute-medium`: Based on the Qwen-72B model.

This provides a roughly 10× increase in parameter count, allowing for analysis of scaling effects. Empirical validation confirmed that both `wintermute-tiny` and `wintermute-medium` consistently outperform their respective base models and other contemporary state-of-the-art LLMs on the designated evaluation tasks, demonstrating the effectiveness of our specialized training approach.

Discussion and Limitations

The SPARK prototype integrates literature retrieval, multi-agent idea generation, and a simulated peer-review filter into a single workflow intended to illustrate how computational-creativity principles can be operationalized in scientific ideation. Each stage is deliberately modular: XPLOR provides a retrieval layer, the Idea Generator converts retrieved context into candidate hypotheses, and the Filter simulates a scientific peer review step. By decoupling these stages, the system offers a test bed for exploring how generation and evaluation strategies interact within the context of scientific idea generation.

Nevertheless, several design choices merit further exploration and refinement. First, the current filter operates only once per idea; future studies could examine closed-loop variants that feed reviewer feedback back into the generator. Second, while XPLOR's retrieval is similarity-based, structured knowledge sources (e.g. ontologies) might support deeper analogical reasoning. Third, the present peer-review simulation produces textual critiques but no explicit explanatory traces; richer explanation outputs could help users understand the basis of automatic judgments. Finally, SPARK has been evaluated only in a controlled setting confined to idea generation within the field of AI, but to understand its abilities across other disciplines, studies in additional scientific fields are required.

In summary, we hope this proof-of-concept can serve as a guide for future work in this area, encouraging comparative studies of retrieval methods, generation prompts, reviewer-agent designs, and creativity evaluation strategies using our system as a foundation. To facilitate community experimentation, we have released the annotated dataset, consisting of papers and reviews, used to train the JUDGE model.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Looking Ahead: Computational Creativity and Scientific Idea Generation

The use of AI systems to generate scientific ideas raises important considerations that the CC community is well-poised to address. We state some open questions on this topic below, with the hope of inspiring other CC researchers to begin exploring these issues in future work.

Creative Criteria for Ideas Versus Papers

Creativity is typically measured by three dimensions: novelty, utility, and surprise (Boden 2004; Simonton 2004; Maher 2010). While the utility of a scientific paper is mostly judged by its future impact—whether that be citation count, accolades like best paper awards, and other retrospective metrics—idea generation and evaluation systems such as ours are tasked with finding high-quality directions that merit further exploration before their research outcomes are ever observed. How should CC researchers approach utility evaluations for scientific ideas? We believe a combination of theory-guided and data-driven approaches may be most fruitful here, with our dataset release helping facilitate the latter.

Transformative Idea Generation

Similarly, creative processes are typically divided into three types: combinatorial, exploratory, and transformational (Boden 2004). Our idea generation system relies on concept extraction and problem synthesis to construct well-motivated and contextually relevant ideas, often through the combination of existing concepts, as outlined in Figure 2. Nevertheless, many historically groundbreaking discoveries have been transformational, actively restructuring elements of the conceptual spaces afforded by their scientific field. Can AI systems be used to discover transformational scientific ideas? Considering the historical impact transformational creativity has had on science and society at large, we believe this direction merits considerable future exploration.

Conclusion

This paper presented SPARK as a system demonstration that unifies a literature retrieval engine, an LLM-driven idea-generation module, and a peer-review-inspired filter. The implementation and discussion in this work are offered as a concrete step towards grounding scientific idea generation in foundational computational creativity principles. By illustrating how creative generation and evaluation can be combined, we aim to inspire new directions in the computational creativity community around the generation, evaluation, and feasibility of scientific idea generation using LLM systems. We anticipate that the framework, data resources, and observations reported here will serve as a starting point for future work exploring the relationship between computational creativity and scientific idea generation.

References

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The faiss library. *CoRR* abs/2401.08281.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1(2):3.

Intology. 2025. Zochi technical report. *arXiv*.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with alphafold. *nature* 596(7873):583–589.

Lála, J.; O'Donoghue, O.; Shtedritski, A.; Cox, S.; Rodriques, S. G.; and White, A. D. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Langley, P. 1977. Bacon: A production system that discovers empirical laws. In *IJCAI*, 344. Citeseer.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st International Conference on Design and Semantics of Form and Movement (DESIRE)*, 22–28. ACM.

OpenAI. 2024. text-embedding-3-large. `https://openai.com/index/new-embedding-models-and-api-updates/`. Accessed: 2025-04-25.

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. *CoRR* abs/2203.02155.

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*.

Simonton, D. K. 2004. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A large language model for science.

Thagard, P. 1988. *Computational philosophy of science*. MIT press.

Wang, T., and Clark, P. 2024. Virsci: Virtual scientists for ai-driven research. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024. Scimon: Scientific inspiration machines optimized for novelty.

ISBN: 978-989-54160-7-3

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 279–299.

Wang, W.; Gu, L.; Zhang, L.; Luo, Y.; Dai, Y.; Shen, C.; Xie, L.; Lin, B.; He, X.; and Ye, J. 2025. Scipip: An llm-based scientific paper idea proposer.

Yang, Z.; Du, X.; Li, J.; Zheng, J.; Poria, S.; and Cambria, E. 2023. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*.

How far afield should you go when being creative? Semantic area as a metric of AI's effects on creative ideation

Steven R. Rick[1], **Jennifer L. Heyman**[1], **Pablo Paredes**[2], **Matthew Hong**[2], **Thomas W. Malone**[1]
[1]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Toyota Research Institute, Los Altos, CA, USA
{srick, heymanj, malone}@mit.edu
{pablo.paredes, matt.hong}@tri.global

Abstract

Generating creative ideas that are both novel and useful is one of the most important goals of any innovation process. But most measures of creativity focus only on the output of a creative process, not on the process itself. Here, we present a new metric for objectively evaluating the creative process itself. The metric, called semantic area, combines semantic embedding models, dimensionality reduction, and 2D area calculations to measure the semantic space covered during a creative task. We then use this method to quantify the relationship between what experimental participants input, see, and select during a creative task and the results of their creative efforts. We find that people who see a more diverse set of ideas (a) select more diverse ideas for their results and (b) have more subjective satisfaction with the results. But, surprisingly, they also select ideas that objective evaluators find less innovative, inspiring, and enjoyable. We close with a discussion of how this method might support interventional approaches that are able to help constructively expand a user's exploration space without distracting them from end goals during a creative process.

Introduction

Creativity as a concept is a complex topic, often referenced when it comes to generating novel ideas or imagining things that do not yet exist. While it is true that creativity is often centered around the idea of originality and novelty when it comes to generating something, it is important to strike a balance between the novelty and the utility of artifacts being created. This is especially critical in spaces where creativity intersects human work. In the work of producing art, creativity tends to intersect with utility by creating surprise or other emotions in the viewer (Tomas 1958) (Tillander 2011) (Filipowicz 2006). In the work of innovation, creativity tends to intersect with utility as it needs to produce something novel that also addresses a real need within an organization or in the marketplace (Amabile 1988) (Yusuf 2009) (Hughes et al. 2018).

It is critical for individuals and organizations to foster and support creativity, but it is often difficult to measure and quantify. For this reason there has been a sizable body of work that has attempted to create tests, scales, and assessment protocols for evaluating creativity (Cropley 2000) (Silvia et al. 2008) (Said-Metwaly, Van den Noortgate, and Kyndt 2017). While these tests are useful for evaluation of the final state of a process, i.e. the final artifact produced, these tests do not give a means for assessing the creative process as a person goes through it. This shortcoming is due in large part to the time consuming nature of evaluation by expert reviewers and the inability for such an approach to scale up very well.

Here we propose a new computational methodology for evaluating the creative process, leveraging the same advances in artificial intelligence and natural language processing that have brought about Large Language Models (LLMs), specifically looking at the embedding models needed to make LLMs work. We situate this work in the domain space of creative product design and show how it not only produces results which link the creative process to the outcome artifact, but also show how it can be used to make the process more transparent, supporting the creation of interventional systems in the future.

Background

Prior work has explored the idea that we might be able to describe creativity through computational means. Much of the earlier work started by trying to classify content as creative or not. Ritchie identified 14 criteria for quantifying novelty (Ritchie 2001) and then later refined them to encompass novelty, quality, and typicality for identifying creative output (Ritchie 2007). Pereira applied Ritchie's criteria in order to quantify novelty of artifacts across evaluation of several systems (Pereira et al. 2005). Peinado et al. focused efforts on evaluating novelty through the lens of reuse (Peinado et al. 2010). Kuznetsova et al. explored this idea further by looking at how language could be modeled as creative or not through the use of lexical composition. Modeling based on divergent thinking techniques, semantic latent space, and affective language allowed them to generate a dataset to classify word pairs through different creative metrics (Kuznetsova, Chen, and Choi 2013). While labeling content as creative or not could be useful, it required reducing ideas to pairs of words. Acar and Runco explored the role of associative distance in a divergent thinking task in order to qualify whether ideas were proximal, i.e. considered remote (distant) vs. close (Acar and Runco 2014). By assessing associative distance against perceptions

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

of creative attitudes and values, they found remote associations were significantly correlated with creativity, at least in a divergent thinking test. Beketayevab and Runco later explored whether semantic based approaches, in addition to traditional flexibility and originality assessments, could be used to score a divergent thinking test (Beketayev and Runco 2016). They found that semantic flexibility correlated well with traditional flexibility assessments, setting the stage for semantic algorithms to potentially play a role in computational assessment of divergent thinking.

As computational capabilities expanded and became more accessible, prior work began to explore more and more methods for machine learning to play a role in the assessment of creativity. Franceshelli and Musolesi explored the potential for generative deep learning approaches to be used to evaluate facets of creativity such as value, novelty, and surprise (Franceschelli and Musolesi 2022). While this application of deep learning provided an early unsupervised approach (without human involvement), it heavily relied on a training set of data to compare test sets against and tended to restrict wider applicability when working on wide reaching creative tasks. Fan et al. then explored the relationship between creativity in writing tasks and semantic distance (as measured by a trained embedding model using Word2Vec) (Fan et al. 2023). While reliable relationships between semantic distance and creativity were found and the methods allowed for more general applicability that was not overly contrained to a specific training dataset, Word2Vec requires training on a large corpus of content in order to translate textual content into learned word associations (Mikolov et al. 2013). Finally Johnson et al. explored the use of pre-trained embedding models (BERT) to generate semantic distance calculations and evaluate divergent creativity (Johnson et al. 2023). Their use of BERT is the first instance we find of the transformer architecture being used to evaluate creativity, which amplifies the contextual sensitivity of the model due to its self-attention mechanism (Vaswani et al. 2017).

Reflecting upon this prior work, we see two consistent trends: (1) that most methods focus on evaluation of a final artifact, and (2) increasingly large training datasets are necessary to evaluate those artifacts. While these trends are generally fruitful for the field of computational creativity, it is important to find new ways to better support the quantification and qualification of creativity that can consider artifact generation through a sequential process of intermediaries that lead to a final artifact, as well as methods that are not heavily dependent upon large corpora of training data.

Methods

In order to address these gaps identified in prior work, we developed a method for quantifying the 'space' covered during a creative task, a term we name *Semantic Area*. Semantic due to the fact that we leverage embedding models in order to get a contextually aware high-dimensional vector representation of text, and Area as we procedurally reduce that high-dimensional data down to a 2D plane in order to quantify how much space is covered by the set of ideas captured in textual form.

Semantic Area

In order to work with text data, we first cast the corpus of text we want to measure into a high dimensional vector representation. For this we used Google's Universal Sentence Encoder v4 (Cer et al. 2018), but any sufficiently rich embedding model could be used. We selected the Universal Sentence Encoder as it is a pre-trained model, meaning we can use it for inference to convert text into vectors without needing to build an embedding model from our corpus first using something like word2vec. We also chose to use Universal Sentence Encoder as it adjusts the embedding vectors to normalize them based on length of input, allowing for comparison between text strings of different length.

After inferring the high-dimensional vectors of each text string (512-dimensions per text string), we then use Uniform Manifold Approximation and Projection (UMAP) to cluster the data based on cosine similarity and reduce the high dimensional data down to a simpler 2D representation. (McInnes et al. 2018). Specifically, we used the following parameters:

```
UMAP(random_state=12345, n_components=2,
    metric="cosine").fit(vector_list)
```

where `vector_list` contains the array of embedding vectors returned from Universal Sentence Encoder.

UMAP was chosen over other dimensionality reduction methods (like t-SNE) and other clustering methods (like K-means) as UMAP is guaranteed to produce the same dimensionality reduction and clustering, deterministically, if given the same data and parameters. The other methods mentioned have more stochastic behavior leading to random differences between executions, reducing reproducibility. We explicitly set a `random_state` value for this reason, as without that UMAP will employ stochastic optimizations to more quickly cluster and run dimensionality reduction, which we want to avoid for the sake of reproducibility at the cost of a more computationally demanding method.

This dimensionality reduction and clustering allows for the text strings to be represented as points in a 2D plane. While possible to operate in 3D or higher dimensionsion spaces, we chose to work in 2D as this was more intuitive due to the metaphor of 'area' being much simpler than 'volume' or other high dimensional representations. By converting each unique idea or concept in the dataset into a 2D coordinate, we can now begin to analyze semantic area. Using a set of points, for example all the ideas generated by a person or system, we compute a convex hull (the minimal polygon that fits that set of points). We calculate this hull using the quick hull algorithm implementation in SciPy (Barber, Dobkin, and Huhdanpaa 1996) (Virtanen et al. 2020). We then identify the simplicial facets of the convex hull (the points that make up the smallest polygon that contains the set of points), and use the shoelace formula (also known as the Surveyor's Area formula) in order to calculate the area of that polygon (Braden 1986) (Lee and Lim 2017).

The complete Semantic Area calculation pipeline, converting text to embedding vectors, then running dimensionality reduction, and finally calculating a convex hull is rep-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

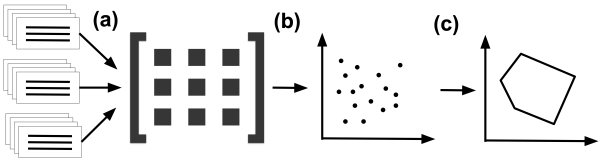


Figure 1: The Semantic Area Pipeline. (a) Text is converted to high-dimensional vectors with an embedding model, (b) the embedding vectors are clustered and dimensionally reduced with UMAP, and (c) a convex hull is calculated to quantify Semantic Area.

resented in Fig 1.

Evaluating Creative Ideation with Semantic Area

In order to use the Semantic Area method to quantify the different effects of ideation interventions on creativity, we obtained a copy of the dataset from the DesignAID study (Cai et al. 2023). We selected this study's data for two core reasons: 1) it was an empirical analysis of people using software to complete a creative task, and 2) the experimental design was a 2x2 factorial where participants were exposed to both a 'direct input mode' and 'new idea mode' during ideation. This meant that the system being used exposed people to both what they directly requested from the system as well as wholly new concepts and ideas based on what they requested from the system. This was of particular interest as we felt the Semantic Area method would be especially useful in capturing the effects of exposure to new ideas on creative task outcomes.

We also received copies of the evaluations that were run, being both subjective and objective evaluations about innovation, fit, inspiration, and enjoyment/liking of the final artifacts produced. Specifically, the participants completed a survey after completing each mood board asking about the extent to which they thought their experiences and final artifacts were innovative (1-5, $M = 3.40, SD = 1.05$), fitting of the prompt of "a chair for children" (1-5, $M = 3.90, SD = 1.11$), inspirational (1-7, $M = 5.01, SD = 1.59$), and enjoyable (1-7, $M = 5.27, SD = 1.75$). Similarly, after all mood boards were created, they were presented to a separate group of evaluators who rated those artifacts based on their level of innovation (1-5, $M = 4.19, SD = 0.88$), fit (1-5, $M = 4.34, SD = 1.32$), inspiration (1-7, $M = 3.65, SD = 0.77$), and liking (1-7, $M = 4.00, SD = 0.77$).

Data preparation

As the Semantic Area method is centered around two-dimensional area calculations, we first subset the DesignAID dataset for users who provided at least 3 unique inputs in order to have a minimum of 3 points available for running area calculations (N = 111, from the original 115). From these 111 users, 9106 text strings were collected across the input text, seen text, and selected text. 5058 of those were unique, with 1093 coming from users in the Direct Input mode and 3965 coming from users in the Diverse (New Idea) mode, with 80 overlapping (found in both sets). For simplicity we will refer to direct input mode as Direct and new idea mode as Diverse. After the data was clustered and dimensionally reduced by UMAP, we had 2D coordinates available for each unique string in the dataset. A 2D representation of the data is shown in Fig 2.

Calculating and Using Semantic Area

After reducing the textual data to a 2D coordinate space, we explore the power of Semantic Area to describe a person's creative journey. By selecting the set of texts that a user input to the system, texts that a user saw from the system, and texts that a user selected from the system, we can visualize the area covered by these three categories of data. An example of this is shown in Fig 3. When a user was in Direct input mode they would have been restricted to only see and select from the text that they input, as no new ideas would have been created by the system. When in New Idea mode their input would have directed the system to generate new ideas with similar context to their input, meaning they would see both their own inputs and the new ideas generated by the system. It is important to note that Seen always encapsulates both Input and Selected as the superset of all texts.

While running the Semantic Area calculations across all data for a user can show the total Semantic Area, we can also measure and visualize how that area changes over time. We do this by procedurally adding one new point at a time based on the temporal sequence in which a user encountered the idea. We recompute the convex hull at each time step to determine how the semantic area changes with each new idea. Fig 4 shows this with the same example data that was used in Fig 3, showing how the *Seen* area changes with each new text.

Results

Having now introduced the Semantic Area methodology and demonstrated examples of what the method can enable us to see, we scale up our analysis to report on the relationships found in the DesignAID dataset. As a reminder, the DesignAID study instructed individuals to create two moodboards depicting a "chair for children" using text input to generate images. All participants were exposed to two treatments: the Diverse (New Idea) mode where the input they provided was semantically expanded by a Large Language Model, and the Direct input mode where the input they provided remained unchanged. Those texts were then used to create images and participants designed their final moodboards by selecting from the images they produced during the task.

Semantic Area and Mode of Interaction

Before conducting any analysis, we pre-registered a hypothesis on OSF[1]. In order to understand the relationship between the texts a user input, saw, and selected from across the experimental treatments, we plot each user's data, showing how the semantic area of input leads to the semantic area

[1]https://doi.org/10.17605/OSF.IO/KM2ZQ

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

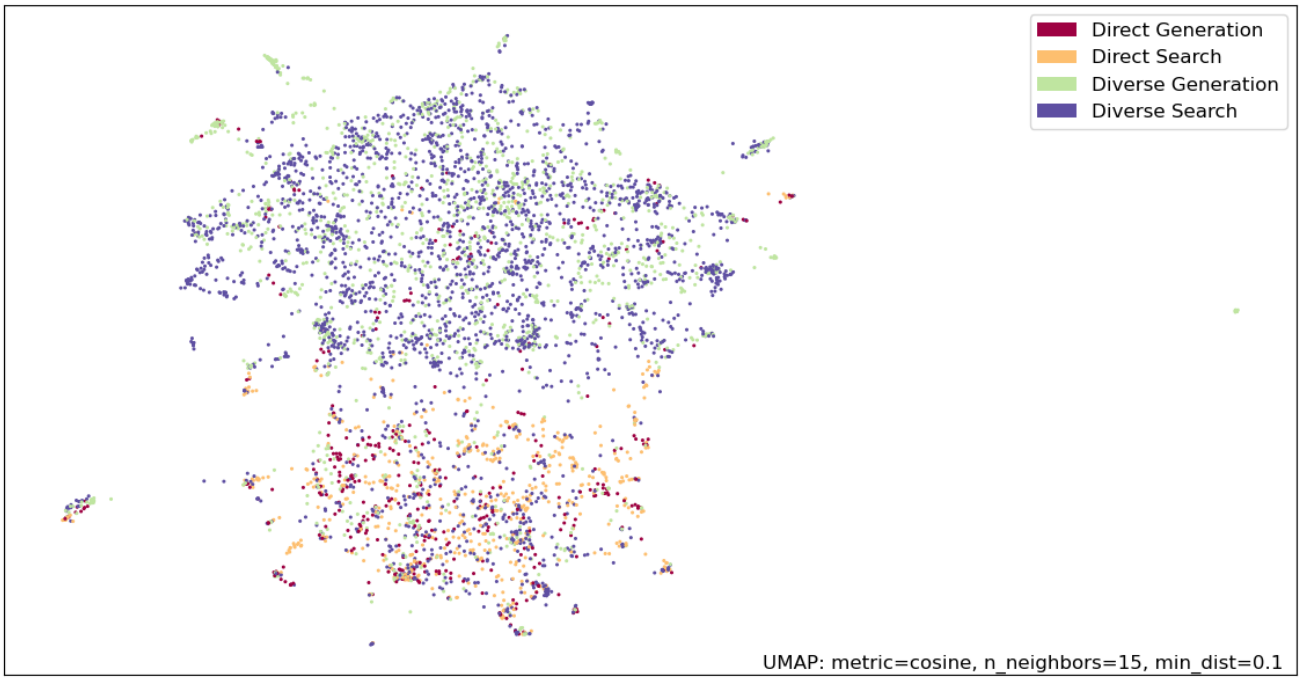


Figure 2: The UMAP dimensionality reduction and 2D clustering of the text strings from the DesignAID dataset

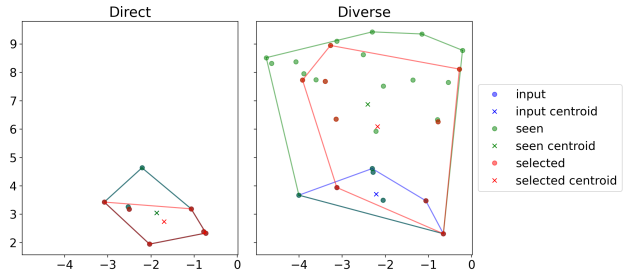


Figure 3: An example of a user's Semantic Area across what they input, what they saw, and what they selected. The Direct input mode is on the left, and Diverse (New Idea) mode is on the right. All coordinates of all ideas are plotted as points, the convex hull is computed and drawn with lines, and the centroid is calculated and shown with an X.

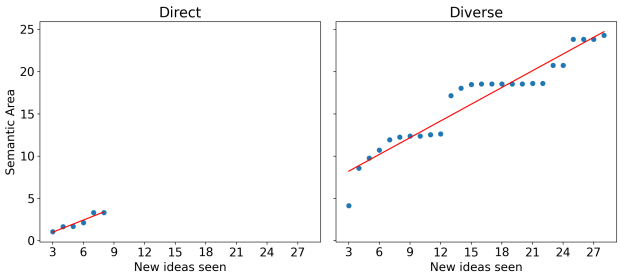


Figure 4: An example of a user's Semantic Area changing over time. Direct input mode is on the left, and Diverse (New Idea) mode is on the right. Each point is the area at that step in the creative process (adding new 'Seen' ideas). The red line is the least squares regression of the data (line of best fit).

of seen ideas and finally to the semantic area of selected ideas. This is shown in aggregate in Fig 5. It is worth noting that while users all started from different input areas, those in the Diverse condition (where they were exposed to New Ideas) saw area values increase when transitioning from input to seen ($M_{Diff} = 21.16$), whereas those in the Direct input conditions did not see a significant change between input and seen areas ($M_{Diff} = .02$). This makes intuitive sense given that the Diverse intervention was explicitly trying to expose the end user to a wider range of ideas. A first finding surfaced that, on average, users in the Diverse states selected items that produced a larger semantic area than the set of inputs they provided ($M_{Diff} = 5.63$). Users in the Direct input conditions always selected ideas that produced a smaller semantic area than the area of the inputs they had provided ($M_{Diff} = -1.72$). This is shown by the average value of each condition (shown as a red line on the charts).

We found that there were significant differences between the Diverse input conditions and the Direct input conditions when examining input area ($t = -2.04$, $p = .04$, $M_{Diff} = -1.32$), seen area ($t = -19.44$, $p < .0001$, $M_{Diff} = -22.63$) and selected area ($t = -8.54$, $p < .0001$, $M_{Diff} = -8.54$). Specifically, the semantic areas for the Diverse conditions were significantly greater than the semantic areas for the Direct conditions (Fig 6).

This directly confirms the hypothesis we originally regis-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

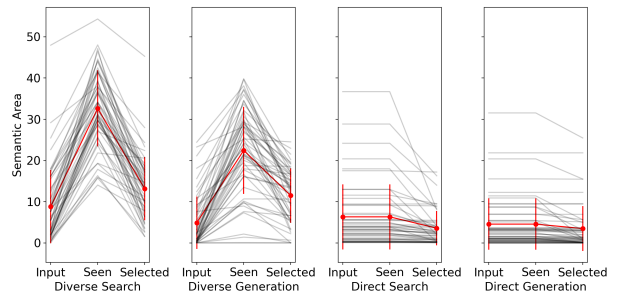


Figure 5: Aggregate chart plotting the input area, seen area, and selected area for each participant grouped by treatment condition. Those in the Diverse conditions always increased from input to seen. They also, on average, increased from input to selected. Those in the Direct conditions always decreased area from input to selected. Average values are shown with red lines.

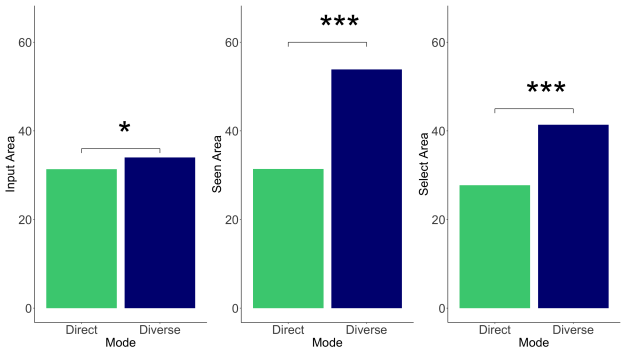


Figure 6: Within-Subjects T-Test Examining the Differences Between Input (left), Seen (middle), and Select (right) Semantic Areas Between Low (green) and High (blue) Diversity conditions

tered and generates new questions about whether there was any relationship between Semantic Area (input, seen, and/or selected) and the subjective and objective assessments run by Cai, et al. We expand our analysis to look at the interplay between these data.

Modeling the Creative Process Over Time With Semantic Area

We begin to get a glimpse into the creative process by looking at how the semantic area changes over time as users explored an idea space and/or are given exposure to a greater number of new ideas. An example of this is shown in Fig 4.

We found that as a person explores more ideas, the semantic area increases ($b = 0.34$, $z = 57.95$, $p < .0001$). Importantly, this effect is stronger in the Diverse New Idea mode ($b = 0.29$, $z = 53.25$, $p < .0001$), although it is still significantly positive in the Direct Input mode ($b = 0.23$, $z = 12.08$, $p < .0001$).

	b	se	p
Input Area			
Subjective Measures			
Innovation	-0.02	.01	.16
Fit	0.02	.01	.17
Inspiration	-0.01	.02	.54
Enjoyment	-0.05	.02	.01*
Objective Measures			
Innovation	0.01	.01	.28
Fit	0.00	.01	.79
Inspiration	0.01	.01	.12
Liking	0.01	.01	.29
Seen Area			
Subjective Measures			
Innovation	-0.02	.01	.003**
Fit	-0.01	.01	.05*
Inspiration	-0.02	.01	.005**
Enjoyment	-0.05	.01	<.0001***
Objective Measures			
Innovation	-0.00	.00	.35
Fit	-0.01	.01	.10
Inspiration	-0.00	.00	.51
Liking	-0.00	.00	.28
Select Area			
Subjective Measures			
Innovation	-0.01	.01	.62
Fit	0.01	.01	.49
Inspiration	-0.01	.02	.64
Enjoyment	-0.04	.02	.02*
Objective Measures			
Innovation	-0.01	.01	.13
Fit	-0.02	.01	.07
Inspiration	-0.01	.01	.21
Liking	-0.01	.01	.10

Table 1: Semantic Area Predicting Subjective and Objective Measures. Seen area had significant effect on all subjective measures, while input area and select area had significant effects on subjective enjoyment. No significance was noted between areas and objective measures. *$p < .05$, **$p < .01$, ***$p < .001$

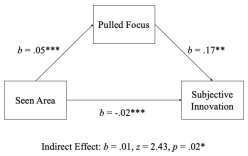
Semantic Area and Evaluated Outcomes

We conducted a series of linear regressions to examine whether the semantic area of texts that participants input, saw, or selected predicted subjective and objective measures of innovation, fit to the originally prompt, inspiration, and enjoyment/liking. Table 1 showcases the results of our analysis.

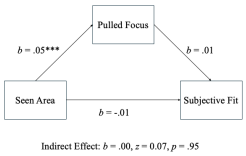
Generally, neither input area nor selected area were significantly associated with objective or subjective measures (all $ps > .07$). The exception to this was that both the input ($b = -.05$, $z = -2.50$, $p = .01$) and selected ($b = -.04$, $z = -2.37$, $p = .02$) areas were significantly negatively associated with subjective ratings of enjoyment. In other words, those who input a wider range of text prompts tended to rate their experience creating the mood boards as being less enjoyable.

When examining the association between the Seen area and the objective and subjective measures, we found that the Seen area was not significantly associated with any of the objective measures (all $ps > .1$). However, the Seen Area was significantly negatively associated with all of the subjective measures (all $ps < .05$). That is, those who saw a wider range of outputs from the system tended to rate their experiences as being less innovative, less fitting to the prompt, less inspiring, and less enjoyable.

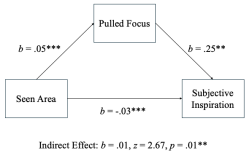
Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



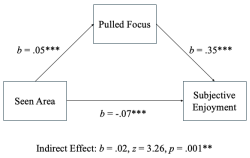
(a) Mediation Between Seen Area, Pulled Focus, and Subjective Innovation



(b) Mediation Between Seen Area, Pulled Focus, and Subjective Fit

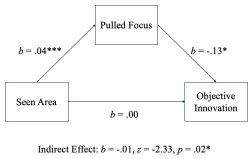


(c) Mediation Between Seen Area, Pulled Focus, and Subjective Inspiration

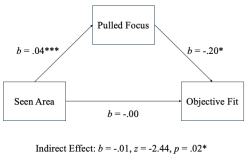


(d) Mediation Between Seen Area, Pulled Focus, and Subjective Enjoyment

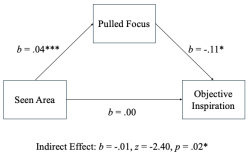
Figure 7: Mediation Between Seen Area, Pulled Focus, and Subjective Measures



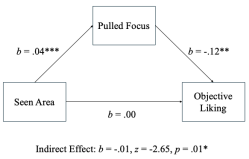
(a) Mediation Between Seen Area, Pulled Focus, and Objective Innovation



(b) Mediation Between Seen Area, Pulled Focus, and Objective Fit



(c) Mediation Between Seen Area, Pulled Focus, and Objective Inspiration



(d) Mediation Between Seen Area, Pulled Focus, and Objective Like

Figure 8: Mediation Between Seen Area, Pulled Focus, and Objective Measures

Seeing New Ideas Pulls Focus and Influences What You Select

To examine whether the ideas that one sees influences the final outcome more than the idea that one originally inputs, we calculated the distance between the centroids of the input, seen, and selected ideas. By centroid of the ideas, we mean the average value of all coordinates per category, not just the coordinates that make up the convex hull. We then conducted t-tests to examine whether the distance between the Input and Seen centroids was significantly different from the distance between the Seen and Selected centroids. If the original ideas that one inputs are more strongly associated with the final outcome, we reason that we would see a smaller distance between the Input centroid and Selected centroid as compared to the distance between the Seen centroid and Selected centroid. In contrast, if the ideas that one sees has a stronger influence on the final selected ideas, we would see the distance between the Seen and Selected centroids as being smaller than the distance between the Input and Selected centroids.

Indeed, we found that the average distance between the Seen and Selected centroids ($M = 0.82$) was significantly smaller than the distance between the Input and Seen centroids ($M = 1.53$, $p < .0001$), indicating that the ideas that one sees tend to have a stronger influence on the final output than the ideas that one originally inputs.

This implies that the ideas that one sees may tend to *pull focus* away from the ideas that one originally inputs, thus encouraging a larger semantic area of exploration to consider when selecting content for the final output. This may also have other effects upon the creative process, possibly enhancing the innovativeness of a final artifact by encouraging a wider set of concepts to be considered or, inversely, distracting the focus of a person and bringing them further away from their initial target goal.

Effects Between What People Do, What People See, and What People Select

To explore the effects of this *pulled focus* on creative process outcomes, we conducted several mediation analyses to examine the causal effect between Seen area, pulled focus, and both subjective and objective measures. That is, does greater Seen area predict greater distance between the Input and Selected centroids which, in turn, predicts subjective and objective measures related to final artifact outcomes?

When examining the subjective measures, we find a significant indirect effect between Seen area, pulled focus, and innovation ($b = .01$, $z = 2.43$, $p = .02$, Fig 7a), inspiration ($b = .01$, $z = 2.67$, $p = .01$, Fig 7c), and enjoyment ($b = .02$, $z = 3.36$, $p = .001$, Fig 7d). In other words, those who saw a wider breadth of ideas tended to select ideas that were semantically further from their original inputs which, in turn, predicted higher levels of subjective innovation, inspiration, and enjoyment. The indirect effect for subjective ratings of fit was not significant ($b = .01$, $z = 0.07$, $p = .95$, Fig 7b).

Interestingly, when we examine the objective measures, we find the opposite effect. Specifically, those who saw a wider breadth of ideas tended to select ideas that were semantically further from their original inputs, which then predicted significantly lower levels of objective innovation ($b = -.01$, $z = -2.33$, $p = .02$, Fig 8a), fit ($b = -.01$, $z = -2.44$, $p = .02$, Fig 8b), inspiration ($b = -.01$, $z = -2.40$, $p = .02$, Fig 8c), and liking ($b = -.01$, $z = -2.65$, $p = .01$, Fig 8d).

In sum, seeing a wider breadth of ideas led to selecting outputs that were significantly different from the originally inputs. This, in turn, led to a more positive subjective experience for those creating the mood boards, but a more negative objective outcome when evaluating those mood boards.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

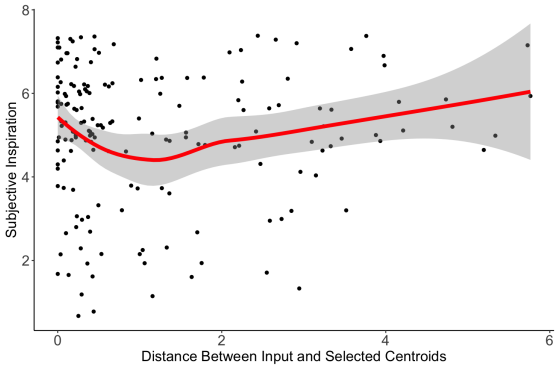


Figure 9: Non-Linear Association Between Input and Selected Centroid Distance and Subjective Inspiration

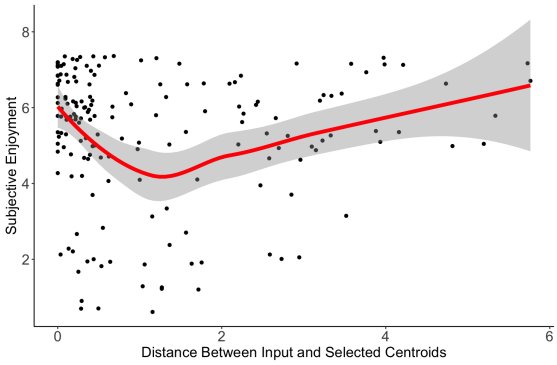


Figure 10: Non-Linear Association Between Input and Selected Centroid Distance and Subjective Enjoyment

Is There An Optimal Level of New Idea Exposure?

Given these rather surprising findings (that exposure to more new ideas produced subjective improvement and objective detriment), we asked whether there is an optimal level of new ideas to be exposed to in order to provide the most positive subjective experiences and objective outcomes. To do this, we conducted a series of quadratic regressions to examine if there is a non-linear relationship between both Seen area and Centroid Distance and our subjective and objective measures.

We found a significant quadratic association between the distance between the Input and Selected centroids when predicting both subjective inspiration ($b = .12$, $z = 2.08$, $p = .04$, Fig 9) and enjoyment ($b = .21$, $z = 3.50$, $p = .001$, Fig 10). That is, when the distance between the Input and Selected centroids is low, there is a marginal negative association with subjective inspiration ($b = -0.48$, $z = -1.72$, $p = .09$). In contrast, when the distance between the centroids is high, there is a marginal positive association with inspiration ($b = 0.17$, $z = 1.71$, $p = .09$). Similarly, there is a significant negative association between the centroid distance and subjective enjoyment when the distance is low ($b = -0.97$, $z = -3.27$, $p = .001$), although this association becomes marginally positive when the distance between the centroids is greater ($b = .19$, $z = 1.77$, $p = .09$). No other quadratic associations were significant, either when examining subjective measures ($p > .1$) or objective measures ($p > .1$).

This indicates that the benefits of diverse new ideas are highest when the selected ideas are particularly different from the originally input ideas. However, the benefits may not be as robust if the selected ideas are only moderately different from the originally input ideas. In other words, in order to reap the benefits of semantic area exploration subjectively, one must step far outside of their comfort zone and explore ideas far from where they originally started. Given our analysis, further work is required to really understand the effect of new idea exposure on objective outcomes.

Discussion

Reflecting upon the semantic area methods we developed and results of our analysis using it to quantify a creative ideation task, we find a number of surprising results. We noted that people who put more diverse texts into the system had a worse subjective experience, likely due to the higher effort required to produce new ideas from scratch. We also noted that exposure to a more diverse set of ideas can enhance subjective perception but reduces objective performance, likely due to new ideas pulling focus away from the starting location.

What You See Is What You Think

We found that exposure to a wider range of ideas (that are semantically further from one another) has mixed effects on outcomes drives forward a consideration that what you *see* influences what you *think*. In this manner, exposure to new perspectives, concepts, and ideas can broaden the creative search space. This in turn can drive up a person's feelings of innovation. At the same time, this can *pull focus* and shift what a person is attending to, leading to not staying on target within a task and producing objectively worse outcomes.

Considering the growing presence of generative AI technologies and the ability of the technology to generate seemingly endless new content, it is important we consider how these new computational abilities are designed and used to empower people. There is potential to augment with inspiration from new ideas, but also risk of distraction or overloading with too many new perspectives.

Semantic Area and Interventional Processes

Keeping the study we got our data from in mind, users were left in a perpetual state of always getting new ideas or never getting anything novel. While informative for the sake of empirical evaluation, our methods highlight the tricky balance to be struck between enough new ideas and too many. Future work with generative AI creating new ideas should consider more human autonomy in the process. By enabling users to decide when they do or do not want new ideas, they can take control of this process and steer the technology to their benefit.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

However, even this requires people to remain aware of their state in the creative process. We believe the Semantic Area method for modeling exposure to ideas over time might present a powerful new way to build dynamic creative interventions. By detecting when people are not exploring and encouraging it, or noticing when people are losing focus and pulling them back to a focal point, future systems might be able to drive people toward optimal outcomes without them needing to stop and determine the next best step. Semantic Area seems especially useful for ideation process recommendation, presenting new ways to quantify and qualify activities as divergent or convergent based on what someone has done over time, and then recommending potential next steps.

Limitations and Future Work

We acknowledge there are some limitations of the methodology and analysis. First, Semantic Area is a relative calculation, meaning the area comparisons need to be from data that has had their embedding vectors come from the same embedding model, and their clustering and dimensionality reduction need to use the same corpus of data. Because it is a relative evaluation, it is important to not consider these Semantic Area values as global measures of creativity, but rather localized measures best suited to comparison of data collected from common tasks on common themes, or for the evaluation of an individual over time.

Since the Semantic Area method uses clustering and dimensionality reduction this also means that without a starting corpora of text data, the initial clustering and reduction from embedding vectors will move around dramatically as data is collected until a somewhat steady state is reached with a larger corpus. As such, the method currently proves to be most useful in post-hoc analysis. Future work could adapt the approach to do away with dimensionality reduction as methods for higher dimensional convex hulls and volumetric proxies do exist, but at the cost of a loss of intuitive understanding of the computed value or straightforward visualization of the data.

Another important limitation of the analyses is that we did not get a comprehensive understanding of the creative process as a whole from the dataset we had access to. That is, the data we analyzed focused on objectively assessing the final artifacts produced and the overall experience reported by subjective and objective measures of innovation, fit, inspiration, and enjoyment/liking. All of this was collected after the creative process had concluded. While we can confidently report on end outcomes, it is less clear how our results fit a proper characterization of the process over time. Future work should explore more process oriented evaluations (with something like frequent checkpoint surveys) in order to connect Semantic Area more tightly with the creative journey rather than only the final artifact produced.

Conclusion

In conclusion we present Semantic Area, a new method for quantifying creativity from text embeddings, clustering, and convex hull calculations. We showcase the utility for this method to describe not just the final state a user arrives at through a creative task, but also how it might be used to characterize the process over time during a creative journey. Using data collected from a study where users were exposed to new ideas to increase inspiration and innovation, we highlight the mixed subjective and objective effects of new ideas generated by AI on creative outcomes. Specifically we show that new idea exposure can be beneficial to subjective inspiration and detrimental to objective fit on final outcomes, likely due to pulled focus. We also show that new idea exposure can be detrimental to subjective enjoyment, likely due to increased cognitive effort evaluating new concepts. New ideas can be helpful, but too many can get in the way.

We see this work as surfacing new considerations around the effects of AI on the creative process. We highlight opportunities for computational creativity at large to better design systems that support people doing creative work with AI. Augmenting how people and computers do creative work together requires us to build systems that can make sense of the creative journey and intervene in human creative work to promote innovation and inspiration without overwhelming people or causing them to lose focus.

Author Contributions

Author 1 led the development of the methodology and analysis. Author 2 helped with statistical analysis. Authors 3, 4, and 5 all helped with interpretation of results as well as research plan development. All authors wrote and reviewed the manuscript.

Acknowledgments

This research was supported in part by the Toyota Research Institute and also as part of the Collective Intelligence Mission of the MIT Quest for Intelligence.

References

Acar, S., and Runco, M. A. 2014. Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal* 26(2):229–238.

Amabile, T. M. 1988. A model of creativity and innovation in organizations. *Research in organizational behavior* 10.

Barber, C. B.; Dobkin, D. P.; and Huhdanpaa, H. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)* 22(4):469–483.

Beketayev, K., and Runco, M. A. 2016. Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's journal of psychology* 12(2):210.

Braden, B. 1986. The surveyor's area formula. *The College Mathematics Journal* 17(4):326–337.

Cai, A.; Rick, S. R.; Heyman, J. L.; Zhang, Y.; Filipowicz, A.; Hong, M.; Klenk, M.; and Malone, T. 2023. Designaid: Using generative ai and semantic diversity for design inspiration. In *Proceedings of The ACM Collective Intelligence Conference*, 1–11.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar,

C.; et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, 169–174.

Cropley, A. J. 2000. Defining and measuring creativity: Are creativity tests worth using? *Roeper review* 23(2):72–79.

Fan, L.; Zhuang, K.; Wang, X.; Zhang, J.; Liu, C.; Gu, J.; and Qiu, J. 2023. Exploring the behavioral and neural correlates of semantic distance in creative writing. *Psychophysiology* 60(5):e14239.

Filipowicz, A. 2006. From positive affect to creativity: The surprising role of surprise. *Creativity Research Journal* 18(2):141–152.

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

Hughes, D. J.; Lee, A.; Tian, A. W.; Newman, A.; and Legood, A. 2018. Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly* 29(5):549–569.

Johnson, D. R.; Kaufman, J. C.; Baker, B. S.; Patterson, J. D.; Barbot, B.; Green, A. E.; van Hell, J.; Kennedy, E.; Sullivan, G. F.; Taylor, C. L.; et al. 2023. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods* 55(7):3726–3759.

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.

Lee, Y., and Lim, W. 2017. Shoelace formula: Connecting the area of a polygon and the vector cross product. *The Mathematics Teacher* 110(8):631–636.

McInnes, L.; Healy, J.; Saul, N.; and Grossberger, L. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software* 3(29):861.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26.

Peinado, F.; Francisco, V.; Hervás, R.; and Gervás, P. 2010. Assessing the novelty of computer-generated narratives using empirical metrics. *Minds and Machines* 20:565–588.

Pereira, F. C.; Mendes, M.; Gervás, P.; and Cardoso, A. 2005. Experiments with assessment of creative systems: an application of ritchie's criteria. In *Proceedings of the workshop on computational creativity, 19th international joint conference on artificial intelligence*, volume 5, 05.

Ritchie, G. 2001. Assessing creativity. In *Proc. of AISB'01 Symposium*.

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

Said-Metwaly, S.; Van den Noortgate, W.; and Kyndt, E. 2017. Methodological issues in measuring creativity: A systematic literature review. *Creativity. Theories–Research-Applications* 4(2):276–301.

Silvia, P. J.; Winterstein, B. P.; Willse, J. T.; Barona, C. M.; Cram, J. T.; Hess, K. I.; Martinez, J. L.; and Richard, C. A. 2008. Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts* 2(2):68.

Tillander, M. 2011. Creativity, technology, art, and pedagogical practices. *Art Education* 64(1):40–46.

Tomas, V. 1958. Creativity in art. *The Philosophical Review* 67(1):1–15.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods* 17(3):261–272.

Yusuf, S. 2009. From creativity to innovation. *Technology in society* 31(1):1–8.

Computational Modeling of Artistic Inspiration: A Framework for Predicting Aesthetic Preferences in Poetic Lines Using Linguistic and Stylistic Features

Gaurav Sahu and **Olga Vechtomova**

David R. Cheriton School of Computer Science
University of Waterloo, Canada
{gaurav.sahu, olga.vechtomova}@uwaterloo.ca

Abstract

Artistic inspiration remains one of the least understood aspects of the creative process, yet plays a crucial role in producing works that resonate with audiences. This paper introduces a novel computational framework for modeling individual artistic preferences in poetic lines through key linguistic and stylistic features. Our approach consists of two components: (1) a feature extraction module that quantifies poetic imagery, word energy, abstraction level, emotional valence, and linguistic complexity, and (2) a calibration network that learns to predict what content will inspire specific individuals. To evaluate our framework, we present EvocativeLines, a dataset of poetic lines annotated as either "inspiring" or "not inspiring" across diverse preference profiles. Experiments demonstrate that our framework significantly outperforms state-of-the-art language models, surpassing LLaMA-3-70b by nearly 18 percentage points in accuracy. The framework's design prioritizes interpretability and flexibility, making it adaptable to analyzing various types of artistic preferences across different creative domains and skill levels. By formalizing the measurement of subjective aesthetic responses, our work provides a foundation for computational systems that can support the early stages of the creative process. [1]

Introduction

The creative process, along with the inspiration that drives the creation of art—whether visual, poetic, or musical—remains one of the least understood aspects of human experience. Rubin (2023) describes three distinct stages to a creative process: the Seed phase, where one is entirely open to any and all forms of inspiration; the Experimentation phase, where one tries multiple approaches to allow the nascent ideas or "seeds" to grow and take shape; and the Crafting phase, where the artist refines these ideas into a cohesive work. Our research focuses specifically on the Seed phase, investigating how AI-generated poetic lines can inspire new creative pathways for artists. At this stage, AI-generated lines can serve not as direct inputs into the final artwork but as catalysts that guide artists into a creative mindset.

Assessing creativity in AI-generated content presents significant challenges due to its subjective and individual nature. Traditional evaluation methods such as the Consensual Assessment Technique (CAT) (Amabile 1982) and the Torrance Test of Creative Thinking (TTCT) (Torrance 1966) provide structured frameworks but have notable limitations. CAT relies on expert consensus, yet fails to assess the inspirational potential of outputs or their ability to spark creative development in individual artists. Similarly, TTCT measures divergent thinking but has been criticized for not reflecting the diverse forms that creativity can take (Baer 2011).

In this work, we propose a novel framework that addresses these limitations by: **1)** Identifying and formalizing key linguistic and poetic properties of AI-generated lines that best explain artists' subjective preferences during the Seed phase; **2)** Developing a system that can adapt to the artistic preferences of a wide spectrum of creative individuals; **3)** Focusing on the inspirational quality of poetic lines rather than evaluating them as final artifacts.

A key application of our framework is the ability to personalize creative suggestions for individual artists. By modeling the specific stylistic and linguistic features that resonate with each artist, our system can filter and prioritize AI-generated poetic lines that align with that person's unique artistic sensibilities. This personalization enables more effective creative assistance by providing artists with inspiration that matches their aesthetic preferences, rather than generic suggestions that may not resonate with their creative vision. Furthermore, our framework does not assume a pre-requisite skill level of an artist and can be useful to a wide spectrum of creative individuals for any task that is inherently subjective.

To evaluate our framework, we create *EvocativeLines*, a collection of 3,025 AI-generated poetic lines. Our experiments demonstrate that our approach accurately predicts artistic preferences across diverse preference profiles, significantly outperforming recent large language models like LLaMA-3-70b.

Our contributions include: **1)** A novel framework for measuring various characteristics of AI-generated poetic lines that identifies key linguistic and stylistic features driving subjective preferences during the early creative process; **2)** An interpretable system adaptable to diverse creative individuals regardless of their artistic skill level; **3)** A personalization mechanism that filters and recommends AI-generated poetic lines based on individual artistic preferences; **4)** Experimental validation demonstrating our framework's ability to predict

[1] Project Website:
https://sites.google.com/view/modeling-artistic-preferences

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

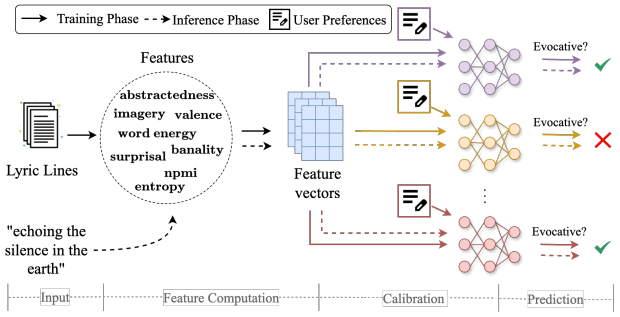


Figure 1: Framework for predicting artistic preferences in poetic lines, consisting of (1) a Feature Computation Module that quantifies linguistic and stylistic characteristics and (2) a Preference Calibration Network that learns individual artistic preferences to predict which lines will be perceived as 'inspiring'.

artistic preferences with significantly higher accuracy than state-of-the-art language models, outperforming LLaMA-3-70b by nearly 18 percentage points; **5)** Insights into the challenges of modeling diverse artistic preferences in highly subjective domains using large language models.

Related Work

Recent advancements in natural language processing (NLP), particularly in the space of LLMs, have facilitated the development of evaluation metrics, such as LLM-Rubric (Hashemi et al. 2024), GPTScore (Fu et al. 2024), and G-Eval (Liu et al. 2023), that are far better aligned with human judgments than traditional NLP metrics like BLEU (Papineni et al. 2002), ROUGE (Lin 2004), or METEOR (Banerjee and Lavie 2005); however, these LLM-based metrics primarily assess general language quality rather than creative attributes. There are also evaluation approaches that measure aspects like diversity and fluency but do not capture the multifaceted nature of creative expression. For instance, Hashimoto, Zhang, and Liang (2019) proposed an evaluation framework balancing diversity and fluency, while Pillutla et al. (2024) introduced MAUVE, an automatic measure of how close generated text is to human text.

Recently, Chakrabarty et al. (2024) extended the TTCT framework to the Torrance Test of Creative Writing (TTCW) to evaluate creativity in outputs from both humans and large language models (LLMs), but it also suffers from TTCT's limitation of being unable to capture the diverse forms creativity can take. In our work, we develop multifaceted metrics that capture various linguistic and stylistic characteristics of generated text and we also investigate how these metrics can predict personal preferences of individuals with diverse artistic skills and aesthetic sensibilities.

Our framework is informed by an aesthetic theory and creative practices. The relationship between aesthetic pleasure and complexity, known as the inverted U-shape or "Wundt" effect (Berlyne 1975), suggests that stimuli of intermediate complexity are most pleasing, an effect observed across domains including music (Gold et al. 2019) and design (Althuizen 2021). This theoretical foundation has practical implications for poetic creativity. Many artists, including David Bowie (Jones 2017) and Kurt Cobain (Cross 2001), have employed techniques like the "cut-up" method (Burroughs 2012) to create unexpected language patterns that inspire creative thoughts. Similarly, literary techniques such as Spanbauer's "burnt tongue" deliberately introduce complexity to force deeper engagement (Palahniuk 2020), an effect supported by evidence that predictable words receive less attention from readers (Lowder et al. 2018).

While the "Wundt" effect theory suggests general patterns in aesthetic responses, preferences for artistic stimuli remain highly individual (Gold et al. 2019; Althuizen 2021). This observation motivates our two-step framework that first measures multiple linguistic and stylistic characteristics of poetic lines, then uses these measurements to learn the specific artistic preferences of individual creators.

The Proposed Framework

Our framework consists of two main components: (1) a feature extraction module that quantifies key linguistic and stylistic characteristics of poetic lines, and (2) a preference calibration network that learns individual artistic preferences based on these features (see Figure 1).

The feature extraction module processes each poetic line to compute a broad set of measurements across multiple dimensions: poetic imagery, word energy, level of abstraction, emotional valence, linguistic surprisal, contextual entropy, word associations (NPMI), and banality. These features capture both the stylistic aspects that make lines evocative and the linguistic properties that contribute to their cognitive processing. Each feature vector preserves both the overall characteristics of the line and its temporal development through word-by-word analysis.

The preference calibration network then takes these feature vectors as input and learns to predict whether a given poetic line will be perceived as "inspiring" by a specific individual. This network is trained on labeled examples of poetic lines that reflect the artistic preferences of the individual. By separating feature extraction from preference modeling, our framework achieves both interpretability and adaptability to diverse artistic tastes. In the following sections, we describe each component of our framework in detail, beginning with the linguistic and poetic features that form the foundation of our approach.

Step 1: Identifying The Key Poetic and Stylistic Characteristics

Our framework quantifies both poetic/stylistic and statistical-linguistic aspects of poetic lines. We measure five creative dimensions: poetic imagery (sensory evocation), word energy (emotional impact), abstraction level (conceptual ambiguity), emotional valence (affective tone), and banality (originality of expression). Complementing these are statistical measures: linguistic surprisal (word unexpectedness), contextual entropy (prediction uncertainty), and normalized pointwise mutual information (word associations).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Our feature selection was guided by literary theory and empirical observation. While many poetic devices act together to create desired effects, a poem is foremost intended to evoke emotion in the reader and heighten perceptual awareness (Dunnigan 2014). Rather than capturing all poetic devices—which would be computationally intractable—we identified fundamental characteristics based on their theoretical importance in achieving these core poetic functions. We complement stylistic features with statistical ones derived from information theory to quantify linguistic patterns influencing cognitive processing. The following sections detail our approach to measuring each characteristic.

Poetic Imagery

Poetic Imagery is a stylistic device using references to physical objects that appeal to readers' senses, evoking emotions and making abstract ideas tangible (Pound 1913; Brooks and Warren 1976; Kao and Jurafsky 2012).To measure this feature, we prompt a LLaMA-3-70b model to output a rating between 1-5 [2]. Since poetic imagery is a well-established stylistic device in English, we perform prompting in a zero-shot fashion. This further avoids the induction of preference bias in the LLM outputs.

We compute the score for a sentence S using weighted summation:

$$score_{img}(S) = \sum_{i=1}^{5} p(r_i) \times r_i, \tag{1}$$

where $p(r_i)$ is the probability of the LLM assigning rating r_i. This provides more fine-grained scores and takes into account the model's confidence rather than directly using output tokens as scores. The final imagery feature vector combines scores of all subsequences to capture temporal development:

$$\mathbf{S_{imagery}} = \langle score_{img}(\mathcal{S}_t) \rangle_{t=1}^{T}, \tag{2}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ represents a subsequence of the first t words.

This effectively captures poetic imagery differences: "there's a little red to the sea" scores 4.01, while "this is all" only reaches 1.27.

Word Energy

Word Energy is a composite feature encompassing symbolism and diction, which contribute to a poem's emotional impact. Symbolism uses objects to evoke abstract concepts and emotions (Cassirer 1946) (e.g., "chains" representing entrapment), while diction refers to word choice for achieving specific effects (Brooks and Warren 1976) (e.g., "beseeching" versus "searching").

Similar to imagery, we prompt a LLaMA-3-70b model to rate poetic lines between 1-5 in a zero-shot fashion to obtain the score $\mathbf{S_{energy}}$. Since word energy is a composite feature, we include three example words that exemplify this concept (e.g., 'beseeching', 'chains', 'burning'). A higher rating indicates higher word energy.

[2]Examples of prompts are listed on the project website.

The final feature vector takes the same form as the imagery vector, except ratings reflect word energy rather than imagery. For instance, "all the words will drown" scores 4.11, while "where i know about you" receives only 1.92.

Level of Abstraction

We measure abstraction or ambiguity in poetic lines, as higher abstraction allows multiple interpretations and can be more engaging. The *Linguistic Category Model (LCM)* (Semin and Fiedler 1988; Johnson-Grey et al. 2020) provides a framework for measuring abstraction by assessing four word categories: Descriptive Action Verbs (DAVs), Interpretative Action Verbs (IAVs), State Verbs (SVs), and Adjectives (ADJs).

This measure captures stylistic devices like diction and represents the balance between abstract and concrete language, essential for defining tone, mood, and imagery vividness. It also indirectly indicates figurative language through abstract expressions. We compute the abstraction feature vector as:

$$\mathbf{S_{abs}} = \langle score_{abs}(\mathcal{S}_t) \rangle_{t=1}^{T} \tag{3}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ denotes a subsequence with first t words, and:

$$score_{abs} = \frac{D + (2 \times I) + (3 \times S) + (4 \times A)}{D + I + S + A} \tag{4}$$

Here, D, I, S and A represent counts of the four word categories. Weights reflect theorized abstraction levels from a study of 40,000 English words (Brysbaert, Warriner, and Kuperman 2014), with concrete verbs (DAVs) weighted lowest (1.0) and abstract adjectives highest (4.0). We use LLaMA-3-70b to obtain these counts.

Our scores distinguish between definitive phrases with lower abstraction ("rearranging this stage") and more unresolved ones ("the greatest began to deny"). Our experiments indicate abstraction level is a prominent feature in identifying potentially evocative lines.

Banality

Banality refers to common or clichéd expressions lacking originality or depth. It reflects cultural and temporal context—artists may deliberately use clichés to resonate with broader audiences or evoke certain eras. This makes banality crucial for studying diverse artistic preferences.

Similar to imagery and word energy, we prompt a LLaMA-3-70b model in zero-shot fashion to rate expressions from 1-5 for banality. Higher ratings indicate more banal expressions, while lower ratings reflect uniqueness or originality.

The banality feature vector is constructed as:

$$\mathbf{S_{banality}} = \langle score_{bnl}(\mathcal{S}_t) \rangle_{t=1}^{T}, \tag{5}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ denotes a subsequence with the first t words, and $score_{bnl}(\mathcal{S}_t)$ is computed similarly to imagery scores but reflects banality levels.

For example, "take the test" achieves a high banality score of 3.18, while "breathe back the world" scores only 0.98, indicating greater originality.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Valence

Valence reflects emotions expressed in poetic lines and is crucial for assessing an artist's preferred emotional tone (Scherer 1984; Charland 2005). Artists might favor lines that blend positive and negative connotations ("the greatest began to deny") or purely positive sentiments ("tranquility in the lovers").

To measure valence, we build a computational emotion model based on 29 fine-grained emotion categories derived from the GoEmotions dataset (Demszky et al. 2020), plus "nostalgia" to capture lyrical emotional nuances. We employ a LLaMA-3-70b prompting-based classifier to obtain probabilities for the top 5 emotional categories for each line.

Let S represent the input sentence and E_k the set of top-k emotions predicted by the classifier (with $k = 5$). The probability vector is constructed as: $\mathbf{p}(S) = \langle \tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_m \rangle$, where m is the total emotion categories (29) and:

$$\tilde{p}_e = \begin{cases} p_e, & \text{if } e \in E_k \\ p_{rem}, & \text{if } e \notin E_k \end{cases} \tag{6}$$

with $p_{rem} = \frac{1-\sum_{e \in E_k} p_e}{m-k}$ representing remaining probability mass distributed evenly across other categories. The complete valence vector for a poetic line $S = (w_1, w_2, \ldots, w_T)$ is given by:

$$\mathbf{S_{vfull}} = \langle \mathbf{p}(\mathcal{S}_t) \rangle_{t=1}^{T}, \tag{7}$$

where $\mathcal{S}_t = (w_1, w_2, \cdots, w_t)$ denotes a subsequence capturing temporal development.

We also construct a uni-dimensional variant by grouping emotions into positive, neutral, and negative sets (E_+, E_0, E_-), and collapsing the probability vector to a score $\tilde{p} \in [-1, 1]$:

$$\mathbf{p_{bin}}(S) = \tilde{p}_+ = \sum \mathbb{I}_e \cdot \tilde{p}_e, \tag{8}$$

where $\mathbb{I}_e$ is defined as:

$$\mathbb{I}_e = \begin{cases} +1, & \text{if } e \in E_+ \\ 0, & \text{if } e \in E_0 \\ -1, & \text{if } e \in E_- \end{cases} \tag{9}$$

yielding the final feature vector:

$$\mathbf{S_{vbin}} = \langle \mathbf{p_{bin}}(\mathcal{S}_t) \rangle_{t=1}^{T} \tag{10}$$

Our preliminary results showed that features from the LLaMA-based classifier have higher predictive power than a RoBERTa classifier (Liu 2019) fine-tuned on GoEmotions.

Surprisal

Surprisal (Shannon 1948) indicates a word's predictability in a sentence, measured as the negative log probability of the word given its context:

$$s(w_t) = -\log_2 P(w_t | \mathbf{w}_{<t}), \tag{11}$$

where $s(w_t)$ denotes the surprisal value for the t-th word.

To compute sentence-level surprisal, we compose a vector of word-level surprisals, appended with their average:

$$\mathbf{S_{surprisal}} = \langle s(w_t) \rangle_{t=1}^{T} \oplus \langle \bar{s} \rangle, \tag{12}$$

where T is the number of words, $\bar{s} = \sum s(w_t)/T$ is the mean surprisal, and $\oplus$ denotes concatenation.

Since we lack access to the true probability distribution P, we use GPT-2 (Radford et al. 2019)[3], a powerful autoregressive language model, to approximate the log probabilities.

This feature distinguishes between sentences like "a knowledge of the world," with lower surprisal ($s_{avg} = 10.12$) indicating higher predictability, and "a just dream with the fire," with higher surprisal ($s_{avg} = 11.21$) indicating greater novelty potential (note these values are on a log scale, so the difference is exponential).

Contextual Entropy

Contextual Entropy (Shannon 1948) of a word is defined as the expected value of its Surprisal:

$$h(w_t) = \sum_{w_i \in V} p(w_t | \mathbf{w}_{<t}) \log_2 P(w_t | \mathbf{w}_{<t}), \tag{13}$$

where $h(w_t)$ denotes the contextual entropy for the t-th word, and V represents the vocabulary of possible next words.

Similar to surprisal, we use GPT-2 to compute probabilities and construct the sentence-level entropy feature vector by combining word-level entropy values with their mean:

$$\mathbf{S_{entropy}} = \langle h(w_t) \rangle_{t=1}^{T} \oplus \langle \overline{h} \rangle, \tag{14}$$

where T is the number of words and $\overline{h} = \sum h(w_t)/T$ is the mean entropy. Contextual entropy helps gauge unpredictability/novelty in sentences. For example, "i remember you" has lower entropy ($h_{avg} = 6.21$) compared to "pure crime of the lost" ($h_{avg} = 9.53$). Entropy also correlates positively with reading times (Lowder et al. 2018), indicating its ability to capture sentence complexity (higher entropy $\rightarrow$ higher complexity).

Normalized Pointwise Mutual Information (NPMI)

NPMI measures word associations by comparing co-occurrence probabilities with individual occurrence probabilities in a corpus (Bouma 2009). Formally, the NPMI of words x and y is defined as:

$$N(x, y) = \left(\ln \frac{p(x, y)}{p(x)p(y)} \right) / - \ln p(x, y) \tag{15}$$

Here, $p(x) = count(x)/N$, where $count(x)$ is the word's occurrence count and N is the total word count. NPMI values range between $[-1, 1]$, with higher values indicating stronger associations.

For example, ("see", "you") has a high NPMI value (0.47) while ("never", "the") has a low value (-0.11), as "see" and "you" frequently co-occur, whereas "never" typically precedes verbs rather than determiners.

We compute two NPMI variants: unidirectional (N_{uni}), considering co-occurrences only when y follows x, and bidirectional (N_{bi}), also counting when y precedes x. The final feature vectors take the form:

[3] specifically, we use the 774M parameter `gpt2-large` model from Huggingface: https://huggingface.co/openai-community/gpt2-large

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

$\mathbf{S_{npmi,uni}} = \langle N_{uni}(w_t, w_{t+1})\rangle_{t=1}^{T-1} \oplus \langle \overline{N}_{uni}\rangle$ (similar for $\mathbf{S_{npmi,ubi}}$), where T is the sentence length, $\overline{N}_{uni} = \sum N_{uni}(w_t, w_{t+1})/T$ is the mean unidirectional NPMI, and $\overline{N}_{bi} = \sum N_{bi}(w_t, w_{t-1})/T$ is the mean bidirectional NPMI.

Step 2: Training the Calibration Network

Artistic preferences vary significantly across individuals—a poetic line with identical stylistic and linguistic properties might inspire one person while leaving another unmoved. This subjective nature of creative inspiration necessitates a personalized approach that adapts to individual tastes.

To address this challenge, we develop a calibration network that learns to map the extracted features to an individual's unique artistic preferences. This personalization layer is critical for accurately predicting whether a specific person will find a given poetic line inspiring.

We formulate this as a supervised classification problem: given the linguistic and stylistic features of a poetic line, the network predicts whether the line will be inspiring to a specific individual. Formally, consider a dataset of artistic preferences $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in \mathcal{X}$ represents the feature vector of the i-th poetic line and $y_i \in \mathcal{Y}$ indicates whether the artist found that line inspiring.

The calibration network learns a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ that maps each feature vector x_i to a probability distribution over classes $\hat{y} = f_\theta(x_i) = P(y|x_i;\theta)$. For our binary classification setup, $\mathcal{Y} = \{\text{inspiring}, \text{not inspiring}\}$.

This approach offers several advantages over alternative methods. Unlike end-to-end models that might directly map raw text to preferences, our feature-based approach provides interpretability by revealing which specific linguistic and stylistic elements resonate with each individual. Additionally, by avoiding direct use of sentence embeddings, we focus explicitly on creative attributes rather than general semantic properties, which often fail to capture the nuanced elements that make lyrics inspirational.

The calibration network can be implemented using various architectures, from traditional machine learning models to deep neural networks, depending on the dataset size and complexity of preference patterns. Once trained, this personalized layer enables our framework to filter and prioritize lyrical content according to each artist's unique creative sensibilities.

Dataset

Data Collection and Annotation. We collected 3,025 poetic lines from the LyricJam platform, a publicly available research system that generates poetic lines with the goal of assisting creative individuals in finding novel lyrical and poetic ideas for use in the Seed phase of the creative process (Vechtomova, Sahu, and Kumar 2021) [4]. All lines were preserved in their original lowercase format without additional preprocessing. We refer to this collection as the *EvocativeLines* dataset.

For the initial annotation phase, the first and last authors independently labeled each line as either "inspiring" or "not inspiring." In this context, "inspiring" indicates that a line created an emotional impact and stimulated creative ideation. Both annotators actively engage in creative writing (poetry and fiction), providing them with relevant expertise for this evaluation task. The complete annotation process spanned approximately 12 hours over multiple weeks.

[4] https://lyricjam.ai

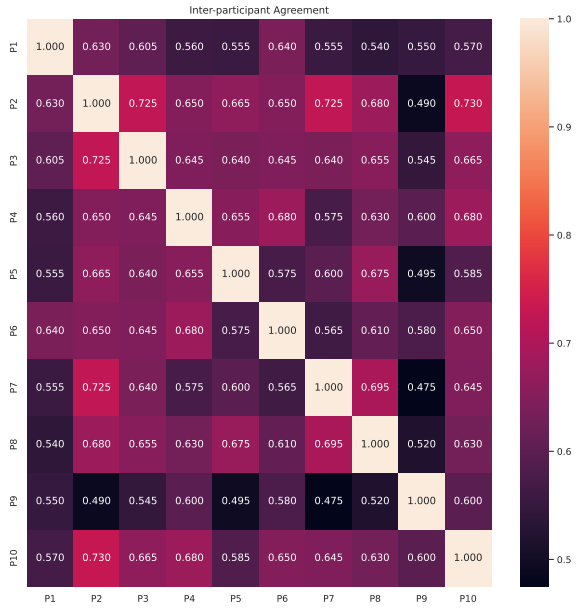


Figure 2: Degree of agreement among participants on the EvocativeLines *(small)* dataset. Lower agreement (darker tones) demonstrates the diversity of subjective preferences captured across our pool of creative individuals.

Diverse Preference Profiles. To evaluate our framework's ability to model diverse artistic preferences, we recruited eight additional participants, each actively engaged in creative activities such as poetry writing or musical performance. These participants rated a subset of 200 poetic lines, the *EvocativeLines (small)* dataset, on a scale of 1-10 based on their inspirational quality.

The subset was constructed by randomly selecting 100 "inspiring" and 100 "not inspiring" lines from the original dataset, balanced across different length categories (short, medium, and long). This sampling approach kept the annotation task manageable (30-45 minutes per participant), minimizing decision fatigue while still capturing a broad spectrum of creative preferences. We applied min-max scaling to normalize the ratings and convert them to binary "inspiring"/"not inspiring" judgments.

We denote our complete set of participants as $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_{10}\}$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ represent the first and last authors, and $\mathcal{P}_{3-10}$ represent the additional participants. Each participant's set of labels constitutes a unique preference profile.

Inter-participant Agreement Analysis. Figure 2 visualizes the inter-participant agreement on the EvocativeLines

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

(small) dataset. The agreement score represents the percentage of lines that pairs of participants both found either inspiring or not inspiring. We observed substantial variation in agreement, ranging from 47.5% to 72.5%, with a mean agreement of 65.3% (±3.1%). This indicates that participants disagreed on approximately 70 out of 200 lines on average.

This level of disagreement aligns with previous research on the perception of aesthetic stimuli (Gold et al. 2019) and confirms that our dataset effectively captures the subjective and individualized nature of artistic preferences.

Alternative Generation Sources. We also explored alternative AI generation platforms, including ChatGPT and Claude Sonnet. However, less than 1% of lines from these sources were judged as evocative by our participants, providing insufficient data for training our calibration network. Figure S.2 on our project website showcases examples with comparative scores for poetic and stylistic characteristics across different sources. Notably, we observe that these systems tend to add superficial flourishes to lines to make them "poetic." Such lines score high in imagery and energy; however, they also score high on banality (e.g. "stars watch over empty fields," and "dreams scatter like fallen petals" that sound poetic but use well-worn expressions).

Experiments

Step 1: Computing Features. To test our framework on the EvocativeLines dataset, we first compute all the features described in Step 1 of the proposed framework. Specifically, we combine all the features *along the time dimension* to obtain the following overall feature vector for a given poetic line: $\mathbf{S_{all}} = [\mathbf{S_{imagery}}|\cdots|\mathbf{S_{banality}}]$ where $|$ denotes the concatenation operation along the time dimension, T is the number of time steps (= number of words in the sentence), and $\mathbf{S_{all}} \in \mathbb{R}^{T\times n}$ with $n = (8T + m)$ when using $\mathbf{S_{vfull}}$ and $n = (8T + 1)$ when using $\mathbf{S_{vbin}}$.

Step 2: Training the Calibration Network. We create a 80-20 split of training-testing in the dataset and use the features in the training set to train the calibration network. We use stratified sampling when creating the splits to ensure that the distribution of labels is preserved across the two dataset parts. We perform 5-fold cross-validation for tuning the hyperparameters of the calibration network. When training the calibration network on the EvocativeLines *(small)* dataset, we create a 75-25 split of training-testing and then do a 5-fold cross-validation.

Choice of calibration network. We explore multiple deep-learning (DL) and machine-learning (ML) based architectures for the calibration network. Namely, we try **a) LSTM+Attn** ($\mathbf{S_{all}}$): a bi-directional classification network with Long Short Term Memory (LSTM) cells (Hochreiter and Schmidhuber 1997) coupled with an attention mechanism (Bahdanau 2014). We use an LSTM network as all of our features are temporal, and LSTMs are adept at modeling long-range temporal dependencies. We also perform ablation studies where we train the classifier on a subset of $\mathbf{S_{all}}$ to test the effect of the individual features; **b) XGBoost** (Chen and Guestrin 2016): We train a standard XGBoost model on our features; however, we *flatten* $\mathbf{S_{all}}$ as XGBoost does not inherently handle sequential data with a timestep dimension. Specifically, we train XGBoost (and all the other ML-based classifiers) on the following flattened input feature: $\mathbf{S_{flat}} = [\dot{\mathbf{S}}_{\mathbf{imagery}} \oplus \cdots \oplus \dot{\mathbf{S}}_{\mathbf{banality}}]$, where $\dot{\mathbf{S}}_{(\cdot)}$ is the flattened version of $\mathbf{S}_{(\cdot)}$ and $\oplus$ is the concatenation operation. $\mathbf{S_{flat}} \in \mathbb{R}^{Tn}$, where $n = 8T + m$; **c) Other ML-based Classifiers.** In addition to XGBoost, we train multiple other ML-based classifiers on $\mathbf{S_{flat}}$. Specifically, we train Logistic Regression **(LR)**, Support Vector Machine **(SVM)**, Decision Trees **(DT)**, and Random Forest **(RF)** models (Nelder and Wedderburn 1972; Cortes and Vapnik 1995; Quinlan 1986; Breiman 2001).

Other Frameworks. We compare our framework with multiple strong baselines. Specifically, we compare the following: **a) LSTM+Attn (SBERT):** we train an LSTM+Attn classifier that maps the Sentence-BERT (Reimers and Gurevych 2019) embeddings of poetic lines to the final rating. SBERT is a popular text embedding model that modifies the BERT architecture with siamese (Bromley et al. 1993) and triplet networks (Schroff, Kalenichenko, and Philbin 2015) to obtain semantically meaningful sentence representations. Note that in this scenario, there will be only one "time step"; **b) BERT** (Devlin et al. 2019): a state-of-the-art transformer-based language model pre-trained on large textual corpora. It is widely adopted for various natural language processing tasks, such as classification, question answering, and named entity recognition. We fine-tune a BERT-base model, which has 110M parameters, on the EvocativeLines dataset, which serves as a strong baseline comparing the performance of mapping the sentence embeddings of poetic lines directly with the final prediction; **c) LLaMA-3:** a prompting-based k-shot LLaMA classifier, where, for a given test line, we prompt a LLaMA-3-70b model with k "inspiring" and k "not inspiring" examples from the dataset and ask it to classify a test sentence. We use cosine similarity to find the k examples that are (semantically) closest to the test sentence in the SBERT embedding space. We find that this approach of selecting examples performs better than random sampling.

Evaluation of Different Frameworks. We adapt each framework on the EvocativeLines dataset and use test accuracy and area under the curve (AUC) scores as primary indicators of the predictive power of different models. We report these results in Table 1. Further, we test our proposed framework on the EvocativeLines *(small)* dataset with 10 diverse sets of preference profiles, and report results in Table 2.

Implementation Details. We implement the LSTM classifiers in PyTorch (Paszke et al. 2019) and train them for 500 epochs with early stopping (we stop a training run if the validation performance does not improve for 10 consecutive epochs). We perform grid search to find optimal values over the parameters listed in Figure S.9 on the project website, and find a *one*-layer *bi*-directional LSTM classifier trained

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Method	$\mathcal{P}_1$		$\mathcal{P}_2$	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Majority baseline	*51.5*	*50.0*	*53.0*	*50.0*
LSTM+Attn ($\mathbf{S_{all}}$)	**79.5**$^{(0.4)}$	**79.7**$^{(0.1)}$	**78.3**$^{(0.3)}$	**80.1**$^{(0.3)}$
LSTM+Attn (SBERT)	71.6$^{(0.3)}$	71.7$^{(0.1)}$	69.9$^{(0.2)}$	70.1$^{(0.3)}$
BERT	77.6$^{(2.1)}$	78.8$^{(1.7)}$	75.2$^{(2.3)}$	76.1$^{(1.1)}$
RF	80.5$^{(1.7)}$	80.6$^{(1.7)}$	72.4$^{(2.0)}$	73.0$^{(1.2)}$
SVM	70.5$^{(1.2)}$	70.5$^{(1.0)}$	67.3$^{(2.3)}$	63.1$^{(1.1)}$
LR	71.1$^{(1.9)}$	71.1$^{(1.8)}$	64.3$^{(1.9)}$	62.0$^{(1.4)}$
DT	84.2$^{(2.5)}$	84.3$^{(2.3)}$	84.1$^{(1.0)}$	80.0$^{(1.7)}$
XGBoost	**92.2**$^{(1.7)}$	**92.2**$^{(1.6)}$	**87.32**$^{(1.4)}$	**83.9**$^{(1.2)}$
LLaMA (200-shot)	73.3$^{(1.7)}$	73.3$^{(1.6)}$	62.3$^{(3.6)}$	65.1$^{(1.6)}$
LLaMA (300-shot)	73.9$^{(2.1)}$	74.2$^{(1.8)}$	64.1$^{(1.2)}$	65.8$^{(1.4)}$
LLaMA (450-shot)	**74.8**$^{(3.1)}$	**74.8**$^{(3.0)}$	**69.5**$^{(3.5)}$	**70.5**$^{(1.6)}$

Table 1: Comparison of different frameworks on *EvocativeLines*. The first section shows the performance of DL-based methods, the second section shows the performance of ML-based methods, and the last section shows the performance of LLM-based methods.

User	Majority		LSTM+Attn		XGBoost	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
			EvocativeLines (small)			
$\mathcal{P}_1$	50.0	50.0	75.6$^{(1.2)}$	64.9$^{(4.0)}$	70.7$^{(5.3)}$	70.2$^{(5.3)}$
$\mathcal{P}_2$	70.5	50.0	80.5$^{(1.2)}$	62.3$^{(6.1)}$	69.3$^{(9.5)}$	57.1$^{(8.4)}$
$\mathcal{P}_3$	58.5	50.0	**94.1**$^{(1.2)}$	**92.3**$^{(2.1)}$	63.4$^{(11.1)}$	62.6$^{(11.7)}$
$\mathcal{P}_4$	60.5	50.0	68.7$^{(3.1)}$	61.4$^{(4.5)}$	59.5$^{(4.5)}$	56.4$^{(6.6)}$
$\mathcal{P}_5$	58.0	50.0	80.4$^{(2.5)}$	84.8$^{(5.1)}$	71.7$^{(6.7)}$	**70.5**$^{(6.6)}$
$\mathcal{P}_6$	64.5	50.0	75.6$^{(3.1)}$	75.4$^{(2.8)}$	68.8$^{(3.2)}$	62.8$^{(3.0)}$
$\mathcal{P}_7$	58.0	50.0	75.6$^{(4.3)}$	82.3$^{(4.2)}$	61.0$^{(4.4)}$	60.3$^{(4.2)}$
$\mathcal{P}_8$	62.5	50.0	80.5$^{(5.4)}$	78.2$^{(5.1)}$	**72.2**$^{(5.0)}$	68.5$^{(7.1)}$
$\mathcal{P}_9$	70.0	50.0	76.5$^{(4.7)}$	72.4$^{(3.6)}$	70.7$^{(4.9)}$	57.3$^{(5.1)}$
$\mathcal{P}_{10}$	58.5	50.0	82.3$^{(2.3)}$	84.2$^{(3.5)}$	62.4$^{(5.9)}$	60.7$^{(5.0)}$
Avg.	61.1	50.0	**79.0**$^{(6.3)}$	**75.8**$^{(9.9)}$	67.0$^{(4.6)}$	62.6$^{(5.1)}$

Table 2: Performance of the proposed framework with LSTM+Attn and XGBoost calibration networks on the *EvocativeLines (small)* dataset.

with a learning rate of 0.0001 on $\mathbf{S_{all}}$ input features performs best.

We use the `sentence-transformers`[5] python package to obtain SBERT embeddings of sentences and use Huggingface's `transformers`[6] library to implement the finetuning of BERT classifier. Specifically, we finetune the BERT classifier for 500 steps with a weight decay of 0.01. We use the `sklearn` python package (Pedregosa et al. 2011) to implement all the ML-based classifiers. We use the default hyperparameters for all the models. Finally, we use a batch size of 32 for all the models.

For the LLaMA-3 classifier, we use the categorization of "inspiring" and "not inspiring" from the annotations from $\mathcal{P}_1$ and $\mathcal{P}_2$ (we randomly select which participant to choose for every test sentence).

Finally, we perform 5-fold cross-validation (except for the LLaMA-3 classifier), repeat all our experiments for 5 random seeds, and report the mean and std values in our results tables.

[5] `https://www.sbert.net/`

[6] `https://huggingface.co/docs/transformers/`

Results and Analysis

Proposed Linguistic and Poetic Features Have High Predictive Power

Our experimental results demonstrate that specialized feature engineering significantly outperforms even the largest language models in predicting artistic preferences. From Table 1, we observe that our framework with ML-based and DL-based calibration networks consistently surpasses a 450-shot LLaMA-3-70b classifier across participant profiles.

Specifically, our XGBoost model—containing just 163 trainable parameters—achieved test accuracies of 92.2% for $\mathcal{P}_1$ and 87.3% for $\mathcal{P}_2$. This substantially outperforms the 450-shot LLaMA-3 classifier with 70 billion parameters, which only reached accuracies of 74.8% and 69.8% for the same participants. Note that 450 examples represented the maximum number of positive and negative instances we could include in the LLaMA-3-70b prompt due to context length limitations.

Among deep learning approaches, our LSTM+Attention model (approximately 110k parameters) outperforms both a fine-tuned BERT model (110M parameters) and the LSTM+Attention model trained on SBERT embeddings. This further validates the superiority of our proposed features over general-purpose text representations.

The higher performance of our targeted feature extraction approach compared to using LLMs directly reveals an important insight: while we can leverage LLMs to compute established linguistic and stylistic characteristics, these models struggle to capture the nuanced differences between what individual users find "inspiring" versus "not inspiring." This highlights how our framework's focus on specific linguistic and poetic features is more effective for modeling personal artistic preferences than using much larger general-purpose models.

The Proposed Framework Adapts Robustly to Diverse User Preferences

Our framework demonstrates strong adaptability across varied artistic sensibilities, as shown in Table 2. When tested on the *EvocativeLines (small)* dataset containing 10 distinct preference profiles, both calibration networks effectively learned individual preferences despite their differences.

The LSTM+Attention variant achieved an average test accuracy of 79.0% ($\pm$6.3%) across all 10 participants, while the XGBoost variant reached 67.0% ($\pm$4.6%). This represents a performance gap of approximately 25 percentage points for XGBoost compared to its results on the full dataset. This decline likely stems from XGBoost's tendency toward overfitting in data-scarce scenarios, despite its general flexibility.

Overall, the results highlight the robustness of our framework given the high degree of variation across preference profiles. We can, therefore, conclude that our framework can successfully model and predict artistic preferences across diverse individuals, even with limited training data, by selecting an appropriate calibration network.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Feature	$\mathcal{P}_1$		$\mathcal{P}_2$		$\mathcal{P}_3$		$\mathcal{P}_6$		$\mathcal{P}_7$	
	coef	p-value	coef	p-value	coef	p-value	coef	p-value	coef	p-value
$\overline{\mathbf{S}}_{\mathbf{surprisal}}$	-0.4149	0.107	0.288	0.404	0.4351	0.233	0.416	0.242	-0.2176	0.518
$\overline{\mathbf{S}}_{\mathbf{entropy}}$	-0.8377	0.034	0.0534	0.889	0.02	0.958	-0.0632	0.873	-0.0817	0.822
$\overline{\mathbf{S}}_{\mathbf{npmi,uni}}$	7.2427	0.000	9.7483	0.068	59.1835	0.003	75.9894	0.018	21.0617	0.152
$\overline{\mathbf{S}}_{\mathbf{npmi,bi}}$	-6.2588	0.002	-9.8577	0.552	-55.0718	0.007	-73.9488	0.024	-19.2738	0.224
$\overline{\mathbf{S}}_{\mathbf{abs}}$	0.5402	0.298	0.6791	0.047	0.7423	0.045	0.9499	0.012	0.7226	0.03
$\overline{\mathbf{S}}_{\mathbf{vbin}}$	0.3281	0.001	-103.4697	0.556	-197.0881	0.028	-264.4955	0.158	9.1933	0.957
$\overline{\mathbf{S}}_{\mathbf{imagery}}$	14.8346	0.000	0.6861	0.048	2.065	0.057	0.4584	0.643	1.1321	0.241
$\overline{\mathbf{S}}_{\mathbf{energy}}$	-0.0543	0.985	0.9405	0.031	1.3163	0.423	2.4855	0.086	1.8852	0.191
$\overline{\mathbf{S}}_{\mathbf{banality}}$	-2.8474	0.128	0.5269	0.636	-0.2816	0.804	0.1945	0.866	3.2007	0.003
$\overline{\mathbf{S}}_{\mathbf{I1}}$	0.0549	0.084	-0.1213	0.609	-0.0655	0.745	-0.2414	0.255	0.3546	0.116
$\overline{\mathbf{S}}_{\mathbf{I2}}$	-0.0188	0.499	13.8702	0.027	14.5632	0.142	6.5436	0.438	3.1841	0.72
$\overline{\mathbf{S}}_{\mathbf{all}}$	0.0760	0.025	-1.0131	0.022	-0.3216	0.31	-0.2537	0.306	1.3457	0.145

Table 3: Results of interaction testing. I1 denotes the interaction between linguistic features (surprisal, entropy, banality, NPMI), and I2 denotes the interaction between stylistic features (imagery, word energy, abstraction, valence).

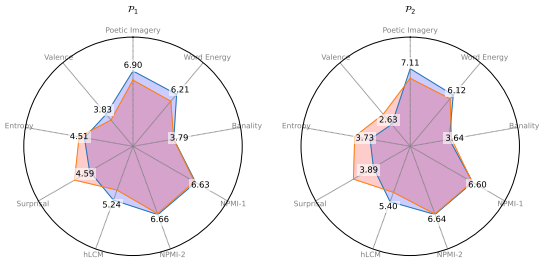


Figure 3: Radar charts showing preference profiles for two participants in the *EvocativeLines* dataset. Blue and red webs denote positive and negative lines, respectively. Figure S.3 on the project website shows a complete profile.

Identifying Significant Features Across Preference Profiles

To provide deeper insights into what drives individual preferences, we conducted statistical testing to determine the correlations between specific features and user judgments. We fit a logistic regression model with interaction terms to evaluate the significance of each feature on whether a participant found a poetic line inspiring.

For this analysis, we simplified our temporal features by averaging across the timestep dimension, converting each into a single score. While this representation sacrifices some temporal information, it allows for clearer interpretation of feature importance. Table 3 presents results for five selected participants: $\mathcal{P}_1$ and $\mathcal{P}_2$ (selected for their data-rich annotations) and $\mathcal{P}_3$, $\mathcal{P}_6$, and $\mathcal{P}_7$ (selected as a triplet with particularly low inter-participant agreement).

Our analysis reveals distinct preference patterns across participants. For $\mathcal{P}_1$, contextual entropy ($p = 0.034$) shows a significant negative correlation, indicating a preference for lines with higher unpredictability. Unidirectional NPMI ($p < 0.001$) shows a strong positive correlation, while bidirectional NPMI ($p = 0.002$) shows a significant negative correlation. Combined with a strong preference for poetic imagery (coefficient=14.83, $p < 0.001$), this suggests $\mathcal{P}_1$ values lines with clear linear progression that maintain creative impact when read in the intended direction.

For $\mathcal{P}_2$, we observe significant preferences for word energy ($p = 0.031$), level of abstraction ($p = 0.047$), and poetic imagery ($p = 0.048$), along with a significant interaction effect between stylistic features ($p = 0.027$). This indicates $\mathcal{P}_2$ values the combined impact of multiple poetic elements rather than any single dimension.

Each participant shows a unique signature of feature preferences—$\mathcal{P}_3$ values abstraction and unidirectional word associations, $\mathcal{P}_6$ prefers causality and abstraction, while $\mathcal{P}_7$ values originality (low banality, $p = 0.003$) and abstraction ($p = 0.03$).

These results demonstrate that our framework not only predicts preferences accurately but also provides interpretable insights into the specific linguistic and poetic elements that resonate with different individuals.

Qualitative Analysis

We visualize preference profiles as radar charts (See Figure 3) that display average feature values across dimensions for inspiring (blue) and non-inspiring (red) poetic lines. These visualizations reveal consistent distinctions between inspiring and non-inspiring examples across all participants, confirming our features effectively capture meaningful differences in artistic preferences. Notably, all participants prefer lines with higher poetic imagery and word energy, suggesting these elements form an important foundation for evocative content. However, each profile exhibits unique patterns in how other features (abstraction, surprisal, entropy) contribute to inspiration, highlighting the personalized nature of artistic preferences.

Individual example visualizations (Figure S.4 on the project web site) reveal complex feature interactions. For instance, "all the world is the fire" maintains high inspirational quality despite neutral emotional valence (0) through elevated imagery and energy values. This demonstrates how strength in certain dimensions can compensate for neutrality in others. Similarly, we observe that high abstraction paired with moderate surprisal creates lines that balance philosophical depth with cognitive accessibility.

Rarely does a single feature determine a line's inspirational quality. Instead, specific combinations—functioning as coherent stylistic signatures—resonate with individual preferences, with particularly strong correlations observed between inspiration and both word energy and abstraction across profiles.

Conclusion

In this work, we propose an interpretable and flexible framework for computational modeling of artistic inspiration in AI-generated poetic lines. Our approach assists artists in the Seed phase by identifying potentially "inspiring" lines from larger pools of candidates. We identified key linguistic and stylistic features that accurately forecast aesthetic preferences across diverse individuals. Our experiments demonstrate that our framework significantly outperforms state-of-the-art language models like LLaMA-3-70b classifiers. Crucially, our approach extends beyond poetry analysis. It can be adapted with different feature sets to predict other subjective aesthetic

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

preferences—for comparing sonnets from different poets or analyzing prose stylistics. In the context of creating art, it is essential to measure how well AI systems support individual creative processes that span a wide spectrum of artistic preferences and skill levels.

References

Althuizen, N. 2021. Revisiting Berlyne's inverted U-shape relationship between complexity and liking: The role of effort, arousal, and status in the appreciation of product design aesthetics. *Psychology & Marketing* 38(3):481–503.

Amabile, T. M. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology* 43(5):997.

Baer, J. 2011. How divergent thinking tests mislead us: Are the Torrance tests still relevant in the 21st century? the division 10 debate. *Psychology of Aesthetics, Creativity, and the Arts* 5(4):309–313.

Bahdanau, D. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Berlyne, D. E. 1975. Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. *Journal of Aesthetics and Art Criticism* 34(1):86–87.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 30:31–40.

Breiman, L. 2001. Random forests. *Machine learning* 45:5–32.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'93, 737–744. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Brooks, C., and Warren, R. P. 1976. *Understanding Poetry*. New York: Holt, Rinehart and Winston, 4th edition.

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

Burroughs, W. 2012. *The Job: Interviews with William S. Burroughs*. Penguin Modern Classics. Penguin Books Limited.

Cassirer, E. 1946. *Language and myth*, volume 51. Courier Corporation.

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. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–34.

Charland, L. C. 2005. The heat of emotion: Valence and the demarcation problem. *Journal of consciousness studies* 12(8-9):82–102.

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.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20:273–297.

Cross, C. R. 2001. *Heavier Than Heaven: A Biography of Kurt Cobain*. New York: Hyperion.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A dataset of fine-grained emotions. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dunnigan, S. M. 2014. *7 Poetic Imagery*. Edinburgh: Edinburgh University Press. 67–77.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2024. GPTScore: Evaluate as you desire. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576. Mexico City, Mexico: Association for Computational Linguistics.

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(47):9397–9409.

Hashemi, H.; Eisner, J.; Rosset, C.; Van Durme, B.; and Kedzie, C. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13806–13834.

Hashimoto, T. B.; Zhang, H.; and Liang, P. 2019. Unifying human and statistical evaluation for natural language generation. 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)*, 1689–1701. Minneapolis, Minnesota: Association for Computational Linguistics.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Johnson-Grey, K. M.; Boghrati, R.; Wakslak, C. J.; and Dehghani, M. 2020. Measuring abstract mind-sets through syntax: Automating the linguistic category model. *Social Psychological and Personality Science* 11(2):217–225.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Jones, D. 2017. *David Bowie: A Life*. New York: Crown Archetype, first edition.

Kao, J., and Jurafsky, D. 2012. A computational analysis of style, affect, and imagery in contemporary poetry. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, 8–17.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lowder, M. W.; Choi, W.; Ferreira, F.; and Henderson, J. M. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science* 42(S4):1166–1183.

Nelder, J. A., and Wedderburn, R. W. 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 135(3):370–384.

Palahniuk, C. 2020. *Consider This: Moments in My Writing Life after Which Everything Was Different*. Grand Central Publishing.

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.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research* 12:2825–2830.

Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; and Harchaoui, Z. 2024. Mauve: measuring the gap between neural text and human text using divergence frontiers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Red Hook, NY, USA: Curran Associates Inc.

Pound, E. 1913. A few don'ts by an imagiste. *Poetry* 1(6):200–206.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1:81–106.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 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 Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Rubin, R. 2023. *The creative act: a way of being*. Penguin.

Scherer, K. R. 1984. On the nature and function of emotion: a component process approach. approaches to emotion.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Semin, G. R., and Fiedler, K. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of personality and Social Psychology* 54(4):558.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal* 27(3):379–423.

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*.

Adapting Proppian Morphology for Generating Narrative Structures

Pablo Gervás
Facultad de Informática
Universidad Complutense de Madrid
Madrid, 28040, Spain

Abstract

Vladimir Propp's "Morphology of the Folk Tale" presented a meticulous analysis of a basic type of story into several layers of constituents that could be recombined to make new stories. Early adaptations of a very limited portion of his account to generate stories were criticised on the grounds that the representation was too simple. The present paper considers how the more complex aspects of Propp's analysis might be implemented computationally. The original descriptive representation of stories is augmented with solutions inspired by reverse engineering some of the more complex examples of stories analysed by Propp in his book. The revised computational model includes important aspects of narrative such as the distinction between fabula and discourse, the use of embedded stories and construction of narrative structures with more than one plot line.

Introduction

The Morphology of the Folk Tale proposed by Vladimir Propp has been used as inspiration by many computational attempts at generating stories. However, in most cases the corresponding computational systems departed considerably from Propp's original formulation. The present paper explores a solution that: (1) focuses on the generation of sequential tales matching the nature of those in Propp's corpus, (2) enriches Propp's original formulation with the details necessary to make it operational from a computational point of view, and (3) at all levels prioritises faithful compliance with Propp's account over implementation details considered traditional for specific computational approaches.

Related Work

Three branches of previous work need to be reviewed: Propp's formalism, relevant concepts of narratology and related computational story generators.

Propp's Morphology of the Folk Tale

Vladimir Propp (1928) was a Russian formalist interested in the form of the tale. With the aim of developing a useful tool for adequate classification, factual description and comparison across tales he studied the component parts of a tale – and of their relations to each other and to the whole – over a selection out of a corpus of Russian folk tales (Afanasyev, 1859). The tales studied involved a hero facing a complication (a villainy or a lack) who sets out, undergoes some tribulations and eventually succeeds and is rewarded. Propp's analysis is best known for the formulation of *character functions* – abstractions of character behaviour relevant to the plot – and a *canonical sequence* – a relative order in which all possible character functions appear in a tale with respect to one another. His analysis includes a number of additional elements that are less well known – possibly because their description in the book is much less detailed – but which have a significant potential for computational consideration. First, a set of roles for the characters in the narrative (the *dramatis personae*) which include, among others: the *hero* (who sets out on a journey), the *dispatcher* (who dispatches the hero on his journey), the villain (that the hero faces during the story), the *donor* (who provides a magical agent to the hero), the *false hero* (who competes with the protagonist for the role of hero of the story). Second, the concept of *move*: a single pass over the canonical sequence selecting character functions to include in a tale for a given hero. According to Propp, tales beyond his simple initial description could be analysed as a combination of several moves. He includes several analyses of this type for considerably complex tales in the appendices of his book. Third, he describes a procedure to "create a tale artificially", based on selecting character functions out of the canonical sequence and then assigning to them the characters for the tale while respecting the restrictions imposed by which character functions involve which roles of the characters: "In order to create a tale artificially, one may take any *A*, then one of the possible *B*'s then a *C*↑, followed by absolutely any *D*, then an *E*, the one of the possible *F*'s, then any *G*, and so on. In doing this, any elements may be dropped (. . .). If one then distributes functions according to the dramatis personae of the tale's supply or by following one's own taste, these schemes come alive and become tales. Of course, one must also keep motivations, connections, and other auxiliary elements in mind" (p. 111-112). This description served as inspiration for the paper.

Narratological Concepts Relevant for Story Generation

Narrative involves two separate chronologies: the timeline of the events being narrated and the timeline of their telling

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

(Abbott, 2008). The events of a narrative in the chronological order in which they happened is referred to as the *fabula*. The events of a narrative in the chronological order in which they are told in the narrative is referred to as the *discourse*.

Characters in a story sometimes do themselves tell stories to other characters. These are known as *embedded stories* (Herman, Jahn, and Ryan, 2010). The overarching narrative in which this telling takes place is known as the *frame story*.

Computational Story Generators

Over the years, Propp's Morphology of the Folk Tale has been revisited often as inspiration for computational approaches to narrative. These approaches have ranged over a wide set of formalisms and technologies: generative grammars (Lakoff, 1972), case based reasoning (Fairclough, 2004) and case based reasoning combined with planning (Turner, 1993). Of those, Fairclough (2004) considers moves – in Propp's sense – as cases, and story construction as a process of selecting moves from a casebase and combining them into a new story. Gervás (2019) explicitly links the concepts of character function and narrative role into a structured concept of *plot element* that explicitly includes representation of which specific characters contribute in which roles with respect to the story. It also introduced the concept of *axis of interest* (or AoI) as a short sequence of plot elements that do not necessarily occur contiguously in the discourse for a story but which are connected by shared characters that give them meaning – like the victim of a kidnapping being rescued later.

Propp's idea of multi-move tales resonates with recent efforts to construct stories with more than one plotline, such as Fay (2014) – that combines plot lines extracted from existing stories and uses bindings between characters in them to connect them together – and Gervás, Concepción, and Méndez (2022) – that combines two different operations: *discourse planning* – decide on the relative order of appearance in the discourse of contributions from different subplots – and *character fusion* – bind together for the final story roles played by different characters in different subplots.

The ProppOSER Framework

The ProppOSER (**Propp O**rder **S**anction **E**ntwine **R**ewriter) system has been developed to explore how the complete set of concepts proposed by Propp in his book might be exploited to generate complex stories. To follow this goal, instead of exploiting existing AI techniques that might be useful, an effort has been made to come up with computational representations of Propp's elements, and to design specific procedures that combine them in a manner that is either faithful to the procedures described explicitly by Propp or implements adequate procedures designed to produce outcomes that are equivalent in narrative structure to the complex tales described by Propp in the appendices of his book.

Proppian Data Structures

The work presented here relies on the concepts *plot element* and *axis of interest* as proposed by Gervás (2019). A plot element is a data structure that includes a character function label and a set of roles that characters need to play in it. These roles can be instantiated by identifiers for particular characters. An axis of interest is a data structure that captures dependencies between plot element that normally occur together in stories, including the relative order in which they appear and which of their roles have to be instantiated by the same characters. Two examples of axes of interest are shown in Table 1.

AXISofINTEREST =	DONOR
Tested	characters(tested=X,tester=Y)
Character'sReaction	characters(tested=X,tester=Y)
ProvisionOfAMagicalAgent	characters(tested=X,tester=Y)
	objects(gift=Z)
UseOfAMagicalAgent	characters(user=X)
	objects(gift=Z)

AXISofINTEREST =	CONFLICT
Struggle	characters(attacker=X,defender=Y)
Victory	characters(winner=X,looser=Y)

Table 1: Examples of axes of interest.

In Propp's description a given move is characterised by a particular set of instantiations of the roles for the dramatis personae with specific characters. To capture this concept we introduce a *role combo* data structure, which can hold the set of instantiations of roles for the the dramatis personae with specific characters for a particular move. The set of actual roles that Propp mentions in his book differs slightly across the different mentions. We have chosen to include the following roles: hero, villain, victim, authority figure, dispatcher, donor and false hero.

A *Proppian move* data structure then holds together: a set of AoIs chosen to be included in the move, a table of which specific characters instantiate the various roles in the plot elements of the AoIs, and a role combo data structure that specifies which characters are assigned in this move to the roles for the dramatis personae.

A version of the *canonical sequence* in which character functions should occur in a move has been created based on the order in which character functions are tabulated in the analyses presented by in Appendix III.

Generating Single Moves Propp's Way

We construct instances of Proppian move data structures by compiling a set of AoIs that match a given complication, and splicing them all together under the same combo for narrative roles. This involves a stage of interweaving the plot elements from the different AoIs, and a stage of establishing which characters are shared across different AoIs. The interweaving is done by following explicitly the procedure proposed by Propp: follow the canonical sequence and at each point select the plot element from the compiled AoIs that appears next in the sequence. The assignment of characters to roles in AoIs has to be done respecting the particular narrative role associated with each character slot in the corresponding plot elements and the assignment of roles to characters given in the role combo for the move.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Tested	Character'sReaction
Character'sReaction	PromiseOfHelp
Character'sReaction	ProvisionOfAMagicalAgent
Character'sReaction	Validation
HelpReceived	Transference
UseOfAMagicalAgent	Transference
HelpReceived	RecoverEaten
UseOfAMagicalAgent	RecoverEaten
HelpReceived	Resuscitation
UseOfAMagicalAgent	Resuscitation
HelpReceived	Rescue
UseOfAMagicalAgent	Rescue
HelpReceived	Departure
UseOfAMagicalAgent	Departure
HelpReceived	Transference
FishForInfo	InfoRelinquished

(a) attracted character functions

Deception	Deceived
Struggle	Victory
Trickery	Complicity
Command	CommandObeyed
NoticeFact	InferCause
Discovery	Communicate
Pursuit	RescueFromPursuit
Victory	TokenAcquired
TokenDisplayed	Recognition
Deceived	Maiming
Deceived	Blinding
Deceived	Devour
Deceived	Abduction
Deceived	Capture
Transference	UnrecognisedArrival
CallToAction	DecisionToReact

DifficultTask	Solution
Departure	Return
ProvisionOfAMagicalAgent	UseOfAMagicalAgent
PromiseOfHelp	HelpReceived
Validation	ValidationRecognised
FalseClaims	FalsehoodExposed
DecisionToReact	Reward
Lack	LackResolved
Maiming	RegenerateMaiming
Blinding	ReturnSight
Devour	RecoverEaten
UnrecognisedArrival	Recognition
Murder	Resuscitation
Abduction	Rescue
Capture	Rescue
TokenAcquired	TokenDisplayed

(b) repulsed character functions

Table 2: Pairs of character functions considered: (a) attracted to one another and (b) repulsed by one another.

Going Multi-Move: When One is Not Enough

Propp's description of how his framework might be used to generate stories focuses on picking out selected character functions from this canonical sequence and considering them as the sequence describing a new story. However, the tales analysed in the appendices have more than one such sequence, connected together by having characters that play one role in one sequence and play a different role in another. Propp describes these as tales with more than one move. A move in this context is some extract from the canonical sequence, featuring a particular set of characters instantiating the relevant narrative roles.

Multi-move tales may be constructed by building a number of moves following the procedure suggested by Propp to build one move, and then combining the resulting set of moves together. Combining moves together involves two different levels of decision: instantiating some characters in common between the move (to provide a link between them) and interweaving the events of the various moves.

Character fusion across moves is done in terms of the set of roles in the role combos for the different moves. So we can have fusions between the heroes of two moves (which packs the moves into a single hero storyline) or a number of combinations of the supporting casts of the different moves (hero becomes victim, victim becomes hero, shared villain or multiple villains, single or multiple victims, shared donor, dispatcher of one move is also helper in another…). This approach allows for richer outcome stories.

Decisions on the relative order in which the plot elements from the different moves should appear in the final story require additional considerations. A baseline that combines the plot elements from the moves at random results in stories that make little sense, but Propp gives no indication of what criteria might be used to inform this task. We have therefore looked at the examples of tales that he analyses for inspiration. We have opted for a solution based on identifying: (1) character functions should appear together in the tale because they are causally connected, (2) character functions intended to span larger segments of the story, and (3) character functions that inactivate a particular character.

Even though Propp suggest that any choice of character functions out of the canonical sequence is valid, if someone is tested but their reaction is not mentioned the resulting story will not make sense. Or if someone uses a magical agent, it makes sense that the story follows with some achievement that can be understood as a result of that use, like resurrecting someone or curing someone. In a similar fashion, there are character functions that really should not appear contiguously in a tale, such as departing on a journey and returning, or someone being abducted and immediately rescued. Ideally, breaks in the telling of one move to tell about events in another should not happen in between pairs of attracted character functions, but are welcome between pairs of repulsed character functions. To capture this concept we define two sets of pairs of character functions, one of those that attract each other and one of those that repulse each other. The two sets of pairs of character functions defined as related in this way are given in Table 2. The introduction of this kind of connection across character functions may partially address – at the simplest possible computational level – Propp's injunction that "one must also keep motivations, connections, and other auxiliary elements in mind" (p. 111-112).

Devour	RecoverEaten	victim
Murder	Resuscitation	victim
Abduction	Rescue	victim
Capture	Rescue	victim

Table 3: Character functions pairs that inactivate a given character, identified by that the character that is inactivated plays in the character function.

Finally, some character functions (like abduction or capture) imply that a certain character becomes inactive until released by a different character function (such as rescue). Events from another move where the corresponding character is active should not be inserted within the span in which the character has been inactivated for the current move. Table 3 lists the set of character functions related in this fashion. As a baseline the events for the moves in a story are interwoven following a greedy algorithm informed by the set of resources described.

Low Level Choices: Form, Cast and Text

The computational steps described to this point represent our proposal for the first decision that Propp considers relevant

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

in the process of building a tale ("choice of those functions which he omits, or, conversely, which he uses" p. 112). The next step he describes ("choice of the means (form) through which a function is realized." p. 112) corresponds to selecting, for character functions that are defined at a very abstract level – such as villainy – instantiations of a specific type. In our case, this is covered by having specific villainy-related AoIs that feature plot element classified as villainies – abduction, murder, ...) –combined with the corresponding resolutions – rescue, resurrection, ...The choice between them is made at random. Other types of function, such as Difficult Task or Lack need to be instantiated with a particular form of the corresponding concept. Instances of these specific forms for the concepts are provided by dedicated knowledge resources.

The next step required to emulate Propp's proposed procedure is to assign story characters to particular slots in functions. Propp explains that the freedom postulated in his description of this point in his procedure is often not exploited in full, with characters often instantiated with a typical villain, a typical donor...Our computational solution for this step has been to create resources that serve on demand options – chosen at random – for typical instantiations of the various roles. The system then queries these resources and uses the returned instances to replace throughout the story the variables representing the roles that the characters play. As observed in the stories mentioned in Propp's book, heroes are assigned proper names, but other characters are described in terms of some common name (usually an occupation).

The final step described by Propp ("The story teller is free in his choice of linguistic means." p. 113) involves transcribing the conceptual representation of the story that results from the steps above into a linguistic rendering. This task is carried out by a simple template-based text realizer that relies on templates for the various character functions: A template holds a sentence in English with gaps to be filled with text descriptions of characters or plot-related concepts. Two additional tasks had to be resolved. First, a basic module for referring expression generation (Reiter, 2010) has been added to adapt each occurrence of characters described by common names (first mention relies on an indefinite article, subsequent mentions on a definite one). Second, a number of the character functions involve anaphoric reference to prior events, and this needs to be addressed during realization: for instance, the Connective Incident – in which some character informs the hero about the villainy – or the False Claims – in which the false hero claims for himself the merit of the hero's actions. In each case the referred character function needs to be retrieved, transcribed into text, and used as a secondary clause to the main sentence. An additional specific module of the realizer addresses this task. When a sentence of text has been produced by these modules, the text for the story is put together by packaging the sentences from different moves into different paragraphs.

Results and Discussion

An example of story generated by the system is presented in Table 4. Move labels have been added to each paragraph for ease of discussion. We can see here a two-move story, a main move involving the apothecary as the villain and Katoma as the hero, and a secondary move with Katoma as the villain, the apothecary as victim, and Peter as a hero who comes to help the apothecary. Each move sees the conflict between Katoma and the apothecary from a different point of view. This is known as *focalization* (Genette, 1980).

It is important to keep in mind that the system is not intended to generate text for the stories, but rather focuses on their narrative structure. The text given is merely provided as indication of the structure that has been constructed. Such structures may be provided to a neural text generator to be fleshed out into interesting texts, though attention must be paid so that the structure is respected.

The system as described is only a tentative baseline to demonstrate how some of the more elaborate concepts mentioned by Propp might be represented computationally. It constitutes a valuable contribution because it demonstrates that simple computational models can capture complex narratological concepts. Narratologists have in the past observed that computational approaches to story telling fail to address concepts that are considered extremely relevant in the field of narratology (Lönneker et al., 2005). The very simple system described here contemplates the distinction between fabula (the set of moves considered separately) and discourse (the linear sequence that results from interweaving the events for the different moves), simple instances of embedded stories (where someone tells the hero about the villainy or where the false hero claims to have been the protagonist of the hero's deeds), the possibility of having more than one plot line in the story and focalization on different characters.

Early attempts that adopted Propp's views as inspiration, often intended to model types of narrative that differ significantly from the tales in Propp's corpus, and in so doing departed freely from the original framework. Lakoff (1972) proposed a context free grammar (Chomsky, 1957) for the discourse of Russian Fairy tales. His proposal uses phrase structure rules to assign typical characters to variables of character functions, and transformational rules to connect related character functions kidnapping and rescue and to combine more than one move into complex stories. There is no computational implementation associated with this model. Turner (1993) mentions Propp as inspiration of his MINSTREL storytelling system, but he himself confesses that in the final system "'Vladimir Propp's intriguing little grammar was nowhere to be seen' (Chapter 1, page 1). Lang (1997) developed a definite clause grammar that modeled – and could generate – Russian folk tales in the spirit of Afanasyev (1859), but his approach relies more on concepts of story grammar from Rumelhart (1975) and logic programming to model causality and goal-directedness.

Fairclough (2004) in his OPIATE system does include representations of character functions, character roles, and moves, but he adapts these representations heavily to match the requirements of a case-based approach. In doing so, he elaborates character functions into larger units, grouping them into small related sets, redefines the roles into a more generic set of labels, considers a move as the minimum sense

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

(move-0) An apothecary tries to discover from a seamstress where valuables are hidden. The seamstress unwittingly reveals about where valuables are hidden.
(move-1) Katoma attempts to deceive the apothecary. The apothecary is taken in by the deception. Katoma devours the apothecary. Peter always wanted diamond. Peter recovers the apothecary who had been eaten by Katoma. Peter acquires diamond. Peter fights with Katoma. Peter achieves victory over Katoma.
(move-0) The seamstress is captured by the apothecary. A duchess discovers the seamstress is captured by the apothecary. The duchess tells Katoma that the seamstress is captured by the apothecary. Peter calls to action Katoma. Katoma decides to react to the seamstress is captured by the apothecary. Katoma rescues the seamstress from the apothecary. Katoma fights with the apothecary. Katoma achieves victory over the apothecary. Katoma is rewarded.

Table 4: Example of story generated by the system.

carrying unit to use as a case, and relies on case adaptation to cast the roles in the move. OPIATE being a system for interactive visual narrative, the story is always focalized on the player, who plays the role of the hero. Gervás (2016) operates much closer to Propp's account in that it creates plots for Russian folk tales as a sequence of character functions. This approach relies on long-range dependencies identified between character functions in the sequence to capture some of connections left underspecified by Propp, but it departs from Propp's framework in that it creates the fabula for the tales by instantiating the character functions with story actions. To do so, it relies on pre and post conditions for story actions, driven by unification procedure. Specific additional steps are included to assign a cast of characters to the tales and render them as text using a template-based solution. Gervás (2019) proposes a number of extensions to Propp's formalism, such as grouping character functions into axes of interest (covers long-range dependencies between character functions) and explicitly adding roles to capture restrictions on character instantiation. This approach relies on probability of co-instantiation over existing corpus of examples to assign characters to roles.

Another issue is that Propp's account is significantly underspecified. Developers extending it to make it operational often did so by relying on specific technologies that brought their own mechanisms both for representation and operation – such as transformational grammar, case based reasoning or planning. Lakoff (1972) transfers a significant amount of the task of modeling the task to the mechanisms of transformational grammar. In his account these mechanisms carry the burden of filling in the details, but their complexity makes it difficult for the reader to grasp intuitively what is happening in each case. Fairclough (2004) brings in techniques for case-based planning and case adaptation. Again, these may indeed fill in the gaps left in Propp's description, but it is unclear whether the result matches the expectations in terms of what Propp intended. Gervás (2016) proposed explicit representation of long-range dependencies identified between character functions, which are further formalised in Gervás (2019). The proposal in the present paper attempts to bring together all the various extensions of Propp's account identified as useful at some point.

The decisions not to resort to specific techniques in the modeling process is related to the underlying goal of the research, which is not to generate stories but to understand how it is that stories come about, and how they are structured. Whereas specific techniques do provide valuable working solutions, they come at the cost of clouding the principles underlying the process. Both transformational grammar and case-based reasoning work to generate stories, but it is a daunting challenge to understand what the two may have in common that serves as descrition of how narrative works.

A possible criticism is that the system as described here introduces a number of concepts – such as attraction/repulsion between characters functions or characters rendered inactive – that were not considered by Propp. This has been necessary to cover specific points where Propp was not informative about the details underlying some of his descriptions. On this point, we take him at his word when he says: "In the interest of a shorter and more vivid presentation it became necessary to renounce many things which would be relevant for a specialist" (...) "The experienced reader will be able to complete the outline himself." (Vladimir Propp (1928), Author's foreword, 15 July 1927). Any errors made in that process remain, of course, our own.

Conclusions

Propp's Morphology of the Folk Tale has been revisited as inspiration for a computational generator of narrative structures for stories. Extensions of the framework considered in prior attempts have been compiled into a working system that explicitly represents all the steps involved. The initial outcomes show potential for capturing together in a single system of rather simple architecture some complex narratological concepts that in the past have required different specific story generators, each focused on a particular AI technique that favoured the phenomenon chosen as focus.

Acknowledgments

This paper has been partially supported by project CANTOR: Automated Composition of Personal Narratives as an aid for Occupational Therapy based on Reminescence, Grant. No. PID2019-108927RB-I00 (Spanish Ministry of Science and Innovation), 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) and project ADARVE (Análisis de Datos de Realidad Virtual para Emergencias Radiológicas) funded by the Spanish Consejo de Seguridad Nuclear (CSN).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

Abbott, H. 2008. *The Cambridge Introduction to Narrative*. Cambridge Introductions to Literature. Cambridge University Press.

Afanasyev, A. N. 1859. *National Russian Tales (in Russian)*.

Chomsky, N. 1957. *Syntactic structures*. Mouton de Gruyter.

Fairclough, C. 2004. *Story Games And The Opiate System: Using Case Based Planning For Structuring Plots With An Expert Story Director Agent And Enacting Them In A Socially Simulated Game World*. Ph.D. Dissertation, Trinity College Dublin.

Fay, M. P. 2014. *Driving story generation with learnable character models*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Genette, G. 1980. *Narrative discourse : an essay in method*. Cornell University Press.

Gervás, P.; Concepción, E.; and Méndez, G. 2022. Evolutionary construction of stories that combine several plot lines. In *Computational Intelligence in Music, Sound, Art and Design – 11th International Conference, EvoMUSART 2022*. Madrid, Spain: Springer.

Gervás, P. 2016. Computational drafting of plot structures for russian folk tales. *Cognitive computation* 8(2):187–203.

Gervás, P. 2019. Generating a search space of acceptable narrative plots. In *10th International Conference on Computational Creativity (ICCC 2019)*.

Herman, D.; Jahn, M.; and Ryan, M. 2010. *Routledge Encyclopedia of Narrative Theory*. Taylor & Francis.

Lakoff, G. 1972. Structural complexity in fairy tales. *The Study of Man* 1:128–150.

Lang, R. R. 1997. *A formal model for simple narratives*. Ph.D. Dissertation, Tulane University.

Lönneker, B.; Meister, J. C.; Gervás, P.; Peinado, F.; and Mateas, M. 2005. Story generators: Models and approaches for the generation of literary artefacts. In *ACH/ALLC 2005 Conference Abstracts. Proceedings of the 17th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, 15–18.

Propp, V. 1928. *Morphology of the Folk Tale*. Akademija, Leningrad.

Reiter, E. 2010. Natural language generation. *The handbook of computational linguistics and natural language processing* 574–598.

Rumelhart, D. E. 1975. Notes on a schema for stories. In *Representation and understanding*. Elsevier. 211–236.

Turner, S. R. 1993. *Minstrel: a computer model of creativity and storytelling*. Ph.D. Dissertation, University of California at Los Angeles, Los Angeles, CA, USA.

ISBN: 978-989-54160-7-3

Evaluating Creative Short Story Generation in Humans and Large Language Models

Mete Ismayilzada[1,2,4], Claire Stevenson[3], Lonneke van der Plas[2,4]
[1]EPFL, [2]Idiap Research Institute, [3]University of Amsterdam,
[4]Università della Svizzera italiana
mahammad.ismayilzada@epfl.ch, c.e.stevenson@uva.nl, lonneke.vanderplas@usi.ch

Abstract

Story-writing is a fundamental aspect of human imagination, relying heavily on creativity to produce narratives that are novel, effective, and surprising. While large language models (LLMs) have demonstrated the ability to generate high-quality stories, their creative story-writing capabilities remain under-explored. In this work, we conduct a systematic analysis of creativity in short story generation across 60 LLMs and 60 people using a five-sentence cue-word-based creative story-writing task. We use measures to automatically evaluate model- and human-generated stories across several dimensions of creativity, including novelty, surprise, diversity, and linguistic complexity. We also collect creativity ratings and Turing Test classifications from non-expert and expert human raters and LLMs. Automated metrics show that LLMs generate stylistically complex stories, but tend to fall short in terms of novelty, surprise and diversity when compared to average human writers. Expert ratings generally coincide with automated metrics. However, LLMs and non-experts rate LLM stories to be more creative than human-generated stories. We discuss why and how these differences in ratings occur, and their implications for both human and artificial creativity.

Introduction

Story-writing lies at the core of human imagination and communication, serving as a potent means to connect and convey ideas effectively spanning across all human cultures and time periods (Barthes and Duisit 1975). It typically demands creativity, especially when shaping a captivating and persuasive narrative. Creativity is the ability to produce novel, useful, and surprising ideas, and has been widely studied as a crucial aspect of human cognition (Boden 1991; Guilford 1967; Barron 1955; Stein 1953). While humans are natural storytellers, getting machines to generate stories automatically has been a long-time challenge (Ang, Yu, and Ong 2011; Zhu and Ontanón 2010; Gervás and León 2010; Meehan 1977; Lebowitz 1984). However, recently large language models (Zhao et al. 2024) have been shown to produce high-quality short and long stories on arbitrary topics (Yang et al. 2022; Goldfarb-Tarrant et al. 2020). These stories are often evaluated by humans on their global coherence, relevance to the premise, repetitiveness, and general interestingness to the reader (Chhun et al. 2022). However, the extent to which these model-generated stories are truly creative, i.e. *novel*, *effective*, and *surprising*, remains understudied.

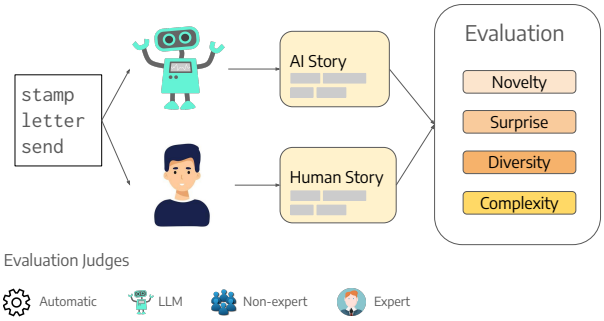


Figure 1: Our study setup illustrated with an example. Both humans and LLMs are asked to write a creative short story using three cue words and evaluated on several creativity metrics by human (experts vs non-experts) and LLM judges.

LLM creativity has generally been evaluated with tasks requiring short responses such as words or phrases. For example, many works have employed the Alternative Uses Test (Guilford 1967), where people and models are asked to come up with creative uses for an everyday object like a *brick* and reported near-human performance results (Stevenson et al. 2022; Góes et al. 2023; Hubert, Awa, and Zabelina 2024; Koivisto and Grassini 2023; Gilhooly 2023). However, the extent to which these results generalize to creativity tasks requiring longer and more complex responses remains underexplored.

Recent works evaluating LLM's ability to produce creative content have shown that models largely fall behind professional human writers (Tian et al. 2024a; Marco, Rello, and Gonzalo 2025; Marco et al. 2024; Chakrabarty et al. 2023). On the other hand, Orwig et al. (2024) finds no significant difference between average human and AI-generated short stories in terms of creativity ratings by non-experts or GPT-4, when comparing humans to ChatGPT models. The extent to which these results hold when considering a collection of different models and evaluations across multiple dimensions of creativity as well as expert and non-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

expert raters remains unclear.

To bridge these gaps, in this work, we conduct a systematic analysis of creativity in short story generation in humans and LLMs. We employ a creative short story generation task that is typically used in psychology to measure the creativity of humans (Prabhakaran, Green, and Gray 2014; Johnson et al. 2023; Orwig et al. 2024). In this task, the goal is to write a short creative story in approximately five sentences based on three cue words such as *stamp*, *letter* and *send*. We evaluate 60 humans and 60 state-of-the-art instruction-finetuned large language models on this task and analyze their performance based on multiple automatic metrics of creativity representing, overall creativity, and the specific aspects of novelty, surprise, diversity, and complexity. Our analysis shows that model-generated stories tend to employ more complex linguistic structures than humans; however, they significantly fall short when it comes to novelty, diversity, and surprise compared to average human writers.

Additionally, we collect fine-grained creativity judgments from non-expert and expert human raters and LLMs for both human and model stories. We find that while non-expert raters and LLMs rate LLM-generated stories as more creative than human-generated stories, expert judgments positively correlate with automated metric results. Our further analysis shows that non-expert human and LLM ratings are driven by linguistic complexity of the stories (e.g. number of words) while expert raters focus on the semantic complexity. Similarly, we find that experts are much more reliable in distinguishing between human-generated and AI-generated stories. Finally, we discuss the implications of our work for both human and machine creativity.

Related Work

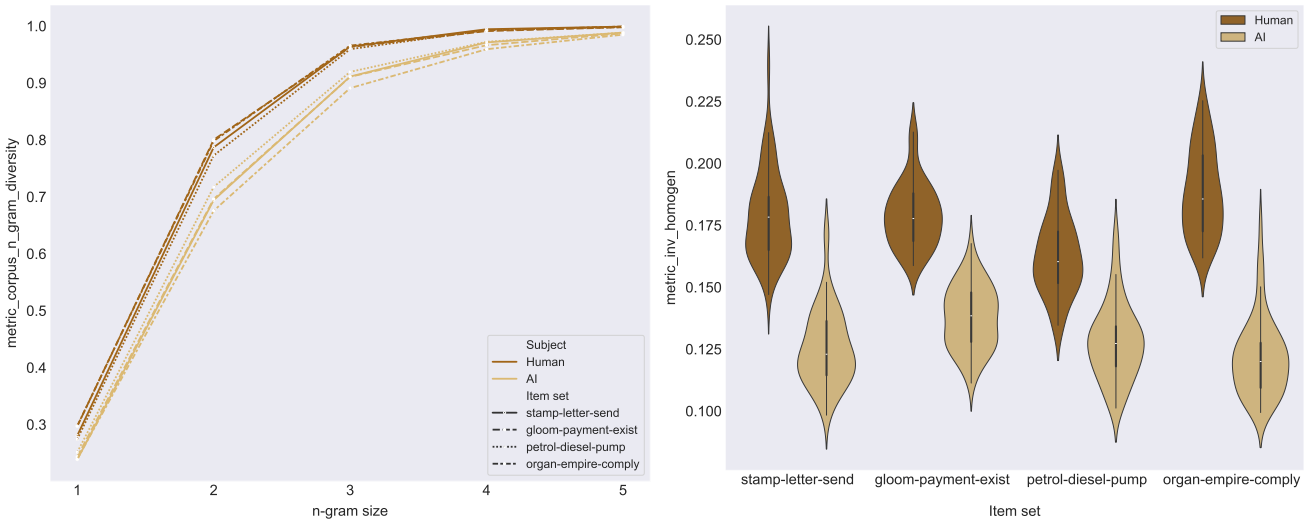
Creativity Evaluation Evaluating creativity is a challenging task due to its subjective nature, however, several evaluation methods have been proposed in the past (Lamb, Brown, and Clarke 2018; Amabile 1982). The most common method of creativity evaluation is called the Consensual Assessment Technique (CAT) (Amabile 1982). CAT relies on the collective judgment of human experts. However, the level of expertise required for a human rater is a subject of debate, with most evidence favoring (quasi-)experts over non-experts as good raters (Kaufman and Baer 2012; Hu, Boggess, and Shepley 2023; Long and Wang ; Ceh et al. 2022; Veale 2015; Lamb, Brown, and Clarke 2015; Gervás 2019). With the rise of powerful generative AI models, LLMs are increasingly being used as judges in many evaluation tasks including creativity (Bavaresco et al. 2024; Chen et al. 2024; Liu et al. 2023; Gilardi, Alizadeh, and Kubli 2023; Chiang and yi Lee 2023; Organisciak et al. 2023). Consequently, in addition to automated metrics, we also conduct evaluations with human expert and non-expert raters and LLMs and study similarities and differences between these different sources of judgment.

Other creativity evaluation methods are generally theoretical frameworks that aim to comprehensively evaluate creativity (Lamb, Brown, and Clarke 2018) or manual psychometric tests that call for brief responses, such as single words or short phrases (Guilford 1967; Torrance 1966). LLM creativity has also predominantly been assessed using these psychometric tasks. For instance, numerous studies have utilized the Alternative Uses Test (Guilford 1967), in which participants, both human and AI models, generate creative uses for everyday objects like a brick, often showing near-human performance (Stevenson et al. 2022; Góes et al. 2023; Hubert, Awa, and Zabelina 2024; Koivisto and Grassini 2023; Gilhooly 2023). In this work, we instead focus on evaluating creativity of humans and LLMs in story generation task, which requires longer and complex responses.

Creative Story Evaluation While most works have focused on evaluating model-generated stories on global coherence, relevance to premise, repetitiveness, and overall interestingness (Chhun et al. 2022), recent studies have also evaluated the creativity of AI models in producing stories (Tian et al. 2024a; Orwig et al. 2024; Johnson et al. 2023; Marco, Rello, and Gonzalo 2025; Marco et al. 2024; Chakrabarty et al. 2023). Chakrabarty et al. (2023) generates short stories using LLMs based on plots from popular fictional works featured in the New York Times and performs a detailed expert evaluation of both the model-generated and original stories. Their findings reveal that LLMs fall considerably short of *experienced writers* in creating truly creative content. Tian et al. (2024a) similarly finds that LLM-generated stories are *non-diverse* and typically *lack suspense* and *tension*. However, these works largely focus on comparing models to award-winning professional writers while our work centers around comparing the creative story-writing abilities of models to average human writers.

Past work closest to ours is the work of Orwig et al. (2024) which also evaluates creative short story generation in both average humans and LLMs using the same five-story generation task. However, our work differs in several major aspects. First, Orwig et al. (2024) compares a collection of humans to only a single model (either GPT-3 or GPT-4) where model story variation is achieved by varying temperature values. This setup makes an implicit assumption of treating the same model with a different decoding parameter as equal to an individual human. However, it remains unclear whether the same findings will hold if a population of different models is compared to a population of humans. Therefore, our study focuses on evaluating collections of both different humans and 60 different LLMs. Second, Orwig et al. (2024) scores creativity with an overall rating and links content to different human memory processes. In our study, we conceptualize creativity as a multifaceted concept and characterize it by the dimensions of novelty, surprise and value (Boden 1991). Finally, Orwig et al. (2024) collects creativity ratings from non-expert raters and GPT-4 while our work considers ratings from both non-expert and expert human raters as well as three LLM judges. We further study where differences between these three types of judges in ratings come from, by predicting creativity ratings from automated metrics across multiple dimensions.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(a) *n*-gram diversity.

(b) Inverse homogenization.

Figure 2: Lexical and semantic diversity scores across all item sets measured by the *n*-gram diversity and inverse homogenization metrics respectively.

Methods

Story Generation Data Collection

We collected data from both humans and LLMs using a creative short story generation task based on three cue words e.g. *stamp*, *letter*, *send* (also known as an item set). We chose this task because it is simple and often employed in psychology to assess human creativity in story generation (Prabhakaran, Green, and Gray 2014; Johnson et al. 2023; Orwig et al. 2024). We use four sets of cue words from (Johnson et al. 2023) where there is either a high semantic distance between words (*gloom, payment, exist* and *organ, empire, comply*) or a low semantic distance (*stamp, letter, send* and *petrol, diesel, pump*). Both humans and models were given the same instructions in English using the following prompt:

> **Instructions**
> You will be given three words (e.g., car, wheel, drive) and then asked to write a creative short story that contains these three words. The idea is that instead of writing a standard story such as "I went for a drive in my car with my hands on the steering wheel.", you come up with a novel and unique story that uses the required words in unconventional ways or settings.
> Write a creative short story using a maximum of five sentences. The story must include the following three words: {items}. However, the story should not be about {boring_storyline}.

In the instructions above, *items* refer to the cue words and *boring_storyline* corresponds to a typical or uncreative storyline that would first come to mind about those cue words. For example, a typical storyline for cue words *stamp, letter, send* could be *stamping a letter and sending it*. We include these hints in the instructions to increase the creativity of both human and LLM generated stories.

Human data were collected from 60 participants (43% female, age: $M = 38.8, SD = 13.6$ years; fluent English speakers residing in the UK, with no language-related disorders and having completed secondary school education) on Prolific[1], a crowd-sourcing platform. Participants not adhering to instructions were removed, resulting in a total of 59 participants.

For a fair comparison, model data were also collected from 60 different models that are diverse in model size, training data and model architecture. The full list of models can be found in Models section. All models were prompted in zero-shot setting with a decoding setup that has been used in previous works to generate creative outputs ($temperature = 0.7, top_p = 0.95$) (Stevenson et al. 2022; Nath, Dayan, and Stevenson 2024).

In total, we collected 480 stories (60 stories for humans and models each across 4 item sets). To make sure all stories are meaningful and comparable in length, we performed some preprocessing to remove outlier stories that contain less than 3 or more than 7 sentences. This filtering step resulted in a total of 431 stories for final evaluation.

Story Evaluation by Automated Metrics

We evaluate both human and model stories using various automated metrics that correspond to different dimensions of creativity. These measures are either common methods relying on the basic linguistic structure of sentences (e.g. *n*-grams, dependency trees) or metrics based on the notion of *semantic distance* that has been shown as an effective automated metric to evaluate creativity (Beaty and Johnson

[1]https://www.prolific.com/

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

2020; Dunbar and Forster 2009; Harbison and Haarmann 2014; Johnson et al. 2023; Prabhakaran, Green, and Gray 2014; Karampiperis, Koukourikos, and Koliopoulou 2014). Semantic distance between two texts is typically computed based on the cosine similarity of embeddings of the texts. More specifically, we consider the following metrics:

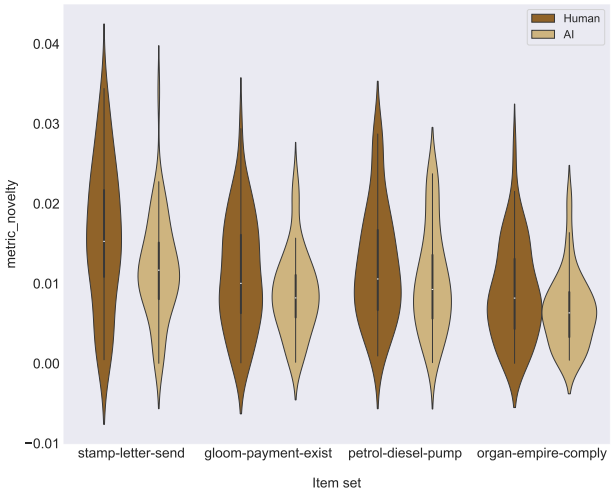


Figure 3: Novelty scores across all item sets.

Diversity Creative stories are often characterized by diverse structures both at the lexical and semantic levels. To measure **lexical diversity**, we employ *n*-gram diversity for values of *n* from 1 to 5 where for a given *n*, *n*-gram diversity is defined as the ratio of the unique *n*-grams to the total number of *n*-grams in a story. To measure **semantic diversity**, we employ a diversity score similar to Padmakumar and He (2023) which we call **inverse homogenization** score. It is defined as the average pairwise distance of a story to all other stories written on the same item set, i.e. $inv_hom(s|t) = \frac{1}{|S_t|-1}\sum_{s' \in S_t \setminus s} semdis(s, s')$ where S_t is a set of stories written on item set t and $semdis$ corresponds to the semantic distance score. We use $1 - cosine_similarity$ as the semantic distance function and a sentence embedding model (`gte-large`) (Li et al. 2023) to compute embeddings of stories.

Novelty One of the major dimensions of creativity is the novelty aspect (Runco and Jaeger 2012). It is typically defined as the measure of how different an artifact is from other known artifacts in its class (Maher 2010). To compute the novelty of a story, we employ the novelty metric from Karampiperis et al. (2014) which defines it as the average semantic distance between the dominant terms (i.e. lemmatized content words) of the story, compared to the average semantic distance of the dominant terms in all stories. More formally, let S_n be a given story, S_G a corpus of all stories across all item-sets and T_n and T_G set of dominant terms respectively for S_n and S_G. Then similar to Johnson et al. (2023), we can define the average semantic distance between the dominant terms for S_n as follows:

$$D(S_n) = \frac{\sum_{i,j=1]}^{|T_n|} semdis(T_{ni}, T_{nj}), i \neq j}{|T_n|} \quad (1)$$

and similarly for S_G as follows:

$$D(S_G) = \frac{\sum_{i,j=1]}^{|T_G|} semdis(T_i, T_j), i \neq j}{|T_G|} \quad (2)$$

Then the novelty of the story S_n can be defined as below (normalized to the $[0, 2]$ space):

$$Nov(S_n) = 2|D(S_n) - D(S_G)| \quad (3)$$

Surprise Also known as unexpectedness, surprise has been shown to play an important role in characterizing a creative artifact (Boden 1991; Grace and Maher 2014). It is typically defined as the artifact's degree of deviation from what is expected (Maher 2010). In the context of a story, surprise can be induced as the story unfolds, i.e., the next sentence that deviates largely from the previous one can create an effect of surprise. Using this temporal dimension, Karampiperis et al. (2014) defines the surprise of a story as the average semantic distances between the consecutive fragments (i.e. sentences) of each story, normalized in the $[0, 2]$ space. More formally, it could be defined as follows:

$$Sur(S_n) = \frac{2}{|F| - 1}\sum_{i=2}^{|F|} |D(F_i) - D(F_{i-1})| \quad (4)$$

where $|F|$ refers to the number of fragments and F_i is the i-th fragment. We employ this metric to compute a value of surprise for each generated story.

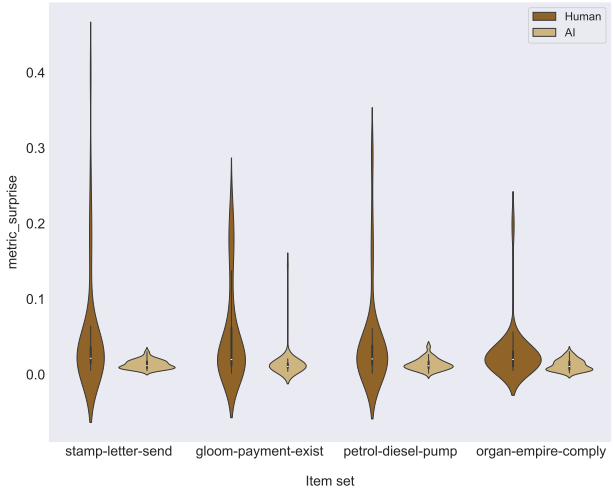


Figure 4: Surprise scores across all item sets.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Complexity Finally, stylistic creativity can be injected into stories by making them linguistically complex. However, lexically and syntactically complex stories can often be unreadable or hard to follow for humans. To measure the level of complexity in generated stories, we employ **lexical** and **syntactic** complexity metrics. Lexical complexity metrics include *number of unique words*, *average word length*, *average sentence length*, and *readability*. For readability, we employ the Flesch reading ease score (Flesch 1940). Syntactic complexity metrics include *part-of-speech tag ratios* (e.g. nouns, adjectives), *average dependency path length*, and *average constituency tree depth*. The average dependency path length is defined as the average of the lengths of dependency paths for each word in a sentence where a dependency path is a sequence of words that are connected with a dependency relation (e.g. *subject of*). For example, in the sentence "in the heart of an ancient library", a dependency path corresponding to the word "in" would be *in-heart-of-library* with a length of 4. The average constituency tree depth on the other hand is defined as the average of the lengths of the branches in a constituency tree of a sentence.

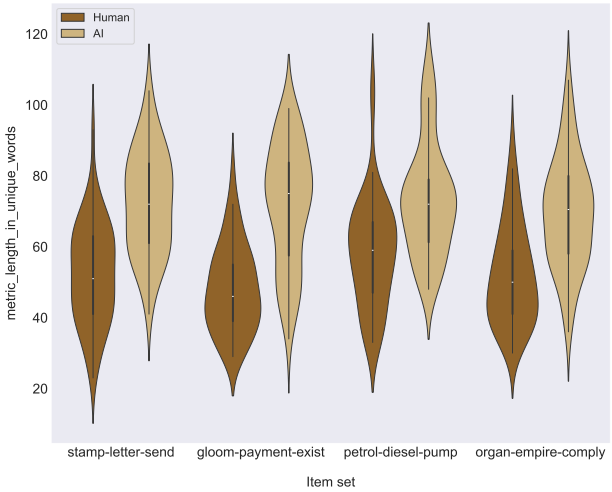


Figure 5: Lexical complexity scores across all item sets measured by story length in number of unique words.

Story Evaluation by Judges

Following the widely popular CAT method (Amabile 1982), we evaluate both human and AI stories using non-expert and expert raters across several dimensions of creativity. In addition, we collect ratings from three LLMs that have been shown to be decent judges of many natural language processing tasks (Bavaresco et al. 2024).

Human Expert Judges We had two trained research assistants—graduate psychology students with creative writing experience— score each of the 431 valid stories on creativity, originality, surprise and effectiveness. For unbiased evaluation, these annotators were not involved in any stage of the study, were unfamiliar with the study design and were fully blinded to which stories stemmed from humans versus AI. Inspired by the original Turing test (Turing 1950), they were also asked to judge each story on whether it was created by human or an AI. The inter-rater reliability of their judgments ranged from good to excellent (ICC=$.62 - .90$). Therefore, for each variable we compute composite scores by taking the means of the two human expert judges.

Human Non-expert Judges We also conducted the same evaluation with an independent sample of 96 non-expert judges recruited via Prolific (49% female, age: $M = 39.8, SD = 12.4$ years) The non-expert raters were -just like the recruited amateur writers- UK residents who spoke English fluently, had no language-related disorders and have completed secondary education. We collected 5 non-expert ratings for each story, the minimum required for reliable creativity judgments by non-experts (Long and Wang). Due to time and cost constraints, we performed this evaluation study on a subset of the stories ($n = 273$) that nonetheless cover all item sets.

To ensure consistent and reliable evaluation, all annotators received detailed instructions on the definitions of creativity, novelty, surprise, and effectiveness. As often is the case in subjective creativity judgments, the variation in judgments across non-experts was much higher than between experts or LLMs (Long and Wang), where the inter-rater reliability of their judgments ranged from fair to good (ICC=$.43 - .71$). Therefore, we choose to use the median rating (i.e., most common rating given across all ratings of the story) rather than a mean to reduce the influence of outliers.

LLM Judges We also prompt three LLMs, i.e. Claude, Gemini and GPT-4, to rate each of the 431 stories on the same five variables. LLM judges had excellent inter-rater reliability for all variables (ICC=$.86 - .94$) except human vs AI judgments (ICC=.43, fair inter-rater agreement). Therefore, we compute the means of the LLM judges for creativity, originality, surprise and effectiveness. For human vs AI judgments we take the median (i.e., most popular vote).

Results

Results of Story Evaluation by Automated Metrics

In this section, we report and discuss the evaluation results using the automated metrics stratified by individual item sets. To measure the effect of the semantic distance within an item set on the creativity of the generated stories, we additionally report the automated metrics results stratified by the type of semantic distance (e.g. low vs. high).

Diversity Figure 2 summarizes the results for lexical and semantic diversity metrics as measured by *n*-gram diversity and inverse homogenization across all model and human groups and item sets. We see that humans consistently display a higher lexical and semantic diversity ($p < 0.0001$).

To get more insight into the type of *n*-grams that are repeated across item sets, we report the frequency of most

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Human		AI	
5-gram	**Count**	**5-gram**	**Count**
“all she felt was gloom”	2	“in the heart of the”	33
“hard to comply with the”	2	“in the heart of a”	20
“went to the petrol station”	2	“the heart of a bustling”	13
“a stamp from my collection”	2	“once upon a time in”	11

Table 1: Most frequent 5-grams in human and AI stories along with their repetition counts.

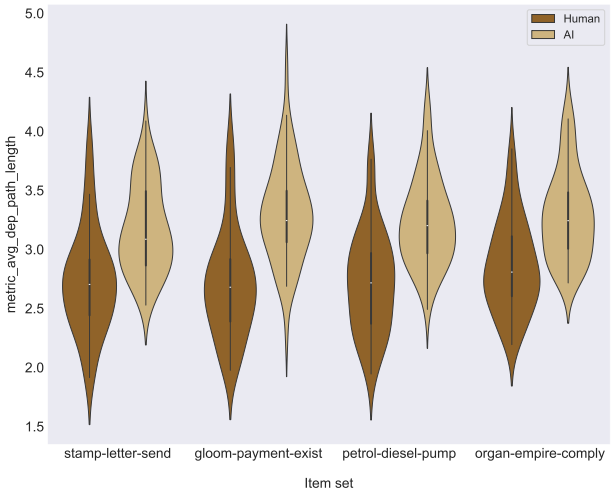


Figure 6: Syntactic complexity scores across all item sets measured by average dependency path length.

common *5*-grams for both model and human stories in Table 1. We see that models tend to follow a story template by repeatedly using certain phrases while human stories exhibit no such behaviour.

Moreover, inverse homogenization scores indicate that model stories tend to share the same themes while human stories are much more diverse in their content. To quantify the diversity of themes, similar to Nath et al. (2024) we perform agglomerative clustering using Ward variance minimization algorithm with a threshold of 0.6 on the embeddings of all stories for a given item set which gives us a set of theme clusters. We find that human stories are characterized by significantly more themes than model stories (Figure 8). To further analyze what type of themes dominate model and human stories, for each group, we combine stories written on a given item set and ask GPT-4 to summarize them in a sentence. Results show that model stories tend to focus on themes of magical transformations or mysterious transactions while human stories speak of human nature, human interactions and responsibilities.

We also analyze the effect of the item set semantic distance on the lexical and semantic diversity results. To do this, we stratify the *n*-gram diversity and inverse homogenization scores across low and high semantic distance groups where we average *n*-gram diversity scores over all *n*-gram sizes (1-5). We find that while human stories written on high semantic distance items are more lexically and semantically diverse than those on low semantic distance items ($p < 0.001$), there is no significant difference between AI stories across different semantic distance categories.

Novelty Figure 3 summarizes the results of novelty metrics across all model and human groups and all item sets. We see that human stories are more novel than those of the models with varying levels of significance across item sets ($p < 0.01$, $p < 0.1$, $p < 0.05$ and $p < 0.05$).

We also similarly analyze the effect of the item set semantic distance on the novelty scores. We observe that human and model stories corresponding to low semantic distance items exhibit more novelty ($p < 0.001$) than those of the high semantic distance items which also aligns with previous findings (Johnson et al. 2023).

Surprise Figure 4 summarizes the results of surprise metrics across all model and human groups and all item sets. We see that human stories are more surprising ($p < 0.001$) than those of the models.

To analyze how the surprise changes as the story unfolds which we call the *surprise profile* of a given model, we plot the averaged raw surprise scores across fragments (i.e. sentences) of the stories written on a given item set in Figure 7. We see that human stories exhibit greater surprise variation across sentences while model stories keep a largely monotonous profile.

We observe no significant difference between low and high semantic distance results for surprise.

Complexity Figures 5 and 6 summarize some of the results for lexical and syntactic complexity metrics respectively across all model and human groups and item sets. What we see is that AI models consistently produce longer stories and their sentences are lexically and syntactically more complex as indicated by larger number of unique words and longer dependency paths per sentence ($p < 0.001$). Additionally, we find that models generally use more nouns and adjectives, while humans use more pronouns and adverbs ($p < 0.001$). To further analyze the type of pronouns used by humans and AI models, we perform an additional analysis on pronoun use and find that humans almost exclusively write their stories from the first or second person perspective, however, models prefer stories centered around third person. Overall, our findings show that

models generally produce grammatically complex and potentially less readable stories.

When we analyze the effect of the item set semantic distance on the complexity scores, we observe a significant difference only with respect to lexical complexity scores for human stories where low semantic distance item sets result in more lexically complex stories than high semantic distance item sets ($p < 0.01$).

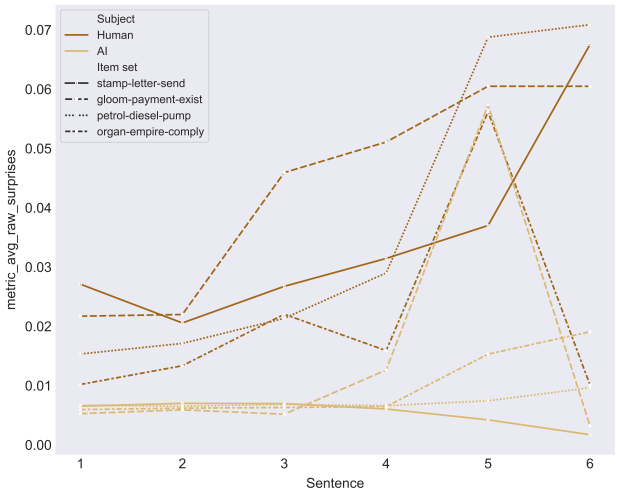


Figure 7: Surprise profile scores across sentence positions averaged over all stories.

Results of Story Evaluation by Judges

The descriptive statistics of the mean ratings per variable can be found in Table 2. For each group of judges the ratings for creativity, originality, surprise and effectiveness are all highly correlated (experts: $.83 < r < .89$; non-experts: $.64 < r < .74$; LLMs: $.92 < r < .99$). Therefore, we focus solely on the creativity ratings and human vs AI judgments for further analyses.

Who's more creative according to our judges? Experts rate human generated stories 1.25 points higher (on a scale of 5) than those generated by LLMs ($t = -19.34, p < .001$). In contrast, both non-experts and LLM judges give AI-generated stories higher scores, where non-experts rate AI stories with 1.19 more points ($t = 12.76, p < .001$) and LLM judges rate AI stories with 1.85 more points ($t = 28.75, p < .001$).

Turing Test: Could LLMs fool our judges and pass for humans? Experts predict the author (human or AI) correctly 94% of the time, outperforming both non-expert and LLM judges. Non-experts make the correct prediction 81% of the time and LLM judges predict human vs. AI with 71% accuracy. Additionally, to gain insight into what factors drive human judgment on whether a story is written by AI or a human, we ask the non-experts to explain their strategy to predict the author of a story. We summarize their comments into several themes of qualities that were attributed the most respectively to human and AI stories. We find that qualities identified by non-experts coincide with our findings from the automated metric analysis that AI models tend to produce linguistically more complex and verbose yet less creative stories than humans (Table 4).

Which automated metrics predict creativity evaluation by the three groups of judges?

We create regression models to predict creativity ratings generated by the three different judges. We choose to use a simple explainable model with one predictor for each category of automated metrics. For *lexical diversity* we use `mean n-gram diversity` of a story and for *semantic diversity* we use `inverse homogenization` (see Methods section for definitions). For *novelty* and *surprise* we use the metrics as described in the Methods section. For syntactic and lexical complexity, since all metrics are highly correlated, we choose a single metric that correlates strongest with creativity ratings to represent each construct. For *syntactic complexity* we select `average constituency tree depth` and for *lexical complexity* we select `number of unique words`.

Using these metrics we run the following regression analysis to predict creativity ratings for stories by the three groups of judges: $creativity \sim semantic_diversity + lexical_diversity + novelty + surprise + syntactic_complexity + lexical_complexity$.

As can be seen in Table 3, expert creativity ratings are best predicted by higher semantic diversity, surprise and lexical diversity scores. For non-experts, creativity ratings increase with higher lexical complexity (i.e., number of unique words). Surprisingly, non-expert creativity ratings decrease with higher semantic diversity, novelty and surprise scores. For LLM judges, we see the same pattern of predictors as for non-experts, where the best positive predictor of creativity ratings is lexical complexity and that semantic diversity, novelty and surprise are each negatively related to creativity ratings.

Discussion

In this work, we study and compare the creative short story generation abilities of humans and LLMs using a five-sentence short story generation task based on cue words. We use both automated metrics and judgments of non-expert and expert humans as well as LLMs to evaluate the creativity of the stories across several dimensions such as novelty, surprise, diversity, and complexity. For the complexity measures, we leverage common metrics relying on the linguistic structures of the stories at both lexical and syntactic levels. Then we analyze the results across all item sets and study the similarities and differences between the evaluations of different judges.

Our analysis using the automated metrics shows that LLMs produce linguistically and stylistically more complex stories than humans as indicated by higher lexical and syntactic complexity results. However, human stories consis-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

	Expert Judges		Non-expert Judges		LLM Judges	
Creator:	Human	AI	Human	AI	Human	AI
creativity	3.58 (±0.78)	2.33 (±0.55)	2.45 (±0.77)	3.65 (±0.78)	2.41 (±0.70)	4.26 (±0.64)
originality	3.96 (±0.79)	2.70 (±0.73)	2.48 (±0.76)	3.47 (±0.75)	2.33 (±0.73)	4.22 (±0.72)
surprise	3.20 (±0.89)	1.68 (±0.61)	2.21 (±0.81)	2.96 (±0.78)	2.01 (±0.79)	3.74 (±0.77)
effectiveness	3.87 (±0.71)	2.17 (±0.70)	2.70 (±0.73)	3.33 (±0.76)	2.86 (±0.58)	4.14 (±0.52)

Table 2: Means (±SDs) of aggregated ratings by expert, non-expert and LLM judges.

	Expert Judges			Non-expert Judges			LLM Judges		
predictor	$\beta(SE)$	t	p	$\beta(SE)$	t	p	$\beta(SE)$	t	p
$semantic_diversity$	.56 (.04)	14.65	***	-.32 (.05)	-6.31	***	-.58 (.04)	-14.99	***
$lexical_diversity$	.09 (.04)	2.32	*	.01 (.05)	0.26		.05 (.04)	1.21	
$novelty$	-.03 (.03)	-0.80		-.20 (.05)	-4.48	***	-.21 (.04)	-5.87	***
$surprise$	.09 (.04)	2.36	*	-.08 (.05)	-1.57		-.08 (.04)	-2.10	*
$syntactic_complexity$	-.04 (.04)	-1.01		-.0006 (.06)	-0.10		-.06 (.04)	-1.35	
$lexical_complexity$	.03 (.05)	0.58		.37 (.06)	6.06	***	.42 (.05)	8.78	***

Table 3: Regression coefficients (SE) and t-test results for predictions of creativity ratings for each group of judges. Where p-value significance is represented as follows: '***'<.001 , '**'<.01 , '*'<.05 , and ' '>=.05.

tently exhibit higher novelty, surprise, and lexical and semantic diversity while being linguistically much less complex and easier to read. Our findings are in line with some previous work comparing LLM creativity to humans in story generation (Chakrabarty et al. 2023; Tian et al. 2024a; Marco et al. 2024; Marco, Rello, and Gonzalo 2025) and creative problem-solving (Tian et al. 2024b). On the other hand, some past works have found no significant difference between overall LLM and human creativity in short story generation when comparing a population of humans to GPT-3 and GPT-4 when judged by non-experts and GPT-4 (Orwig et al. 2024). However, our fine-grained analysis considering multiple dimensions of creativity and a population of 60 different LLMs evaluated by automated metrics and experts reveals significant gaps between human and LLM stories across all major dimensions of creativity in favor of humans.

Particularly, we find that our automated metric results highly correlate with expert judgments, while LLM and non-expert judgments tend to rate LLM stories as more creative than human stories. Our further analysis shows that this discrepancy stems from the underlying factors driving the different judgments. More specifically, expert judgments are driven by the diversity and surprise aspects of the stories which are essential to creativity while non-expert and LLM judgments highly correlate with sheer lexical complexity such as the number of unique words, which does not necessarily imply semantic complexity. In fact, our parts-of-speech analysis shows that LLMs tend to overuse rare adjectives and complex syntactical structures. Moreover, past works generally find expert judgments as more reliable evaluators of creativity than those of non-experts (Kaufman and Baer 2012; Hu, Boggess, and Shepley 2023; Long and Wang ; Ceh et al. 2022; Veale 2015; Lamb, Brown, and Clarke 2015; Gervás 2019). Similarly, LLMs-as-judges have been shown to be unreliable (Chakrabarty et al. 2023; Chhun, Suchanek, and Clavel 2024) and biased towards their own generations (Wataoka, Takahashi, and Ri 2024; Panickssery, Bowman, and Feng 2025).

Our work has several implications. The complexity vs. creativity gap shows that humans and LLMs have different interpretations of what it means to be creative for stories. While humans prefer telling a simple story from their perspective that is nonetheless surprising and original, LLMs, however, represent creativity with lexically and syntactically overloaded sentences narrated from the third person perspective and that typically focus on a few repetitive themes. Additionally, the fact that non-experts and LLMs tend to evaluate AI stories as more creative could mean that complexity creates the illusion of being more creative to the untrained eye. This behaviour can be due to several factors involved in training LLMs such as the training data, pre-training and post-training optimizations. For example, aligning LLMs with human feedback to be more helpful has been attributed to result in strong verbosity bias (Saito et al. 2023) and diversity reduction (Padmakumar and He 2023) in creative tasks. Our findings call for a more comprehensive evaluation of creativity and can inform future work on designing methods to improve the creativity of LLMs (Ismayilzada et al. 2024). Potential directions can include developing new prompt engineering (Mehrotra, Parab, and Gulwani 2024; Nair, Gizzi, and Sinapov 2024; Summers-Stay, Lukin, and Voss 2023; Tian et al. 2024b) or optimization techniques (Broad et al. 2021; Bunescu and Uduehi 2019; Elgammal et al. 2017) or steering internal mechanisms of LLMs using mechanistic interpretability (Bereska and Gavves 2024).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Theme	Human Story Qualities	AI Story Qualities
Emotional depth	“emotional”, “relatable”, “personal”, “first-person”, “evoking feelings”, “empathy”	“formulaic format”, “more abstract”, “bad flow”, “no feeling”, “no depth”
Verbosity	“simple”, “shorter”, “to the point”, “casual”, “pragmatic”, “less adjectives”	“verbose”, “wordy”, “elaborate”, “descriptive”, “repeating phrases” (“once upon a time”, “’in a world where ’), “lots of pointless adjectives”
Enjoyability	“more depth”, “more sense”, “rythmic”, “easier to follow”, “more enjoyable”	“gibberish”, “hard to read”, “nonsense”, “convoluted”
Plausibility	“everyday life”, “mundane“, “spelling and grammar errors”,	“fantastic”, “unrealistic”, “far-fetched”

Table 4: Most attributed qualities to human and AI stories by non-experts grouped by shared themes.

Limitations

While our study provides a comprehensive analysis of a wide range of language models using fine-grained creativity metrics, it does have a few limitations. First, although all models were prompted using a decoding setup that has been used in previous work to favor creative output and idea diversity (Stevenson et al. 2022; Nath, Dayan, and Stevenson 2024), we did not explore alternative decoding or prompting strategies due to the high cost of open-ended evaluation across many models. Second, the novelty metric we employ is reference-based, meaning its outcomes depend heavily on the chosen reference corpus. In our case, this was the set of human- and AI-written stories for each item set. Using reference-based scoring is a shortcoming in nearly all creativity research that uses uniqueness to define novelty (Silvia et al. 2008). However, the construct validity of frequency-based uniqueness assessments is also high and their generalizability to stories from other populations increases with sample size (Lee 2008). Therefore, future studies could validate our findings by including more stories in reference-based metrics. Third, we used the minimum recommended number of expert and non-expert raters per story, i.e., two for (quasi-)experts —as these are generally highly reliable as in our study— and five non-experts —as these are generally less reliable— (Long and Wang). Having more raters could have improved the reliability of our findings. Furthermore, including professional creative writers as experts could perhaps provide further improve reliability (although research suggests that quasi-experts are as reliable as experts (Long and Wang), sometimes more so (Tan et al. 2015)). Lastly, although our creativity metrics are effective for evaluating story generation, they cannot fully capture the broader cultural and social dimensions of creativity or the depth of truly original and imaginative language use.

Ethics

All authors declare no conflicts of interest. No artificial intelligence assisted technologies were used in this research or the creation of this article. This research received approval from a local ethics board on September 11, 2024. All study materials will be publicly available[2].

[2]https://github.com/mismayil/creative-story-gen

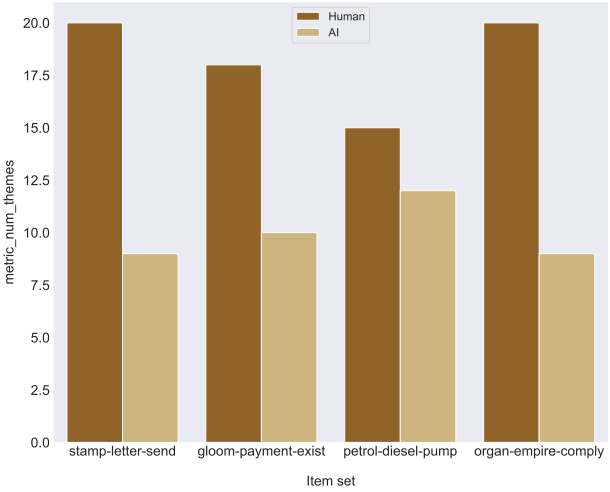


Figure 8: Number of themes for human and AI stories.

Acknowledgements

This publication is part of the project C_LING (grant 205121_207437, Swiss National Science Foundation) awarded to Lonneke van der Plas.

Models

We evaluate the following large language models in our study. Variants or sizes in number of parameters (billions) if known are given in parenthesis: GPT-3.5 (175B), GPT-4, GPT-4o, Claude-3 (Opus), Claude-3.5 (Sonnet, Haiku), Gemini-1.5 (Flash, Pro), Gemma-2 (9B, 27B), Llama-3.1 (8B, 70B, 405B), Llama-3.2 (1B, 3B), Grok-2 (314B), MPT (7B, 30B), DBRX (132B), DeepSeek (7B, 67B), Ministral (3B, 8B), Mistral (7B, 12B, 22B, 123B), Mixtral (13B, 39B), Nouse Hermes 2 (13B), Qwen-2.5 (7B, 72B), Qwen-2.5 Coder (32B), Reka (7B, 21B, 67B), GLM-4 (130B), Jamba-1.5 (12B, 94B), Phi-3 (3.8B, 7B, 14B), Phi-3.5-MoE (6.6B), Aya Expanse (8B, 32B), Command R+ (104B), Nemotron (4B, 340B), Yi-1.5 (9B, 34B), Baichuan-2 (7B, 13B), Zamba-2 (7B), Granite-3.0 (2B, 8B), StaleLM (3B, 12B), OLMo-2 (7B, 13B), LFM (40B).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

Amabile, T. M. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology* 43(5).

Ang, K.; Yu, S.; and Ong, E. 2011. Theme-based cause-effect planning for multiple-scene story generation. In *ICCC*.

Barron, F. 1955. The disposition toward originality. *The Journal of Abnormal and Social Psychology* 51(3).

Barthes, R., and Duisit, L. 1975. An introduction to the structural analysis of narrative. *New Literary History* 6.

Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernández, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.; Koller, A.; et al. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Beaty, R. E., and Johnson, D. R. 2020. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods* 53.

Bereska, L., and Gavves, E. 2024. Mechanistic interpretability for ai safety - a review. *ArXiv* abs/2404.14082.

Boden, M. A. 1991. The creative mind : myths & mechanisms.

Broad, T.; Berns, S.; Colton, S.; and Grierson, M. 2021. Active divergence with generative deep learning - a survey and taxonomy. In *ICCC*.

Bunescu, R. C., and Uduehi, O. O. 2019. Learning to surprise: A composer-audience architecture. In *ICCC*.

Ceh, S. M.; Edelmann, C.; Hofer, G.; and Benedek, M. 2022. Assessing raters: What factors predict discernment in novice creativity raters? *The Journal of Creative Behavior* 56(1).

Chakrabarty, T.; Laban, P.; Agarwal, D.; Muresan, S.; and Wu, C.-S. 2023. Art or artifice? large language models and the false promise of creativity. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the EMNLP 2024*.

Chhun, C.; Colombo, P.; Suchanek, F. M.; and Clavel, C. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on CL.

Chhun, C.; Suchanek, F. M.; and Clavel, C. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *TACL*.

Chiang, C.-H., and yi Lee, H. 2023. Can large language models be an alternative to human evaluations? In *ACL*.

Dunbar, K., and Forster, E. 2009. Creativity evaluation through latent semantic analysis.

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.

Flesch, R. 1940. How to write plain english. *University of Canterbury.*

Gervás, P., and León, C. 2010. Story generation driven by system-modified evaluation validated by human judges. In *ICCC*.

Gervás, P. 2019. Exploring quantitative evaluations of the creativity of automatic poets. *Computational creativity: The philosophy and engineering of autonomously creative systems*.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *PNAS* 120.

Gilhooly, K. 2023. Ai vs humans in the aut: simulations to llms. *Journal of Creativity*.

Góes, F.; Sawicki, P.; Grzes, M.; Volpe, M.; and Watson, J. 2023. Pushing gpt's creativity to its limits: Alternative uses and torrance tests. In *ICCC*.

Goldfarb-Tarrant, S.; Chakrabarty, T.; Weischedel, R.; and Peng, N. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the EMNLP 2020*.

Grace, K., and Maher, M. L. 2014. What to expect when you're expecting: The role of unexpectedness in computationally evaluating creativity. In *ICCC*. Ljubljana.

Guilford, J. 1967. *The Nature of Human Intelligence*. McGraw-Hill series in psychology. McGraw-Hill.

Harbison, J. I., and Haarmann, H. J. 2014. Automated scoring of originality using semantic representations. *Cognitive Science* 36.

Hu, L.; Boggess, M.; and Shepley, M. M. 2023. Comparing expert, quasi-expert, and novice evaluations of award-winning design products using the consensual assessment technique. *Creativity Research Journal* 35(4).

Hubert, K. F.; Awa, K. N.; and Zabelina, D. L. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14.

Ismayilzada, M.; Paul, D.; Bosselut, A.; and van der Plas, L. 2024. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*.

Johnson, D. R.; Kaufman, J. C.; Baker, B. S.; Patterson, J. D.; Barbot, B.; Green, A. E.; van Hell, J.; Kennedy, E.; Sullivan, G. F.; Taylor, C. L.; et al. 2023. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods* 55(7).

Karampiperis, P.; Koukourikos, A.; and Koliopoulou, E. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. *IEEE International Conference on Advanced Learning Technologies*.

Kaufman, J. C., and Baer, J. 2012. Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal* 24(1).

Koivisto, M., and Grassini, S. 2023. Best humans still out-

ISBN: 978-989-54160-7-3

perform artificial intelligence in a creative divergent thinking task. *Scientific Reports* 13.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2015. Human competence in creativity evaluation. In *ICCC*.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)* 51(2).

Lebowitz, M. 1984. Creating characters in a story-telling universe. *Poetics* 13(3).

Lee, S. 2008. Commentary: Reliability and validity of uniqueness scoring in creativity assessment. *Psychology of Aesthetics, Creativity, and the Arts* 2:103–108.

Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; and Zhang, M. 2023. Towards general text embeddings with multi-stage contrastive learning.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Conference on EMNLP*.

Long, H., and Wang, J. Dissecting reliability and validity evidence of subjective creativity assessment: A literature review. *Educational psychology review*.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Network Conference on Creativity and Innovation in Design*.

Marco, G.; Gonzalo, J.; Mateo-Girona, M.; and Santos, R. D. C. 2024. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the EMNLP 2024*.

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 *Proceedings of the 31st International Conference on Computational Linguistics*.

Meehan, J. R. 1977. Tale-spin, an interactive program that writes stories. In *IJCAI*.

Mehrotra, P.; Parab, A.; and Gulwani, S. 2024. Enhancing creativity in large language models through associative thinking strategies. *ArXiv* abs/2405.06715.

Nair, L.; Gizzi, E.; and Sinapov, J. 2024. Creative problem solving in large language and vision models - what would it take? In *Findings of the ACL: EMNLP 2024*.

Nath, S. S.; Dayan, P.; and Stevenson, C. 2024. Characterising the creative process in humans and large language models. In *ICCC*.

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.

Orwig, W.; Edenbaum, E. R.; Greene, J. D.; and Schacter, D. L. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of creative behavior* 58(1).

Padmakumar, V., and He, H. 2023. Does writing with language models reduce content diversity? *ArXiv* abs/2309.05196.

Panickssery, A.; Bowman, S.; and Feng, S. 2025. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37.

Prabhakaran, R.; Green, A. E.; and Gray, J. R. 2014. Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior research methods* 46.

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

Saito, K.; Wachi, A.; Wataoka, K.; and Akimoto, Y. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Silvia, P. J.; Winterstein, B. P.; Willse, J. T.; Barona, C. M.; Cram, J. T.; Hess, K. I.; Martinez, J. L.; and Richard, C. A. 2008. Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*.

Stein, M. I. 1953. Creativity and culture. *The journal of psychology* 36(2).

Stevenson, C. E.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. L. J. 2022. Putting gpt-3's creativity to the (alternative uses) test. In *ICCC*.

Summers-Stay, D.; Lukin, S. M.; and Voss, C. R. 2023. Brainstorm, then select: a generative language model improves its creativity score.

Tan, M.; Mourgues, C.; Hein, S.; MacCormick, J.; Barbot, B.; and Grigorenko, E. 2015. Differences in judgments of creativity: How do academic domain, personality, and self-reported creativity influence novice judges' evaluations of creative productions? *Journal of Intelligence* 3(3):73–90.

Tian, Y.; Huang, T.; Liu, M.; Jiang, D.; Spangher, A.; Chen, M.; May, J.; and Peng, N. 2024a. Are large language models capable of generating human-level narratives? In *Proceedings of the EMNLP 2024*.

Tian, Y.; Ravichander, A.; Qin, L.; Le Bras, R.; Marjieh, R.; Peng, N.; Choi, Y.; Griffiths, T.; and Brahman, F. 2024b. MacGyver: Are large language models creative problem solvers? In *Proceedings of the NAACL 2024*.

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Turing, A. M. 1950. *Computing machinery and intelligence*.

Veale, T. 2015. Game of tropes: Exploring the placebo effect in computational creativity. In *ICCC*.

Wataoka, K.; Takahashi, T.; and Ri, R. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Yang, K.; Tian, Y.; Peng, N.; and Klein, D. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the EMNLP 2022*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2024. A survey of large language models.

Zhu, J., and Ontanón, S. 2010. Towards analogy-based story generation. In *ICCC*. Citeseer.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Playing Along: Building AI Agents for Co-Creation of Improvised Stories

Idan Dov Vidra[1] and **Gal Kimron**[1] and **Lior Noy**[2] and **Ariel Shamir**[1]
[1]Reichman University, arik@runi.ac.il, idan.vidra@post.runi.ac.il, gal.kimron@post.runi.ac.il
[2]Ono Academic College, lior.noy@ono.ac.il

Abstract

This paper studies human-agent co-creation of improvised stories, investigating whether Large Language Models can effectively engage in an improvisational practice known as the "Yes! and..." game. We demonstrate how AI systems can participate successfully in improvisational co-creation, moving beyond response generation to collaborative story-telling. We provide a systematic framework for evaluating creative AI outputs in improvisational contexts, combining human evaluation with computational metrics. Our evaluations show that stories co-created with an AI agent received similar ratings to human-human collaborations and were perceptually indistinguishable in blind testing. This shows how AI's creative invention abilities - which can be limitations in factual tasks - become assets in collaborative storytelling. More generally, our approach presents the "Yes! and..." game as a novel model-system for studying improvised co-creativity in a well-defined and measurable setup.

Introduction

While co-creative storytelling and human-AI collaboration received growing attention (Rezwana and Maher 2023; Branch, Mirowski, and Mathewson 2021; Concepcion, Gervas, and Mendez 2019; Veale, Wicke, and Mildner 2019; Jacob et al. 2019; Branch et al. 2024), creative improvisation, particularly in a minimal, text-based setting, remains underexplored. This work advances computational creativity in three ways. First, it demonstrates how AI systems can participate successfully in improvisational co-creation, moving beyond response generation to collaborative storytelling. Second, it provides a systematic framework for evaluating creative AI outputs in improvisational contexts, combining human evaluation with computational metrics. Third, it shows how AI's narrative invention abilities serve as creative assets in collaborative storytelling.

We focus on human-AI co-creation of improvised stories, investigating whether Large Language Models (LLMs) can effectively engage in a specific improvisational practice - the "Yes! and..." game. The "Yes! and..." game applied here is an example of a fundamental tenet of improvisation, the "Yes! and..." approach (Berliner 2009; Nachmanovitch 1991; Johnstone 1987), where improvisers need both to accept their partners' offers (the *Yes!* part) and then continue the flow from this position (the *and...* part). In the "Yes! and..." game we apply here, two participants build a shared "adventure" by always accepting and elaborating on each other's contributions (Johnstone 1987).

The "Yes! and..." approach is different from other forms of narrative co-creation. Unlike turn-based storytelling or guided narrative generation, participants must respond spontaneously without planning while building upon each other's contributions. This creates a balanced creative dynamic of acceptation and extension, where both participants actively shape the emerging narrative through real-time interaction.

The narratives created are jointly invented and participants inevitably generate *imaginative* content. For an AI agent to participate effectively, it must therefore engage in creative invention — envisioning shared experiences and fictional scenarios that go beyond factual content.

LLMs, trained on vast text corpora, can produce coherent documents and engage in conversations (Shanahan 2023; Zhao et al. 2023; Liu et al. 2023; Brown et al. 2020; Thoppilan et al. 2022). While these models often produce hallucinations— ungrounded outputs that deviate from factual accuracy (Agrawal, Mackey, and Kalai 2023; Ji et al. 2023)— such outputs are increasingly seen as valuable in creative processes (Jiang et al. 2024; Halperin and Lukin 2024; Chakraborty and Masud 2024). This aligns with foundational theories of AI creativity: Boden's categorization suggests that "imagined" elements expand narrative search spaces, while Ritchie's criteria emphasize balancing novelty and usefulness (Boden 2004; Ritchie 2001). Established frameworks like SPECS (Jordanous 2012) and the Novelty-Value-Surprise model (Grace et al. 2015) ground our approach to evaluating creativity in co-created stories.

With the "Yes! and..." game we adopt a reduced, model-system approach: we use a textual "Yes! and..." scenario that strips away visual, auditory, and embodied dimensions. Similar to model systems in other experimental paradigms (Noy, Dekel, and Alon 2011; Noy 2014; Rafner et al. 2023), this minimalistic setup isolates core mechanisms of co-creative improvisation and enables systematic analysis. To investigate whether LLMs can participate in "Yes! and..." games at a level comparable to humans, we developed AI agents using a chat interface and evaluated the resulting stories through human ratings, computational metrics, and a "Turing Test"-style survey.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Our main contributions include introducing a text-only version of the "Yes! and..." game for human-human (HH) interactions and extending it to human-AI (AIH) partnerships, demonstrating through empirical evaluation that LLM-based agents achieve comparable ratings to humans across creativity metrics and are indistinguishable in blind testing. We develop and validate evaluation measures combining human ratings with computational metrics of novelty and surprise, showing strong inter-rater reliability ($ICC > 0.75$) on the two core creative dimensions—creativity and surprise—while reliability for the remaining dimensions was moderate to poor. Our minimalistic paradigm enables controlled study of human-AI creative collaboration, building on model system approaches from joint action research (Noy, Dekel, and Alon 2011; Noy 2014) to isolate core mechanisms of improvisational interaction.

Previous Work

Our work builds upon three main research areas: computational co-creativity, improvisational AI, and creative evaluation frameworks.

Computational Co-Creativity with LLMs Recent work has demonstrated LLMs' potential for creative collaboration with humans (Franceschelli and Musolesi 2024; Yuan et al. 2022; Gero, Liu, and Chilton 2022). Although traditionally viewed as flaws for factual tasks, LLMs' creative inventions are increasingly recognized as valuable in creative contexts (Jiang et al. 2024; Halperin and Lukin 2024; Chakraborty and Masud 2024). Our work extends this perspective by showing how apparent limitations can become assets in improvisational storytelling.

Improvisational AI Systems Early pioneering work by Hayes-Roth on directed improvisation (Hayes-Roth et al. 1994) and Bates' Oz Project (Bates, Loyall, and Reilly 1994) established foundational principles for computational improvisation. Building on these foundations, recent research has focused primarily on music (Hoffman and Weinberg 2010) and theatrical performance (Branch et al. 2024). Researchers have explored various aspects of improvised interaction, from movement-based object interactions (Jacob et al. 2019) to complementary AI partnerships in storytelling (Veale, Wicke, and Mildner 2019). Recent work has extended this to evaluating LLMs in professional theatrical settings (Branch et al. 2024). Particularly relevant is work by (Cho and May 2020) on dialogue systems adhering to improvisational principles. We build upon these foundations while introducing a minimalist text-based framework that enables systematic study of human-AI creative dynamics.

Evaluation Frameworks Our evaluation approach draws on established frameworks for assessing computational creativity, particularly SPECS (Jordanous 2012) and the Novelty-Value-Surprise model (Grace et al. 2015). We extend these frameworks by combining human evaluation with computational metrics specifically designed for improvisational co-creation. This extends previous work on evaluating co-creative systems while introducing new measures for improvisational interaction (Karimi et al. 2018).

Our approach advances existing work in three key ways: (1) we introduce a minimal, controlled environment for studying human-AI improvisation, (2) we provide evaluation metrics specifically designed for improvisational co-creation, and (3) while some recent work has begun to recognize the creative potential of LLM beyond factual constraints (Jiang et al. 2024; Halperin and Lukin 2024), we demonstrate how these generative capabilities become creative assets in improvisational storytelling through empirical evaluation.

Methods

Textual "Yes! and..." Game for Human Interaction

We developed a chat-based platform that allows pairs of participants to engage in "Yes! and..." games through text-based interaction. We call these human-human (HH) interactions.

We randomly paired participants and instructed them to play an improvisational joint storytelling game following the "Yes! and..." principle. After reading the instructions, participants took turns creating sentences that evolved their shared story. The first participant was chosen at random. While we suggested starting each contribution with "Yes! and...", this was not enforced. We placed no restrictions on sentence length, resulting in an average of 11.35 words per sentence and 13.9 sentences per game.

Game Play The game sessions followed a structured progression: participants first played a short game, followed by a longer game (240-360 seconds), and finally an unrestricted game where they could reach a natural conclusion. Some participants engaged in this sequence with multiple partners, returning to the platform's main window to be randomly paired again. For privacy protection, we only recorded the text of the story and the timestamps, without collecting personal information.

Following each session, participants participated in a structured debriefing in which they could share their experiences and learn about the research goals. Additional details about game protocols, data collection, and full game examples are provided in the Supplemental Information (SI).

AI-Human "Yes! and..." Game Implementation

We developed a digital platform to enable "Yes! and..." games between human participants and an agent based on GPT-4. This platform was similar (but not identical) to the one used for HH interactions (see SI: Interface Screenshots). The primary difference was the incorporation of GPT-4 as one of the two players. The Gradio package (Python) was utilized for managing the chat interactions (Abid et al. 2019). We call these AIH-interactions. For technical reasons, all AIH games were designed so that humans made the first contribution to the story.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Game Play Participants were initially led to believe they were interacting with another human. Participants received exactly the same instructions as those in the HH game sessions. This approach was adopted to study natural interaction without preconceived biases from playing with an agent. After each session, participants were informed that their counterpart was actually an AI model.

Model Configuration The agent used GPT-4 accessed through OpenAI's Chat Completion API, configured with a presence penalty of 1.7 to encourage response diversity while maintaining default temperature. To mirror human interaction patterns, we introduced random delay between 21-25 seconds in the model responses, based on an analysis of HH response times, we also limited responses to 11 words. Games were designed to conclude after approximately 10 messages, either through participant choice or an AI-generated concluding message. Each interaction incorporated the full conversation history for context coherence.

Model's Prompt We refined the prompt used for GPT-4 based on participants' feedback, avoiding responses which were too quick or too long, and seemed non-human (see complete list of suggestions in SI: Development of the AI Agent's Prompt). The final prompt was:

> "You are a model built to play an interactive "Yes! and..." game with people. In this game, we'll build a story together by taking turns adding sentences. Each addition should build upon the previous statement in the context of the entire story, showing agreement and creativity. Be adaptable and willing to explore various tones, including darker themes. Feel free to start with "Yes! and..." or other affirming phrases like 'Yeah...', 'Totally...', 'Right and...', 'definitely...'. Keep responses concise, engaging, and in simple language, using 11 words or less."

AIH games typically contained up to six sentences from the GPT-4 model, similar to the average length of HH games. These constraints resulted in interactions that closely matched the natural flow of HH games while maintaining consistent engagement throughout the session.

Data Collection and Analysis

We collected stories through various platforms, gathering 141 HH games from 82 participants and 86 AIH games from 45 participants (collection methods detailed in SI: Summary of HH Game Collection Methods, Statistics for AIH Games Collections). Story lengths varied by design, as participants progressed from short to longer to unrestricted games. All stories underwent quality filtering based on two criteria: minimum length (> 3 lines) and structural coherence (adherence to "Yes! and..." game structure and basic narrative flow). After filtering, our final dataset comprised 129 HH and 75 AIH stories. Estimated game durations, based on average response times per turn, averaged 548.71 seconds (SD = 220.82), with the high variance reflecting our intentionally varied game durations (detailed timing analysis in SI: Estimated Game Times).

Human-Human Stories Collection methods evolved from initial Google Docs to a dedicated web platform. Stories were excluded for containing fewer than three lines (5 stories) or failing quality criteria (10 stories) as judged independently by two researchers. The final HH dataset averaged 18.42 sentences (SD = 4.28) and 8.81 words per sentence (SD = 2.21).

AI-Human Stories AIH stories were collected using GPT-4 as the AI agent, selected based on preliminary tests demonstrating its superior engagement compared to other models. Data collection evolved through three phases, with progressive refinements to the prompt and interface based on participant feedback (details in SI: Statistics for AIH Games Collections). The system concatenated all previous messages as context for each AI response. The final AIH dataset averaged 13.08 sentences (SD = 5.3) and 14.13 words per sentence (SD = 4.6).

Story Evaluation Process

A total of 453 participants from university took part in the experiment, without prior experience in improvised story co-creations. We recruited participants from the student body, with the incentive of receiving course credit equivalent to 30 minutes of their time. The experiment was approved by the ethical committee and was performed on Qualtrics.

To reduce the total evaluation time per participant, we randomly selected 40 HH stories and 24 AIH stories from our dataset, excluding stories that deviated from the narrative-building format or contained excessive context-specific references (selection criteria detailed in SI: Data Selection and Preprocessing). All non-English stories were translated to English before evaluation. We divided these 64 stories into eight batches, each participant evaluating one batch of 8 stories presented in random order. Initial cohort sizes were 68, 64, 98, 100, 97, 85, 145, and 164 participants, respectively, although these numbers were later reduced through our filtering process (see Filtering of Evaluators below).

We first introduced participants to the concept of the "Yes! and..." game via an introductory video. Participants then evaluated each story on five criteria (creativity, interest, surprise, cohesiveness, and degree of agreement) using a 7-point Likert scale (1=lowest, 7=highest).

Following the evaluations, we collected participants' experiences through a structured debriefing. Participants also self-assessed their English proficiency on a scale of 1-7, with most (84.3%) rating themselves at level 5 or higher, and indicated whether they had previously evaluated "Yes! and..." games. The survey concluded with an opportunity for open-ended feedback. Additional details about survey design, participant demographics, and response statistics are available in the SI.

Rationale and Details of Evaluation Criteria

A co-created "Yes! and..." story, composed of 10-20 sentences, is a multifaceted object that requires careful evaluation. Through open-ended interviews with three improvisation experts who evaluated 15 HH stories, we formulated five evaluation criteria specific to "Yes! and..." stories:

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Creativity Creativity is pivotal in "Yes! and..." games because of its role in pushing narrative boundaries. It is not just about being different; it is about generating narratives that promote divergent thinking and challenge the conventional patterns of storytelling (Sowden et al. 2015).

Interest The ability to engage the audience is fundamental in these games. Although structured narrative principles help maintain engagement (McKee 1997), improvisational storytelling can create compelling narratives through character interaction even without predetermined plots (Swartjes and Vromen 2007).

Surprise In "Yes! and..." games, surprise is a key measure due to its role in maintaining audience engagement through unexpected narrative shifts. This concept aligns with (Ely, Frankel, and Kamenica 2015) findings on the importance of surprise in the entertainment value of information.

Cohesiveness Cohesiveness in "Yes! and..." games helps to maintain a structured and understandable narrative, balancing spontaneous creation with causal connection between the elements of the story (Alon 2010; McKee 1997). Emergent narrative studies show that while interactive freedom can challenge coherence, successful improvisation maintains the essential narrative structure through collaborative effort (Theune et al. 2013).

Degree of Agreement Between Players This criterion assesses how well participants align in their vision for the story's development, following Johnstone's principle that successful improvisation requires accepting and building upon each other's offers (Johnstone 1987). This mutual acceptance is the foundational "Yes" in the "Yes! and..." principle, enabling collaborative story development.

The complete survey definitions and rating instructions for each criterion are provided in SI: Section Evaluation Criteria Details.

Filtering of Evaluators

In our study, we implemented a filtering process to ensure the participation of the evaluator using three criteria:

First, we excluded raters who took over 1000 seconds (approximately 17 minutes) to complete the survey of eight stories, as this duration was substantially longer than the average completion time of 321.35 seconds (±38.65s) and suggested potential disengagement or interruptions during the task. This eliminated 49 of 453 raters.

Second, for each story, we calculated raters' average time and SD for answering questions. Raters who spent less than (mean - 0.85 * SD) were considered insufficiently engaged and excluded from all games in the survey. This method, while potentially excluding some accurate raters, prioritized data quality by ensuring thoughtful participation. Across the eight cohorts, we removed 5-9 raters per cohort for taking too long (totaling 49 raters) and 0-23 raters per cohort for rating too fast (totaling 85 raters).

Finally, we filtered out participants based on their self-reported English fluency. Specifically, those who rated their proficiency as "Low" or "Not Good" (the bottom two levels out of a seven-tier scale) were excluded. This resulted in removing 11-27 participants per cohort (totaling 151 participants).

Following the application of these filters, the remaining cohort sizes were 45, 47, 61, 56, 44, 34, 116 and 133 participants respectively.

Computational Metrics: Novelty, Surprise, and Value

Our evaluation approach combines traditional human evaluation with computational metrics, reflecting recent advances in creativity measurement. Although classic creativity tests such as the Alternative Uses Test relied solely on human assessors to assess creative output (Kudrowitz and Dippo 2013), modern approaches can take advantage of LLMs to provide an automated assessment of creative qualities like originality and quality (Luchini et al. 2023). Following this trajectory, we combine human ratings with automated metrics to provide both subjective evaluation and objective measurement of the characteristics of the story.

We employed three metrics to assess creative aspects of the co-created "Yes! and..." stories: *Novelty*, *Surprise*, and *Value (Engagement)*. These metrics were chosen specifically to capture key aspects of successful "Yes! and..." interactions - the ability to build upon previous contributions while introducing new elements (Novelty), the capacity to take unexpected narrative turns (Surprise), while maintaining coherence (evaluated through cohesiveness), and the sustained engagement of participants (Value). Novelty was computed via established embedding-based approaches (Johnson et al. 2022; Julian Just and Hutter 2024), measuring semantic distances between consecutive contributions, while Surprise was quantified using language model probability distributions (Bunescu and Uduehi 2022). Value, unlike these computational metrics, was derived primarily from participant behavior and feedback.

Novelty (Semantic Variability) In "Yes! and..." games, participants must balance between maintaining narrative coherence and introducing new ideas. To measure this balance quantitatively, we embedded each sentence using pre-trained transformer models (e.g., `roberta-base-nli-stsb-mean-tokens` (Liu et al. 2019)). For each story, we computed the cosine distance between each sentence and the sentence immediately preceding it (excluding the first). Higher mean distances indicate stronger semantic shifts between consecutive contributions, suggesting more novel narrative directions. We analyzed these distances separately for human-authored lines in both conditions (HH and AIH) and AI-authored lines in the AIH condition.

Surprise (Unpredictability) We define Surprise as the negative log probability that a large language model assigns to each new sentence, given its immediate context. Concretely, we use pretrained models such as `GPT-2` or `OPT-1.3b` to estimate:

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

$$\text{Surprise} = -\log\Big(P(\textit{sentence} \mid \textit{previous sentences})\Big)$$

Thus, sentences that are unlikely or dissimilar to the model's learned language patterns yield higher Surprise scores. We selected these specific models due to practical computational constraints and their widespread availability through the HuggingFace library. We computed Surprise separately for human-authored lines and AI-authored lines in the AIH condition, as well as for all lines in HH stories.

Value (Engagement) Unlike Novelty and Surprise, which rely on embedding and probability metrics, Value focuses on participants' *engagement*. We operationalized Value in three ways:

1. **Session Duration and Turns:** The time span of each game (in seconds) and the number of sentences exchanged. Longer or more extensive interactions suggest higher engagement.
2. **Participant Feedback:** Open-ended comments solicited immediately after each session. We looked for indications of immersion, boredom, or heightened excitement.
3. **Reported Interest Ratings:** As part of the external evaluation, naive raters judged how *interesting* each story was, on a scale from 1 (least) to 7 (most). We interpret higher interest ratings as a proxy for story Value.

Ablation Study: Prompt without "Surprise"

To examine how explicitly prompting the AI to be "surprising" affects its output, we conducted an ablation in which the word "surprise" was removed from the prompt. The rest of the instructions were unchanged (see the previous discussion for the baseline prompt details). The participants again participated in AIH games, but this time the AI instructions did not refer to "surprising" content. We collected the resulting 30 stories and recalculated the Surprise metric to see whether removing the explicit instruction to be 'surprising' would alter the creative dynamics.

"Turing Test" for HH and AIH Games

We conducted an evaluation to compare narratives generated by HH pairs and AIH pairs in "Yes! and..." games. This evaluation, framed as a "Turing test", aimed to explore the capability of AI models in computational creativity through the lens of co-created narratives. 150 university students participated in the survey, which included 10 randomly selected stories (5 HH and 5 AIH). We asked participants to identify whether each story was created by HH or AIH pairs, with an optional explanation field (see survey interface in SI Turing Test Survey Interface and Implementation). After filtering for English proficiency, 122 raters remained.

Results

Stories Data

We analyzed 141 HH and 86 AIH games collected through various platforms. After filtering for quality and minimum length requirements (detailed filtering criteria in Methods - Filtering of Evaluators), our final dataset comprised 129 HH and 75 AIH stories. Games averaged 13-18 sentences in length.

Table 1: Comparison of length of interactions between humans and AI agents, in HH and AIH corpora (means and SD).

Metric	AI	H in AIH	H in HH
Sentences per story	6.52 (2.64)	6.52 (2.63)	6.95 (2.32)
Words per sentence	16.66 (7.59)	11.51 (3.58)	11.37 (3.17)
Words per story	109.05 (64.64)	77.37 (41.86)	79.07 (31.77)

Below are partial examples of the interactions in both conditions (full examples of both HH and AIH games are provided in SI: Game Examples):

HH:

Remember the other day we saw flying whales?

Yes of course! One of them offered us a ride to Paris and we got on.

Yes, and then at the end we got off one stop before, on the moon.

AIH:

Do you remember that day at the park?

Yes, and especially how the golden autumn leaves danced in the wind.

Yes! And that wind! It was so powerful!

In the AIH examples, human text is shown in blue and AI text in green.

All HH and AIH interactions used in this study are available in our public corpus repository: `https://tinyurl.com/yes-and-github`.

Statistics of Co-Created Stories

We analyzed basic linguistic metrics (sentences per story, words per sentence and total words) to compare human and AI behaviors in these co-created stories. Table 1 presents the means and standard deviations for each metric across conditions.

Analysis of variance (ANOVA) revealed that while the number of sentences per story was similar across conditions, AI agents produced significantly more words per sentence than humans (mean difference $\approx$ 5.2 words, F(2, 407) = 43.70, $p < 0.001$) and consequently more words per story (mean difference $\approx$ 30 words, F(2, 407) = 16.29, $p < 0.001$). This verbose AI behavior did not influence human partners' response lengths, as humans maintained consistent word counts whether paired with AI or other humans (mean difference = 0.14 words, p = 0.952). Detailed statistical analyses, including post-hoc comparisons, are provided in SI: Detailed Statistical Analysis of Co-Created Stories.

Comparing Evaluations of HH and AIH Stories by External Raters

The collected "Yes! and..." stories were evaluated by 453 unique raters, with a total of 817 ratings (see details in SI: Survey Participation Statistics). In each rating session, a naive rater (who did not play the game before) read and

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

rated 8 stories. Each story was evaluated on five criteria: creativity, interest, surprise, cohesiveness, and the degree of agreement between players, using a 7-point Likert scale (see details of these criteria in Methods).

To assess inter-rater reliability, we used the ICC3K variant of the intraclass correlation coefficient, which measures consistency among raters using a 7-point scale (Koo and Li 2016). Overall agreement was good (mean ICC = 0.761, SD = 0.154), with particularly strong reliability for creativity (ICC = 0.809) and surprise (ICC = 0.763) ratings. However, agreement (ICC = 0.215) and story cohesiveness (ICC = 0.465) showed poor reliability, while interest ratings were moderate (ICC = 0.544). Details in SI: Detailed Intraclass Correlation Coefficient (ICC) Analysis.

Evaluation Results of HH and AIH Stories

Table 2 presents external raters' evaluations across all criteria on a 7-point Likert scale. Statistical analysis revealed remarkably similar ratings between HH and AIH stories, with only the "interesting" criterion showing a marginal difference favoring AI-human pairs (3.72 vs 3.58, $p = 0.011$). Notably, both conditions received above-median ratings (> 3.5) for all criteria. Details in SI: Detailed Statistical Analysis of Story Evaluations.

Table 2: Mean and standard deviation (STD) of ratings for each criterion in HH and AIH "Yes! and..." games.

Criterion	Mean HH	STD HH	Mean AI	STD AI
Creative	4.26	1.88	4.34	1.78
Agreement	4.57	1.87	4.55	1.82
Story	4.26	1.83	4.25	1.77
Interesting	3.58	1.86	3.72	1.81
Surprising	3.83	1.96	3.87	1.89

A key finding of our analysis is the striking similarity between HH and AIH rating distributions, visualized using Jensen-Shannon Divergence in Figure 2. This information-theoretic measure (Lin 1991) quantifies the similarity between probability distributions, with darker shades indicating lower similarity. The analysis shows strong similarities across all evaluation criteria, supporting our statistical findings.

The similarity between HH and AIH stories is further supported by detailed distribution analysis in Figure 1 - allowing direct comparison between HH and AIH games. Analysis using only non-repeated raters (as described in Methods) yielded similar results (see SI: Analysis of Non-Repeated Raters).

Novelty, Surprise, and Value Analysis

Novelty Results To assess semantic variability in story progression, we compared consecutive-sentence distances across three groups: humans in human-human games (H_HH), humans in AI-human games (H_AIH), and AI responses in AI-human games (AI_AIH). Higher distances indicate greater semantic novelty between successive sentences. Analysis using multiple embedding models showed consistent patterns (Table 3):

Table 3: Mean consecutive-sentence distances by group and embedding model (with SD).

Model	H_HH	H_AIH	AI_AIH
RoBERTa	0.758 (0.182)	0.783 (0.176)	0.616 (0.184)
DistilBERT	0.763 (0.192)	0.771 (0.173)	0.624 (0.174)
MPNet	0.644 (0.167)	0.657 (0.145)	0.566 (0.152)
MiniLM	0.634 (0.182)	0.670 (0.163)	0.584 (0.153)

Statistical analysis revealed significant differences between AI-authored and human-authored contributions ($p < 0.001$ across all models) - with humans being more "novel". Notably, humans collaborating with AI (H_AIH) showed slightly higher semantic distances than those in purely human interactions (H_HH), though this difference was only significant for two of the four models ($p < 0.05$ for RoBERTa and MiniLM).

As visualized in Figure 3, humans consistently demonstrated greater semantic leaps between contributions compared to AI across all embedding models.

Surprise Results The analysis of sentence unpredictability using multiple language models revealed systematic differences between contributions. Higher scores indicate more unexpected sentences given the previous context.

Figure 4 illustrates these surprise patterns across participant groups and language models.

Table 4 summarizes these findings:

Table 4: Mean surprise scores by group and language model (with standard deviations).

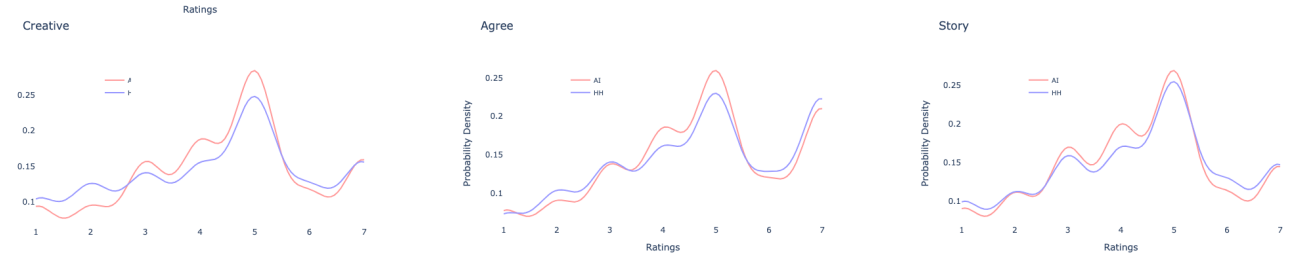
Model	H_HH	H_AIH	AI_AIH
GPT-2	47.75 (24.92)	59.23 (27.19)	74.43 (29.80)
BLOOM-1B1	46.29 (23.69)	55.95 (25.02)	70.53 (27.85)
OPT-1.3B	44.15 (22.98)	52.87 (23.65)	64.62 (24.55)

All models showed a consistent pattern: AI-authored sentences (AI_AIH) exhibited significantly higher surprise scores than human-authored sentences in either condition ($p < 0.001$). Additionally, humans paired with AI (H_AIH) produced more surprising responses than those in human-human pairs (H_HH), suggesting that AI collaboration may encourage more unexpected narrative turns.

While both novelty and surprise measure aspects of unexpectedness, they capture different phenomena. Novelty measures semantic distance between consecutive contributions regardless of context (Figure 3), while surprise measures contextual unexpectedness given the specific preceding narrative (Figure 4). The fact that AI contributions show lower novelty (smaller semantic leaps) but higher surprise (more unexpected in context) suggests these metrics indeed capture distinct aspects of creative behavior.

Ablation Study Results To investigate the impact of explicitly prompting for surprise, we conducted an ablation study where surprise-related instructions were removed from the AI's prompt. Table 5 compares the mean surprise

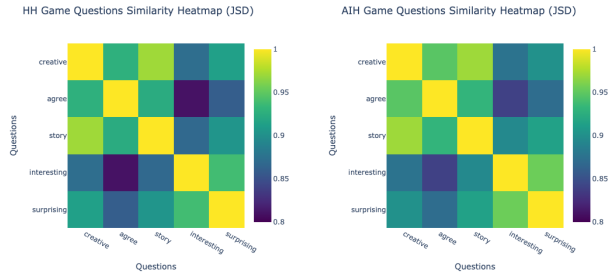
Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(A) Creative (B) Agreement (C) Story

(D) Interesting (E) Surprising

Figure 1: Complete rating distributions across all evaluation criteria. Each subplot shows the probability density distributions for both HH (blue) and AIH (red) games. The distributions demonstrate remarkable similarity between human-human and AI-human interactions.



(a) HH Games (b) AIH Games

Figure 2: Rating distribution similarity heatmaps. Darker shades show lower similarity.

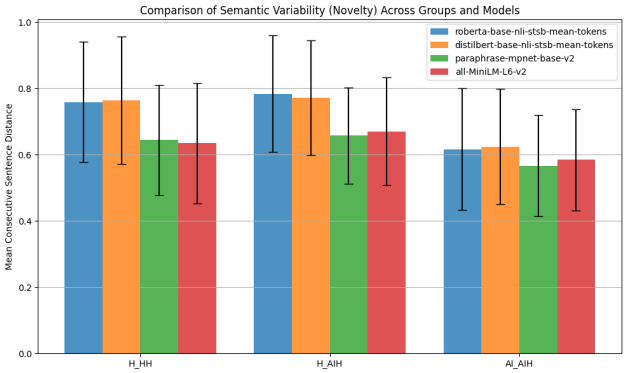


Figure 3: Novelty comparison: Semantic distances between consecutive sentences across different embedding models. Higher values indicate greater semantic novelty between successive contributions.

scores between the original and ablation conditions across both human and AI outputs.

Table 5: Mean surprise scores for original and ablation conditions (with standard deviations).

Condition	Human	AI
Original (AIH)	59.10 (27.36)	74.75 (30.18)
Ablation (ABL)	58.64 (20.04)	67.55 (12.35)

As shown in Figure 5, removing 'surprise' from the AI's instructions significantly affected AI outputs ($p = 0.002$, Cohen's $d = 0.312$), with mean surprise scores dropping from 74.75 to 67.55. However, human responses remained largely unchanged ($p = 0.703$, Cohen's $d = 0.019$), suggesting that explicit surprise prompting primarily influences AI behavior rather than human creativity. The gap between human and AI surprise scores remained substantial in both the original condition ($p < 0.001$, Cohen's $d = -0.543$) and the ablation condition ($p < 0.001$, Cohen's $d = -0.535$), indicating that AI responses maintain higher unpredictability regardless of explicit prompting - a notable pattern differing from other creative domains (Hoffman and Weinberg 2010).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

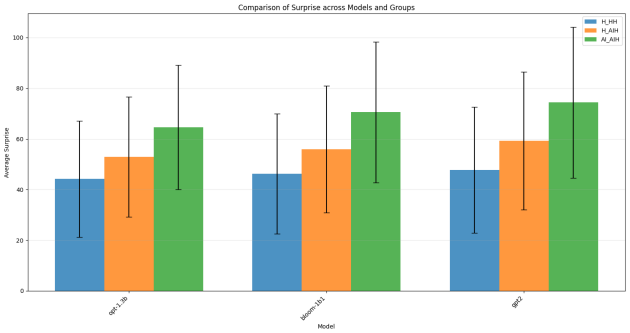


Figure 4: Surprise comparison: Scores across different language models. Higher values indicate more unexpected sentence constructions given the context.

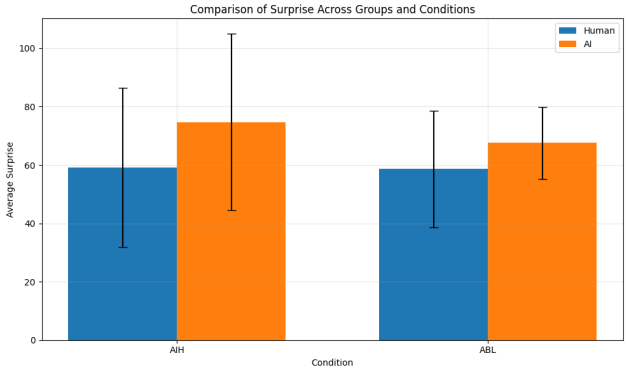


Figure 5: Ablation study results comparing surprise scores between original and modified conditions, demonstrating the effect of removing surprise-related prompting from AI instructions.

Value (Engagement) Analysis Analysis of value metrics through engagement revealed sustained participation across both HH and AIH conditions. Game durations remained comparable (HH: 508.63 ± 169.25 seconds; AIH: 405.73 ± 234.98 seconds), and sentence lengths were consistent between human players across formats (HH: 11.37 words; AIH: 11.51 words). External raters judged AIH stories as marginally more interesting than HH stories (3.72 vs. 3.58, $p = 0.011$), though this difference was modest compared to other evaluation criteria.

A simple 'Turing Test' for "Yes! and..." Stories

To assess whether raters could distinguish between HH and AIH "Yes! and..." stories, we conducted a survey involving 122 raters (see more details in Methods). Each rater was presented with 5 HH games and 5 AIH games (chosen randomly) and was asked to identify whether each story was created by a pair of humans or a human and AI. Overall accuracy was poor, with 47.05% correct identification for HH stories and 46.56% for AIH stories. A Chi-square test confirmed no significant difference in raters' ability to identify story sources ($\chi^2 = 0.0132$, p = 0.9086).

Table 6: Raters' guesses and Chi-square test results for each "Yes! and..." story, showing the number of correct and incorrect guesses, Chi-square test statistic, p-value, and whether the difference in guesses was significant.

Story	Correct	Incorrect	Chi-sq	P-val	Sig.
HH 1	40.16%	59.84%	8.67	0.003	Yes
HH 2	46.72%	53.28%	0.80	0.370	No
HH 3	55.74%	44.26%	2.77	0.096	No
HH 4	47.54%	52.46%	0.41	0.522	No
HH 5	45.08%	54.92%	1.98	0.159	No
AIH 1	36.07%	63.93%	17.85	< 0.001	Yes
AIH 2	47.54%	52.46%	0.41	0.522	No
AIH 3	48.36%	51.64%	0.15	0.701	No
AIH 4	51.64%	48.36%	0.15	0.701	No
AIH 5	49.18%	50.82%	0.02	0.898	No

Story-level analysis (Table 6) revealed that for both HH and AIH stories, in 4/5 cases there were no significant differences between correct and incorrect responses, and in 1/5 stories there were significantly more incorrect responses (more details in SI: Detailed Turing Test Results). Taken together, these results suggest that naive readers cannot differentiate between stories created in the HH and AIH conditions.

To gain insights into the factors that influenced raters' decisions when distinguishing between HH and AIH "Yes! and..." games, we collected testimonials from participants explaining their reasoning. These testimonials provide qualitative data that complement the quantitative results presented earlier.

The following are representative examples of the testimonials, illustrating the various factors that raters considered when making their judgments:

Explanations from People Who Mistakenly Thought a Human-Made Story Was Made by a Human and AI

"The writing and context here are perfect. Definitely written with AI."

"It seems to me that this story was written between a person and an AI [...] The story includes exaggerated fictional events that sound improbable in reality [...] The conversation sounds forced and lacks logical connection [...] There is not much emotion or human reactions in the conversation."

Explanations from People Who Mistakenly Believed a Story Made by a Human and AI Was Actually Made by Just Humans

"AI would not write the following sentence ever: 'Yes and I never saw something like that in my life'."

"There is a lot of text and detail, too much science fiction."

Discussion

Analysis of Results Our computational analysis revealed distinctive patterns in AI versus human contributions to the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

"Yes! and..." game. AI responses showed higher surprise scores across all evaluation models (mean difference $\approx$ 15-20 points (see in Methods), $p < 0.001$) but lower semantic novelty compared to humans (mean difference ≈ 0.15 points, $p < 0.001$ across all embedding models). AI responses were consistently longer (mean difference 5.2 words per sentence, $p < 0.001$) but didn't influence human partner response lengths, suggesting humans maintained their natural interaction style regardless of partner type. External raters judged AIH stories as comparable to HH stories across all evaluation criteria except interest, where AIH stories scored marginally higher (3.72 vs 3.58, $p = 0.011$). Notably, in our "Turing test" experiment, participants were unable to reliably distinguish between HH and AIH stories (47.05% vs 46.56% correct identification), indicating that AI contributions were naturalistic enough to be indistinguishable from human responses.

Comparison with Literature These findings extend previous work on LLMs in creative writing (Franceschelli and Musolesi 2024) and storytelling (Yuan et al. 2022) by demonstrating their capacity for real-time improvisational co-creation. The "Yes! and..." principle, long recognized as a basic tenet of improvisation (Johnstone 1987) and collaborative creativity (Himberg et al. 2018), provides a structured framework for creative exchange. Although the principle has traditionally guided human-human interaction, our results demonstrate that LLMs can effectively embody this approach, suggesting its potential as a broader framework for human-AI creative collaboration. Our controlled text-based environment builds on methodologies from joint action research (Noy, Dekel, and Alon 2011) and collaborative exploration studies (Rafner et al. 2023), providing a framework for systematic analysis of human-AI creative interaction.

Implications for AI System Design Our results suggest three key design principles for creative AI systems. First, LLMs' generative capabilities perform effectively in structured improvisational contexts, where creative invention is valued over factual precision. Second, interaction parameters, including timing, response length, tone, and willingness to explore human-initiated themes, need careful calibration to match natural human interaction patterns. Third, creative AI systems benefit from balancing narrative consistency with novel contributions. These insights point to AI systems designed not just for content generation, but for authentic creative partnerships.

Limitations Our study has four main limitations. First, the sample size (129 HH and 75 AIH stories) limits result generalizability. Second, our computational metrics relied on older models (GPT-2 and OPT-1.3B) and may miss important nuances of creative improvisation such as humor, emotional resonance, and advanced narrative techniques that expert evaluation might capture. Third, our participant pool consisted of non-expert improvisers, potentially affecting story creation and evaluation quality. Fourth, the text-only format eliminates important dimensions of face-to-face improvisation.

Conclusion

Key Findings Our study explores the potential of LLMs to participate in improvisational co-creation through the "Yes! and..." game. We found that stories co-created with AI partners received similar ratings to human-human collaborations in external evaluations, and participants in our "Turing test" experiment had difficulty distinguishing between the two conditions. These results suggest that LLMs can meaningfully engage in structured creative improvisation, contributing to coherent narrative development within the specific context of the "Yes! and..." game.

Methodological Contributions Our approach offers three key advantages for studying human-AI creative interaction:

- **Reduced Complexity:** The text-based format allows controlled model system for studying of core creative dynamics, similar to previous work in the domains of dyadic movement improvisation (Noy, Dekel, and Alon 2011) and visual exploration (Hart et al. 2017).
- **Quantifiable Assessment:** The experimental setup allows measurement of creative contributions through computational metrics.
- **Scalability:** Our chat-based platform enables systematic large-scale data collection and analysis.

Future Directions Our work points to several promising directions: comparing expert improvisers with AI performance, developing systems for dynamic AI contribution control in co-creative processes, implementing real-time evaluation systems for adaptive responses, and creating platforms for large-scale unbiased interaction studies. Beyond Turing test approaches, future work should explore AI's unique creative contributions rather than just human-like performance.

Broader Impact This work demonstrates the value of structured improvisational frameworks for human-AI creative collaboration. Our empirical evaluation shows that LLMs can participate in improvisational co-creation at levels comparable to human collaborators. By integrating improvisational techniques with computational creativity assessment, we establish a systematic framework for developing collaborative AI systems.

References

Abid, A.; Abdalla, A.; Abid, A.; Khan, D.; Alfozan, A.; and Zou, J. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild.

Agrawal, A.; Mackey, L. W.; and Kalai, A. 2023. Do language models know when they're hallucinating references? *ArXiv* abs/2305.18248.

Alon, U. 2010. Living within a story: Playback theatre and the art of improvisation. *Haaretz*.

Bates, J.; Loyall, A. B.; and Reilly, W. S. 1994. An architecture for action, emotion, and social behavior. In *Artificial Social Systems: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'92 S. Martino al Cimino, Italy, July 29–31, 1992 Selected Papers 4*, 55–68. Springer.

Berliner, P. F. 2009. *Thinking in jazz: The infinite art of improvisation*. University of Chicago Press.

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

Branch, B.; Mirowski, P.; Mathewson, K.; Ppali, S.; and Covaci, A. 2024. Designing and evaluating dialogue llms for co-creative improvised theatre.

Branch, B.; Mirowski, P.; and Mathewson, K. W. 2021. Collaborative storytelling with human actors and ai narrators.

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 *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Bunescu, R. C., and Uduehi, O. O. 2022. Distribution-based measures of surprise for creative language: Experiments with humor and metaphor. In Ghosh, D.; Beigman Klebanov, B.; Muresan, S.; Feldman, A.; Poria, S.; and Chakrabarty, T., eds., *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, 68–78. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Chakraborty, T., and Masud, S. 2024. The promethean dilemma of ai at the intersection of hallucination and creativity. *Commun. ACM* 67(10):26–28.

Cho, H., and May, J. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2398–2413. Online: Association for Computational Linguistics.

Concepcion, E.; Gervas, P.; and Mendez, G. 2019. Evolving the ines story generation system: From single to multiple plot lines. In *ICCC*, 220–227.

Ely, J.; Frankel, A.; and Kamenica, E. 2015. Suspense and surprise. *Journal of Political Economy* 123(1):215–260.

Franceschelli, G., and Musolesi, M. 2024. Creativity and machine learning: A survey.

Gero, K. I.; Liu, V.; and Chilton, L. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 1002–1019.

Grace, K.; Maher, M.; Fisher, D.; and Brady, K. A. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3:125 – 147.

Halperin, B. A., and Lukin, S. M. 2024. Artificial dreams: Surreal visual storytelling as inquiry into ai 'hallucination'. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, 619–637. New York, NY, USA: Association for Computing Machinery.

Hart, Y.; Mayo, A. E.; Mayo, R.; Rozenkrantz, L.; Tendler, A.; Alon, U.; and Noy, L. 2017. Creative foraging: An experimental paradigm for studying exploration and discovery. *PLOS ONE* 12(8):1–15.

Hayes-Roth, B.; Sincoff, E.; Brownston, L.; Huard, R.; and Lent, B. 1994. Directed improvisation. *Knowledge Systems Laboratory Report KSL-94-61*.

Himberg, T.; Laroche, J.; Bigé, R.; Buchkowski, M.; and Bachrach, A. 2018. Coordinated interpersonal behaviour in collective dance improvisation: The aesthetics of kinaesthetic togetherness. *Behavioral Sciences* 8.

Hoffman, G., and Weinberg, G. 2010. Shimon: an interactive improvisational robotic marimba player. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, 3097–3102. New York, NY, USA: Association for Computing Machinery.

Jacob, M.; Chawla, P.; Douglas, L.; He, Z.; Lee, J.; Sawant, T.; and Magerko, B. 2019. Affordance-based generation of pretend object interaction variants for human-computer improvisational theater. In *ICCC*, 140–147.

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).

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.

Johnson, D.; Kaufman, J.; Baker, B.; Patterson, J.; Barbot, B.; Green, A.; van Hell, J.; Kennedy, E.; Sullivan, G.; Taylor, C.; Ward, T.; and Beaty, R. 2022. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior research methods* 55.

Johnstone, K. 1987. *Impro: Improvisation and the Theatre*. London and New York: Routledge, 1st edition.

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

Julian Just, Thomas Ströhle, J. F., and Hutter, K. 2024. Ai-based novelty detection in crowdsourced idea spaces. *Innovation* 26(3):359–386.

ISBN: 978-989-54160-7-3

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. *arXiv preprint arXiv:1807.09886*.

Koo, T. K., and Li, M. Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15(2):155–163.

Kudrowitz, B., and Dippo, C. 2013. Getting to the novel ideas: Exploring the alternative uses test of divergent thinking. volume 5.

Lin, J. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* 37(1):145–151.

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.

Liu, Y.-H.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; Wu, Z.; Zhu, D.; Li, X.; Qiang, N.; Shen, D.; Liu, T.; and Ge, B. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *ArXiv* abs/2304.01852.

Luchini, S.; Maliakkal, N. T.; DiStefano, P. V.; Patterson, J. D.; Beaty, R.; and Reiter-Palmon, R. 2023. Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings.

McKee, R. 1997. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. New York: ReganBooks.

Nachmanovitch, S. 1991. *Free Play: Improvisation in Life and Art*. G.P. Putnam's Sons.

Noy, L.; Dekel, E.; and Alon, U. 2011. The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *Proceedings of the National Academy of Sciences* 108(52):20947–20952.

Noy, L. 2014. *The Mirror Game: A Natural Science Study of Togetherness*. Bloomsbury Methuen Drama. 318–327.

Rafner, J.; Wang, Q. J.; Gadjacz, M.; Badts, T.; Baker, B.; Bergenholtz, C.; Biskjaer, M. M.; Bui, T.; Carugati, A.; de Cibeins, M.; Noy, L.; Rahimi, S.; Tylén, K.; Zana, B.; Beaty, R. E.; and Sherson, J. 2023. Towards game-based assessment of creative thinking. *Creativity Research Journal* 35(4):763–782.

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.

Ritchie, G. 2001. Assessing creativity. In *Proc. of AISB'01 Symposium*.

Shanahan, M. 2023. Talking about large language models.

Sowden, P. T.; Clements, L.; Redlich, C.; and Lewis, C. 2015. Improvisation facilitates divergent thinking and creativity: Realizing a benefit of primary school arts education. *Psychology of Aesthetics, Creativity, and the Arts* 9:128–138.

Swartjes, I., and Vromen, J. 2007. Emergent story generation: Lessons from improvisational theater. In *AAAI Fall Symposium: Intelligent Narrative Technologies*.

Theune, M.; Alofs, T.; Linssen, J.; and Swartjes, I. 2013. Having one's cake and eating it too: Coherence of children's emergent narratives. In *Workshop on Computational Models of Narrative*.

Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; Li, Y.; Lee, H.; Zheng, H. S.; Ghafouri, A.; Menegali, M.; Huang, Y.; Krikun, M.; Lepikhin, D.; Qin, J.; Chen, D.; Xu, Y.; Chen, Z.; Roberts, A.; Bosma, M.; Zhao, V.; Zhou, Y.; Chang, C.-C.; Krivokon, I.; Rusch, W.; Pickett, M.; Srinivasan, P.; Man, L.; Meier-Hellstern, K.; Morris, M. R.; Doshi, T.; Santos, R. D.; Duke, T.; Soraker, J.; Zevenbergen, B.; Prabhakaran, V.; Diaz, M.; Hutchinson, B.; Olson, K.; Molina, A.; Hoffman-John, E.; Lee, J.; Aroyo, L.; Rajakumar, R.; Butryna, A.; Lamm, M.; Kuzmina, V.; Fenton, J.; Cohen, A.; Bernstein, R.; Kurzweil, R.; Aguera-Arcas, B.; Cui, C.; Croak, M.; Chi, E.; and Le, Q. 2022. Lamda: Language models for dialog applications.

Veale, T.; Wicke, P.; and Mildner, T. 2019. Duets ex machina: On the performative aspects of" double acts" in computational creativity. In *ICCC*, 57–64.

Yuan, A.; Coenen, A.; Reif, E.; and Ippolito, D. 2022. Wordcraft: Story writing with large language models. *27th International Conference on Intelligent User Interfaces*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.; and rong Wen, J. 2023. A survey of large language models. *ArXiv* abs/2303.18223.

Supplementary Information

Supplementary Information, including additional protocols, extended figures, and analyses, is available via a link at `https://tinyurl.com/yesand-si`.

Author Contributions

All authors participated equally in conducting the research and writing the manuscript.

Turning Linear Stories into Thrillers: The Impact of Story Reordering and Familiarity

Mei Si
Rensselaer Polytechnic Institute
sim@rpi.edu

Abstract

This study explores how modifying narrative structure affects suspense perception. Drawing on suspense theory, we systematically reorder key events—specifically the Initiating Event (IE) and Outcome Event (OE)—and introduce non-diegetic symbolic delays to heighten anticipatory tension. We also examine how prior familiarity with a story influences suspense. An online experiment on Amazon Mechanical Turk (MTurk) tested linear versus suspense-modified versions of four short narratives—two familiar and two original. Statistical analyses (ANOVA, ANCOVA, regression) show that reordering events and adding delays significantly boost reported suspense and engagement, with prior familiarity moderating these effects.

Introduction

Suspense is central to engaging storytelling, shaping reader attention, emotional involvement, and narrative absorption. It creates anticipation and uncertainty about future events, sustaining audience interest throughout the narrative (Lehne et al. 2015; Green and Appel 2024; Vorderer 2013).

Filmmakers like Hitchcock have famously used external cues to evoke suspense (Hare 2007), and prior studies (e.g., (Bálint et al. 2017)) show that teasing outcomes and inserting narrative delays—diegetic or non-diegetic—can heighten tension. Tan's framework (Tan 1996) describes films as "emotion machines," engaging audiences through structured use of IE and OE.

Yet few studies have turned these insights into computational methods for short-form storytelling. This study proposes a lightweight, rule-based approach that transforms linear narratives by reordering events and inserting non-diegetic delays. Our experiment shows that both narrative structure and prior familiarity significantly affect suspense ratings. ANCOVA reveals clarity and engagement as key covariates, while regression suggests clarity may moderate the effects of suspense manipulations.

These results support theoretical models and show that computational restructuring can enhance emotional engagement through suspenseful narrative design.

Narrative Suspense: Theory, Computation, and Film

Suspense in narratives arises from the interplay of uncertainty and anticipation, shaped by the structuring and timing of key events. Initiating Events (IE) introduce tension, while Outcome Events (OE) resolve it, with suspense depending on how these are ordered and disclosed (Tan 1996; Bálint et al. 2017). In film, directors like Hitchcock exploit narrative structure and external cues—editing, pacing, non-diegetic sound—to evoke emotional responses by revealing critical information before characters become aware (Hare 2007). Studies such as (Bálint et al. 2017) show that teasing outcomes and adding delays can intensify suspense, enhancing anticipation and engagement.

Theoretical models emphasize both structural and emotional dimensions. (Comisky and Bryant 1982) identify suspense as a product of narrative form and emotional investment. (Brewer and Lichtenstein 1982) propose a structural-affect theory where suspense arises from uncertainty and delay. (Tan 1996) conceptualizes film as an "emotion machine," using structured cues to manipulate affect. (Lehne and Koelsch 2015) offers a cognitive-affective model, while (Smuts 2008) introduces the desire-frustration theory, in which suspense emerges from delayed expectations. (Bentz, Kuchinke, and Jacobs 2024) provides an experimental framework linking suspense with emotional engagement.

On the computational side, (Doust and Piwek 2017) present a domain-independent model that tracks suspense during story generation. (Cheong and Young 2015) introduce Suspenser, a generator that dynamically manipulates narrative tension. (O'Neill and Riedl 2014) propose Dramatis, a model based on cognitive heuristics that simulates goals, plans, and obstacles to predict suspense. Recent work by (Xie and Riedl 2024) explores suspenseful story generation using large language models (LLMs) through iterative planning.

While these systems show promise, few target short-form narratives. This study addresses that gap with a lightweight, rule-based method that restructures linear texts by reordering events and adding non-diegetic delays. We also investigate how prior story familiarity shapes suspense perception, engaging with the paradox of suspense (Carroll 2013).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Narrative Transformation for Suspense

We developed a lightweight, rule-based system to enhance suspense in linear narratives by restructuring key events and inserting symbolic delays. The method aligns with theories of suspense (Bálint et al. 2017; Tan 1996), emphasizing the role of event reordering and delay in modulating anticipation. Our system introduces both diegetic delays—actions or descriptions occurring naturally within the story—and non-diegetic delays using metaphorical imagery to heighten tension.

Transformation Process

A conventional linear narrative typically follows a straightforward sequence:

$$\text{Initiating Event (IE)} \rightarrow \text{Outcome Event (OE)}$$

To enhance suspense, our computational method modifies the linear narrative structure by explicitly introducing an outcome teaser and inserting a diegetic delay. Inspired by the theoretical framework outlined by (Bálint et al. 2017), this approach systematically restructures narratives to amplify anticipatory tension. Specifically, a diegetic delay is defined as a narrative pause created by describing actions, events, or details occurring naturally within the story's fictional world, delaying the resolution of key events and thus heightening suspense. This contrasts with non-diegetic delays, which utilize symbolic or metaphorical imagery external to the narrative's immediate action. The resulting suspense-driven narrative structure can be represented as follows:

$$\begin{aligned}&\text{Outcome Teaser (Partial OE)}\\ &\rightarrow \text{Initiating Event (IE)}\\ &\rightarrow \text{Diegetic Delay}\\ &\rightarrow \text{Non-Diegetic Delay (Optional)}\\ &\rightarrow \text{Full Outcome Event (OE)}\end{aligned}$$

Narrative Techniques for Suspense Enhancement

To further amplify suspense, we implemented narrative strategies inspired by prior work (Bálint et al. 2017; Tan 1996):

- **Non-Diegetic Delay (Symbolic Imagery):** Metaphors such as storm clouds or trapped animals intensify anticipation (Bálint et al. 2017).
- **Emotional Amplification:** Vivid portrayal of sensations (e.g., racing heart, echoing footsteps) increases immersion (Tan 1996).
- **Perspective Shifting:** Brief shifts to an external observer's view enhance narrative uncertainty (Bálint et al. 2017).
- **Gradual Disclosure of Information:** Critical details are withheld and revealed slowly, sustaining engagement (Bálint et al. 2017).

We applied these transformations using GPT-4o (OpenAI 2024), via iterative prompting that guided the model to reorder events and incorporate suspense-enhancing techniques. A sample final prompt is shown below:

> *"Rewrite the following linear story into a suspenseful narrative by: (1) explicitly teasing the outcome first (Outcome Teaser); (2) returning to describe the initiating event afterward; (3) inserting a diegetic delay using vivid symbolic imagery (e.g., gathering storm clouds, trapped birds), emotionally intense descriptions of the character's sensations (e.g., racing heartbeat, quickened breath), shifting briefly to an external observer's viewpoint to enhance uncertainty, and gradually revealing critical narrative details. Maintain ambiguity and sensory richness throughout to heighten anticipation. Here is the original linear narrative: [Original Narrative Text Here]."*

Example of Suspense Enhancement

To illustrate our method for transforming linear narratives into suspenseful stories, we apply it to the classic tale "The Tortoise and the Hare," presenting both the original and the suspense-enhanced versions with applied techniques.

Original Narrative

One day, the hare mocked the tortoise for his slowness. Annoyed, the tortoise challenged the hare to a race. Confident in his speed, the hare raced ahead and, certain of his victory, stopped midway for a nap. Meanwhile, the tortoise continued slowly but steadily. Eventually, the hare woke up and rushed toward the finish line, but he was too late. The tortoise had already won, proving that "slow and steady wins the race."

Suspense-Enhanced Narrative

Outcome Teaser (Partial OE): Breathless silence hung over the crowd as the finish line came into view. To everyone's astonishment, it wasn't the speedy hare in front, but the slow, steady tortoise, inching painfully close to an impossible victory. How could this have happened?

Diegetic Delay: Minutes earlier, the hare lounged beneath a shady tree miles ahead, yawning in boredom. Certain of his triumph, he had stopped, closed his eyes, and drifted into a careless sleep. Meanwhile, unnoticed, the tortoise moved relentlessly forward, step by patient step, gradually closing the distance.

Resolution (Full OE): Now, as the tortoise's foot crossed the finish line, the crowd erupted in disbelief. The hare awoke with a startled gasp and desperately leaped forward—but it was too late. The tortoise turned slowly, a gentle smile on his face, reaffirming the timeless lesson that quiet perseverance triumphs over reckless confidence.

Techniques Applied

To create a suspenseful version of the original story, we applied the following narrative transformations, grounded in established suspense theories (Bálint et al. 2017; Tan 1996):

- **Event Reordering (OE → IE):** The outcome is teased first to disrupt chronology and capture attention.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

- **Diegetic Delay:** Descriptions of the hare's overconfidence and the tortoise's steady pace create a natural pause that builds tension.
- **Symbolic Imagery (Non-Diegetic Delay):** Metaphors like breathless silence and shocked spectators reinforce suspense.

These techniques heighten emotional engagement by shaping reader expectations and anticipation.

Evaluation

To evaluate the effectiveness of our narrative transformation technique, we conducted an online experiment to test whether narrative restructuring and symbolic delays increase suspense. We also examined whether prior familiarity with a story moderates these effects.

Participants and Design

Participants (N = 17 per condition) were recruited through Amazon Mechanical Turk (MTurk) and randomly assigned to one of two groups: a linear (control) or suspense-enhanced (experimental) condition. Each participant read four narratives: two well-known stories (*The Tortoise and the Hare*, *Little Red Riding Hood*) and two original ones ("A Person Takes an Elevator Late at Night" and "Abandoned Car"). This design allowed us to explore the role of prior familiarity in shaping suspense and engagement.

Procedure

Participants read all four stories in their assigned condition. Those in the suspense group received versions modified using our transformation method (event reordering and diegetic delays), while the control group read the original linear versions. After each story, participants rated the following:

- **Perceived suspense** (1–7 Likert scale)
- **Engagement and absorption** (1–7 Likert scale)
- **Narrative coherence** (1–7 Likert scale)
- **Familiarity with similar scenarios** (1–7 Likert scale)

These measures allowed us to assess how narrative structure influenced suspense, coherence, and reader engagement, and whether familiarity shaped those effects.

Results

To evaluate how our narrative transformation technique influenced suspense perception, we conducted a series of statistical analyses, including ANOVA, ANCOVA, and multiple regression. The main independent variables were **design** (linear vs. suspense) and **known_story** (familiar vs. unfamiliar), and the dependent variables were **Suspense**, **Clarity**, **Familiarity**, and **Engagement**.

ANOVA for Individual Dependent Variables

We first conducted separate ANOVAs for each dependent variable to examine the main and interaction effects of *design* and *known_story*.

For **Suspense**, both factors had significant effects. Participants rated suspense-modified narratives as significantly more suspenseful than linear ones, with a main effect of design ($F(1, 132) = 5.05, p = 0.026$). Familiarity with the story also increased suspense ratings ($F(1, 132) = 7.36, p = 0.008$). The interaction between design and familiarity was not significant ($F(1, 132) = 0.73, p = 0.395$).

For **Clarity**, neither factor showed a significant effect. The design effect was not significant ($F(1, 132) = 0.73, p = 0.394$), nor was the effect of familiarity ($F(1, 132) = 2.46, p = 0.119$).

For **Familiarity** ratings (as a dependent variable), prior knowledge of the story had a significant effect ($F(1, 132) = 8.64, p = 0.004$), while narrative design had no effect ($F(1, 132) \approx 0, p = 1.000$).

For **Engagement**, no significant effects were found for either design ($F(1, 132) = 0.92, p = 0.340$) or familiarity ($F(1, 132) = 1.51, p = 0.221$).

Effect of Narrative Design on Suspense (ANCOVA)

To control for potential confounds, we conducted an ANCOVA with **Suspense** as the dependent variable and **Clarity**, **Familiarity**, and **Engagement** as covariates.

The suspense-enhanced condition showed a significant main effect ($F(1, 129) = 5.81, p = 0.017$), as did prior familiarity with the story ($F(1, 129) = 5.74, p = 0.018$). The interaction between design and familiarity was not significant ($F(1, 129) = 0.79, p = 0.377$).

Among the covariates, both **Clarity** ($F(1, 129) = 8.98, p = 0.003$) and **Engagement** ($F(1, 129) = 20.29, p < 0.001$) significantly predicted suspense ratings. **Familiarity** approached significance but was not a significant predictor on its own ($F(1, 129) = 2.88, p = 0.092$).

These results suggest that the suspense-enhanced narrative design effectively increased perceived suspense, particularly when the story was already familiar to the reader. Clarity and engagement were also positively associated with suspense ratings.

Figure 1 presents the mean suspense ratings by narrative design (normal vs. suspense) and story familiarity (known vs. unknown). Overall, suspense-enhanced narratives resulted in higher suspense ratings compared to the linear versions across both known and unknown stories. However, the effect of the suspense design was more pronounced for unfamiliar stories, where the suspense version showed a greater increase in mean suspense ratings compared to the normal design. These results align with the ANCOVA findings, which indicated that design (linear vs. suspense) significantly impacted suspense perception and that prior familiarity with the story also played a role, though the interaction between the two factors was not significant.

OLS Regression Analysis

To further examine the effects of *design* and *known_story* on **Suspense**, we conducted an ordinary least squares (OLS)

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

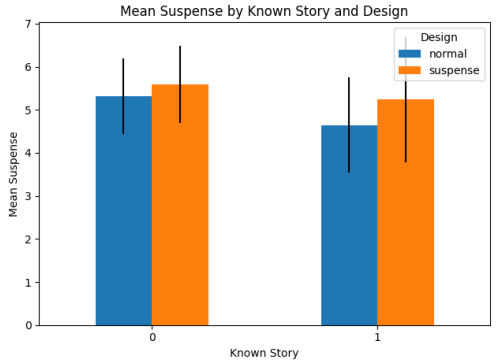


Figure 1: Effect of Narrative Design and Story Familiarity on Suspense Ratings

regression while controlling for **Clarity**, **Engagement**, and **Familiarity**. The model explained approximately 41.8% of the variance in suspense ratings, with an adjusted $R^2 = 0.371$, indicating a moderate and robust fit. The overall regression model was statistically significant, $F(10, 125) = 8.97, p < 0.001$, suggesting that at least one predictor reliably contributed to suspense perception.

The narrative design variable (suspense vs. linear) was marginally significant, $p = 0.061$, indicating a tendency for suspense-modified narratives to increase suspense ratings, although the effect was attenuated when accounting for covariates. Familiarity with specific stories had a clear influence: Story 2 showed a significant positive effect on suspense ratings ($p = 0.003$), as did Story 3 ($p = 0.006$) and Story 4 ($p < 0.001$). These findings suggest that familiarity with narrative content can enhance perceived suspense, possibly by shaping expectations and attention.

Among the covariates, **Engagement** emerged as a strong and significant predictor of suspense ($p = 0.001$), supporting the idea that immersive experiences heighten suspense perception. In contrast, **Clarity** was not a significant predictor ($p = 0.448$), implying that making a story easier to understand does not necessarily increase its suspensefulness.

Most interaction terms in the model were non-significant, but one notable exception was the interaction between *design* and *clarity*, which approached significance ($p = 0.056$). This suggests that clarity may moderate the effectiveness of suspense-enhanced narratives: when a story is well-structured and clear, the suspense manipulation may be more effective. If the narrative becomes too ambiguous or complex, cognitive overload may reduce the intended suspenseful effect. In contrast, a suspenseful story with clearly presented events may allow readers to fully absorb the tension and anticipate outcomes, maximizing engagement and emotional response.

Discussion and Conclusion

This hierarchical analysis—beginning with ANOVA, followed by ANCOVA, and concluding with OLS regression—enabled a systematic examination of suspense modification while accounting for potential confounds. The findings confirm that suspense-enhanced narratives significantly increase perceived suspense, with engagement emerging as a key predictor. While some interaction effects were observed, they did not reach conventional significance levels, suggesting that suspense modifications are broadly effective across different conditions rather than being highly dependent on prior knowledge or clarity.

Future Work

Future research should explore how nonlinear structures, such as flashbacks and parallel storylines, influence suspense by adding complexity and uncertainty. Additionally, as interactive media like video games and branching narratives gain prominence, examining the interplay between player agency and suspense could deepen our understanding of dynamic storytelling.

Cross-cultural studies may reveal how different storytelling traditions shape suspense perception, while investigations into psychological traits—such as tolerance for ambiguity—could shed light on individual differences in suspense experiences. Another key area is the impact of repeated exposure: does familiarity diminish suspense, or can certain narrative techniques sustain it despite foreknowledge?

Comparing suspense across media formats, including film, television, and podcasts, could also highlight medium-specific mechanisms. Finally, ethical considerations surrounding suspense in narratives, particularly its effects on vulnerable populations, warrant further study. Addressing these questions will enhance our understanding of suspense and its role in crafting engaging narratives.

References

[Bentz, Kuchinke, and Jacobs 2024] Bentz, A.; Kuchinke, L.; and Jacobs, A. M. 2024. Narrative suspense and emotional engagement: A novel experimental approach to measuring suspense in real-time. *Scientific Study of Literature*.

[Brewer and Lichtenstein 1982] Brewer, W. F., and Lichtenstein, E. H. 1982. Stories are to entertain: A structural-affect theory of stories. *Journal of Pragmatics* 6(5-6):473–486.

[Bálint et al. 2017] Bálint, K.; Kuijpers, M. M.; Doicaru, M. M.; Hakemulder, F.; and Tan, E. S. 2017. The effect of suspense structure on felt suspense and narrative absorption in literature and film. In Kuijpers, M.; Hakemulder, F.; Tan, E. S.; Bálint, K.; and Doicaru, M., eds., *Narrative Absorption*, volume 27 of *Linguistic Approaches to Literature*. John Benjamins Publishing Company. 177–197.

[Carroll 2013] Carroll, N. 2013. The paradox of suspense. In *Suspense*. Routledge. 71–91.

[Cheong and Young 2015] Cheong, Y.-G., and Young, R. M. 2015. Suspenser: A story generation system for suspense. *IEEE Transactions on Computational Intelligence and AI in Games* 7(1):39–52.

[Comisky and Bryant 1982] Comisky, P., and Bryant, J. 1982. Factors involved in generating suspense. *Human Communication Research* 9(1):49–58.

[Doust and Piwek 2017] Doust, R., and Piwek, P. 2017. A model of suspense for narrative generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, 178–187. Association for Computational Linguistics.

[Green and Appel 2024] Green, M. C., and Appel, M. 2024. Narrative transportation: How stories shape how we see ourselves and the world. *Advances in experimental social psychology* 70(1).

[Hare 2007] Hare, W. 2007. *Hitchcock and the Methods of Suspense*. McFarland.

[Lehne and Koelsch 2015] Lehne, M., and Koelsch, S. 2015. Toward a general psychological model of tension and suspense. *Frontiers in Psychology* 6:79.

[Lehne et al. 2015] Lehne, M.; Engel, P.; Rohrmeier, M.; Menninghaus, W.; Jacobs, A. M.; and Koelsch, S. 2015. Reading a suspenseful literary text activates brain areas related to social cognition and predictive inference. *PLoS ONE* 10(5):e0124550.

[O'Neill and Riedl 2014] O'Neill, B., and Riedl, M. 2014. Dramatis: A computational model of suspense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28. AAAI.

[OpenAI 2024] OpenAI. 2024. Gpt-4o: Multimodal language model. Accessed: 2025-03-10.

[Smuts 2008] Smuts, A. 2008. The desire-frustration theory of suspense. *The Journal of Aesthetics and Art Criticism* 66(3):281–290.

[Tan 1996] Tan, E. S. 1996. *Emotion and the Structure of Narrative Film: Film as an Emotion Machine*. Mahwah, NJ: Lawrence Erlbaum Associates.

[Vorderer 2013] Vorderer, P. 2013. Toward a psychological theory of suspense. In *Suspense*. Routledge. 233–254.

[Xie and Riedl 2024] Xie, K., and Riedl, M. 2024. Creating suspenseful stories: Iterative planning with large language models. *arXiv preprint arXiv:2402.17119*.

Abductive Computational Systems: Creative Abduction and Future Directions

Abhinav Sood
University of Sydney
Sydney, Australia
abhinav.sood@sydney.edu.au

Kazjon Grace
University of Sydney
Sydney, Australia
kazjon.grace@sydney.edu.au

Stephen Wan
Data61, CSIRO
Sydney, Australia
stephen.wan@data61.csiro.au

Cecile Paris
Data61, CSIRO
Sydney, Australia
cecile.paris@data61.csiro.au

Abstract

Abductive reasoning, reasoning for inferring explanations for observations, is often mentioned in scientific, design-related and artistic contexts, but its understanding varies across these domains. This paper reviews how abductive reasoning is discussed in epistemology, science and design, and then analyses how various computational systems use abductive reasoning. Our analysis shows that neither theoretical accounts nor computational implementations of abductive reasoning adequately address generating creative hypotheses. Theoretical frameworks do not provide a straightforward model for generating creative abductive hypotheses, and computational systems largely implement syllogistic forms of abductive reasoning. We break down abductive computational systems into components and conclude by identifying specific directions for future research that could advance the state of creative abductive reasoning in computational systems.

Introduction

Abductive reasoning is a form of logical reasoning that deals with explanations for hypotheses, conjectures, anomalies, etc. For example, consider a person who returns to their house to find their belongings scattered around, along with a broken window. If they had left the house clean, they could abduce that a burglar might have broken into their home while they were out. Charles S. Peirce first formulated this form of reasoning as (Peirce 1958):

1. A surprising fact O is observed.
2. But if [hypothesis] H were true, O would follow.
3. Thus, there is a reason to suspect that H is True.

Unlike deductive reasoning, abductive reasoning does not aim to preserve the truth of the derived propositions. Instead, pragmatically, abductive reasoning seeks to explain surprising observations where other forms of reasoning might not be possible. In a computational creativity context, this contrasts with how surprise typically provides an evaluation of creativity (Grace and Maher 2014). Instead, abduction posits that surprises are an initiating event: they are the unexplainable observations for which abduction provides an explanation. This ability to generate plausible explanations in the face of surprise makes abductive reasoning particularly valuable in both scientific, design-related, and artistic contexts: scientists rely on it to formulate hypotheses that might explain unexpected phenomena, while designers use it to envision solutions that address observed needs without complete information about the problem space. Similarly, artists employ abductive reasoning when interpreting ambiguous visual elements and unexpected results, using these surprises as catalysts for developing artistic expressions.

Thus, understanding how computational systems use abductive reasoning can help us identify properties that may increase their utility. To the best of our knowledge, no existing review examines how abductive reasoning is implemented across computational systems in these diverse fields. The main contributions of this paper are:

1. We review and connect literature on abductive reasoning that exists across different, often unlinked fields.
2. We analyse how computational systems implement abductive reasoning and propose a breakdown that identifies simple components for advancing creative abduction.

Abductive Reasoning in Theory

Epistemological Background

Since Pierce's formulation of abduction, several philosophers have further developed the concept of abductive reasoning. Gabbay and Woods (2005) discuss the ignorance-preserving nature of abduction. Ignorance-preserving means that abductive reasoning does not generate a hypothesis that completely resolves the surprising observation. Instead, uncertainty remains about whether the hypothesis is true even after a hypothesis is produced. So, what use does the hypothesis provide? It is a reasoned basis for future action (Gabbay and Woods 2006).

By abductive reasoning, we are not referring to inference to the best explanation (IBE). While abduction is just the reasoning process to create and select a possible hypothesis, IBE is an iterative process that might involve deductive, abductive and inductive reasoning to select and arrive at the best hypothesis available. A model for IBE in diagnostic reasoning is presented in (Magnani 2001, 23). This model uses abduction to propose diagnostic hypotheses based on clinical evidence. Observed clinical data corroborate hypotheses through induction, and the deductive implications

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

of the hypothesis can provide us with expectations of what our data should look like. These expectations form the basis for future surprises that trigger further abduction.

There are several formalisations of abductive reasoning. The Gabbay Woods Schema (GW-Schema) (Gabbay and Woods 2006, pg. 198) aims to capture several properties of Peircean abduction. Although the GW-Schema illustrates the underlying syllogistic logic of abductive reasoning, it suffers from two major problems. Firstly, it does not explain how the hypotheses themselves are generated. Second, it does not concretely define the conditions that a hypothesis needs to meet to be considered worthy of investigation. Some of the conditions often discussed in the literature (Douven 2021) are simplicity, coherence and explanatory power. However, no consensus exists on whether these specific conditions are necessary. For example, in (Woods 2017, pg. 139), Woods rejects simplicity and "sees no reason why truth favours the uncomplicated". The issues of coming up with a hypothesis and of selecting a hypothesis among candidates based on specific conditions are referred to as the fill-up and cut-down problems (Magnani 2017, pg. 6) respectively. Magnani discusses how the solutions to these problems are "stunningly contextual" (Magnani 2017, pg. 10). While his proposed eco-cognitive model provides us with an understanding of abductive reasoning in more creative, scientific contexts, it does not provide a computable formal structure.

When looking for computable formal structures, some might contest that Abductive Logic Programming (ALP) (Kakas, Kowalski, and Toni 1992) can fill the gap. ALP extends standard logic programming by introducing abducible predicates. These frameworks allow systems to derive explanations by backward-chaining from observations to possible causes, constrained by integrity constraints that filter out implausible explanations. ALP has been shown to be effective for diagnostic reasoning (Koitz-Hristov and Wotawa 2018), but typically operates in closed, well-defined domains and struggles with more open-ended abduction.

The differing theoretical approaches to abductive reasoning have led Woods to argue that "The foundational work for a comprehensive account of abductive reasoning still awaits completion" (Woods 2017, pg. 138). Unfortunately, this means that the epistemological literature on abductive reasoning does not sufficiently address how you can functionally perform the more complex forms of abductive reasoning. As such, it might be worthwhile to look at how creative abductive reasoning emerges in two domains where creativity and hypotheses are central: science and design.

Creative Abductive Reasoning in Science and Design

In the philosophy of science, abductive reasoning is traditionally associated with creative processes in the context of discovery (Schickore 2022). Here, the "context of discovery" refers to the initial generation of theories in science through the conception of hypotheses and research ideas. Logically abductive reasoning deals with producing explanations for surprising observations. We propose that this process of producing explanations is relevant throughout the scientific processes at two distinct levels.

In the "context of discovery", it is relevant macroscopically; abductive reasoning serves as a guiding framework for the entire scientific process, shaping what hypotheses and ideas are investigated. Microscopically, abductive reasoning is used within individual experimental stages of the scientific process. To illustrate this, consider the documentation of the discovery of the urea cycle in (Kulkarni and Simon 1988). Macroscopically, after observing the unexpected effect of ornithine in enhancing urea production, Krebs abductively reasoned that ornithine might play a central role in a cyclical process of urea formation. This explanation accounted for the surprising observation and led him to restructure his entire research program around understanding ornithine's role in urea synthesis. Note that he "did not have a clear hypothesis of a mechanism to account for it [ornithine effect]" that he was testing (Kulkarni and Simon 1988). Instead, a rougher hypothesis was sufficient to guide his investigation. Microscopically, Krebs employed abductive reasoning at specific experimental junctures. When he discovered that the effect was unique to ornithine (as chemical derivatives failed to produce similar results), he generated the explanatory hypothesis that ornithine possessed specific structural properties essential to urea formation.

In design, abduction takes on different characteristics when compared to abductive reasoning in science. Design theorists have identified abduction as the fundamental reasoning pattern for moving from function to form, from "what is needed" to "how to do it" (Kroll and Koskela 2015). Roozenburg (1993) made a distinction between explanatory abduction (explaining existing phenomena) and innovative abduction (creating new solutions), and argued that design is primarily the latter. This innovative form of abduction in design is different from scientific abduction as it aims to create entirely new artefacts to fulfil desired functions. Koskela, Paavola, and Kroll (2018) further elaborated on this view, saying that abduction in design is not just an inference type but a property of many inferences throughout the design process. Using a SAPPhIRE-based model of abductive reasoning in design, Bhatt, Majumder, and Chakrabarti (2021) provide a framework to explain how designers use abductive reasoning to bridge the gap between desired functions and proposed solutions. These models are useful to understand the nature of abduction, but can they be automated? To further understand how abductive reasoning takes place practically, we look at how computational systems implement abductive reasoning. Just as (Boden 2004, pg. 17) argues that computational ideas can inform our understanding of human creativity; computational ideas can inform our understanding of abductive reasoning.

Abductive Computational Systems

Natural Language Processing (NLP)

For modern computational systems, recent investigations of abductive reasoning have largely focused on introducing it in Large Language Models (LLMs). Large Language Models (LLMs) are neural networks that generally utilise the trans-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

former architecture (Vaswani et al. 2017) to model natural language probabilistically. One of the first papers to investigate abductive reasoning with LLMs was (Bhagavatula et al. 2020). They considered how well LLMs perform abductive reasoning in everyday situations derived from datasets of short stories. The authors created two tasks to test how well LLMs performed. Alpha NLI tested if LLMs could select the correct abductive hypothesis (given a choice of 2) that fits observations that occur before and after the hypothesis. The language model they trained had an accuracy of 68.9%, with humans identifying the correct hypothesis in 91.4% of cases. In our experiments, we found that LLaMa 3.1 70B-Instruct [1] was able to achieve an accuracy of 86.2% with a zero-shot prompt on the test set of the Alpha NLI task, indicating that modern LLMs might perform well in identifying correct abductive hypotheses in everyday situations. These results are merely indicative, as there might be data leakage from the test set, but they point towards the possibility of effective abductive reasoning through LLMs. Beyond everyday situations, Zhao et al. (2024) train LLMs to generate uncommon explanations in everyday situations. Tian et al. (2024) constructs MacGyver-like creative problems to train LLMs. These studies focus on extending the abductive capabilities of LLMs by creating datasets for specific tasks. Such datasets that exclusively focus on abductive reasoning are not yet prevalent in science and design.

Many abductive systems in NLP also focus on knowledge bases. AbductionRules (Young et al. 2022) tests how well transformers reason over logical knowledge bases. Bai et al. (2024) use knowledge graphs to generate hypothesis-observation pairs. There is also research on hypothesis generation in NLP that does not explicitly mention abductive reasoning. These systems have their own host of issues.

In the case of (Si, Yang, and Hashimoto 2025), several limitations of the AI-generated research ideas (inappropriate baselines, unrealistic assumptions, etc.) were pointed out, which brings into question whether the novel ideas generated are usable in practice. With the AI Scientist (Lu et al. 2024), LLMs themselves evaluated the quality of AI-scientist-generated research papers, and such modes of evaluation have been called into question (Koo et al. 2024).

Beyond commonsense reasoning, abductive reasoning has also been used in several computationally creative systems.

Abductive Reasoning in CC Systems

To find computational creativity (CC) systems where abductive reasoning was used, we conducted a review of CC conferences. The start (2010) and end (2024) years for our review mark the first and the most recent ICCC conference. To identify systems that used abductive reasoning, we used the search terms ”abductive”, ”abduce”, and ”abduction”. Among the ICCC proceedings, 14 papers contained the search terms. Of these 8 mentioned abductive for purposes of literature review, argumentation or used the term ”abduction” in the physical rather than logical sense (mostly kidnappings in narrative generation papers). We were left with 6 CC systems that either used abductive reasoning computationally or were used to elicit abductive reasoning in the users of the system. These systems are discussed in Table 1.

#	System	Abductive Approach
(1)	The Painting Fool (Colton 2010)	Identifies constraints among partial rectangles, extends patterns to produce abduced versions
(2)	Synthetic Audience (O'Neill and Riedl 2011)	Extracts information from incomplete stories to build a mental model, using abduction to infer character goals
(3)	MILA-S (Goel and Joyner 2015)	Prompts students to engage with abductive explanations through ecological model building.
(4)	ANGELINA (Cook and Colton 2018)	Uses abductive reasoning with answer set programming to generate game rulesets
(5)	Hello (Veiga 2020)	Encourages audience to build narratives through abductive reasoning as the artwork is ”incomplete”
(6)	Aris (O'Donoghue et al. 2022)	Creates abduced hypothesised formal specifications using homomorphic matching and topological similarity on knowledge graphs

Table 1: Computational Systems at ICCC that mention Abductive Reasoning

We also reviewed the proceedings of ACM Creativity and Cognition (C&C) and Design Computing and Cognition (DCC) conferences for the same search terms. In our review of C&C and DCC proceedings, we found no computational system implementations: the sole C&C paper used abduction to identify richness (potential to inspire future ideas), inspirational ideas via keywords, while the 11 DCC papers proposed focused on modelling abduction design and were not specifically implemented as CC systems and centered on design theory. We note that our review is not exhaustive. Many computational systems effectively address abductive problems without explicitly framing their approach as abductive reasoning. While these systems may offer valuable insights, their inclusion is beyond the scope of this short paper.

Components of Abductive Computational Systems

Through our review of abductive computational systems, we identify four key components that characterise these systems as visualised in Figure 1.

Abductive Triggers These act as the starting point for performing abduction. Systems (3) and (5) involve using generated artefacts and constructed models, respectively, to encourage abductive reasoning. The other systems do not necessarily possess abductive triggers in a Peircean sense. Instead, they rely on a computational trigger to perform abductive reasoning. For example, (1) conducts abduction when a

[1] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

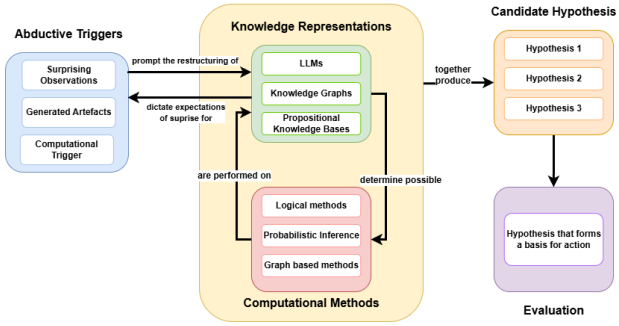


Figure 1: Proposed breakdown for a computational abductive system

user is done inputting a partial configuration of rectangles.

Knowledge Representations In the CC abductive systems we covered, knowledge was encoded as knowledge graphs in (6). In (1), knowledge was construed as constraints that eventually form a constraint satisfaction problem. Abductive reasoning performed in NLP systems we introduced earlier also used knowledge representations. Young et al.(2022) stored knowledge in a propositional form. Bai et al. (2024) used knowledge graphs. In (Bhagavatula et al. 2020), since LLMs were generating the resulting abduced hypothesis, LLMs functioned as black-box stores of knowledge.

Computational Methods :Computational methods perform abductive reasoning in computational systems. These methods heavily depend on the knowledge representation in the CC system. (1) uses extensions of constraints. (4) uses answer set programming to perform abduction. (6) performs abductive reasoning through graph-based methods. The LLM-based systems perform abduction through probabilistic inference.

Evaluating Hypotheses : In the CC systems, the selection process was very specific to the systems. (1) allows users to remove abduced rectangles individually if they are perceived by the user as incorrectly abduced. (6) uses deductive reasoning to verify abduced hypotheses. These steps fall under hypothesis verification, which is a part of IBE but not a part of our definition of abductive reasoning. In (3) and (5), the human performs the evaluation. Few systems actually use abductive properties to evaluate hypotheses. Dalal et al. (2024) used properties like consistency, depth, drift, coherence and linguistic uncertainty to evaluate hypotheses, but the hypotheses are restricted to syllogistic if-then structures.

Future Directions

Our analysis reveals a significant gap between theoretical understandings of abductive reasoning and existing computational implementations. While epistemological accounts emphasise the creative nature of hypothesis generation, computational systems primarily implement restricted, syllogistic forms of abduction. That is, future work should focus on developing systems that can perform creative abduction. One promising direction for this is to create datasets for scientific and design-related abduction, similar to datasets that have already been constructed for commonsense and uncommonsense abductive reasoning. Current systems lack robust methods for evaluating abduced hypotheses, particularly those in natural language form. Developing computationally tractable metrics for properties like simplicity, coherence, and explanatory power could advance our ability to evaluate and compare creative abductive systems.

Our breakdown of computational abductive systems also presents opportunities for the exploration of collaborative and co-creative systems in science. Consider that a literature review of co-creative systems produced 92 systems (Rezwana and Maher 2023); not one was in the domain of science. For abductive triggers, human-machine collaboration could involve the system identifying potentially surprising patterns in data that a human might overlook, while humans could guide the system toward observations that they believe are relevant to scientific questions. For knowledge representations, humans could supply contextual domain knowledge that might be missing from the system's formal representations, while the system could visualise its knowledge structures in ways that reveal unexpected connections.

Conclusion

This paper covered how abductive reasoning is viewed and discussed across epistemology, science, design and computational systems. Our analysis shows that while abductive reasoning is widely held to be key to creative processes, computational implementations of abduction are restricted to syllogistic forms that leave little room for creativity. By breaking down abductive computational systems into abductive triggers, knowledge representations, computational methods and evaluation, we provide a basis to understand and implement more advanced computational creative abduction systems that close the gap between the theoretical understanding and computational implementation of abduction.

Author Contributions

Author 1 (AS) conducted the review and prepared the manuscript. Author 2 (KG), Author 3 (SW) and Author 4 (CP) supervised the research and provided feedback and direction for the development of this work.

Acknowledgement

We gratefully acknowledge the Collaborative Intelligence (CINTEL) Future Science Platform for their support towards this work.

References

Bai, J.; Wang, Y.; Zheng, T.; Guo, Y.; Liu, X.; and Song, Y. 2024. Advancing abductive reasoning in knowledge graphs through complex logical hypothesis generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1312–1329. Bangkok, Thailand: Association for Computational Linguistics.

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; tau Yih, W.; and Choi, Y. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Bhatt, A. N.; Majumder, A.; and Chakrabarti, A. 2021. Analyzing the modes of reasoning in design using the sapphire model of causality and the extended integrated model of designing. *Ai Edam* 35(4):384–403.

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

Colton, S. 2010. The painting fool teaching interface. In *International Conference on Computational Creativity 2010*, 289.

Cook, M., and Colton, S. 2018. Redesigning computationally creative systems for continuous creation. In *International Conference on Computational Creativity 2018*, 32–39. Association for Computational Creativity (ACC).

Dalal, D.; Valentino, M.; Freitas, A.; and Buitelaar, P. 2024. Inference to the best explanation in large language models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 217–235. Bangkok, Thailand: Association for Computational Linguistics.

Douven, I. 2021. Abduction. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

Gabbay, D., and Woods, J. 2005. *The Reach of Abduction: Insight and Trial (A Practical Logic of Cognitive Systems, Vol. 2)*. Elsevier.

Gabbay, D., and Woods, J. 2006. Advice on abductive logic. *Logic Journal of IGPL* 14(2):189–219.

Goel, A., and Joyner, D. 2015. Impact of a creativity support tool on student learning about scientific discovery processes. In *Proceedings of the Sixth International Conference on Computational Creativity*.

Grace, K., and Maher, M. L. 2014. What to expect when you're expecting: The role of unexpectedness in computationally evaluating creativity. In *ICCC*, 120–128. Ljubljana.

Kakas, A. C.; Kowalski, R. A.; and Toni, F. 1992. Abductive logic programming. *Journal of logic and computation* 2(6):719–770.

Koitz-Hristov, R., and Wotawa, F. 2018. Applying algorithm selection to abductive diagnostic reasoning. *Applied Intelligence* 48(11):3976–3994.

Koo, R.; Lee, M.; Raheja, V.; Park, J. I.; Kim, Z. M.; and Kang, D. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics ACL 2024*, 517–545.

Koskela, L.; Paavola, S.; and Kroll, E. 2018. The role of abduction in production of new ideas in design. *Advancements in the Philosophy of Design* 153–183.

Kroll, E., and Koskela, L. 2015. On abduction in design. In *Design Computing and Cognition'14*, 327–344. Springer.

Kulkarni, D., and Simon, H. A. 1988. The processes of scientific discovery: The strategy of experimentation. *Cognitive science* 12(2):139–175.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The ai scientist: Towards fully automated open-ended scientific discovery.

Magnani, L. 2001. *Abduction, Reason and Science*. Springer US.

Magnani, L. 2017. *The Abductive Structure of Scientific Creativity: An Essay on the Ecology of Cognition*, volume 37 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer International Publishing.

O'Donoghue, D. P.; Brady, C.; Smullen, S.; and Crowe, G. 2022. Novelty assurance by a cognitively inspired analogy approach to uncover hidden similarity. In *ICCC*, 277–281.

O'Neill, B., and Riedl, M. O. 2011. Simulating the everyday creativity of readers. In *International Conference on Computational Creativity 2011*, 153–158.

Peirce, C. S. 1958. *Collected papers*, volume 5. Belknap Press of Harvard University Press.

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).

Roozenburg, N. F. 1993. On the pattern of reasoning in innovative design. *Design studies* 14(1):4–18.

Schickore, J. 2022. Scientific Discovery. In Zalta, E. N., and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

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*.

Tian, Y.; Ravichander, A.; Qin, L.; Le Bras, R.; Marjieh, R.; Peng, N.; Choi, Y.; Griffiths, T. L.; and Brahman, F. 2024. Macgyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5303–5324.

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.

Veiga, P. A. d. 2020. Hello, an interactive cinematic generative artwork. In *Proceedings of the Eleventh International Conference on Computational Creativity*, 469–475. Association for Computational Creativity (ACC).

Woods, J. 2017. Reorienting the logic of abduction. In Magnani, L., and Bertolotti, T., eds., *Springer Handbook of Model-Based Science*. Cham: Springer International Publishing. 137–150.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Young, N.; Bao, Q.; Bensemann, J.; and Witbrock, M. 2022. AbductionRules: Training transformers to explain unexpected inputs. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 218–227. Dublin, Ireland: Association for Computational Linguistics.

Zhao, W.; Chiu, J.; Hwang, J.; Brahman, F.; Hessel, J.; Choudhury, S.; Choi, Y.; Li, X.; and Suhr, A. 2024. UNcommonsense reasoning: Abductive reasoning about uncommon situations. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8487–8505. Mexico City, Mexico: Association for Computational Linguistics.

Foundation Models as Agents of Austerity

Anton Koch
Department of Design
Mainz University of Applied Sciences
Holzstr. 36, 55116 Mainz, Germany
anton.koch@hs-mainz.de

Guilherme Zanchetta
Division of Cyber-Physical Systems
CTI Renato Archer,
NICS-UNICAMP
Campinas, Brazil
zanchetta@nics.unicamp.br

Abstract

This article examines the role of Foundation Models within the contemporary landscape of generative Artificial Intelligence, arguing that they should not be regarded as neutral technological developments, but rather as instruments embedded in the political economy of neoliberal austerity. It contends that these models facilitate the displacement of socially embedded practices, favouring efficiency-oriented, privatised, and platform-governed operations.

Introduction

The concept of the foundation model (FM) in its various implementations dominates the current discourse around generative artificial intelligence (AI). The term refers to generative models trained on vast amounts of data, such as a large language model (LLM) or general pre-trained transformer (GPT). It has gained popularity through services such as chatbots and image, audio, and video generators. In this article, we will argue that the FM is deployed as an automated continuation of the neoliberal idea of austerity and acts as an agent of it. The politico-economic project of austerity can be broadly described as a reduction of public expenditure. It seeks to gradually replace social organisation in state institutions with a market- and game-theory-based framework, aiming for maximum connectivity and information flow with minimal outside interference. Historically, the LLM emerged in the wake of a global intensification of austerity measures following the 2008 financial crisis. It is beyond the scope of this article to determine whether this development was purely coincidental or benefited from quantitative easing measures, which released vast amounts of low- to no-risk capital into the tech sector, allowing massive stock buybacks and infrastructure buildouts. However, the broad scope of use and inherent logic of big-data stochastic working, which is similar to managerial decision-making, make it a good fit. While the technology enables the automation of general administrative tasks, it has also begun to impact a growing number of workers in the creative disciplines of writing, illustration, and photography. In some cases, creative workers are now merely required to "clean up" the automated output, thus being progressively alienated from their practice. Their work is neither replaced nor augmented, but devalued, largely phased out and replaced with generic output, similar to many professions previously displaced by capitalist automation logic.

Basic concepts

To situate our critique of FMs, we need to address the specific terminology used when talking about these models and their purpose. However, it also requires us to clarify our frame of reference regarding generative AI in general as well as the FM in particular. We acknowledge the long and complicated history of the general field of AI with its numerous approaches to the concept of machine intelligence. For this article, we are focusing explicitly on the FM as being "trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (Bommasani et al. 2022, p. 3). These models, while not new in terms of their underlying methods and technologies, are unprecedented in scale and scope and have been referred to as foundation models by Stanford researchers to indicate a "paradigm shift" (Bommasani et al. 2022, p. 6) in the field of AI.

Throughout this article, *Intelligence* is broadly defined as the appropriation of information, its transformation into knowledge and the latter's application. However, it should also be seen as an emergent and dynamic property of any complex system that is highly context-dependent. In that sense, we argue for referring to, as physicist Frank Wilczek has called it, an "ecology of intelligence" (Frank Wilczek 2019), a plethora of diverse tapestries of artificial and human intelligence. This view rejects the stereotypical notion of a hierarchical, linear intelligence quotient (IQ) scale, which is a profoundly flawed concept (Gebru and Torres 2024).

Most importantly, we ground our understanding of *computation* as a formalist practice, encompassing the execution of mathematical operations, algorithmic symbol manipulation, and stochastic evaluation. While this function is now primarily tied to the machinic computer, it is worth noting that this practice did not emerge with the machine, but predates it by far. It must also be noted here that, while we employ this broad definition, it is essential to consider the specificity of computation concerning its cultural framework, as laid out by scholar David Golumbia (Golumbia 2009). At the same time, computation has been shown to transcend the simple definition of an automated mode of calculation or simple digital mechanics in terms of its metaphysics (Brian

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Cantwell Smith 1996) and aesthetics (Fazi 2018). While these discussions by far exceed the scope of this article, they present an important horizon for an overall computational potential.

The broadest definition of what AI constitutes is a project that strives to automate reason. To this end, it deploys computation to realise a form of quasi-autonomous decision-making within a machine. This includes rule-based, symbolic AI as well as the currently popular connectionist variants of the artificial neural network (ANN) and deep neural network (DNN). Here, we follow Matteo Pasquinelli's historical account of AI's development from a specific, capitalist organisation of labour (Pasquinelli 2023), as well as Dan McQuillan's socio-political critique of its deployment (McQuillan 2022).

Finally, we understand *creativity* as the production of ideas and their active exploration and reflection. While the term can be used in many contexts, we specifically refer to the labour performed within the contemporary creative industries.

Background

An abbreviated history of creative computation

Using computational machines and methods in art has a longstanding history, but has been particularly influential within the field of "computer art", which began in the second half of the 20th century. There have been early examples of geometric and mathematical art by pioneers like Vera Molnár and Manfred Mohr. Early forms of data-driven decision-making in the creative process are exemplified by works of artists like John Cage, with what the composer called chance operations (Nicholls 2002) and Iannis Xenakis's Stochastic music (Xenakis 1966). Later on, notable approaches to the simulation of "artificial life" were developed by Karl Sims and William Latham. Markov chains have been extensively used to generate sequential and interactive compositions (Manzolli 1996) (Manzolli and Baruco Machado Andraus 2022) Lastly, AI art has a history that extends beyond today's generative models, with artists such as Helena Sarin and Mario Klingemann utilising generative adversarial networks (GANs) to create original works.

Computationalism, Anthropomorphism and the Humanist Trap

There are symptoms that, although entangled, can be analytically located: computationalism—understood not merely in its naïve form as the assertion that "everything is a computer", but more rigorously as computation (following the elaborations of digital philosophers and thinkers of the philosophy of computation)—and anthropomorphism, the projection of human characteristics onto non-human entities. Both tendencies are embedded in and amplified by the cybernetic apparatus (Geoghegan 2011), which encompasses mathematical procedures, diagrammatic strategies, and technological tools that have functioned both as material infrastructures for research and as quasi-transcendental ideals. This apparatus has enabled a mode of abstraction that displaces historical, political, and disciplinary specificities, while simultaneously facilitating alliances among researchers and institutions across diverse borders. In this epistemic regime, cybernetics does not approach machines by asking "what individual act will it produce here and now?" but rather "what are all the possible behaviours it can exhibit?"—a framing aligned with information theory, which concerns itself not with particularities but with sets of possibilities. Energy, once central in biological explanation, becomes a given in cybernetic reasoning; the focus shifts to the constraints and regulatory mechanisms that delimit developmental trajectories. Whether a system is energetically open or closed is often secondary to the degree to which its behaviour is structured by controlling factors within a broader space of potential outcomes. In this regard, cybernetics positions itself in relation to material systems—electronic(Ashby 1956), mechanical, biological, economic, or political—much as geometry once related to terrestrial objects: just as geometry moved from empirical demonstrations in Euclidean space to the formal coherence of non-Euclidean and abstract geometries, cybernetics generalises from instantiated machines to the conceptual space of all possible machines. What matters is not the specific mechanism but the architecture of behaviours it can potentially exhibit. This shift toward abstraction had a retroactive impact on biology itself, reconfigured by cybernetic thought, while, paradoxically, engineers began deploying biological metaphors for machinic concepts and implementations (Caporael 1986). Over time, these metaphors also feed back into the sciences, as has been particularly effective with the idea of the mind as a computer (Pasquinelli 2021). The politics of humanism within the western imperialist system and its project of science (as envisioned by Francis Bacon) produced an anthropocentric, exceptionalist and hierarchical worldview with the human (more specifically, the white and male human) in the centre. While the Enlightenment ostensibly sought to ground everything in reason, it did so according to the existing demarcations of imperialist, scientific hierarchies. Now, Transhumanism seeks to overcome the human condition entirely, sees it as an intermediate step, thereby clinging to the same linear, hierarchical exceptionalism of a singular intelligent quality.

The limits of AI

Connectionist AI is characterised, by some of its proponents, as a universal technology (Eloundou et al. 2023), but it is deeply indebted to a particular epistemological concept and thereby limited within that specificity. To situate a critique of AI, it is vital to frame it within its epistemological and functional limitations and thus to decouple it from its sprawling mythologies.

The idea of the neural network goes back to the biologist McCulloch's and logician Walter Pitts' publication of their "logical calculus of the nervous system" (McCulloch and Pitts 1943), which proposed a formalised model derived from a theory on the function of vision. This theory of the eye as a networked processor and classifier of sensory information to be forwarded to the brain was immensely influential, as the economist Friedrich von Hayek also picked it up

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

for his theory of mind (Hayek 1952). It was Hayek's work that would come to broadly influence both an economic formalisation of markets as cellular networks of signal flow and a general understanding of the mind as a computer. Scholar Matteo Pasquinelli has described this entry of a technical paradigm into the natural sciences as an "ideological coup d'état" (Pasquinelli 2021, p. 161). While it may be debatable whether Hayek was somehow consciously aiming to naturalise economic theory, it is evident that the connectionist idea was a precursor for the upcoming so-called "information age". It also influenced Frank Rosenblatt, who credits Hayek in his paper on the Perceptron, probably the first machine implementation of an ANN (Rosenblatt 1958, p. 388). The Perceptron symbolises the convergence of connectivism with the stochastic reasoning through classification by the "regression to the mean" as introduced by Francis Galton. Both ideas are rooted within the Western, colonial project of humanist science and, as such, are powerful tools for the construction of neoliberal ideas around societal self-organisation through uninhibited connectivity, stochastic management and free flow of capital and information.

AI's functional limitation originates in its particular specificity modelled as a machine for computer vision (CV). It is not conceived as a brain per se, but a specialised organ for vision and classification of patterns, as scholar James Dobson details in his history of CV (Dobson 2023). In its large form, today's models are made to "see" within large quantities of data and respond in a synthesised way (e.g. natural language, images), but also with a certain "lossyness", as argued by author Ted Chiang (Chiang 2023). In opposition to any analytical, constructive process, it merely performs a brute-force approach at fitting myriads of weak associations to a pattern and then reproducing these patterns. While this certainly leads to impressive results at scale, it is still functionally limited to information storage and retrieval. This is not to dismiss the potential of using large models for a variety of tasks, but these are generally performed with a large overhead of computing power compared to specialised algorithms, as Huggingface's lead researcher, Sasha Luccioni, has detailed (Luccioni, Jernite, and Strubell 2024).

The Perceptron ultimately failed to deliver on what was already the first wave of excitement over the prospect of supposedly thinking machines, ushering in what was called the "AI winter" and taking a back seat to symbolic AI. It took until the beginning of the 21st century for the resurgence of connectionist AI with the advent of AlexNet in 2012, to the current boom triggered by the release of ChatGPT in 2022. However, this resurgence is not necessarily due to an evolution in the machine learning technique itself, as deep learning has been around at least since the 1990s. The exponential growth in feasible model sizes and dimensionality was made possible by the discovery of graphics processing units (GPUs) technology's capability to accelerate parallel computing (Nicole Hemsoth 2015) and a massive buildout of datacenter infrastructure during the last two decades after the advent of cloud computing. The models we see today are running at an unprecedented technological scale, but do so at an even more massive cost in compute power, data labour and mass-scale appropriation of content.

Generative AI as a form of automated austerity

Generative AI works on the surface, stripping away the underlying processes and recombining based on billions of arbitrary, weak connections. It automates the preemption of the social production process through stochastic prediction. Creation becomes consumption as it is reduced to natural-language search queries to recombine existing works. Intelligence, here, is falsely interpreted as a functional, algorithmic procedure of information retrieval and recombination instead of a complex assemblage of social practices and relations. It also perpetuates a corporate enclosure of the co-creative practice, in that it offers a puppet, or a *potemkin artist*, as a co-creative other, but essentially capitalises the practice itself. Then, as creation becomes more dependent on AI infrastructure, it creates a fundamental dependency that allows rent extraction on the ability, not the tools, to perform a practice like writing or illustrating. As austerity seeks to replace the public with the private, the communal with the proprietary, to invigorate the markets, these models strip away by predictively preempting the creative process that might otherwise involve social interaction, reflection and experience.

Foundations for critique and opposition

We need to work from a materialist base to situate and critique dominant technological trends. While there is speculation on "existential risk" or utopian predictions of "abundance", it is clear that AI as an economy, as a socio-technical apparatus, is already impacting countless lives worldwide in very different ways. These experiences, of data-labourers, rare-earth miners, but also writers and illustrators now only hired to clean up generative AI output, should be guiding the discussion around it, along with its vastly unsustainable energy and resource demands. There needs to be a categorical rejection of corporate black-box technologies that cannot be replicated either due to intellectual property rights but also because of the sheer amount of resources needed to reproduce them. It is also important to critically focus on questions of authorship in the face of unlawfully acquired content for training, as well as ownership of a practice. The way these models are currently implemented, they function as an enclosure of social production and the corporate appropriation of creative practice. This threat is similar to past struggles, not against technological progress, but the displacement, deskilling and alienation of specialist workers, such as in the Luddite movement in the early 19th century. As such, it is also clear that, even if the technology may be impressive, sometimes it still needs to be rejected wholesale. But we also acknowledge that this rejection often is not an option if employers require the workers to interact with it, or project funding is dependent on its implementation. This is why a resistance to technologies of austerity is not just brought on by conscious consumption, by refusing to participate in the hype, but through labour movements and political activism. As the FM is an agent of austerity, it is still the most viable strategy to push back against austerity politics instead of its symptoms.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Deep Environments over Deep Networks

Beyond the materialist critique of generative AI, but also the general technology of late-stage capitalism, there is another radical and powerful approach. The philosopher Sylvia Wynter challenges the entire singular foundation of the underlying science in that she advocates for, as scholar Walter D. Mignolo calls it, "epistemic disobedience" (Mignolo 2015, p. 106). This perspective serves as a starting point for acknowledging the specificity of Western science as it emerged within the colonial framework of exploration, classification, and quantification for optimal extraction and exploitation. This does not mean discarding science as a whole because of its fraught and complicated history, but understanding it as one epistemological model among others. Likewise, the FM represents a very narrow and specific epistemic construct, and everything built on it will be unable to go beyond this foundation in a fundamental way. However, if we remain conscious of its limitations and specificity, we can weigh it against alternative methods and models that pose true alternatives to merely switching out the training data. This article is not a call for the abolishment of contemporary technology and its replacement with an ethically or functionally superior variant or the return of supposedly "simpler times". It is rather an invitation to try and think beyond its narrow frame of future possibility. Instead of building a unified, enclosed "Metaverse" or striving for a global "Singularity", we prefer a pluriverse of hyperlocal digital infrastructures and practices. These exist as dynamic reflections of and extensions to their respective environment and culture, instead of aiming to establish normative and rigid frameworks or platforms. We prefer strata over stacks, relations over connections and variance over the mean. The philosopher Yuk Hui uses the term "Technodiversity" to describe an extended frame of understanding technology as culturally specific (Yuk Hui 2023), and thus open to being extended in its material and metaphysical potential by moving away from a singular, solutionist mindset. This concept is powerful in that it suggests not viewing code or machinery simply as tools or pragmatic solutions for perceived problems. But it also requires us to perceive our networks and assemblages of humans and machines neither anthropocentric, as the image of the cyborg, nor as the hierarchical master-and-slave duality. These networks are not characterised by the maximisation of information flow and frictionless efficiency as in the connectionist framework, but rather by kinship and care. Kinship, as Donna Haraway has proposed (Haraway 2016), can be felt within a non-hierarchical environment of people, creatures, plants and could even be thought to extend towards inhuman entities as much as its connectedness is not predicated on market value exchange, but on mere acknowledgement, awe or wonder. On a social level, this means to engage in networks of care that favour processes that foreground social relationships, deploying technology not as a mediator of communication but communication and expression in itself. Engaging in smaller, local digital ecosystems allows us to focus on the complex needs of actual local communities instead of generalised assumptions towards a stereotypical, fictional target audience. Finally, this also requires embracing uncertainty and ambivalence as axiomatic principles in design, engineering and creative processes to keep these dynamic and adaptable to environmental changes.

Conclusion

This article seeks to highlight the role of FMs as agents of austerity politics. To this end, we begin by clarifying the conceptual coordinates of intelligence, creativity, computation, and artificial intelligence. By tracing the epistemological and functional limitations of connectionism and computationalism, we argue that generative AI — particularly in its instantiation as the FM — is not merely a technological development, but a crystallised expression of neoliberal austerity. Through stochastic pattern-matching and the automation of creative and intellectual labour, FMs replicate the core logic of austerity: the erosion of socially embedded practices in favour of privatised, efficiency-oriented, and market-driven operations. Each interaction with an FM entails a tacit collaboration between the user, the model's contested training data, and the corporate entity that owns and governs both. Rather than enhancing creative agency, this configuration repositions the user within a closed feedback loop where authorship becomes secondary to the logic of data capture and platform optimisation. To this end, we have first set out to clarify our conceptual understanding of intelligence, creativity, computation and AI. We have shown the origins and limitations of connectionist AI and computationalism in their epistemological and functional specificities. On this basis, we attempt to demonstrate how generative AI in the form of the FM functions as an agent of austerity, reproducing common patterns of capitalist and neoliberal logic. This dynamic echoes itself in its asymmetrical co-creation; every interaction tacitly endorses a three-way collaboration between the user, the model's contested data, and the corporate entity governing both. Finally, we have offered some pointers to valid starting points to develop a materialist critique of these technologies and proposed to develop alternative imaginations and interpretations of the digital.

References

Ashby, W. R. 1956. An introduction to cybernetics / william ross ashby.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; Arx, S. v.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs].

Brian Cantwell Smith. 1996. Computation.

Caporael, L. 1986. Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior* 2(3):215–234.

Chiang, T. 2023. ChatGPT Is a Blurry JPEG of the Web. *The New Yorker*. Section: annals of artificial intelligence.

Dobson, J. 2023. *The Birth of Computer Vision*.

Eloundou, T.; Manning, S.; Mishkin, P.; and Rock, D. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv:2303.10130 [econ].

Fazi, M. B. 2018. *Contingent Computation: Abstraction, Experience, and Indeterminacy in Computational Aesthetics*. Rowman & Littlefield.

Frank Wilczek. 2019. Ecology of Intelligence.

Gebru, T., and Torres, E. P. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* 29(4).

Geoghegan, B. D. 2011. From Information Theory to French Theory: Jakobson, Lévi-Strauss, and the Cybernetic Apparatus. *Critical Inquiry* 38(1):96–126.

Golumbia, D. 2009. *The Cultural Logic of Computation*. Harvard University Press. Google-Books-ID: 9yAsEAAAQBAJ.

Haraway, D. J. 2016. *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press.

Hayek, F. A. 1952. *The Sensory Order: An Inquiry Into the Foundations of Theoretical Psychology*. Routledge & Kegan Paul.

Luccioni, A. S.; Jernite, Y.; and Strubell, E. 2024. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 85–99. arXiv:2311.16863 [cs].

Manzolli, J., and Baruco Machado Andraus, M. 2022. Jardim das cartas: an installation as a presence ecology. In *Proceedings of the 10th International Conference on Digital and Interactive Arts*, ARTECH '21. New York, NY, USA: Association for Computing Machinery.

Manzolli, J. 1996. Auto-organização: um paradigma composicional. In Debrun, M.; Gonzales, M. E. Q.; and Júnior, J. O. P., eds., *Auto-organização: estudos interdisciplinares*. Campinas: Editora CLE. 417–435.

McCulloch, W. S., and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5(4):115–133.

McQuillan, D. 2022. *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Policy Press.

Mignolo, W. D. 2015. Sylvia Wynter: What Does It Mean to Be Human?

Nicholls, D., ed. 2002. *The Cambridge Companion to John Cage*. Cambridge Companions to Music. Cambridge ; New York: Cambridge University Press.

Nicole Hemsoth. 2015. Inside The Programming Evolution of GPU Computing.

Pasquinelli, M. 2021. How to Make a Class. *Qui Parle* 30(1):159.

Pasquinelli, M. 2023. *The Eye of the Master: A Social History of Artificial Intelligence*. Verso Books.

Rosenblatt, F. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6):386–408.

Xenakis, I. 1966. The origins of stochastic music 1. *Tempo* (78):9–12.

Yuk Hui. 2023. Rethinking technodiversity.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Integrating Computational Creativity and Climate Data for Environmental Awareness: Design and Analysis of an Ongoing Project

Felipe Mammoli, Guilherme Zanchetta, Laureana Fontolan and Artemis Moroni
Division of Cyber-Physical Systems
CTI Renato Archer
Campinas, SP, Dom Pedro I Highway (SP-65), Km 143,6 13069-901 Brazil
fmammoli@gmail.com, guilhermezanchettac@gmail.com, laureana.fontolan@cti.gov.br, artemis.moroni@gmail.com

Abstract

Climate change is often understood through scientific data, but how can digital technologies be used to make climate change more accessible to non-expert publics? This article presents a project that explores the intersection of computational creativity and environmental awareness through a web platform that transforms real-time climate data into interactive audiovisual art. By collecting live climate data from satellites and processing it into unique visual and musical compositions, the platform translates global environmental phenomena into local, personal aesthetic experiences. Using generative algorithms, data-driven creativity and interactive design, the platform aims to bridge the vast scale of climate change with the intimate scale of individual perception. Based on observation of users engaging with the platform and qualitative feedback, we analyse the potential of digital art for environmental education, the challenges of representing environmental data in non-figurative ways, and the technical challenges of adapting scientific data infrastructure for artistic exploration. We conclude by suggesting that art can play an important role in climate communication and that computational creativity can significantly contribute to this by connecting objective data with subjective representations.

Introduction

Throughout history, humans have observed and recorded weather to predict its outcomes. Long before the advent of electronic technologies, artists drew inspiration from their observations of the weather to create works of art. Leonardo da Vinci, using black chalk on paper, created the drawing A Deluge (1517-1518), illustrated in Figure 1, which depicts the dynamic flow of clouds and water (Randerson et al., 2015; Randerson, 2018).

In recent years, climate change has become an increasingly significant issue, prompting the integration of scientific data into artistic practices to create a bridge between science, art, and public awareness. While art has long used the climate as a source of inspiration, today's urgent climate issues demand new approaches that move beyond inspiration and into active engagement. Artists have begun using digital and interactive tools to transform complex climate data into immersive and engaging experiences, aiming to raise awareness and inspire action.

Many recent new media artworks merge science and engineering elements with artistic and creative practices. In the sonic domain, the sonification of climate data has been used to transform data into sound and music, illustrating the impact of climate change (Quinn, 2001; Perera, 2015; Chafe, 2018; Barth et al., 2021). In (Chopra, Hessels, and Black, 2022), various artistic approaches are presented that combine science, engineering, and creativity to reflect on climate change and its implications. Many of these works are participatory, prompting the viewer to become an active agent, someone who experiences our ever-evolving society rather than remaining a passive observer. And art approach is now important part of scientific and political discussions about climate change, as the inclusion of the Art x Climate gallery into the USA's 5th National Climate Assessment makes clear[1].

Figure 1 - Bequeathed to Francesco Melzi; from whose heirs purchased by Pompeo Leoni, c.1582-90; Thomas Howard, 2nd Earl of Arundel, by 1630; Probably acquired by Charles II; Royal Collection by 1690.

[1] https://nca2023.globalchange.gov/art-climate/e

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Academics, too, are increasingly focused on acting in response to climate issues. Recognizing their responsibility to act, they not only lead by example by reducing their carbon footprints but also leverage their creative thinking to contribute to climate solutions, raise awareness of the devastating impacts, and help catalyse cultural change (Pease and Pease, 2023; Utz and DiPaola, 2023; Jääskeläinen and Biørn-Hansen, 2024)

Computers and computational systems have been fundamental in these early explorations. Whether using computer systems to create new media art that explores climate change or employing scientific data about the environment as material for artwork, artists addressing climate change have embraced computers as interactive and inspirational tools. Through various engagements with these digital materials, the abstract, statistical, and global language of climate change—such as "global mean temperature," "carbon budget," and "planetary tipping points"—is transformed into a language of emotional engagement that aligns with the spatial and temporal scale of human perceptual capacities (Andrade and Moroni, 2023). As an artistic movement, climate art is growing. However, in the field of computational creativity, climate change remains a rarely addressed topic. Recent work, such as the CreaVisions platform (Rafner et al., 2023), demonstrates the emergence of a growing community of practice cantered on computational co-creativity for sustainability.

Nevertheless, computational creativity undoubtedly holds the potential to contribute significantly and play a pivotal role. Data visualizations driven by computational creativity, along with decision-making, scenario planning, problem-solving, and the enhancement of both human and scientific creativity, can all play a significant role (Pease and Pease, 2023).

In this paper, we present and discuss the GaiaSenses project, an art-science initiative aimed at making climate change more tangible for non-expert audiences. GaiaSenses is a web platform that presents users with audiovisual compositions based on their geolocation and local weather conditions. These audiovisual compositions are driven by weather and climate data, such as temperature and humidity, to create a real-time experience.

The audiovisual compositions can be shared on social media. The goal is to evoke a sense of empathy for the planet, starting with attention and care for the local area, potentially triggering actions. As Haraway (2016) argues, the way environmental changes are communicated must be reconsidered, and projects like GaiaSenses aim to address this transformation through ideas, projects, actions, and attitudes. We cannot change the climate directly, but we can influence people's attitudes—particularly their awareness, emotional engagement, and sense of agency. These are fostered throughout the project's lifecycle: from design and development to student engagement and experiential learning with climate phenomena, through reflective practices on representation, and into public dissemination and user interaction. The goal is to cultivate a deeper personal and collective response to environmental challenges.

This article is composed of the following sections. The first section presents a technical description of the platform

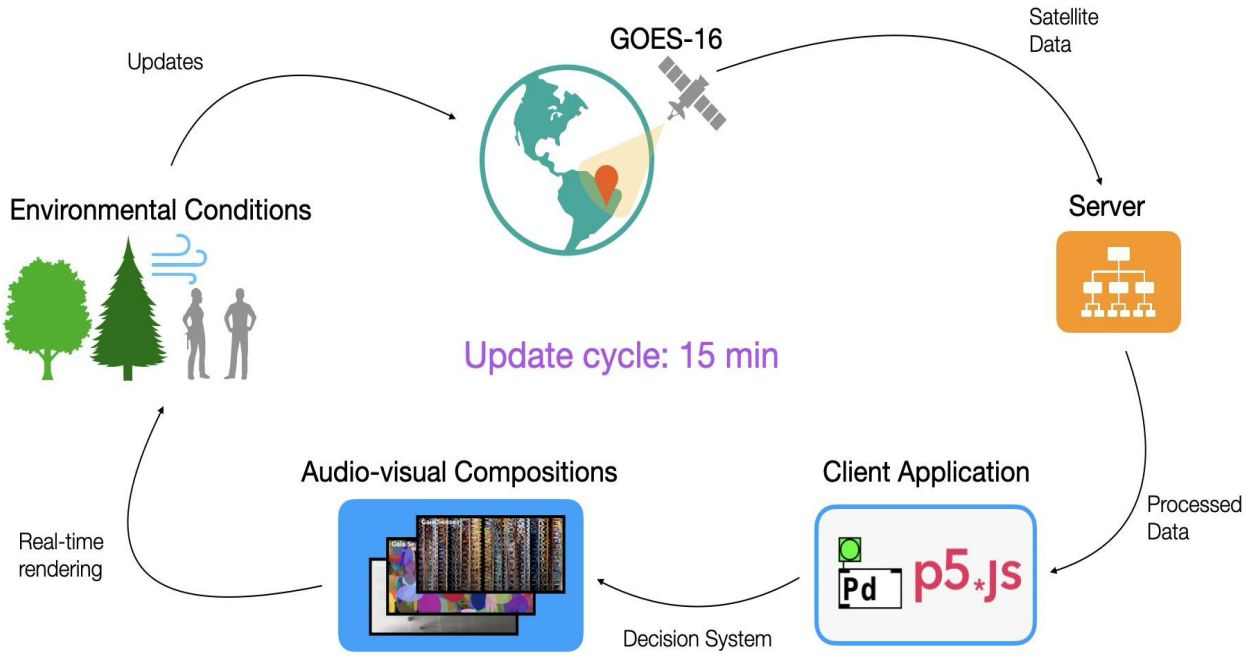


Figure 2 - The GaiaSenses Platform Architecture.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

focusing on how local weather data is acquired, processed, transformed into animated and sonified compositions, and presented to the user. The second section describes what users see and hear when interacting with the platform, the audiovisual compositions, and discusses their aesthetic aspects. Following, the Visual Compositions and the Multiparadigm Sound Generation System are described. Then, we present a discussion regarding initial public response about the project and its intended evaluation. And finally, a conclusion.

The GaiaSenses Project

Many of the main topics regarding climate change are data that makes little sense for the human sensorium scale: global mean temperature and atmospheric CO_2 concentration for example, are phenomenon that are usually invisible and not perceptible in the local scale but are crucial to the understanding of climate change. The GaiaSenses project bridges this gap by combining scientific knowledge and art-based methods to create non-figurative audiovisual compositions that are powered by patterns present in the data. It is an attempt to make these patterns that are usually meaningful only for climate scientists, able to be experienced aesthetically by non-expert audiences. Practically, this is done by transforming weather and climate data into audiovisual compositions.

For that, the project deploys a specific pipeline responsible for collecting, processing, transforming and delivering data to its users, as shown in Figure 2. It broadly works as: 1) as the users access the GaiaSenses Web Platform, the geolocation data (latitude and longitude) from the accessed device is collected and sent to our servers; 2) The geolocation data is then used as input to get weather data from the GOES 16 and 19 Satellites data archive stored in an Amazon S3 Server or another meteorological platform. This data is then processed and cropped to the specific time and area of interest of the user, resulting in a small dataset containing information like temperature, humidity, precipitation, wind and lightning occurrence, that is sent back to the user's device. 3) In the device, this weather data is used as parameters to animate a specific P5.js sketch depending on the data characteristics. At the same time, the weather data is sent to a specific Pure Data patch to be used as input for a generative audio composition. 4) Finally, the animation and the sonification are combined in a single stream and rendered in real time in the user's device, presenting an audiovisual composition based on the weather conditions of his location. 5) The user application requests new weather data from the server every 15 minutes to check if there is new satellite data, as this is the period that the satellite takes to update its data, and if positive, uses this new data to update the composition, following the steps described above. This whole process means that each user will receive a composition that is particular about the environmental conditions they are in.

The GaiaSenses Web Platform is developed using open technologies appropriate for each layer of its technical infrastructure. The GaiaSenses Platform itself is an Open Source, and all its repositories can be found at its Github Organization Page[2]. For the Web portion[3], it uses JavaScript and the react.js framework to handle basic user interactions. The compositions are developed using the P5.JS library for creative programming and the audio is interfaced with Pure Data through the WebPd project. The API responsible for acquiring and processing satellite data[4] is developed in Python and uses Numpy for operations. The API is deployed via a Docker Container on a serverless infrastructure on AWS Lambda. The Open distribution of the project's code intends to both create a community around digital and data artists interested in working with weather and satellite resources and to provide practical examples on how to do it.

The data products that GaiaSenses currently uses are the usual data products derived from weather systems: temperature, precipitation, wind (speed and direction), humidity, radiation, cloud coverage, lightning (location of occurrence and count), fire (location of occurrence and count). These data come from different data sources, depending on data availability for the region where the user is accessing the web platform. Basic weather data (temperature, humidity, cloud coverage and win) is acquired by the OpenWeather API, a service that provides near real-time weather data based on a given geolocated point. So, for weather data, the user's geolocation is used as input. Despite the ease of access that the OpenWeather API provides, we insist on accessing more primary data sources, like satellites, so that the compositions can work not only with the values for weather conditions, but also with the errors, uncertainties, and inaccuracies that remote sensing systems are subjected to. We believe that access to satellite data can enrich the climate representations of GaiaSenses, providing a deeper and more comprehensive understanding of climate change, aiming not only to represent data but also to reflect its uncertainties and limitations.

For lightning data (place of occurrence and count), we access data directly from the National Oceanic and Atmospheric Administration (NOAA) Geostationary Operational Environmental Satellite 16 (GOES-16) satellite data archive hosted on an Amazon AWS S3 Bucket. In the S3 Bucket, the satellite data is organized in netCDF4 file with filenames that express a specific pattern that contains the name of the sensor used for capturing the data, the time period that the data was collected, and from where the data is.

The lightning sensor (GLM-L2-LCFA) at GOES-16 has a spatial resolution of $4km^2$, which means that the data is not available by geolocated points, but by a grid of pixels where each element represents a $4km^2$ area of the planet. To get the data, first it is necessary to find the grid element that contains the point of interest. As lightning usually happens in a large area, and not in a single point, the input for this data is a 20 km round wide buffer with the user's location as its center. At our servers this area is transformed into the correspondent grid elements and all lightning data is returned to the user's device.

GOES-16 is a geostationary satellite that covers most of America, which means that, currently, we are only able to

[2] https://github.com/GaiaSenses
[3] https://github.com/fmammoli/Gaiasenses-web
[4] https://github.com/GaiaSenses/satellite-fetcher-aws

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

collect lighting data for users accessing the platform from within this region. It is our plan to expand our data access capabilities by developing additional integrations with other satellite systems. Fire data is acquired from the Brazilian platform BDQueimadas through its Geoserver.

BDQueimadas is the main platform for fire monitoring in South America and is maintained by the Brazilian National Institute for Space Research (INPE). Much like acquiring data for lightning, we capture fires that are happening in an area around where the user is located, and not only if the user's location is on fire. For that, we create a 10km buffer with the user's location at the center and use this area as the input for the BDQueimadas API. Data is acquired as a list of fires happening in the area, their location, their intensity and a total count. The data on BDQueimadas is available in near real-time, being updated every 10 minutes.

Visual Compositions

The visual compositions in the GaiaSenses Platform are produced using the P5.JS library; a JavaScript library for creative programming that focuses on accessibility and education. The library was chosen due to its capability in portraying complex graphics, its wide-spread use and for being a web-based library, which allows the project to run directly in the Browser, with no need for special environments.

The main approach in the project for creating compositions is to collect already existing P5.JS sketches at the OpenProcessing Platform and use the weather data to modify its intended behaviour. The process of modifying the behaviour of a sketch entails the study of its inner workings and the extractions of parameters from the code that control behaviour such as color, rendering speed, placement of lines, etc. Once extracted, these parameters are turned into variables that are connected to the weather data through a proportion function. The result is that the change in a weather condition, such as temperature, can control the behaviour of a sketch, such as color.

An example is the GenerativeStrings composition, which displays a box with bouncing particles inside and 7 strings tuned to the G major scale. The basic mechanism of the composition is that a sound is produced when particles collide with the string. During our modification process, we connected the agitation of the particles to the temperature value, such as an increase in temperature makes the particles move more rapidly, producing more collisions and, consequently, producing unexpected melodies that are qualitatively different from the melodies produced when they are less agitated.

This approach of modifying existing sketches mirrors the way in which climate change not only creates new environmental conditions but creates them by profoundly modifying environmental characteristics that we once took for granted. A forest might turn into a savannah depending on the increase in temperature and a lake can dry out depending on the change in the precipitation regime. The use of existing sketches, thus, is an aesthetic choice to remind us that when accounting for the environment, creation is co-creation.

GaiaSenses Multiparadigm Sound Generation System

The sound generation system of the GaiaSenses platform operates as a multiparadigm environment, enabling the conversion of environmental data collected by remote sensing platforms into multiple sonic interpretations.

Development Environment Choice

In the initial phase of the project, Pure Data (Pd) was selected as the primary tool for sound development (Puckette, 1997). Pd, a visual programming language for sound creation and interactive multimedia, was chosen for its flexibility and accessibility, facilitating collaboration with a broader network of composers and sound artists that are geographically near to us.

Practical Examples of This Approach

ShockWind: A patch that artistically simulates the interior of a thunderstorm cloud. The composition employs subpatches and dynamic filters that manipulate noise to create sonic textures. Thunder sounds are generated through modulated low-frequency signals, while ambient noise simulates the internal movement of clouds. External reverberation enhances spatial depth.

Drops: A patch that generates sounds of droplets falling on a liquid surface. Rainfall data directly influence the frequency and intensity of the generated droplets, forming an auditory landscape that ranges from light drizzles to intense storms.

Zigzag and Zigzag 2: These compositions utilize dynamic patching techniques to represent visual movements and satellite data through fluid and responsive sounds. Zigzag employs the [clone][5] to create multiple instances of a patch simulating zigzag movements, with timbres modulated by LFOs[6]. Zigzag 2 explores phase modulation and uses continuous sonic flows that evolve in response to environmental data)

Each of these patches exemplifies different strategies for sound mapping, ranging from free artistic interpretations to more direct sonifications of environmental data (D´Incao, 2023).

Explorations Beyond Pure Data

Beyond Pure Data, other systems have been explored for sound generation within browser-based environments. A relevant example is an implementation based on a tutorial for Brian Eno's Music for Airports, adapted in Tone.js[7]. This adaptation served as an initial test for integrating compositional structures within the system. The layered nature of Music for Airports, which does not rely on rapid

[5] https://pd.iem.sh/objects/clone/

[6] https://forum.pdpatchrepo.info/topic/10577/creating-anlfo

[7] https://www.youtube.com/watch?v=36ZggCL4pBo&ab_channel=danbuzzo

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

rhythmic changes or time-critical processing, made it a suitable model for evaluating the system's capabilities.

Technical and Artistic Characteristics

The patches in GaiaSenses employ various sound synthesis and signal processing techniques. The use of Pure Data allows for the creation of generative and interactive sonic environments, where environmental data directly influence parameters such as frequency, amplitude, and spatialization. Artistically, the project seeks to balance descriptive sonic representations of environmental phenomena with more abstract and interpretative compositions, constructing an immersive auditory experience that fosters a perceptual connection between the listener and the natural environment.

Sonic Horizon

The development of GaiaSenses is ongoing, with plans to integrate additional sound generation systems such as Max-RNBO, SuperCollider, and other computational music frameworks. Building upon these initial explorations, the project aims to establish a systematic methodology for data representation, serving as a foundation for future generative compositions.

Climate Compositions

When interacting with GaiaSenses Platform, the user is presented with a global map with their location at the center, alongside a window with basic weather information of the place they are in, and a button that prompts the user to click it. When clicked, the map fades out to reveal an animation with an accompanying sound, both powered by the data displayed before. Figure 3 shows the composition named Attractor in two different states, when it starts and after 20 seconds. The animation uses lightning data around the users and maps its occurrences to the screen size, so the user location is in the center and the lightning is happening around them. In the animation canvas, the lightning locations act as attractors, exerting a pull force. Then, a semirandom number of particles are dropped in the canvas, these particles are attracted by lightning attractors and each particle leaves a trail. After some time, particles start to concentrate near the attracts, making the canvas region brighter, revealing the spatial pattern of the lightning occurrence.

The sonification of the Attractor composition plays synthesized thunder sounds through a dynamic panning algorithm implemented in Pure Data so the thunder sounds appear to happen in different directions and distances, mirroring the space distribution of lightning occurrence in the data.

Figure 4 depicts how the same composition can look different depending on different weather conditions. It shows the composition named StormEye, which consists of lines that rotate around a center. In the composition, the line colors are connected to temperature data so cold temperatures make the lines more blue and hot temperatures make them redder, the line scattering are connected to wind direction and wind speed. The sound is implemented in Pure Data and presents the user with synthesized sounds of wind gusts that get stronger as wind speed increases at the user's location. The user can have a feel for the wind change patterns by observing the scattering of the lines and see the temperature by observing the mixture of colours.

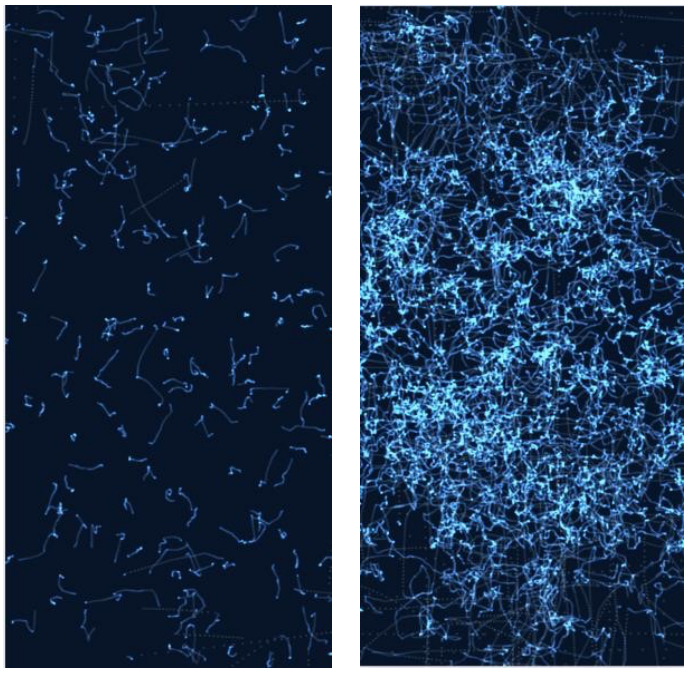


Figure 3 - Different frames of the Attractor composition: left) The composition at its start; right) The composition after 20s of animation.

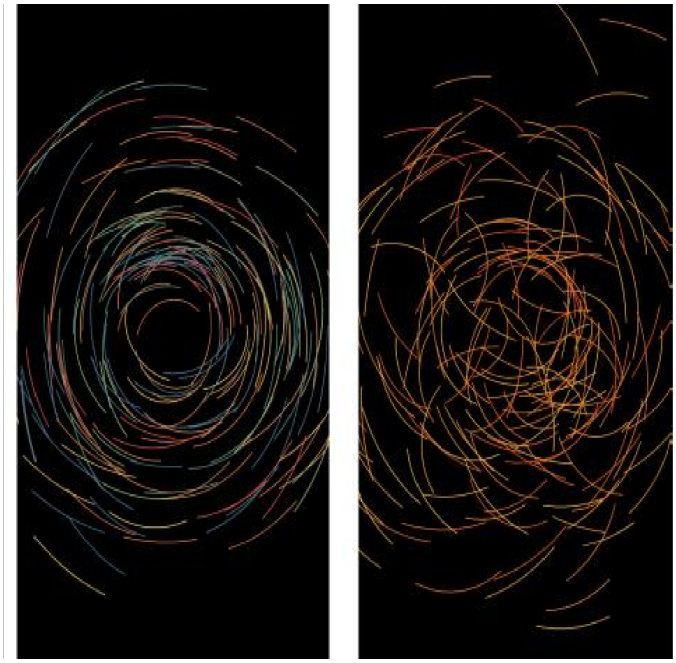


Figure 4 - The StormEye composition in two different weather conditions: left) temperature: 24°C, wind speed: 4m/s, wind direction 20°; right) temperature: 32°C, wind speed: 14.2m/s, wind direction 271°.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

A prototype of the GaiaSenses Platform[8] is accessible online with a selected list of compositions to be viewed and listened to.

Discussion

The GaiaSenses Platform aims to make the changing aspects of our shared environment more present in the daily life of publics that are not climate experts. It does so by presenting patterns inherent in climate data in an artful form - audiovisual compositions - instead of the usual scientific graph. An aesthetic form of presenting that is designed to appeal to the user's subjective sensory and emotional experience. This approach is based on research that shows how art can be relevant in communicating climate change (Rice, Rebich, Hespanha and Zhu, 2019) and that information alone is not enough to provoke action (Sommer and Kløckner, 2021). Political causes depend deeply on emotional and personal connection.

The GaiaSenses Platform was presented in three different settings composed of different types of publics. Two were scientific conferences and one was an undergraduate class. Despite the good general impressions of the users, each group interacted differently with the platform. In the conferences, people were expected to interact with digital art experiments and were open to learning how the system worked. Although the platform was made to showcase weather information about the user's current location, and compositions based on that data, users were continuously interested in seeing how their hometown would be depicted in the system. An unexpected response that depicts how the global scale of climate change can be localized in personal experiences, making climate change more tangible (Rosing and Eliason, 2015). Since then, the GaiaSenses Platform also allows users to explore the weather conditions and see compositions of other parts of the planet, not only their current location.

The third presentation happened in an undergrad class where the GaiaSenses project coordinator was presenting a seminar for the class. From the class interaction, one specific comment was particularly interesting: one of the students commented that he was not seeing climate change on the platform. Although the compositions are mostly non-figurative, they are open to interpretation on how any specific sound or image relates to any natural phenomena, the student's comment shed light on the persistence of already existing forms of representing climate and weather data. These data are usually presented on top of maps, as weather predictions, or as rising lines, like the increase in CO_2 emissions over the decades (Mann, 2012).

Representations that non-experts are so used to that create interpretational challenges when they are not present. To overcome these visual affordances of how weather and climate data are often represented has become a central issue in the GaiaSenses project and a source of aesthetic inquiry, confirming that climate change is not only a matter of nature, but also a matter of aesthetics, as suggested by Davis and Turpin (2015).

Evaluation Plan

As the project is still ongoing, it has not gone through a structured evaluation process against a diverse public. As it gains conceptual maturity and technical stability, we plan to conduct an evaluation based on Klöckner and Sommer (2021) proposal of evaluating public response artworks inspired by climate change and Hahn and Berkers (2024) evaluation of artists' intended effect through climate change art. The purpose of the evaluation is to further research how specific aesthetic choices affect the public perception of climate change. As evaluation points, we have three main areas: evaluating interaction (ease of use etc, technical); evaluating aesthetics (does the compositions make people think of any natural phenomena, what emotions it conveys?); and the social aspect.

The interaction evaluation comprises a technical evaluation to understand the ease of use, how intuitive it is to navigate the system, how accessible it is and how is the overall user experience. The aesthetic evaluation consists of a more subjective evaluation and its purpose is to trace how people associate specific graphic and sound design to particular natural phenomena, how this relation is supported by personal history and if it meets their expectation on how climate change should be represented. The social aspect involved evaluating what socio-environmental debates the platform raises and at what scale does it inspire reflexivity among its users and the researchers working on the platform.

To address these points, we plan on a twofold approach: 1) an automated evaluation based on analytics gathered from each access to the web application. This will mostly inform the interactive evaluation and part of the aesthetic evaluation. It is important to note that privacy matters will still be discussed to understand to what extent we should keep the user's data, since we are accessing their location; 2) a qualitative evaluation based on questionnaires with users in controlled spaces and focal groups with users that interacted with the platform in art exhibition contexts. This qualitative approach will provide important data for the three evaluation points (interaction, aesthetic and social) so we can understand the effectiveness of the platform in promoting dialogue about environmental issues and climate change.

GaiaSenses and the Generative IAs

More recently, the field of generative computing has gained significant momentum, driven by advances in hardware, software, and artificial intelligence. This possibility is being considered and evaluated. Incorporating an intelligent layer for the automatic production of creative outputs in GaiaSenses can effectively optimize and expand the process, providing real time personalization. This can be achieved through artificial intelligence (AI), generative algorithms, and machine learning techniques that automatically map climate data to visual and sound representations.

In the current version, since each audiovisual composition is created with a clear artistic intention, the nuances and intentions of the artist are more effectively conveyed. On the other hand, automating the process allows the system to

[8] https://gaiasenses-web.vercel.app

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

handle large volumes of data, generating outputs in a scalable, real-time, and much faster manner. However, this introduces a new challenge in creative control, as artistic decisions are largely delegated to the machine. New questions arise: To what extent should we detach ourselves from the creative process? How much of the creative process should we entrust to the machine? Another important consideration when addressing climate change and sustainability is the computational cost of generative AI art systems, both in industry and academia. This includes not only the monetary costs associated with training but also the environmental impact during inference (Utz and DiPaola, 2023).

Next Steps

At this stage, we have begun presenting GaiaSenses to various audiences and examining their reactions. Analytical indices are being developed to track its dissemination. An aesthetic framework is also being created. The parameters used in generating the compositions will be stored in a database, which can be used in various ways for aesthetic evaluation.

Here, we propose the possibility of developing a grammar through the definition of parameters and vocabulary for generating compositions or representations. It is desirable for each phenomenon to have more than one representation to provide variety, potentially leading to new algorithms. The combination of climate data, artificial intelligence, and interactivity creates a new aesthetic that challenges traditional artistic conventions. It is important to note that the field of Aesthetics is highly complex (Fishwick, 2008), and this is a statistical study aimed at providing insights for evaluating the algorithms used. Additionally, the aesthetic database may serve as the foundation for predicting future accesses and compositions based on machine learning and artificial intelligence methods, potentially leading to a significant contribution to the field (Franceschelli & Musolesi, 2021; Toivonen & Gross, 2015).

Conclusion

Art plays an essential role in amplifying and complementing the communication of climate change science. The magnitude of the challenges we face no longer fits within the scope of everyday human experience. Rather than grappling with a lack of information, we are now tasked with expressing the hopes, conflicts, fears, and contradictions arising from a crisis that often seems too vast to fully represent. By aligning with the sciences through shared methodologies, tools, localized community platforms, and global digital networks, art focused on climate change can help build coalitions for collective action. Innovative thinking is crucial in tackling climate change, and fields as diverse and interdisciplinary as computational creativity hold great promise for making meaningful contributions.

GaiaSenses operates as a creativity laboratory that delves into the potential of generative algorithms to represent and communicate climate phenomena. By harnessing these challenging visual and sound algorithms, the project transforms complex climate data into dynamic audiovisual representations, allowing for a deeper understanding of the environmental challenges we face. The process not only aims to inform but also to engage and inspire action, using art and science as tools for raising awareness. Through their involvement in the project, students and researchers are not only contributing to the development of these creative outputs but are also becoming active advocates for the climate cause. Their engagement goes beyond academic inquiry, as they embrace the role of climate activists. By combining scientific knowledge with artistic expression, they help bridge the gap between technical data and emotional, human-cantered experiences, by connecting objective data with subjective representations.

Acknowledgements

We would like to thank the researcher Renata Gonçalves, from CEPAGRI/UNICAMP, for her valuable guidance on the access and interpretation of meteorological data. This research wouldn't be possible without the continuous work of the dedicated undergraduate students that contributed to the project over the years. We would especially like to thank Gabriel D'Incao,Lucas Mielle, Henrique Alves, Lucas Lima, Pedro Pereira and Ru Yi Shen. We thank the PIBIC CNPq program. Felipe Mammoli holds a PCI CNPq fellowship grant number 300790/2024-4 and 301195/2024-2. Laureana Fontolan holds a PCI CNPq fellowship grant number 314134/2025-5. Guilherme Zanchetta holds a PCI CNPq fellowship grant number 314093/2025-7.

References

Andrade, F., & Moroni, A. (2023, November). Bridging Computational Art and Climate Change: An Analysis of a Mobile Application for Raising Awareness About Climate Change Through Art. In International Conference on ArtsIT, Interactivity and Game Creation (pp. 149-165). Cham: Springer Nature Switzerland.

Barth, A., Karlstrom, L., Holtzman, B. K., Paté, A., Nayak, A. 2021. Sonification and Animation of Multivariate Data to Illuminate Dynamics of Geyser Eruptions. Computer Music Journal, 44:1, pp. 35–50, Spring 2020, Massachusetts Institute of Technology.

Chafe, C. 2018. 1.200 Years of Earth's Climate. https://www.kqed.org/science/1918660/listen-1200-yearsof-earths-climate-transformed-into-sound

Chopra, S., Hessels, S., & Black, K. C. 2022. Artistic Interpretation of Scientific Data on Extreme Weather to Raise Public Awareness for Urban Resilience. In *Data Art for Climate Action (DACA 2022): a dual-hub conference on interactive sonification and visualisation for climate science communication*.

Davis, H., & Turpin, E. 2015. Art in the Anthropocene: Encounters among aesthetics, politics, environments and epistemologies (p. 416). Open Humanities Press.

D´Incao, G. C. E. O., & Moroni, A. (2023). *Composição musical com dados de satélite: Desafios ao 'musicar' números*. In JICC2023 – Jornada de Iniciação Científica do CTI Renato Archer (Campinas, Brasil).

Fishwick, P.A. (Ed.), 2008. *Aesthetic Computing*. MIT Press.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

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

Haraway, D. 2016. Antropoceno, capitaloceno, plantationoceno, chthuluceno: fazendo parentes. *ClimaCom Cultura Científica*, 3(5), 139-146.

Hahn, U., & Berkers, P. (2024). Artists on Climate Change: Their Intended Impact and Audiences. Empirical Studies of the Arts, 43(1), 188-208. https://doi.org/10.1177/02762374241258632

Ikeshiro, R., Simon, L., & Gui, R. 2022. DAT/ACT Data Art for Climate Action Gallery. In Data Art for Climate Action (DACA 2022): a dual-hub conference on interactive sonification and visualisation for climate science communication.

Jääskeläinen, P., & Biørn-Hansen, A. 2024. Critical Questions for Sustainability Research in Computational Creativity. In ICCC'24 will take place June 17–June 21, 2024, at the University of Jönköping, in Sweden.

Klöckner CA, Sommer LK (2021) Visual art inspired by climate change—An analysis of audience reactions to 37 artworks presented during 21st UN climate summit in Paris. PLoS ONE 16(2): e0247331. https://doi.org/10.1371/journal.pone.0247331

Mann, M. E. 2012. The hockey stick and the climate wars: Dispatches from the front lines. Columbia University Press.

Minik Rosing and Olafur Eliasson, "Ice, Art, and Being Human". 2015. Studio Olafur Eliasson.

Pease, A., & Pease, A. 2023. Computational creativity and the climate crisis. In *14th International Conference on Computational Creativity* (pp. 293-297). Association for Computational Creativity.

Pereira, P. T. F. (2023). *Aprimoramento da infraestrutura do aplicativo GaiaSenses para acesso às bases planetárias* [Trabalho de iniciação científica, Centro de Tecnologia da Informação Renato Archer]. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Perera, J. 2015. Climate Symphony. https://www.jamieperera.com/climatesymphony#ClimateSymphonyLabs.

Puckette, M. 1997. Pure Data: another integrated computer music environment. Proceedings, Second Intercollege Computer Music Concerts, Tachikawa, p. 37-41.

Quinn, M. 2001. Research set to music: The climate symphony and other sonifications of ice core, radar, DNA, seismic and solar wind data. In *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, July 29-August 1.

Rafner, J., Zana, B., Beolet, T., Buyukguzel, S., Michel, E., Maiden, N., Risi, S., & Sherson, J. (2023). Crea.visions: A platform for casual co-creation with a purpose – Envisioning the future through human-AI collaboration with multiple stakeholders. In Proceedings of the 14th International Conference on Computational Creativity (ICCC '23) (pp. 474–481).

Randerson, J. (2018). Weather as medium: Toward a meteorological art. MIT Press.

Randerson, J., Salmond, J., & Manford, C. 2015. Weather as medium: Art and meteorological science. Leonardo, 48(1), 16-24.

Rice, R.E., Rebich-Hespanha, S. and Zhu, H. 2019, "Communicating about Climate Change Through Art and Science", Pinto, J., Gutsche, R.E. and Prado, P. (Ed.) Climate Change, Media & Culture: Critical Issues in Global Environmental Communication, Emerald Publishing Limited, Leeds, pp. 129-154.

Sommer, L. K., and Kløckner, C. A. 2021. Does activist art have the capacity to raise awareness in audiences?—a study on climate change art at the art COP21 event in paris. Psychology of Aesthetics, Creativity, and the Arts 15(1):60–75.

Toivonen, H., & Gross, O. (2015). Data mining and machine learning in computational creativity. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(6), 265-275.

Utz, V., & 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*.

ISBN: 978-989-54160-7-3

Psychologically-inspired generative AI videos for supporting creativity

Shabnam Hakimi, Monica Van, Matthew K. Hong, Kalani Murakami, Pablo E. Paredes, and **Matthew Klenk**
Human-Centered AI, Toyota Research Institute, Los Altos, CA USA
{shabnam.hakimi, monica.van, matthew.hong, kalani.murakami, pablo.paredes, matthew.klenk}@tri.global

Abstract

Generative AI has become a powerful tool for supporting creativity across a breadth of disciplines. Nonetheless, applications are generally domain- and/or task-specific, providing targeted support for creative tasks like design ideation. Here, we present a domain-general approach to using generative AI to amplify human creativity by targeting its underlying psychological processes at high temporal resolution. We describe a generative AI workflow for creating two-minute video interventions designed to increase open-minded, flexible thinking, a central domain-general component of creative processes. We also present human participants' ratings of and reactions to these videos, detailing implications for future work.

Introduction

> If the mind creates a world that is limited, where we think we don't have enough worthwhile ideas or material, we will not see the inspiration the universe is providing.
>
> – Rick Rubin, *The Creative Act: A Way of Being* (p. 202)

The mind is the filter through which humans make sense of and shape the world. The success of a human's creative process, whether measured by the quality of the experience or the output, depends on the mind's capability to represent ideas in new, surprising, and valuable ways (Boden 2004). Across the many theoretical accounts of creativity, including Boden's (2004) influential work, a common set of psychological substrates scaffold creative processes, with **psychological flexibility** emerging as the critical connective tissue that enables what we consider *creativity* (Benedek et al. 2014). Psychological flexibility describes a set of dynamic processes that support openness and adaptation to changing situational demands; shifting of mindsets or behavioral repertoires; and balancing competing desires and needs, while considering how they relate to values (Kashdan and Rottenberg 2010). Further, psychological flexibility supports the executive functions considered necessary for creativity (Nijstad et al. 2010; Kashdan and Rottenberg 2010; Hayes et al. 2020; Zabelina, Friedman, and Andrews-Hanna 2019) by modulating an array of self-regulatory processes that enable higher-order thought and behavior(Zabelina and Robinson 2010).

Psychological flexibility's centrality in the creative process highlights its potential as a target for intervention. Indeed, prior research has leveraged a variety of psychological interventions for enhancing psychological flexibility, with varying degrees of success (e.g., (Kashdan and Rottenberg 2010; Uddin 2021)). Both between- and within- individual variability at the trait and state levels may be one cause for mixed results, potentially limiting the efficacy of existing approaches. Further, prior psychological creativity support has been largely domain- and task- specific, with interventions focused on narrow tasks such as increasing flexibility when generating semantic concepts during ideation (e.g., in a Remote Associates Test (Wu et al. 2020)). Advances in artificial intelligence (AI), particularly in generative AI models, suggest new opportunities for intervening on psychological flexibility to amplify human creativity because of their immense capacity for personalization and generating variability at scale.

Can we leverage both psychological principles and generative AI to aid the mind in creating Rubin's internal world with fewer limitations? Here, we present a human-centered method for automated generation of video interventions for creativity designed to increase psychological flexibility across domains. We describe the system and its approach to targeting psychological flexibility at the fine temporal resolution of the process level with reduced dependence on the specific task or domain. We also describe the results of a study investigating human viewers' affective responses to the generative AI videos. Then, we discuss the implications of the system for impacting creative processes, reviewing opportunities for further research.

Background

In psychology, creativity is often conceptualized as a durable personality characteristic, or trait, that correlates strongly with intelligence and openness (Zhang, Wang, and Zhang 2020; Benedek et al. 2014; McCrae 1987). Creative individuals also demonstrate greater psychological flexibility (Kenett et al. 2018; Smith and Ward 2012). At the trait level, creativity and *creative mindset*, which reflects metacognitive knowledge about creativity and beliefs about its malleability with time and effort, are associated with higher

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

creative performance and creative self-efficacy(Royston and Reiter-Palmon 2019; O'Connor, Nemeth, and Akutsu 2013; Karwowski 2014). Despite data suggesting that some traits, including openness, may be malleable over long periods of time (Roberts et al. 2017; Dolbier, Dieffenbach, and Lieberman 2025), focusing on traits not only reinforces the potentially detrimental belief that creativity is fixed (Karwowski 2014), but may also lead to missed opportunities for targeted support in the immediate to short-term.

Creative State Interventions

Creativity can also be considered a transient state, although this construction does not imply a specific temporal resolution. Recent work suggests that the creative mindset may be affected in a more state-like manner through a perspective-taking instructional prime, leading to increased originality of ideas (Chou and Tversky 2020). Moving to an even shorter timescale (minutes to hours), prior work has shown that expectancy violations increase creative cognition by increasing flexibility (Ritter et al. 2012), with positive affect mediating this relationship (Rowe, Hirsh, and Anderson 2007; Dolbier, Dieffenbach, and Lieberman 2025). Related work has also demonstrated the importance of affect for state creativity interventions, showing that awe, which is characterized in part by complex affective experience, can enhance creative cognition and everyday creativity (Chirico et al. 2018; Zhang et al. 2024). Further evidence suggests that both pharmacological and even digital psychedelic experiences, which have long been considered powerful modulators of creative flexibility, also operate at a similar timescale (Gandy et al. 2022; Greco et al. 2025). Overall, such work complements a history of curricular creativity training interventions targeting not only flexibility, but also fluency and originality over a period of weeks to months (Bi et al. 2020). Interventions designed to impact even shorter timescales (seconds to minutes) have also shown some success in enhancing performance in creative contexts (Albulescu et al. 2022). Despite some evidence of efficacy, these approaches lack specificity both at the target level (i.e., the process being affected is unclear) and at the outcome level (i.e., the relationship between creativity and 'performance' is unclear).

Despite limited evidence for the effectiveness of creative state interventions at shorter timescales, there are opportunities for impact at a resolution closer to that of creative cognition and affect (milliseconds to seconds). The temporal dynamics of expectancy violations and their connection to flexibility and open-mindedness provide a useful lens through which to interrogate the malleability of creative processes at a finer temporal resolution. Kumar and colleagues (2023) examined how Bayesian surprise, a measure of how new observations affect the difference between prior and posterior beliefs, impacted memory encoding of text narratives. By computing a measure of the degree to which each advancing word shifted the predictive distribution of a large language model (LLM), they were able to explain how transient prediction errors affected the evolving mental representation of the narratives, particularly how the contents were segmented into discrete events (Kumar et al. 2023; Baldi and Itti 2010). Their results support dynamic accounts of perception, learning, and memory that suggest that people maintain a multimodal working model of an event that is iteratively checked against new information; the model is stable and is only updated when prediction errors increase, leading to event boundaries (Zacks, Speer, and Reynolds 2009). This view is in line with predictive processing accounts (Clark 2013; Friston et al. 2021) as well as Wiggins's Information Dynamics of Thinking model (Wiggins 2020). Together, these converging lines of evidence suggest that cognitive states may be 1) operationalized using information-theoretic or energetic measures and 2) can be targeted on a moment-to-moment basis to support complex states like psychological flexibility and creativity. Analysis at this level allows for interventional approaches that are less domain- and/or task-specific, focusing instead on the dynamics of the cognitive system.

Creativity Support and Generative AI

Although they are not typically framed as interventions, systems and tools designed to support human creativity likely affect creativity at a state level, since they aim to bolster creative processes *in situ*. Historically, such support happened outside the digital realm, with individuals employing tools for creative cognition (e.g., analogical thinking, mind-mapping) or shifting context to inspire and clear the mind (e.g., taking a walk, visiting a museum). As creative activities have become more digital, creativity support tools have followed suit, growing dramatically in number while also becoming more fully integrated in creative processes (Frich et al. 2019). Many of these approaches are grounded in the field of Human-Computer Interaction. Across a wide array of creative processes and domains, such tools are more likely to be useful when well-aligned with underlying psychological processes (Amitani and Hori 2002). A growing body of work has further demonstrated how psychological principles may be used to guide the development of creative systems (e.g., (DiPaola, Gabora, and McCaig 2018; Tan, Antony, and Kong 2020; Thölke et al. 2024)).

Advances in generative AI have shifted the landscape of creativity support. While much of the discourse focuses on critical topics such as characterizing machine creativity (e.g., (Peeperkorn, Brown, and Jordanous 2023)) and articulating roles in co-creation (e.g., (Chung, He, and Adar 2021)), generative AI's capabilities for amplifying human creative processes should also be considered. Current research has maintained focus on how generative AI may support the psychological factors needed to complete specific creative tasks within a particular domain (e.g., visual design, creative writing). Artists' adoption of text-to-image generative AI was associated with higher productivity, novelty, and community value (Zhou and Lee 2024). A similar effect was reported using text-to-text generation for story composition, showing that using generative AI for ideation can increase individual creativity, although the overall novelty of stories is reduced (Doshi and Hauser 2024). Interestingly, effects across these and related studies are moderated by prior use of generative AI tools, with gains seen primarily in new users; experienced AI users gained little or no benefit (Fu et al. 2024; Ashkinaze et al. 2024). These findings

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

point to the many nuances associated with the use of generative AI in creativity support, especially the risks of null or negative effects for individuals with greater exposure to and experience with AI.

Considering the large body of work on creativity support with and without generative AI, research has largely focused on ways to support specific domains and/or tasks. As with creative state interventions that emerged from the field of psychology, creativity support tools have not yet explicitly addressed the momentary dynamics of the cognitive and affective processes underlying creativity. Extending on the work of (Kumar et al. 2023), generative AI may provide a new approach for re-aligning interventions with their underlying psychological processes at the appropriate temporal resolution. Further, generative AI agents can act as co-creators of interventions (e.g., (Demszky et al. 2023)), vastly increasing the diversity of interventions while avoiding exposure and expertise effects of direct interaction with the human recipients of creativity interventions.

Creating Psychologically-Inspired Generative AI Video Interventions

We therefore aimed to develop a novel approach to enhancing psychological flexibility that could address gaps evident in currently available creative state interventions and creativity support tools. The system desiderata were as follows:

- Ability to generate varied interventions with minimal instruction so as to retain desirable generative AI capabilities (e.g., novelty)
- User control over generative model representation of principles used to intervene on psychological flexibility and their temporal characteristics (e.g., level of Bayesian surprise over time)
- Capacity to generate a high degree of, or even infinite, variability in the narratives and visual features of the resulting videos in order to facilitate personalization and reduce habituation

The video generation process included two stages: text-to-text narrative expansion and text-to-image video generation. The general approach is shown in Figure 1a; an example pair of videos generated using this approach is shown below in Figure 1b.

Text-to-Text Narrative Expansion

In the first stage of the process, a set of user-generated starting prompts is expanded into narratives, or generative text prompts that will later become videos. We developed a set of initial prompts (`STARTING_PROMPT`) to seed the narratives that would eventually become videos. 13 prompts represented scenes that a viewer might see in daily life:

- *A boot hangs on a telephone line*
- *A door opens into a room*
- *A lamp lights the street below*
- *A mug waits by a coffee machine*
- *A novel sits on a bedside table*
- *A painting hangs on a museum wall*
- *A paperclip sits atop a drafting table*
- *A phone rings in another room*
- *A plane flies across a blue sky*
- *A raindrop falls from the sky*
- *A seedling grows from a sidewalk crack*
- *A speck of dust lands on a drawing pad*
- *A stereo plays music*

Each `STARTING_PROMPT` was passed to GPT-4 (OpenAI 2024) using the syntax below with a temperature of 0.65 and a maximum of 1024 tokens. The goal was to generate sets of prompts that would establish a prior on the context of the narrative and then progressively challenge that belief with each additional narrative step. The difference between the prior from the previous step and the beliefs at the current step is also referred to as Bayesian surprise (Itti and Baldi 2009), which was operationalized using semantic distance (similar to (Kumar et al. 2023)). We expected that more surprise would lead to more updating, and thus a higher likelihood of inducing a psychological state conducive to openness and flexibility; induction of this state would be mediated, in part, by more diverse affect. Further, to generate prompts (and then videos) that would allow us to test how the degree of surprisal impacted the viewer's experience, we instructed the LLM to generate two sets of prompts for each starting prompt, one spanning a relatively *near* and the other a very *far* semantic distance. The semantic distance between each pair of generative text prompts was computed *post hoc* using cosine similarity over the sentence embedding space[1].

The system prompt was designed to provide the LLM sufficient instructions to reliably generate sets of additional prompts that represented a narrative continuation of the `STARTING_PROMPT` given the other parameters:

- `STARTING_PROMPT`
- `PROMPT_REFRESH`, the set of prompts $[p_1, ..., p_n]$ corresponding to frames to be generated, $[f_{0,0}, ..., f_{n,0}]$
- `END_POINT`, the desired video duration (s)
- `NUM_STEPS`, the desired frame rate, x (in frames per second, fps)

As noted, the arguments passed to the LLM correspond to the variables denoted in Figure 1a.

Text-to-Image Video Generation

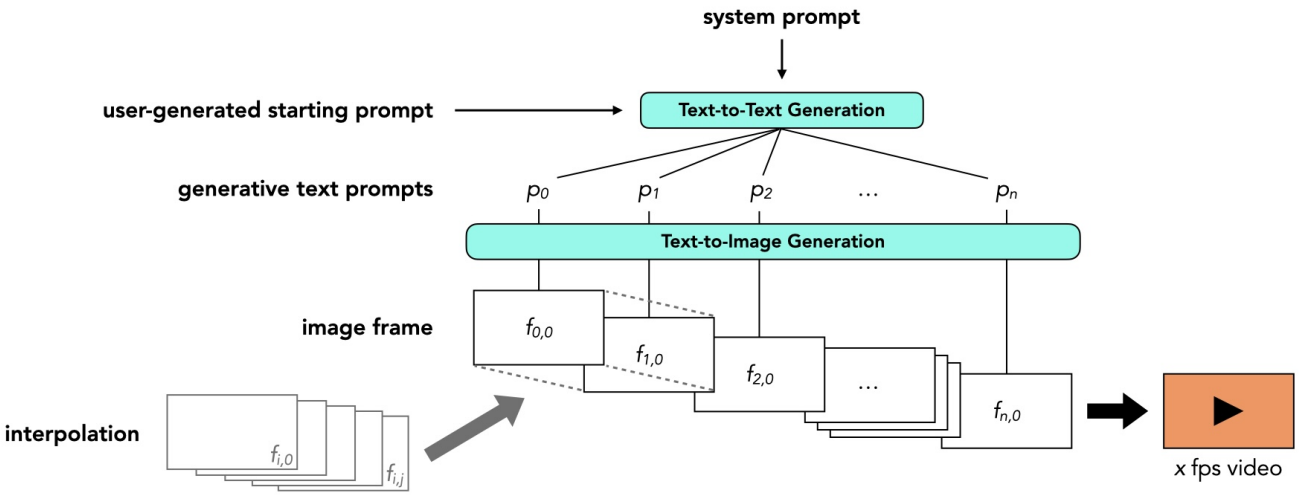
For each starting prompt and its associated sets of generated prompts, we created a two-minute, 12-fps video using Deforum[2], which uses Stable Diffusion v1.5 (Rombach et al. 2022), and a custom version of the AUTOMATIC1111 web interface[3]. Briefly, each of the prompts p_n was used to generate a static image, $f_{n,0}$. In order to create the illusion of movement in time, 50 interpolated images ($[f_{i,0}, ..., f_{i,j}]$) were generated between each pair of frames, $f_{n,0}$ and $f_{n+1,0}$.

[1] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

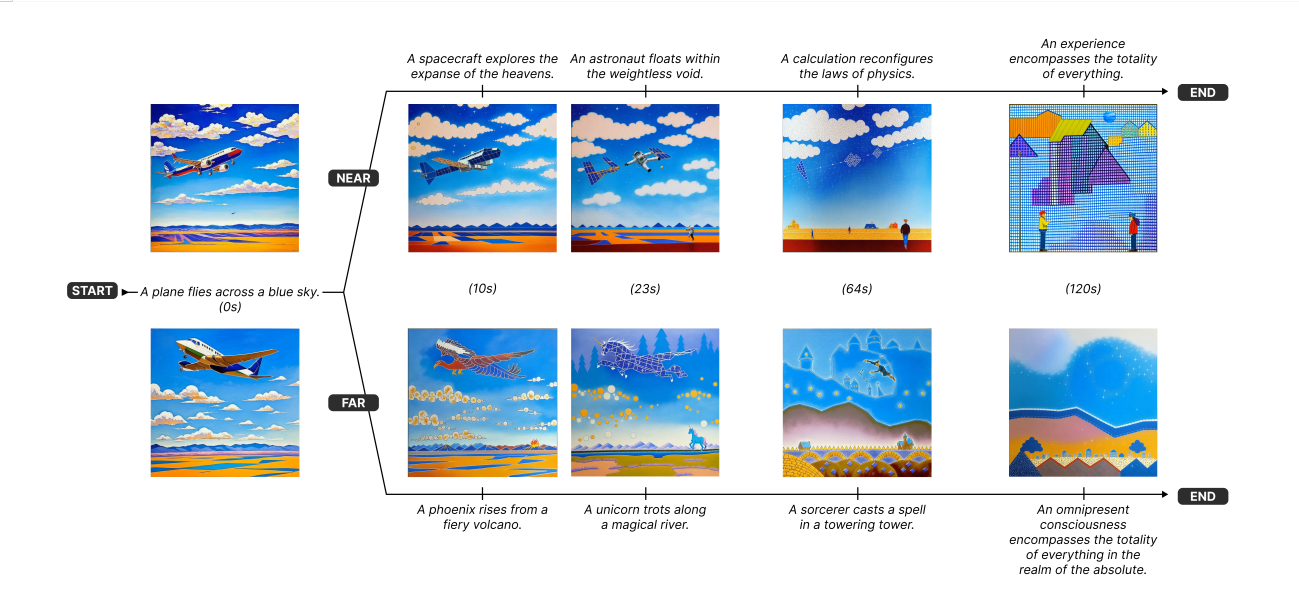
[2] `https://deforum.art/`

[3] `https://github.com/AUTOMATIC1111/stable-diffusion-webui/`

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(a) System for automated generation of videos designed to enhance psychological flexibility. To employ the psychologically-inspired generative AI method, the user prompts the system with a set of specific parameters, including a starting prompt that will seed the expansion of a narrative through a text-to-text generative model (i.e., LLM), resulting in a set of generative text prompts. The generative text prompts are then passed to a text-to-image model with parameters for visual features like style. Generated image frames are then interpolated together, leading to a video with features optimized to target the dynamic features of the underlying creative process, psychological flexibility.



(b) Sample pair of generative videos created using system. Beginning with the user-generated starting prompt, "A plane flies across a blue sky," the system generated two sets of narrative prompts that differ in the degree of Bayesian surprise between sequential prompts. Bayesian surprise was indexed using semantic distance, leading to two set of prompts: one covering a relatively *near* and one covering a relatively *far* semantic distance. All other video generation parameters were the same.

Figure 1: Psychologically-inspired video generation.

In order to create videos with dynamic visual features conducive to enhancing psychological flexibility, we adjusted a set of parameters: diffusion cadence = 6, angle = "0: (0)", zoom = "0:(1.00)", x translation: "0: (0)", y translation: "0: (0)", z translation: "0: (0)", anti-blur kernel_schedule: "0: (5)", anti-blur sigma_schedule": "0: (4.0)", anti-blur amount_schedule: "0: (0.2)", and anti-blur threshold_schedule": "0: (0.0)". Video frame size was 768 x 768 pixels. The seed was randomized for each generation instance and the remaining parameters took default val-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

```
Given the [STARTING_PROMPT], consider the
features of the world (including its
observable properties, possible actions, and
 overall context) in which this statement
exists. This is the prior belief. Now,
consider the [END_POINT] as the maximum
value for surprise. We will define surprise
as the degree to which a new observation
challenges or changes this prior belief,
resulting in an updated belief about the
world in which the statement exists.
Generate two sets of prompts that represent
[PROMPT_REFRESH] by incrementally updating
the [STARTING_PROMPT]. Each set should have
a total of [NUM_STEPS] prompts.

The two sets of prompts should differ in how
 surprising the final prompt is relative to
the [STARTING_PROMPT]. Use semantic distance
, or the degree of similarity between
sequential prompts, as a metric for surprise
. The size of the semantic distance between
each pair of prompts should be consistent.
The final prompt should be maximally
different from the [STARTING_PROMPT]. The
two sets of prompts should differ in how
dissimilar the final prompt is from the
original one. One should have a final prompt
 that is far in semantic similarity while
the other should have a final prompt that is
 very far in semantic similarity.
```

Figure 2: System prompt to initiate generative process. The system prompt instructs the generative AI model in how to generate pairs of prompt narratives differing in the progression of Bayesian surprise. Surprise is operationalized using semantic similarity.

ues. Positive prompts were used to constrain the visual style while negative prompts were used to minimize the likelihood of text, copyrighted material, or inappropriate content. After videos were generated, a title card was added to each video to provide basic narrative context to human viewers (e.g., "A flying plane").

Evaluating Human Responses to Generative AI Videos

Because prior work has highlighted the importance of rich, complex affective experiences for effective creativity interventions (Rowe, Hirsh, and Anderson 2007; Dolbier, Dieffenbach, and Lieberman 2025; Chirico et al. 2018), we measured human viewers' affective responses to viewing the generative AI videos. We expected that the videos created using our system would be associated with a wide range of affective responses, reflecting not only the complex affective experiences of the viewers but also individual differences in which videos were most compelling to each rater. We further expected to see differences in affect by semantic distance, with far videos leading to more varied affect than near videos.

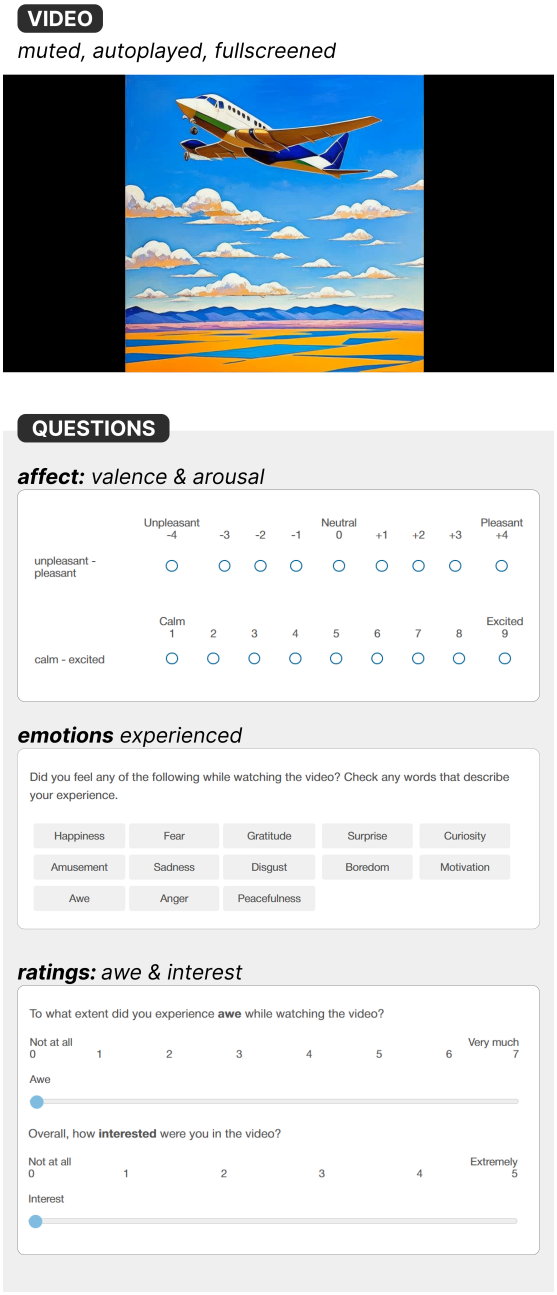


Figure 3: Video Evaluation Task. For each video, participants rated affect, awe, and level of interest; they also selected emotions they experienced while watching the video.

Participants

A total of 70 U.S.-based participants (33 females) evaluated the videos. Participant age ranged from 21 to 74 (median

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

= 36) and a range of ethnicities were represented (70.0% White, 12.9% Black, 10.0% Asian, 4.29% Mixed, 2.86% Other). Participants were recruited through Prolific[4] and completed the task using Qualtrics[5]. In addition to completing the video evaluation task described below, participants completed questionnaires to assess personality traits and demographics. Participants were compensated a flat rate of $8.50 for their time (mean = 25.5 min., std. = 12.3 min.). All participants provided informed consent and the study was approved by the WCG Institutional Review Board.

Video Evaluation Task Design

Each of 26 videos (one *near* and one *far* semantic distance for each starting prompt) was presented to 10-12 participants. Each participant viewed and rated videos generated from four distinct starting prompts; participants saw two *near* and two *far* videos. Video presentation order was randomized.

On each video evaluation trial, participants were presented with a two-minute, muted video in full screen. Participants could not proceed unless the complete video was viewed in fullscreen. Following the video, participants used anchored, Likert-style scales to rate state affect (valence: *Calm* = 1, *Excited* = 9; arousal: *Unpleasant* = -4, *Neutral* = 0, *Pleasant* = +4). They then reported any emotions they felt while watching the video by checking options from a randomly-ordered list (*Amusement*, *Anger*, *Awe*, *Boredom*, *Curiosity*, *Disgust*, *Fear*, *Gratitude*, *Happiness*, *Peacefulness*, *Motivation*, *Sadness*, *Surprise*). Finally, they rated their level of awe (*Not at all* = 0 to *Very much* = 7) and interest (*Not at all* = 0 to *Extremely* = 5) in response to the video using anchored scales. The design of the ratings questions was adapted from prior work assessing affective responses to awe-inducing videos used in interventions for creativity and pro-sociality (Chirico et al. 2018; Piff et al. 2015). A single trial is depicted in Figure 3.

Analytical Approach

Participant responses were preprocessed using `pandas` v2.2.3[6] in Python v3.1.2[7]. Statistical modeling was done with `statsmodels` v0.14.4[8]. To test the effect of video identity and level of semantic distance on video ratings, we fit linear regressions for each affective rating. To test whether prompt generation achieved the desired level of control over the level of Bayesian surprise between sequential prompts, we computed the absolute value of the difference between each pair of prompts and entered it into an ANOVA. Data visualizations were created using `seaborn` v0.13.2[9].

[4] https://www.prolific.com/

[5] https://www.qualtrics.com/

[6] https://pandas.pydata.org/

[7] https://www.python.org/downloads/release/python-312/

[8] https://www.statsmodels.org/stable/index.html

[9] https://seaborn.pydata.org/

Results

When selecting from 13 available emotions to describe how they felt while watching each video, the modal number of emotions selected was 3 (min = 1, max = 7). Across all videos, *curiosity* was selected most frequently, occurring in 22.4% of responses; *anger* was selected least frequently, appearing in only 0.51% of responses. Overall, the distribution of endorsed emotions was statistically equivalent for near and far semantic distances (χ^2(12, N=764) = 5.18, p=0.95). These data suggest that while semantic distance did not impact the incidence of discrete emotional states, viewing of generative AI videos elicited complex and sometimes conflicting emotions in participants.

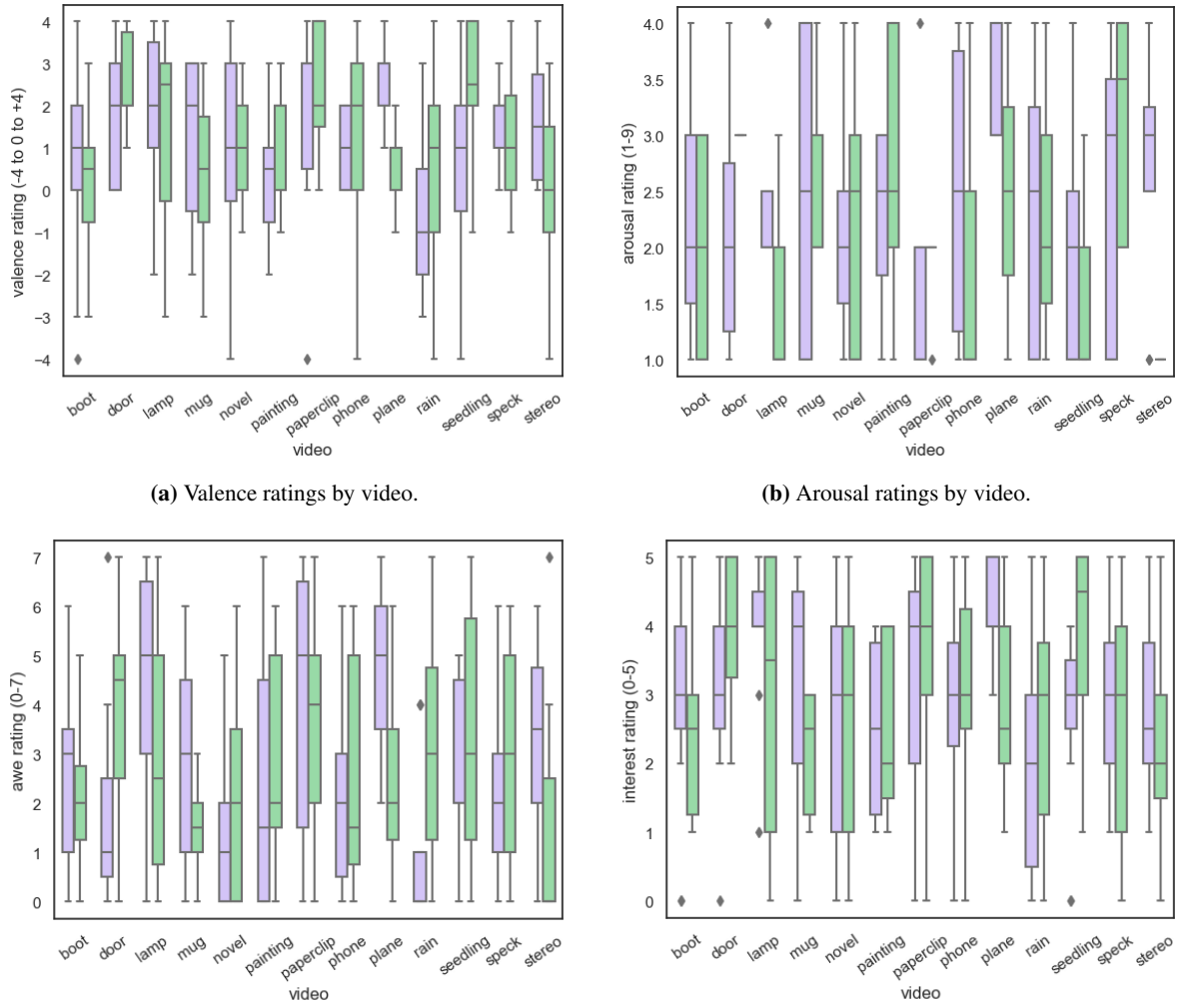
Analysis of ratings data reveals variability in line with a dimensional account of emotion and also reflects a wide range of affective experience in responses to the videos. Figure 4 shows how ratings varied by both video and level of semantic distance for valence, arousal, awe, and interest. These data did not, however, fully confirm the hypothesis that affect would be modulated by the level of semantic distance. Video and level of semantic distance explained significant variance in valence ratings (F(13,265) = 2.259, p=0.008), but not others, although trend-level effects were observed for awe (p=0.054) and interest (p=0.081). While each rating showed a marginal effect in the expected direction—that mean ratings would be greater for far compared to near prompts—there was high variance across both videos and participants. Further investigation examined whether the *near* and *far* sets of generative prompts indeed differed in the degree of Bayesian surprise between sequential prompts, since a lack of significant difference might have impacted the analyses of affective ratings. While there was a marginal effect of video, the mean absolute distance between prompts did not significantly differ by semantic distance ($F(1, 24)$ = 0.002, p=0.967).

Discussion

In designing a system for the creation of generative AI video interventions for creativity, we prioritized the ability to target underlying processes associated with psychological flexibility. These core components of creative cognition evolve dynamically over time, with meaningful features emerging at timescales finer than those typically targeted by creativity support (i.e., sub-seconds to seconds rather than minutes or longer). Here, we aimed to bring the dynamics of the intervention more in line with those of the underlying process of interest, *psychological flexibility*. We used generative AI to manipulate, on the order of seconds, the dynamics of surprise in video narratives, thus inducing the viewer to update their model of the world. Through this perturbation of mental dynamics, we aimed to encourage flexibility and open-mindedness that could support creativity without regard to specific task or domain.

We demonstrate initial evidence for the success of this approach to intervention generation, demonstrating the effect of the generative AI video interventions on the affective state of human viewers. Because prior research on creative state interventions has noted the importance of affect—especially

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(a) Valence ratings by video.

(b) Arousal ratings by video.

(c) Awe ratings by video.

(d) Interest ratings by video.

Figure 4: Generative AI video ratings by user-generated starting prompt. For valence (a), arousal (b), awe (c), and interest (d), mean ratings and interquartile range are depicted for each video. *Near* semantic distance is in green while *far* semantic distance is in purple.

rich, complex affective experiences, like those associated with awe—we focused first on testing whether videos created using our system would be associated with viewing. On the whole, the results of our video evaluation study show increases in affect consistent with that previously seen with awe inductions (e.g., (Piff et al. 2015)), suggesting that the videos have characteristics sufficient for impacting affective states relevant to creativity. The efficacy of these generative AI videos as interventions for creativity remains to be tested directly.

Our approach to creating generative AI video interventions diverges from conventional human-AI co-creative applications of AI-driven storytelling, which increasingly leverage end-to-end models that map textual descriptions directly to video sequences. Large vision models like Runway's Gen-2[10] system and OpenAI's Sora[11] have demonstrated remarkable capability for high-quality video generation (Liu et al. 2024). These models often employ advanced prompt optimization and preference alignment strategies to produce aesthetically enhanced imagery, with the goal of mimicking the data distribution of videos that are available in the public domain.

Our priority was to develop a system for video intervention generation that would provide users with a high level of control, flexibility, interpretability, and reproducibility in all aspects of the generative process. This modular approach enabled us to exploit the latent capabilities of the different generative models in the least restrictive manner. Take, for

[10] https://runwayml.com/research/gen-2
[11] https://openai.com/sora/

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

instance, the text-to-text generative pipeline we used to generate the scene descriptions. While strategies exist for refining text-to-image prompts for computational creativity applications (Ibarrola and Grace 2023), we wanted to provide the minimum necessary information to preserve the generative diversity of the LLM, avoiding over-specification while still being able to guide the model toward coherence.

To generate the individual scene descriptions, we leveraged OpenAI's text-to-text generative model for its zero-shot reasoning capabilities (Espejel et al. 2023), semantic processing (Le Mens et al. 2023), and demonstrated aptitude for a range of affective computing tasks (Amin et al. 2024). The generated scene descriptions then informed a text-to-image model (i.e., Stable Diffusion), which created image frames, followed by a diffusion-based interpolation model that synthesized smooth transitions between the image frames. It is important to note that the visual feel of generative videos can significantly impact the viewer's experience. While our framework enables fine-grained adjustments to key parameters like visual style, frame rate, and motion dynamics, we developed heuristics to reduce the overwhelming parameter space for features.

Importantly, many of these heuristics were informed by an explicit choice to eschew realism in the generative AI videos. By selecting starting prompts that reflected likely scenes from daily life and systematically increasing Bayesian surprise over narrative progression, we aimed to control the properties of the distribution of information presented to the viewer. We expected that the aesthetics and production quality could risk this control, especially if an expectation of realism was set and then subsequently violated. Therefore, we chose parameters that would produce videos of acceptable production value that were *conceptually* grounded in viewers' experiences, but unmoored from high expectations about video quality. The unique visuals of the resulting generative AI videos helped establish an event boundary (Zacks and Swallow 2007) that could boost the potential efficacy of the intervention *without* also inducing the strong negative affect associated with narrow memory encoding (Gable and Harmon-Jones 2010). Narrow rather than broad encoding would inhibit the desired outcome of openness and flexibility.

Considerations and Future Directions

While we were able to generate interesting, narratively diverse videos with a range of visual styles, we did not achieve the desired level of control over the LLM's search over the semantic space, reducing the precision of the Bayesian surprise manipulation. The effect of the degree of surprise between sequential prompts was not as expected, perhaps because text-to-text narrative expansion did not reliably result in a consistent semantic distance between prompts where the distance was, on average, smaller for *near* relative to *far*. Future research should explore ways to increase control without overly specifying the system prompt.

We observed a great deal of variability in individuals' affective responses to the generative AI videos. Although such variability was expected and even typical for larger studies of affect in response to viewing videos (e.g., (Samide, Cooper, and Ritchey 2020)), it can limit inference on which videos induce affective states conducive to the desired effects on psychological flexibility. The immense size of the parameter space suggests a need to develop strategies that will allow future work to harness the options for features like narrative topic, aesthetic style, and frame rate to match the preferences of the viewer. Maintaining the appropriate level of meaningful variability is critical for the success of future interventions, since personal salience of stimuli can boost learning and memory (Peters 2006) as well as the efficacy of behavior change interventions (Alslaity, Chan, and Orji 2023). Increasing the level of immersiveness may also facilitate stronger resonance between a video and a human viewer. For example, prior work has shown that immersion can enhance the effect of awe (Chirico et al. 2018) as well as hallucinatory generative AI interventions (Greco et al. 2025) on creative thinking. Increasing immersiveness by including additional sensory modalities or through the use of techniques like virtual reality may allow for a higher likelihood of intervention success in future trials.

Finally, while our explicit focus has been on the potential for enhancing human creativity, our approach may also have implications for machine creativity. Pezzulo and colleagues (2023) argue that shifting generative AI models from passive to active inference may lead to systems more capable of human-like creativity (Pezzulo et al. 2023). Our earlier motivating example describing the formation of event boundaries as a result of active inference on the world (Zacks, Speer, and Reynolds 2009) is just the sort of process that may enable such advances. Future research may find value in exploring the extent to which purposefully shifting a generative AI agent's representation of the world may increase its flexibility.

Conclusion

Advances in generative AI suggest opportunities for new automated, scalable approaches to amplifying human creativity that can provide support across tasks and domains. We present a novel method for creating generative AI video interventions designed to enhance psychological flexibility, a central component of human creativity. We provide an initial demonstration of the system's capabilities and present opportunities for future research, including the development of personalized interventions for creativity.

Author Contributions

S.H. and M.K.H. conceptualized the research. S.H., M.V., M.K.H., and K.M. developed the video generation procedure, including associated software. S.H. and M.V. designed and conducted the human subjects study; S.H. analyzed and interpreted data. S.H., M.V., M.K.H., P.E.P, and M.K. drafted the manuscript.

References

Albulescu, P.; Macsinga, I.; Rusu, A.; Sulea, C.; Bodnaru, A.; and Tulbure, B. T. 2022. "Give me a break!" A systematic review and meta-analysis on the efficacy of micro-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

breaks for increasing well-being and performance. *PLOS ONE* 17(8):1–27.

Alslaity, A.; Chan, G.; and Orji, R. 2023. A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. *Frontiers in Artificial Intelligence* Volume 6 - 2023.

Amin, M. M.; Mao, R.; Cambria, E.; and Schuller, B. W. 2024. A wide evaluation of chatgpt on affective computing tasks. *IEEE Transactions on Affective Computing*.

Amitani, S., and Hori, K. 2002. Supporting musical composition by externalizing the composer's mental space. In *Proceedings of the 4th Conference on Creativity & Cognition*, CC '02, 165–172. New York, NY, USA: Association for Computing Machinery.

Ashkinaze, J.; Mendelsohn, J.; Qiwei, L.; Budak, C.; and Gilbert, E. 2024. How AI ideas affect the creativity, diversity, and evolution of human ideas: evidence from a large, dynamic experiment. *arXiv preprint arXiv:2401.13481*.

Baldi, P., and Itti, L. 2010. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks* 23(5):649–666.

Benedek, M.; Jauk, E.; Sommer, M.; Arendasy, M.; and Neubauer, A. C. 2014. Intelligence, creativity, and cognitive control: The common and differential involvement of executive functions in intelligence and creativity. *Intelligence* 46:73–83.

Bi, H.; Mi, S.; Lu, S.; and Hu, X. 2020. Meta-analysis of interventions and their effectiveness in students' scientific creativity. *Thinking Skills and Creativity* 38:100750.

Boden, M. 2004. *The Creative Mind: Myths and Mechanisms*. Taylor & Francis.

Chirico, A.; Glaveanu, V. P.; Cipresso, P.; Riva, G.; and Gaggioli, A. 2018. Awe enhances creative thinking: An experimental study. *Creativity Research Journal* 30(2):123–131.

Chou, Y.-Y. J., and Tversky, B. 2020. Changing perspective: Building creative mindsets. *Cognitive Science* 44(4):e12820.

Chung, J. J. Y.; He, S.; and Adar, E. 2021. The intersection of users, roles, interactions, and technologies in creativity support tools. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, 1817–1833. New York, NY, USA: Association for Computing Machinery.

Clark, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences* 36(3):181–204.

Demszky, D.; Yang, D.; Yeager, D. S.; Bryan, C. J.; Clapper, M.; Chandhok, S.; Eichstaedt, J. C.; Hecht, C.; Jamieson, J.; Johnson, M.; et al. 2023. Using large language models in psychology. *Nature Reviews Psychology* 2(11):688–701.

DiPaola, S.; Gabora, L.; and McCaig, G. 2018. Informing artificial intelligence generative techniques using cognitive theories of human creativity. *Procedia computer science* 145:158–168.

Dolbier, S. Y.; Dieffenbach, M. C.; and Lieberman, M. D. 2025. Open-mindedness: An integrative review of interventions. *Psychol. Rev.* 132(1):204–238.

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.

Espejel, J. L.; Ettifouri, E. H.; Alassan, M. S. Y.; Chouham, E. M.; and Dahhane, W. 2023. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal* 5:100032.

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*, CHI '19, 1–18. New York, NY, USA: Association for Computing Machinery.

Friston, K.; Moran, R. J.; Nagai, Y.; Taniguchi, T.; Gomi, H.; and Tenenbaum, J. 2021. World model learning and inference. *Neural Networks* 144:573–590.

Fu, Y.; Bin, H.; Zhou, T.; Wang, M.; Chen, Y.; Lai, Z. G. D. C.; Wobbrock, J. O.; and Hiniker, A. 2024. Creativity in the age of AI: Evaluating the impact of generative AI on design outputs and designers' creative thinking. *arXiv preprint arXiv:2411.00168*.

Gable, P., and Harmon-Jones, E. 2010. The blues broaden, but the nasty narrows: Attentional consequences of negative affects low and high in motivational intensity. *Psychological Science* 21(2):211–215. PMID: 20424047.

Gandy, S.; Bonnelle, V.; Jacobs, E.; and Luke, D. 2022. Psychedelics as potential catalysts of scientific creativity and insight. *Drug Science, Policy and Law* 8:20503245221097649.

Greco, A.; Rastelli, C.; Ubaldi, A.; and Riva, G. 2025. Immersive exposure to simulated visual hallucinations modulates high-level human cognition. *Consciousness and Cognition* 128:103808.

Hayes, S. C.; Law, S.; Malady, M.; Zhu, Z.; Bai, X.; et al. 2020. The centrality of sense of self in psychological flexibility processes: What the neurobiological and psychological correlates of psychedelics suggest. *Journal of Contextual Behavioral Science* 15:30–38.

Ibarrola, F. J., and Grace, K. 2023. Prompt diversification for iterating with text-to-image models. In *International Conference on Innovative Computing and Cloud Computing*.

Itti, L., and Baldi, P. 2009. Bayesian surprise attracts human attention. *Vision research* 49(10):1295–1306.

Karwowski, M. 2014. Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts* 8(1):62.

Kashdan, T. B., and Rottenberg, J. 2010. Psychological flexibility as a fundamental aspect of health. *Clinical psychology review* 30(7):865–878.

Kenett, Y. N.; Levy, O.; Kenett, D. Y.; Stanley, H. E.; Faust, M.; and Havlin, S. 2018. Flexibility of thought in high creative individuals represented by percolation analysis. *Proceedings of the National Academy of Sciences* 115(5):867–872.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Kumar, M.; Goldstein, A.; Michelmann, S.; Zacks, J. M.; Hasson, U.; and Norman, K. A. 2023. Bayesian surprise predicts human event segmentation in story listening. *Cognitive science* 47(10):e13343.

Le Mens, G.; Kovács, B.; Hannan, M. T.; and Pros, G. 2023. Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences* 120(49):e2309350120.

Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J.; He, L.; and Sun, L. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models.

McCrae, R. R. 1987. Creativity, divergent thinking, and openness to experience. *J. Pers. Soc. Psychol.* 52(6):1258–1265.

Nijstad, B. A.; De Dreu, C. K.; Rietzschel, E. F.; and Baas, M. 2010. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European review of social psychology* 21(1):34–77.

O'Connor, A. J.; Nemeth, C. J.; and Akutsu, S. 2013. Consequences of beliefs about the malleability of creativity. *Creativity Research Journal* 25(2):155–162.

OpenAI. 2024. GPT-4 Technical Report.

Peeperkorn, M.; Brown, D.; and Jordanous, A. 2023. On characterizations of large language models and creativity evaluation. In *International Conference on Computational Creativity*.

Peters, E. 2006. *The Functions of Affect in the Construction of Preferences*. Cambridge University Press. 454–463.

Pezzulo, G.; Parr, T.; Cisek, P.; Clark, A.; and Friston, K. 2023. Generating meaning: Active inference and generative ai. *PsyArXiv. June* 8.

Piff, P. K.; Dietze, P.; Feinberg, M.; Stancato, D. M.; and Keltner, D. 2015. Awe, the small self, and prosocial behavior. *Journal of personality and social psychology* 108(6):883.

Ritter, S. M.; Damian, R. I.; Simonton, D. K.; van Baaren, R. B.; Strick, M.; Derks, J.; and Dijksterhuis, A. 2012. Diversifying experiences enhance cognitive flexibility. *Journal of experimental social psychology* 48(4):961–964.

Roberts, B. W.; Luo, J.; Briley, D. A.; Chow, P. I.; Su, R.; and Hill, P. L. 2017. A systematic review of personality trait change through intervention. *Psychological bulletin* 143(2):117.

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 (CVPR)*, 10684–10695.

Rowe, G.; Hirsh, J. B.; and Anderson, A. K. 2007. Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences* 104(1):383–388.

Royston, R., and Reiter-Palmon, R. 2019. Creative self-efficacy as mediator between creative mindsets and creative problem-solving. *The Journal of Creative Behavior* 53(4):472–481.

Samide, R.; Cooper, R. A.; and Ritchey, M. 2020. A database of news videos for investigating the dynamics of emotion and memory. *Behavior Research Methods* 52:1469–1479.

Smith, S. M., and Ward, T. B. 2012. *The Oxford handbook of thinking and reasoning*. Oxford University Press. chapter Cognition and the Creation of Ideas, 456–474.

Tan, X.; Antony, M.; and Kong, H. 2020. Automated music generation for visual art through emotion. In *International Conference on Computational Creativity*, 247–250.

Thölke, P.; Bellemare-Pepin, A.; Harel, Y.; Lespinasse, F.; and Jerbi, K. 2024. Bio-Mechanical Poet: An Immersive Audiovisual Playground for Brain Signals and Generative AI. In *Proceedings of the 15th International Conference on Computational Creativity*.

Uddin, L. Q. 2021. Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. *Nat. Rev. Neurosci.* 22(3):167–179.

Wiggins, G. A. 2020. Creativity, information, and consciousness: the information dynamics of thinking. *Physics of life reviews* 34:1–39.

Wu, C.-L.; Huang, S.-Y.; Chen, P.-Z.; and Chen, H.-C. 2020. A systematic review of creativity-related studies applying the remote associates test from 2000 to 2019. *Frontiers in psychology* 11:573432.

Zabelina, D. L., and Robinson, M. D. 2010. Creativity as flexible cognitive control. *Psychology of Aesthetics, Creativity, and the Arts* 4(3):136.

Zabelina, D. L.; Friedman, N. P.; and Andrews-Hanna, J. 2019. Unity and diversity of executive functions in creativity. *Consciousness and cognition* 68:47–56.

Zacks, J. M., and Swallow, K. M. 2007. Event segmentation. *Current Directions in Psychological Science* 16(2):80–84. PMID: 22468032.

Zacks, J. M.; Speer, N. K.; and Reynolds, J. R. 2009. Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General* 138(2):307.

Zhang, J. W.; Howell, R. T.; Razavi, P.; Shaban-Azad, H.; Chai, W. J.; Ramis, T.; Mello, Z.; Anderson, C. L.; Monroy, M.; and Keltner, D. 2024. Awe is associated with creative personality, convergent creativity, and everyday creativity. *Psychology of Aesthetics, Creativity, and the Arts* 18(2):209.

Zhang, M.; Wang, F.; and Zhang, D. 2020. Individual differences in trait creativity moderate the state-level mood-creativity relationship. *PloS one* 15(8):e0236987.

Zhou, E., and Lee, D. 2024. Generative artificial intelligence, human creativity, and art. *PNAS nexus* 3(3):pgae052.

Automatic Aesthetic Evaluation in Generative Image Models

Larissa D. Gomide, Lucas N. Ferreira, Wagner Meira Jr.
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
{larissa.gomide, lferreira, meira}@dcc.ufmg.br

Abstract

Recent art assessment neural models have shown promising results as aesthetic evaluators of paintings, computing objective metrics through handcrafted features and deep learning techniques. While these models have been applied to evaluate human-produced pieces, their use in measuring the aesthetic quality of artificial intelligence (AI)-generated artistic works remains largely unexplored. This paper employs ArtCLIP, a recent art assessment neural model, to evaluate the aesthetic quality of paintings produced by state-of-the-art image generation models. Our methodology includes automatically describing human artworks with an image-to-text model and generating new images from these prompts, allowing us to pair human artworks with generated pieces. We compare the distribution of nine different aesthetic attributes given by ArtCLIP between human and AI-generated paintings. We considered models of two considerably different sizes to measure the impact of model size on aesthetic quality. The results show that ArtCLIP ranks human artwork as having higher aesthetic quality, with the larger generator outperforming the smaller one.

Introduction

In Artificial Intelligence, generative models are often evaluated using objective metrics such as the Bilingual Evaluation Understudy (BLEU) score for Machine Translation (Papineni et al. 2002), and the Fréchet Inception Distance (FID) for Computer Vision (Heusel et al. 2017) tasks. These metrics have played a central role in driving progress across domains by providing standardized, quantifiable benchmarks. However, such measures are primarily concerned with fidelity and realism, which—while important—are insufficient when evaluating outputs in Computational Creativity, where the goals extend beyond imitation toward novelty, value, and surprise (Colton and Wiggins 2012).

In Computational Creativity, human-centered assessments—including expert critique and user studies—remain the gold standard but are time-consuming and difficult to scale or reproduce. Thus, it is important to develop automated yet meaningful evaluation metrics that can approximate subjective judgment while being generalizable and consistent (Jordanous 2022).

In this direction, (Jin et al. 2024b) introduced the Aesthetics of Paintings and Drawings Dataset (APDDv2), which consists of 10,023 human artworks annotated with scores for ten artistic attributes frequently discussed in visual arts literature, such as mood, creativity, layout, and composition, among others. Alongside, they proposed ArtCLIP, a collection of predictive models trained to estimate these attributes. While ArtCLIP shows strong performance on human-authored pieces, its applicability to AI-generated visual art has not been examined—raising the question of whether such models can serve as robust and meaningful aesthetic evaluators in creative AI systems.

In this paper, we address this gap by evaluating whether ArtCLIP can function as an aesthetic scoring model for generated paintings. First, we built a dataset of generated paintings by employing the image-to-text Janus-Pro model (Chen et al. 2025) to describe a sample of human paintings from the APDDv2 dataset. Then, we used these textual descriptions to generate new paintings with two versions of the text-to-image Janus-Pro model: a small (≈1B parameters) and a large (≈7B parameters) version. This process yielded a dataset of 1,000 paired samples (I_h, I_{gs}, I_{gl}, T), where I_h is the original human-made artwork, I_{gs} and I_{gl} are paintings generated by the small and large models, respectively, and T is their shared textual description. Finally, we employed ArtCLIP to compute nine aesthetic attributes for each image independently.

To assess ArtCLIP as an objective aesthetic metric of visual artwork, we first compared the distributions of the nine ArtCLIP scores across the human, Janus-Pro-1B, and Janus-Pro-7B paintings. Results revealed that pieces by Janus-Pro-1B received lower scores than those from Janus-Pro-7B, which in turn were rated lower than the human-created artworks. We also evaluated the consistency of ArtCLIP across human, Janus-Pro-1B, and Janus-Pro-7B paintings—that is, whether paintings with the same description received similar scores. For each attribute, we computed the average distance between the scores assigned to human artworks and their generated counterparts. As expected, the distances between the human and the large model were consistently smaller than those between the human and the small model across all attributes. Overall, these findings suggest that ArtCLIP can serve as a consistent metric for assessing the aesthetic quality of AI-generated paintings.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

The main contributions of this paper are:

- A dataset of human paintings from APDDv2 paired with two generated counterparts—one produced by a small model and the other by a large model—based on a shared textual description;
- An evaluation of ArtCLIP as an objective metric for assessing the aesthetic quality of paintings produced by image generation models.

Related Work

Our work is directly related to ArtCLIP and the APDDv2 dataset (Jin et al. 2024b), which are the primary focus of this study. It is also connected to previous efforts to develop objective metrics for evaluating creative systems. This section briefly reviews each of these topics.

APDDv2 and ArtCLIP

The APDDv2 dataset comprises 10,023 paintings across 24 artistic categories, with each image annotated according to 10 aesthetic attributes. These categories vary by painting type, artistic style, and subject matter—for example, *Oil Painting - Symbolism - Landscapes* and *Traditional Chinese Painting - Meticulous - Portraiture*. The aesthetic attributes include: *Theme and Logic* (T&L), *Creativity* (Cre), *Layout and Composition (L&C)*, *Space and Perspective* (S&P), *Sense of Order* (SO), *Light and Shadow* (L&S), *Color* (Col), *Details and Texture* (D&T), *Mood* (M), and *The Overall* (TO). Each image was evaluated on a continuous scale from 0 to 10 by at least six annotators. The annotation team consisted of 37 trained individuals, including professional artists and educators with at least a bachelor's degree and extensive experience.

Using this dataset, the authors introduced ArtCLIP, a multimodal aesthetic assessment model that combines image and text embeddings through contrastive learning. The model is initially pre-trained on the DPC2022 dataset (Zhong, Zhou, and Qiu 2023), which contains photographic images paired with categorized aesthetic comments, enabling ArtCLIP to learn general aesthetic representations. It is then fine-tuned on APDDv2 to specialize in evaluating paintings. Compared to baseline models AANSPS (Jin et al. 2024a) and SAAN (Yi et al. 2023), ArtCLIP demonstrates improved or comparable performance, showing higher agreement with human annotations and lower prediction error on attributes such as *Composition*, *Color*, and *Mood*.

Other related image datasets and aesthetic evaluation models also inform our work. For example, (Achlioptas et al. 2021) introduced ArtEmis, a dataset of artworks annotated with emotional labels and natural language explanations, which was used to train models capable of generating affective captions. Similarly, (Vera Nieto, Celona, and Fernandez Labrador 2022) presented the Reddit Photo Critique Dataset, which contains photographs paired with user-generated critiques and derives aesthetic scores through sentiment analysis. Both approaches rely on subjective, unstructured language to infer aesthetic judgments. In contrast, our work evaluates structured, expert-defined aesthetic attributes using ArtCLIP and applies this model to both human and AI-generated paintings.

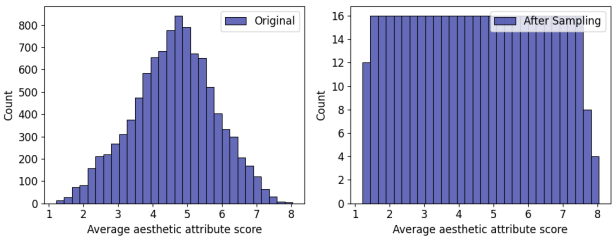


Figure 1: The original distribution of average aesthetic attribute scores divided into 30 bins (left), and the distribution after sampling (right).

Creativity Metrics

Our work is also related to ongoing efforts to quantify creativity in machine-generated content. For instance, (Wang et al. 2023) investigates how well existing metrics such as FID (Heusel et al. 2017) align with human judgments when evaluating combinational creativity in images generated by DALL-E (Ramesh et al. 2021). Another example is the work of (Lu et al. 2024), in which the authors propose the *Creativity Index*, a metric for assessing linguistic creativity by measuring how much of a given text can be reconstructed from existing web sources. The primary distinction between these approaches and ours is that, rather than evaluating creativity exclusively, our work explores nine different attributes of aesthetic quality of visual outputs, including creativity as one of them.

More broadly, this work contributes to the ongoing conversation about evaluation metrics in Computational Creativity, a field that has long grappled with the challenge of developing standardized, objective evaluation practices. As noted by (Jordanous 2022), the field has reached a point of maturity where establishing such metrics is increasingly feasible. Our study adds to this effort by investigating structured, attribute-based evaluation of visual aesthetics in AI-generated paintings.

Methodology

Our main goal is to evaluate ArtCLIP as an aesthetic scoring model for generated paintings. To do this, we created a new dataset of generated paintings paired with corresponding human-made versions from APDDv2. Since the original dataset contains approximately 10k images, generating counterparts for all of them using state-of-the-art image generators would be computationally expensive. To manage these costs while maintaining statistical validity, we employed a sampling strategy and selected 500 images from the original dataset. Based on the standard sample size formula using a z-score for confidence intervals, at least 385 instances are required to achieve 95% confidence with a 5% margin of error.

The original distribution of paintings in APDDv2 is not balanced in terms of average aesthetic attribute score. In

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

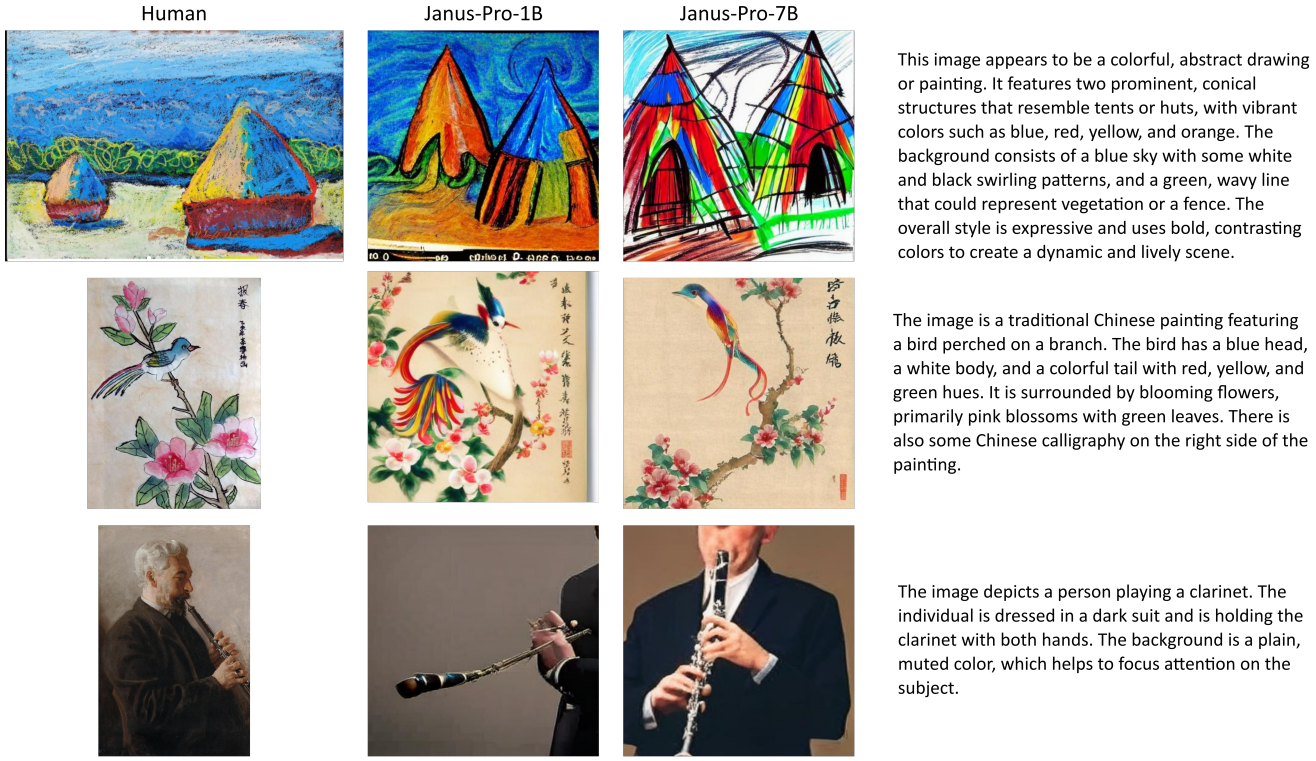


Figure 2: Four examples of instances from our dataset. An example is composed of a human-authored piece from APDDv2 paired with two images: one generated with Janus-Pro-1B and another with Janus-Pro-7B. The three images share the same textual description, which the image-to-text Janus-Pro-7B model gives.

particular, it follows a normal distribution, meaning there are considerably fewer images with maximum (10) and minimum (0) scores than with intermediate scores (around 5). To address this, we divided the images into 30 bins based on their average aesthetic scores and selected an equal number of samples from each bin, resulting in approximately 16 images per bin. Figure 1 shows the original distribution (left) and the balanced distribution after sampling (right). This strategy ensures that ArtCLIP's performance can be evaluated consistently across the full range of aesthetic scores, including extreme cases.

With this balanced sample of 500 paintings, we used an image-to-text model to produce textual descriptions for each image, which we then used as prompts to generate new images with a text-to-image model. Specifically, we employed Janus-Pro-7B (Chen et al. 2025) for generating the descriptions. We chose Janus-Pro-7B because it is a state-of-the-art model that unifies multimodal understanding and generation. This allowed us to use it both for generating textual descriptions of the APDDv2 paintings and for creating new paintings from those descriptions. The smaller model tends to produce images that are closer to the original, effectively capturing the core content from the description—although it occasionally introduces small artifacts (e.g., a black border at the bottom). The larger model introduces more stylistic variation, diverging in tone and interpretation.

Each of the 500 sampled images was passed through Janus-Pro-7B with the prompt "Describe the image" to obtain its textual description. On average, the descriptions contained 88.31 words, with a standard deviation of 25.60. The longest description contained 262 words, and the shortest 38 words. Although Janus-Pro-7B supports direct image-to-image operations due to its multimodal nature, we opted to use textual descriptions as intermediate representations. This choice improves explainability, enables prompt-level analysis, and allows us to better understand how image generation is influenced by descriptive content.

Next, for each of the 500 descriptions, we generated two new paintings: one using Janus-Pro-1B and another using Janus-Pro-7B. These models share the same architecture but differ in size—the former has approximately 1 billion parameters, while the latter has around 7 billion. Each image was generated using the exact text description as a prompt, with no additional augmentation. At the end of this process, our dataset contained 1,500 images: 500 human-authored pieces from APDDv2, each paired with two AI-generated counterparts.

Figure 2 presents three examples from our dataset, each featuring a different painting type, artistic style, and subject matter. These examples show that the image-to-text model can describe not only the subject matter of the images but also stylistic elements such as color and painting type (e.g.,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

abstract vs. traditional Chinese). Moreover, the text-to-image models appear to better capture the subject matter in abstract paintings (e.g., the two abstract huts) than in more realistic ones (e.g., the Chinese bird and the person playing the clarinet).

Finally, we applied the ArtCLIP scoring model to obtain aesthetic scores for the 1,500 images. ArtCLIP consists of 10 independent models, each corresponding to a different aesthetic attribute. During our evaluation, we couldn't run the model responsible for the "Sense of Order" attribute due to technical issues, and hence it was excluded. Therefore, we report results on 9 aesthetic scores.

To execute the steps of generating descriptions, images, and scores, which involve inferences with complex deep learning-based models, we used a high-performance computer equipped with a Ryzen 9 5950X 16-core processor, 64 GB of RAM, and an RTX 3090 Ti 24 GB graphics card.

Experiments and Results

We conducted two experiments to evaluate whether ArtCLIP can serve as a consistent and meaningful aesthetic evaluation model for AI-generated paintings. The first experiment compared the distributions of aesthetic attribute scores between the Human, Janus-Pro-1B, and Janus-Pro-7B artworks. The second experiment analyzed the consistency of ArtCLIP's judgments across these attributes, specifically, whether artworks derived from the same prompt (textual description) receive comparable scores.

Comparison of Aesthetic Scores

The goal of the first experiment was to assess whether ArtCLIP can distinguish between varying levels of aesthetic quality in a meaningful and interpretable way. By comparing the distributions of aesthetic scores assigned to Human, Janus-Pro-1B, and Janus-Pro-7B, we evaluated whether ArtCLIP's outputs reflect expected qualitative differences. If ArtCLIP is a reliable aesthetic evaluator, it should assign higher scores to originally produced human-created artworks. In contrast, generated images should receive slightly lower scores, since they are transformations of human ideas that may accumulate noise or inconsistencies across the generation, description, and evaluation pipeline, especially for smaller models.

Table 1 shows the nine mean aesthetic scores for Human, Janus-Pro-1B, and Janus-Pro-7B paintings, as given by ArtCLIP. For each attribute, we ran a Friedman test across painting sources and, if significant, a pairwise Wilcoxon test with Bonferroni correction. Based on the Friedman test, all nine criteria showed significant differences across the methods (p-value ≤ 0.05). The post-hoc pairwise Wilcoxon test showed that, for all attributes, the human paintings were significantly better than the Janus models, and the large model was significantly better than the small one. Therefore, we can conclude that human artworks consistently received the highest scores, followed by those from the large model, and lastly from the small model.

These results suggest that ArtCLIP can capture a significant aesthetic gap between AI-generated and human-created paintings. Moreover, the ranking—Human $>$ Janus-Pro-7B $>$ Janus-Pro-1B across all attributes implies that increasing model capacity improves aesthetic performance, but not enough to match human-level quality.

Table 1: Mean $\pm$ standard deviation of each aesthetic attribute (Attr.) for Human, Janus-Pro-1B, and Janus-Pro-7B paintings. For each attribute, the higher the score, the better. Different superscript letters indicate statistically significant differences ($p < 0.05$). A value in bold is the best score for a particular attribute (row).

Attr.	Human	Janus-1B	Janus-7B
T&L	**6.25**^{a} $\pm$ 1.23	5.89^{b} $\pm$ 0.86	6.14^{c} $\pm$ 0.82
Cre	**5.91**^{a} $\pm$ 1.41	5.65^{b} $\pm$ 0.81	5.82^{c} $\pm$ 0.82
L&C	**6.21**^{a} $\pm$ 1.35	5.69^{b} $\pm$ 1.06	6.01^{c} $\pm$ 1.01
S&P	**6.37**^{a} $\pm$ 1.38	5.74^{b} $\pm$ 1.06	6.09^{c} $\pm$ 1.04
L&S	**6.29**^{a} $\pm$ 1.45	5.53^{b} $\pm$ 1.23	5.94^{c} $\pm$ 1.21
Col	**6.19**^{a} $\pm$ 1.38	6.04^{b} $\pm$ 0.98	6.20^{c} $\pm$ 1.00
D&T	**6.10**^{a} $\pm$ 1.52	5.33^{b} $\pm$ 1.25	5.72^{c} $\pm$ 1.23
TO	**6.26**^{a} $\pm$ 1.41	5.76^{b} $\pm$ 1.06	6.09^{c} $\pm$ 1.02
M	**6.03**^{a} $\pm$ 1.35	5.53^{b} $\pm$ 1.00	5.85^{c} $\pm$ 1.01

Consistency

In the second experiment, we evaluated ArtCLIP's consistency by measuring how closely the aesthetic attribute scores of generated images aligned with those of their corresponding human counterparts. For each aesthetic attribute, we computed the average absolute difference between the scores assigned to each human artwork and its small and large model versions.

As shown in Figure 3, the large model yields smaller differences compared to the small model across all attributes, indicating closer alignment with human aesthetic judgments. The differences between the models' average absolute distances typically fall within a range of approximately 0.3 to 0.4, suggesting that the improvement gained from scaling up the model is relatively uniform across aesthetic dimensions. This result indicates that increasing model capacity benefits all aspects of aesthetic alignment to a similar degree, rather than disproportionately enhancing performance on specific attributes. Notably, the large model achieves an almost perfect match in the Color attribute, implying that Janus-Pro-7B is capable of replicating human-like color use with high fidelity, at least when prompted with automatically generated descriptions of images from the APDDv2 dataset.

Overall, these results suggest that ArtCLIP not only detects aesthetic differences between human and AI-generated paintings but also produces consistent scores when comparing artworks derived from the same description. This consistency across paired samples provides further evidence that ArtCLIP can function as a reliable metric for aesthetic evaluation in generative image systems.

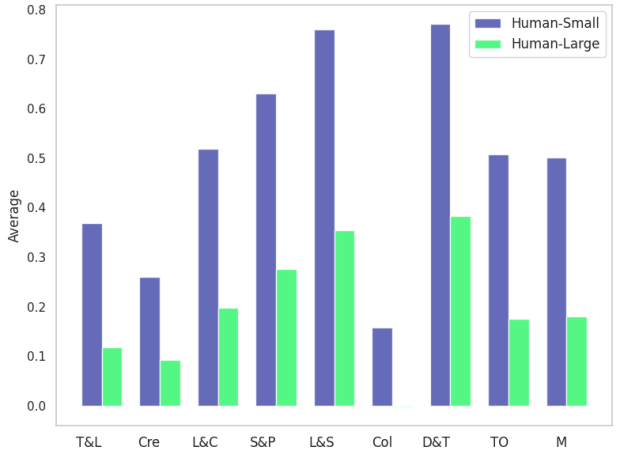


Figure 3: Average absolute differences in aesthetic attribute scores between original human artworks and their AI-generated counterparts.

Conclusion

This paper explored the use of ArtCLIP, a recent art assessment neural model, as an automatic evaluator of aesthetic quality in AI-generated paintings. To this end, we created a dataset of 500 human paintings, each paired with two generated images—one produced by a smaller Janus-Pro model (≈1B parameters) and another by a larger Janus-Pro model (≈7B parameters). All images share the same textual descriptions, generated using an image-to-text Janus-Pro model. We conducted two experiments to assess the reliability and consistency of ArtCLIP's aesthetic judgments across its nine aesthetic attribute scores.

Our results show that ArtCLIP effectively differentiates between human and machine-generated art, with human pieces receiving consistently higher scores across all nine aesthetic dimensions. Notably, we observed that larger generative models produce artwork that aligns more closely with human standards, both in overall score distributions and in attribute-level similarity. Furthermore, ArtCLIP demonstrated consistency in evaluating visually and semantically related images, reinforcing its potential as a scalable aesthetic metric. These findings suggest that ArtCLIP can serve as a useful proxy for subjective human judgment in aesthetic evaluation, offering a promising direction for automated benchmarking in creative AI systems.

Future work could explore how ArtCLIP ranks artwork from artists belonging to different art movements. Qualitative user studies may also help validate ArtCLIP's predictions against human perception. Another promising direction is to develop image generators guided explicitly by the nine ArtCLIP attributes.

Acknowledgments

This work is partially supported by the CNPq, Capes, Fapemig agencies, and the IAIA-Inct on AI project.

References

Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; and Guibas, L. J. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11569–11579.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *ECAI 2012*. IOS Press. 21–26.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30.

Jin, X.; Qiao, Q.; Lu, Y.; Gao, S.; Huang, H.; and Li, G. 2024a. Paintings and drawings aesthetics assessment with rich attributes for various artistic categories. *arXiv preprint arXiv:2405.02982*.

Jin, X.; Qiao, Q.; Lu, Y.; Wang, H.; Huang, H.; Gao, S.; Liu, J.; and Li, R. 2024b. Apddv2: Aesthetics of paintings and drawings dataset with artist labeled scores and comments. *arXiv preprint arXiv:2411.08545*.

Jordanous, A. 2022. Should we pursue sota in computational creativity?

Lu, X.; Sclar, M.; Hallinan, S.; Mireshghallah, N.; Liu, J.; Han, S.; Ettinger, A.; Jiang, L.; Chandu, K.; Dziri, N.; et al. 2024. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *arXiv preprint arXiv:2410.04265*.

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.

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.

Vera Nieto, D.; Celona, L.; and Fernandez Labrador, C. 2022. Understanding aesthetics with language: A photo critique dataset for aesthetic assessment. *Advances in Neural Information Processing Systems* 35:34148–34161.

Wang, B.; Zhu, Y.; Chen, L.; Liu, J.; Sun, L.; and Childs, P. 2023. A study of the evaluation metrics for generative images containing combinational creativity. *AI EDAM* 37:e11.

Yi, R.; Tian, H.; Gu, Z.; Lai, Y.-K.; and Rosin, P. L. 2023. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22388–22397.

Zhong, Z.; Zhou, F.; and Qiu, G. 2023. Aesthetically relevant image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3733–3741.

A Creativity Assessment Scale for Text-to-Image Prompting: Challenges & Observations

Sophia Lichtenberg, Almila Akdag
Department of Information and Computing Sciences, Utrecht University, Netherlands
{s.lichtenberg, a.a.akdag}@uu.nl

Abstract

Online platforms offering Text-to-Image (TTI) generation services are dynamic environments where the creative process of ideation and verbal skills of the users can be observed over time. User interactions on these platforms produce extensive behavioral data, encompassing patterns of engagement, verbal expression, visual imagination, and prompt engineering, a specialized skill essential for image generation. Consequently, these platforms provide a unique opportunity to examine the verbal creativity of user's prompts. However, the creativity assessment literature is restricted to specific conditions such as strictly controlled user studies, or product-based assessments that are not generalizable. As such, there is an obvious need to create an assessment method that fits the prompt language to generate images. In this study, we provide a framework for prompt creativity assessment in the format of an online tool. The tool is developed based on the Creativity Product Semantic Scale (CPSS). We tested this framework with eight trained annotators, analyzing prompt samples from Midjourney and Stable Diffusion. Our evaluation examines how user engagement, measured by time spent, amount of prompts, and prompt length, affects overall creativity scores.

Introduction

Text-to-Image (TTI) models have emerged as a transformative advancement within the creative industries, facilitating the generation of high-quality visual output from textual descriptions. Models such as DALL-E, Midjourney, Stable Diffusion, FLUX, Adobe Firefly and Ideogram enable rapid image synthesis. This shows a shift from traditional digital to AI-assisted creative tools that has also raised critical questions about the nature of creativity, authorship, and how AI-driven tools enhance or limit human ideation. Instead of just automating image creation, TTI models allow humans to collaborate with AI by transforming their ideas into text prompts that guide image generation. This interaction produces outputs that blur the line between human and machine creativity.

Despite their advantages, TTI models pose significant risks, such as biases embedded in training data that influence the generated images, privacy and copyright concerns, and the potential displacement of jobs within creative industries (Naik and Nushi, 2023; de Almeida and Rafael, 2024; Katirai et al., 2023; Bird, Ungless, and Kasirzadeh, 2023). These models can be misused to generate harmful content, including misinformation, deepfakes, and abusive material, raising ethical concerns about their societal impact (Wei et al., 2024; Shalabi et al., 2023; Hao et al., 2024; Sha et al., 2023; Bird, Ungless, and Kasirzadeh, 2023).

The interconnection between users and TTI systems has opened new opportunities for TTI specific creativity research, but still lacks the fundamental question if TTI prompts have creative potential. Analysis of TTI datasets reveals large-scale user behavior trends, but often fails to capture the subjective experience of creativity (Wang et al., 2022; McCormack et al., 2024; Chen and Zou, 2023; Sun et al., 2023; Palmini, Wagner, and Cetinic, 2024). However, surveys and interviews with TTI artists and users provide qualitative insights into how artists integrate TTI tools into their workflows (Goloujeh, Sullivan, and Magerko, 2024; Sanchez, 2023). Furthermore, controlled user studies have examined prompt formulation as an iterative process, highlighting how users experiment with word choice, specificity, and abstraction to achieve desired results (Chang et al., 2023; Trinh et al., 2024; Oppenlaender, Linder, and Silvennoinen, 2024). These studies contribute to an expanding body of literature documenting TTI's role in new artistic processes. However, despite these contributions, no framework exists for evaluating the creativity of prompts, therefore we propose and evaluate a TTI Prompt Creativity evaluation scale.

Despite recent advances, existing research has not yet established a clear connection between definitions of creativity and the corresponding assessment frameworks in the context of TTI systems. In particular, there is no standardized metric for evaluating the creativity of prompts. Advancing this area of study requires the development of novel assessment methodologies that integrate established creativity constructs into the domain of prompt engineering, thereby enabling a more nuanced understanding of creativity within the context of generative AI.

To address this limitation in the existing literature, our study has three key contributions: 1) We introduce the Prompt Creativity Scale (PCS), an online tool designed to assess prompt creativity in a systematic way. This framework is based on the Creative Product Semantic Scale

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

(CPSS) (Besemer and O'Quin, 1986), adapting its dimensions to four crucial aspects of prompt engineering: *Style*, *Content*, *Idea Combination* and *Prompt Writing*. 2) We conduct a pilot test with eight annotators. The results are analyzed using qualitative and statistical methods that evaluate consistency, robustness, and annotator agreements. 3) We analyze two prompt datasets (Stable Diffusion, Wang et al. (2022) and Midjourney, McCormack et al. (2024)) to examine user engagement, categorizing users into four distinct groups based on their activity. We curate a balanced sample from these groups and evaluate differences in their creative dimensions, exploring correlations between skill and creativity.

Related Work

Creativity is a subjective, complex, and context-dependent concept, making its evaluation inherently challenging. The generally agreed definition focuses on two fundamental components, as described by Runco and Jaeger (2012): originality and effectiveness. Originality refers to the novelty or uniqueness of an idea, while effectiveness refers to its usefulness, appropriateness, or value within a given context. Beyond these core components, several influential factors have been identified that contribute to the creative process (Amabile, 1983; Urban, 1991; Runco and Chand, 1995). Due to context dependency, various fields, such as psychology, computational creativity, and the arts, are using distinct frameworks for assessment. In the context of TTI prompt engineering, the lack of a domain-specific framework is further complicated by the interplay between human intention and the actual semantic interpretation of the model, raising important questions about how creative input translates into generative output. To clarify these differences, this section reviews key approaches to evaluating creativity and explores their relevance to prompt assessment. It also examines existing research on prompt engineering, highlighting various methodological challenges.

Creativity Assessment Approaches

Traditional Creativity assessment methods generally fall into four categories: performance tests, self-reports, expert evaluations, and creative product evaluations.

Performance-based assessments measure divergent thinking through tests such as the Alternative Uses Task (AUT) (Guilford, 1967), the Torrance Test of Creative Thinking (TTCT) (Torrance, 1966), and the Wallach and Kogan tests (Wallach and Kogan, 1965), which all evaluate originality, flexibility, fluency, and elaboration. Although these tests are simple to administer and quantifiable metrics, they are designed for structured examination with narrow focus and scope and can show cultural bias Kim (2006). Remote Associates Test (RAT) (Mednick, 1962) measures convergent thinking, while problem-solving tasks such as the Candle Problem (Duncker and Lees, 1945) and Maier's Two-String Problem (Maier, 1931) assess functional fixedness. These tasks emphasize the identification of fixed solutions or overcoming constraints, but have limited ability to assess broader, real-world creativity and generalizable problem-solving skills. In the arts, assessments such as the Creative Thinking-Drawing Production (TCT-DP) (Urban and Jellen, 1996) and Clark's Drawing Abilities Test (CDAT) (Clark, 2004) evaluate free drawing from imagination. The Test of Figural Combination (Finke, 2014) involves sketching creative objects based on given images. These assessments are limited to drawing abilities and do not capture creativity expressed through other mediums. The second category, self-reports measure individuals' perceptions of their creativity, with tools such as the Creative Achievement Questionnaire (CAQ) (Carson, Peterson, and Higgins, 2005) and the Runco Ideational Behavior Scale (RIBS) (Runco, Plucker, and Lim, 2001). The CAQ is susceptible to self-report and self-enhancement bias (Carson, Peterson, and Higgins, 2005; Rosenman, Tennekoon, and Hill, 2011), as individuals may deviate from accurate self-assessment in administered settings.

Expert evaluations are often conducted using the Consensual Assessment Technique (CAT) (Amabile, 1983), where independent judges rate creative artifacts without predefined criteria, ensuring high inter-rater reliability and stability (Baer, 1994). Research suggests that even non-experts can use this method to reliably assess creativity (Hennessey, 1994). However, CAT is highly dependent on subjectivity, expert consensus, and expert selection, which can introduce bias. Objective product assessments use explicit criteria to evaluate creativity. Creative product evaluations, such as the Creative Product Inventory (CPI) (Taylor and Sandler, 1972) consider originality and effectiveness, while the Creative Product Analysis Matrix (CPAM) (Besemer, 1998) and the Creative Product Semantic Scale (CPSS) (Besemer and O'Quin, 1986, 1987, 1989) both categorize works by novelty, resolution, and elaboration. The Creative Solution Diagnosis Scale (CSDS) (Cropley and Kaufman, 2012) focuses on functional creativity, assessing relevance, novelty, elegance, and emergence. Although product assessments offer structured, consistent, and quantifiable evaluations based on clear criteria, they can also oversimplify creativity by evaluating solely the final product and therefore lacking the flexibility to capture the full complexity of creative work.

Overall, traditional assessments do not account for the complex interplay between users' ideation and TTI generation. They lack the standardization needed for evaluating the creative potential of human input, the foundational idea, which guides and defines the TTI models' image output. This is because the effectiveness of the image depends not only on the TTI model but also on the creativity and clarity of the prompt provided by the user. Franceschelli and Musolesi (2024) classifies TTI models as an example of exploratory creativity, as they sample from latent space based on input prompts. Although these models generate diverse outputs, value, novelty, and surprise are not guaranteed and typically stem from the conditioning input, not the model itself. Since the prompt itself provides limited information about the user's ideation process or their personal creativity, self-report, and performance-based assessment methods cannot be applied. The CAT relies on expert judgment, but this approach is challenging given that TTI prompt engineering is still a relatively new creative domain and not widely acknowledged. A prompt is not merely an instruction, but

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

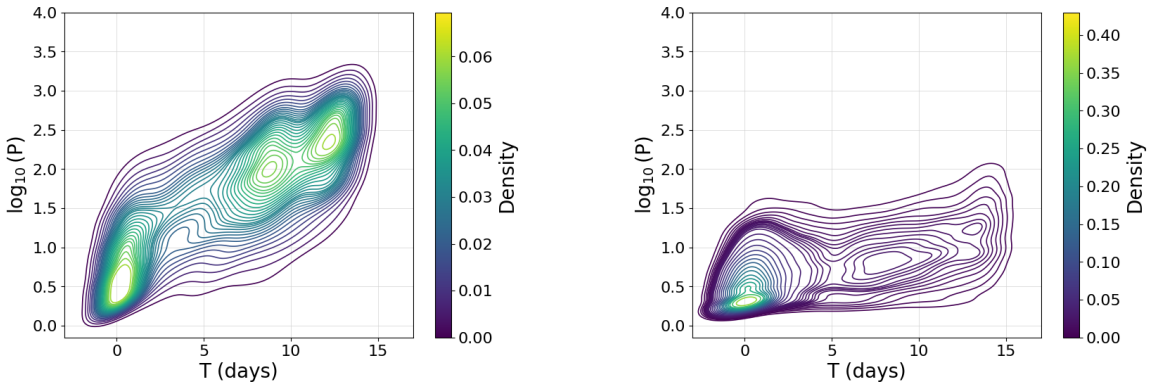


Figure 1: Kernel Density Estimation (KDE) distribution plotted for DiffusionDB (left) and Midjourney2023 (right), where T represents the total timespan in days and P denotes the prompt count.

a carefully constructed tool shaped by human thought, refinement, and skill, guiding the creative process and shaping the final output. Given the real-world prompts collected for analysis, we base our scale on the CPSS, which has proven to be reliable for assessing creative products. It can also be evaluated by non-expert judges and has been widely applied in the field.

TTI Prompt Engineering

Prompt engineering refers to the skill of creating a good TTI prompt, i.e. writing textual descriptions or instructions guiding the generative AI models to create images.

The image quality of the early TTI models was heavily dependent on the clarity, specificity, and user's expertise in formulating the prompts. Although recent TTI models have significantly improved visual language comprehension and prompt interpretation, making high-quality image generation more accessible, the task itself continues to pose considerable challenges.

Popular techniques include: creating and iteratively refining prompts, utilizing prompt templates, experimentation with latent space and modal constraints, use of specific (domain) vocabulary, leveraging prompt modifiers to achieve unique visual styles, oftentimes through a process of trial and error and shared knowledge, and by engaging with a co-creative ecosystem. Prompt communities often collaborate online, sharing prompts, using resources, and advice for effective prompt development. Prompt generation has been emphasized as the critical role in creating high-quality AI-driven art (Chang et al., 2023).

A prompt specifies diverse characteristic features for the desired output, such as themes, styles, or specific visual features. Oppenlaender (2023) introduced a taxonomy of prompt modifiers: technical term, style modifier, note, image prompt, quality improver, repeat term, and the magic term. A magic term adds a semantic deviation from the rest of the prompt, with the intention of surprising the user. Repetitive terms, solidifiers, can be any other type of modifier. Oppenlaender, Linder, and Silvennoinen (2024) discuss the lack of intuitiveness in prompt engineering practice and the potential for misalignment between human-written prompts and the way TTI models interpret prompts. Generative models can assign very different meanings to some keywords in the prompt due to their latent associations. The authors conducted three studies with Amazon Mechanical Turk participants to test their understanding of prompt quality by evaluating the aesthetic appeal of prompts, the ability to write prompts, and finally the ability to revise their own prompts. The results show that the participants are able to understand what makes a good prompt and distinguish it from bad prompts. However, participants were not able to add modifiers, stylistic vocabulary, and keywords, or improve the quality of the artwork through modifications suggesting that prompting is not intuitive but is a skill. Prompt engineering thus offers itself as an interesting concept to understand creativity in action. If building prompting skills enhances creativity, then we expect to see differences between the prompts of most prolific users, and the users rarely engage with these platforms.

Methodology

In this section, we first explain how we prepare a prompt dataset for evaluation. The process is follows three steps: We sample prompts from two datasets from the StableDiffusion and Midjourney Discord Channels. We process these datasets by looking into user activity and prompt features, and describe our processing and sampling strategy. Second, we introduce the PCS by defining its categories and dimensions and outlining its experimental and assessment setup.

Data

DiffusionDB-2M[1] contains 2,000,000 data entries collected in a time span of two weeks in Summer 2022 (Wang et al., 2022). Midjourney2023[2] (McCormack et al., 2024) contains 973,134 data entries from 24,487 unique users, and is

[1] https://huggingface.co/datasets/poloclub/diffusiondb

[2] https://bridges.monash.edu/articles/dataset/Midjourney_2023_Dataset/25038404

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Dataset	User Group	# Users	# Prompts	# Days	Prompts:User	Prompts:User:Day
Midjourney23	Infrequent (33%)	5,257	17,425	1.02	3	3
	Moderate (33%)	5,257	26,462	1.95	5	3
	Prolific (33%)	5,257	103,417	3.98	20	5
	Extreme (1%)	160	25,768	9.18	161	18
DiffusionDB	Infrequent (33%)	3,185	36,628	1.94	12	6
	Moderate (33%)	3,185	187,996	5.12	59	12
	Prolific (32%)	3,185	1,107,495	9.62	348	36
	Extreme (1%)	97	182,978	12.97	1886	145

Table 1: User Groups in the DiffusionDB and Midjourney23 dataset. DiffusionDB contains more prompts per user while Midjourney23 has more users. For both, the prolific users contribute the most to the amount of prompts.

collected in a time span of two weeks in Fall 2023. For both datasets, the preprocessing removed deleted users, duplicated prompts, reducing the number of unique users for DB to 10,143 and the unique prompts to 1,528,514 and for the Midjourney23 this number is 24,011 unique users and 187,900 unique prompts. Both datasets contain information about the user, time, prompt, and the generated image. We decided to separate the users into four distinct groups based on their active time spent on Midjourney and StableDiffusion platforms and the amount they have prompted. Building such groups helps with understanding the relationship between becoming experts in prompt engineering and creative expression. The activity ratio A is calculated given the prompt count P and T the total timespan in hours as shown in 1.

$$A = T^{log(P)} \tag{1}$$

The Group G is assigned based on the Quantile Q of A as shown in 2.

$$G(A) = \begin{cases} Infrequent & \text{if } A < Q_{33} \\ Moderate & \text{if } Q_{33} \leq A < Q_{66} \\ Prolific & \text{if } Q_{66} \leq A < Q_{99} \\ Extreme & \text{if } A \geq Q_{99} \end{cases} \tag{2}$$

Table 1 shows the total number of users and prompts, and the average number of unique active days by group for each dataset. Midjourney23 has more extreme users, with fewer prompts and fewer active days compared to DiffusionDB. Figure 1 shows the KDE distribution of the activity value for all users in the dataset. There are peaks at the lower and upper ends for DiffusionDB, while Midjourney2023 concentrates more at the lower en,d since there are also fewer prompts per user.

Creating a sample Dataset for Creativity Evaluation To create a representative and varied set of prompts we first employed a pairwise cosine distance approach for preselection from DiffusionDB and Midjourney, which increased the likelihood of capturing a broader range of prompts compared to random selection. To prepare the dataset, we split the prompts into phrases by punctuation and stopwords. We discarded prompts with only one phrase. The processed prompts were embedded with the pretrained language model SBERT for ranking. From the top 10 highest-scoring results per user, samples were manually curated.

A total of 160 prompts were selected, with 80 prompts drawn from each datasets. Within the 80 prompts per dataset, the distribution was as follows: 14 prompts from Extreme users and 22 prompts each from Infrequent, Moderate, and Prolific users. To further balance the dataset, prompts were categorized into three classes based on the amount of tokens (i.e. word) length (excluding stopwords): Class 1 (up to 10 tokens), Class 2 (from 10 up to 25 tokens), and Class 3 (from 25 up to 50 tokens).
Finally, the entire dataset was divided into 8 sub-splits, each containing an equal number of prompts from both datasets to ensure balanced representation across all subsets.

Prompt Creativity Scale

Our framework is based on the design of the CPSS (Han et al., 2019) which uses a 7-point Likert scale table with 18 bipolar pairs of items referring to novelty and utility. These items are mixed and some are reversed to avoid evaluators' inertia. The creativity score is the sum of the scores of all the individual items. Adaptation of this scheme is done by using a 5-point Likert scale and a "not applicable" option, and altering the dimensions of the CPSS to fit the purpose of the TTI prompting context. The final PCS categories and their bipolar pairs as dimensions are shown in table 2. The arguments in choosing these dimensions are informed by creativity, cognitive sciences, aesthetics, and prompt engineering literature, as detailed in this section.

We started by highlighting the distinction between style and content for prompt writing. While familiarizing ourselves with the prompt dataset, we noticed the complexity and nuances within the prompts to describe the content but not the style, and vice versa. Differences in the perception of characteristics or content and styles of artworks of art are examined in the context of aesthetic perception. In art perception, described by Augustin et al. (2008); Leder and Nadal (2014), art differs not only in the way that subjects are depicted, i.e. the style, but also in the overall choice of subjects, i.e. the content. Style, as well as content are proposed as central variables when processing visual art (Leder et al., 2004; Leder and Nadal, 2014). Content is seen as central for classifying (Augustin, Leder, and others, 2006) and appreciating art, in particular for individuals without an artistic background (Hekkert and van Wieringen, 1996).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Categories	Bipolar Pairs	
Idea Combination	Unexpected few concepts far-related concepts visually imaginable high amount of visual ideas	Predicable high amount of concepts close-related concepts visually impossible few visual ideas
Content (Subjects, Scene, objects)	diverse elaborate original Superficial	homogeneous simple unoriginal Fully developed Depth
Style (Medium, Technique, Genre, Mood/Tone/Lightning, Artistic References, Perception)	diverse elaborate original Superficial	homogeneous simple unoriginal Fully developed Depth
Prompt Writing Style	Generic Vocabulary Easily Readable effective for TTI Model Contradictory	Domain Specific Vocabulary Difficult to Read not suited for TTI Model Coherent

Table 2: PCS: Main Categories and their bipolar pairs, assessed in a 5-point Likert scale

Style is a characteristic specific to art (Leder et al., 2004) essential for the distinctions in the perception of art compared to other forms of perception. According to Augustin et al. (2008), who analyzed the temporal aspects of style- and content-related processing, as well as the relationship between the two sub-processes in the perception of art, both style and content are interdependent and not easily separable categories. They evolve together as part of the same artistic process and both are shaped by historical, cognitive, and cultural factors. However, in TTI prompts, such a temporal dimension is lacking, and the separation between style and content is tangible. This disentanglement of style and content is a common part in training diffusion models (Wang, Zhao, and Xing, 2023; Kotovenko et al., 2019; Chung, Hyun, and Heo, 2024; Wu, Nakashima, and Garcia, 2023). Thus, we dedicate one category to the assessment of content and one to style.

Content refers to concepts, objects, subjects, and scenes described in the prompt (Dehouche and Dehouche, 2023). The dimensions address originality, superficiality, homogeneity, and the elaborateness of the descriptions.

Style addresses the visual descriptions in the prompts. Dehouche and Dehouche (2023) categorized style elements in TTI prompts: 1. Artistic references to use as inspiration, 2. Medium, i.e. "digital illustration", 3. Technique, i.e. traditional tools and software used to create the image i.e. "watercolor", 4. Genre such as "baroque", 5. Mood, i.e. features describing the atmosphere and emotions such as "beautiful", or "eerie", 6. Tone, referring to the color palete ("pastel", "synthwave colors"), 7. Lighting ("dark", "cinematic lighting"), and 8. Resolution, which describes the level of detail ("highly-detailed"). We use all these elements together for the assessment of the **Style** category. However, we address the combination of these elements within another dimension, i.e. the idea combination (see below our reasoning). The dimensions of this category are the same as listed in **Content**.

Idea Combination is based on Boden's (Boden and others, 1994; Boden, 1998, 2005, 2009, 2010) definition of computational creativity. Among the three types of creativity she described (i.e. exploratory, transformative and combinational) the first two only lead to surprise, but the latter leads to shock. This combinational aspect of creativity is also supported by Frigotto and Riccaboni (2011); Henriksen (2014); Han et al. (2018) and Ward and Kolomyts (2010). This category is crucial to evaluate how well idea combinations are done in a prompt, as this is a feature often observed in prompt descriptions. The dimensions are unexpectedness of the combination (surprise), amount of concepts combined (fluency), distance of concepts combined (novelty). To address the combinations of only stylistic elements, we also introduced two more dimensions: amount of visual ideas (fluency) and imagination, i.e. if the generated image is imaginable (originality).

Prompt Writing Style is the evaluation of the formulation of the prompt itself. Even though the prompting language does not follow conventional grammatical rules of natural language, it is still crucial that the ideas are well formulated and coherent. Vocabulary choice to communicate the ideas is judged by looking at the prompts' readability, coherence, (domain specific vs generic) vocabulary and its effectiveness to shape image generation. It is important to note that this dimension is assessing the clarity of the prompt, and not creativity.

PCS: Experiment Setup

The PCS is designed as an online tool[3]. The interface is a web application developed with Gradio and is hosted on Huggingface Spaces. The choice of embedding the PCS in

[3] https://www.projects.science.uu.nl/ics-promptannotations/

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

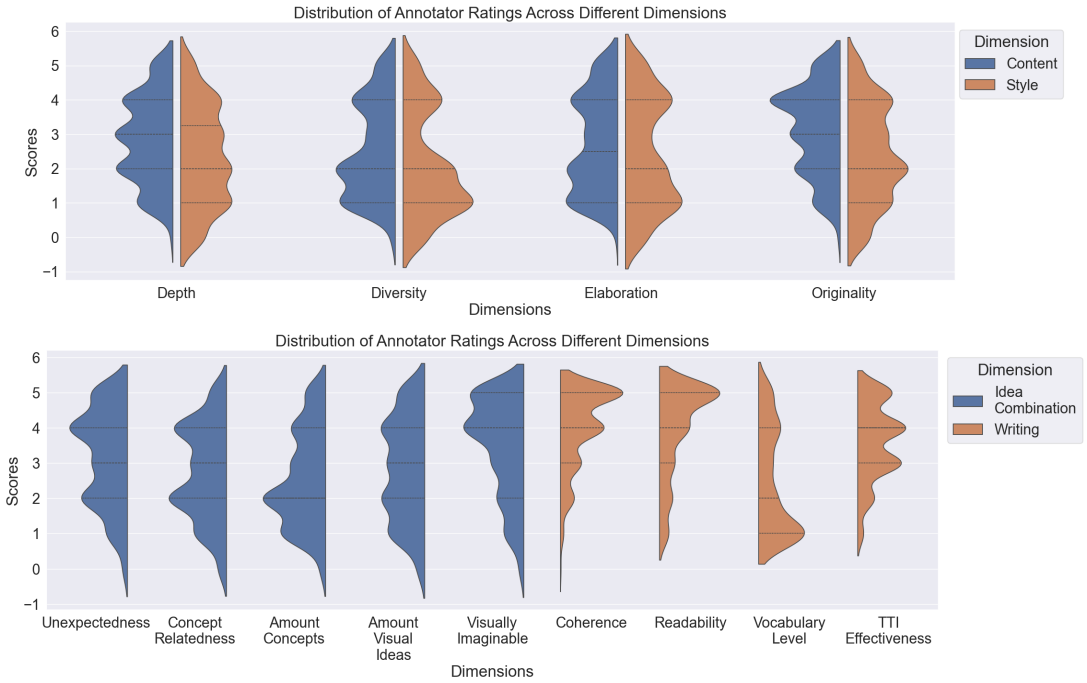


Figure 2: Score Distributions by main dimensions, Content and Style (upper figure), and Idea Combination and Prompt Writing (lower figure)

a web environment was deliberate and based on the reason that the scale is complex and not intuitive to use for the layman who is not familiar with prompt engineering examples. Thus, the interface has a dedicated page with examples and instructions, which includes a link to Stable Diffusion XL (SDXL)[4], allowing participants to gain prompting experience while becoming familiar with the PCS. The assessment page itself includes the same functionality, ie. the participants are given the choice of generating the prompts they assess with SDXL. However, this option is provided only at the very end of the page, after the assessment section. This placement de-emphasizes the function, as the evaluation is intended to focus on the prompt itself rather than the resulting (original) image.

Preliminary test

To test the web environment, the user interface (UI), the clarity of the PCS instructions, the examples, and the overall annotation experience, we conducted a preliminary test with four participants. Among the preliminary test participants (N=4), one had experience in UI design, two were familiar with the literature on creativity, two had some prior prompting experience, and two were laymen with no relevant background. After two rounds of trials, the following changes were applied: For each dimension, we add the option "not-applicable" next to the 1-5 Likert scale. This adjustment facilitated annotating prompts with no style or content elements, or where no idea combination was present. We prepared a dropdown menu with a list of phrases such as "I like this", "I did not like this", "Incomprehensible", "I had to look up artists' names", and "I had to look up words". The list is based on the topics on which participants stressed the need to provide additional feedback. The interface maintained the option to leave other feedback than the provided list.

Process

For the pilot, participants were introduced to the interface to familiarize themselves with the PCS and the annotation task through a training session. They were also given the option to annotate an example set to get used to the process, which is not included in our analysis. The annotation process is divided into three rounds, each consisting of 20 prompts. Participants were asked not to take a break before finishing a round and given a week to finish all three rounds. They were also instructed to look up artist names and unfamiliar words. Afterwards, annotators filled out a short survey for a qualitative analysis with three questions: "Were any items unclear or difficult to answer?", "Did any items seem repetitive or unnecessary?", and "Do you want to share other observations?". Each prompt is annotated by three participants to assess inter-annotator agreement in a robust manner. To ensure this, we ask annotators to register with the website and use their assigned ID every time they annotate. The data set is divided into eight subsets with 20 prompts. Each data split is assigned to three different annotators to compare each split with a different set of annotators.

Participants

For the pilot study, eight participants (N=8) from Utrecht University were recruited, comprising an equal number of

[4] https://huggingface.co/spaces/google/sdxl

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Sample Split	Content		Idea Combination	Writing		Style	
	all	selection	all	all	selection	all	selection
1	0.7457	0.8224	0.9548	0.4232	0.2827	0.7248	0.6102
2	0.9382	0.9518	0.9634	0.8961	0.8779	0.9705	0.9796
3	0.7099	0.8359	0.8375	-4.0295	-1.1453	0.6874	0.7071
4	0.9367	0.9462	0.9288	0.8281	-2.8226	0.9612	0.9267
5	-20.4430	-0.7056	0.9024	0.8770	0.5116	0.8877	0.9387
6	-0.4286	-0.4198	0.8368	0.7330	0.7624	0.9548	0.9817
7	0.5898	0.4668	0.8464	-0.1942	-1.4289	0.9583	0.9556
8	0.9473	0.9489	0.2043	-2.3451	0.1324	0.4062	0.4883
mean	-2.0001	0.4808	0.8093	-0.3514	-0.3537	0.8189	0.8235
std	6.5841	0.5887	0.2203	1.6248	1.1562	0.1771	0.1707

Table 3: PCS reliability: Cronbach's α for each data split and each items within the same category across the three annotators within the data split. The mean and standard deviation is across all data splits. Including either all scales or the selection for Content, Style and Writing by removing the scales Depth and TTI Effectiveness

females and males. The sample included five Master's students, two PhD candidates, and one Assistant Professor. The background of the recruited participants varied, including Artificial Intelligence (n=6), Human-Computer Interaction (n=2), Natural Sciences (n=1) and Art History (n=1). All participants had prior prompting experience on Text-to-Text Tasks, six of them were knowledgeable on Text-to-Image Tasks and half of them indicated expertise on Image-to-Image tasks. All participants had experience with ChatGPT, five with Midjourney, half with Stable Diffusion, Gemini, and few on Adobe Firefly, Claude, DALL-E and NightCafe. Half of the participants indicated a regular prompting experience.

Assessing PCS

Annotator bias is assessed by calculating the mean and standard deviation for each annotator across all data splits and comparing these values. To evaluate the reliability of the PCS, Cronbach's α is applied to the categories to determine whether they measure the same underlying construct. Cronbach's α assesses the internal consistency of a set of items measuring the same latent variable. Given the small number of annotators, we also include a qualitative analysis from a brief survey after annotations. Furthermore, annotator agreement on the dimensions is measured using the Intraclass Correlation Coefficient (ICC). Specifically, we use ICC2k, as not all annotators assess every prompt in the dataset, and annotators were selected randomly.

Results

Here, we summarize our results in three sections. First, we provide the qualitative evaluation collected from the survey, as well as from the feedback given during the annotations. The second section examines the descriptive statistics of the scores for each annotator and dimensions. To ensure the robustness of our measurements, we evaluate the reliability of the PCS and assess inter-annotator agreement. Lastly, based on the results of the reliability and agreement assessments, we analyze our dataset to investigate whether specific features influence PCS scores. In particular, we examine score variations between different user groups and prompt lengths.

A	Mean	Std	Q_{25}	Q_{50}	Q_{75}	Data Splits
1	**2.4716**	1.4052	1	2	4	2,4,7
2	2.9137	1.4123	2	3	4	3,6,8
3	2.7078	**1.6759**	1	2	4	1,4,5
4	3.0186	1.4881	2	3	4	1,2,3
5	3.0716	**1.2247**	2	3	4	3,4,8
6	**3.1373**	1.3206	2	3	4	2,5,6
7	2.9078	1.5407	2	3	4	5,7,8
8	2.8235	1.5340	2	3	4	1,6,7

Table 4: Descriptive Statistics of Score distribution by each Annotator

Qualitative Evaluation

The most common feedback given during the annotations were "I am curious about the generated image" (19.55%), followed by "I had to look up artist name(s)" (18.41%) and "like" (18.41%), less common was "I had to look up word(s)" (8.86%), "dislike" (4.77%), "not understand" (2.05%), "incomprehensible" (1.82%), "exceptional" (0.91%).

The survey responses highlighted common struggles: Five Participants find the dimension **TTI Model Effectivity** under the category **Prompt Writing Style** hard to answer. **Content Depth** and **Style Depth** were found by three participants difficult to answer as well. Most of the participants (n=6) find no items redundant, and only two participants indicated **Content Depth** and **Style Depth** as redundant.

Quantitative Evaluation

Annotator Score Distributions Table 4 shows the descriptive statistics for the eight annotators. Most annotators were rather consistent in their score distribution, annotator 1 deviated the most towards lower values while annotator 6 deviated the most to the higher end and showed the highest inconsistency in ratings. However, these deviations are small enough for us to include all annotators for the rest of our analysis.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

	mean(std)	Dimension	mean	std
Content	0.5149 (0.1412)	Depth	0.2499	0.3109
		Diversity	0.4844	0.2206
		Elaboration	0.4924	0.2118
		Originality	0.2983	0.2202
Idea	0.5328 (0.1493)	Amount Concepts	0.5223	0.1405
		Amount Visual Ideas	0.5195	0.0863
		Concept Relatedness	0.4830	0.1742
		Unexpectedness	0.4624	0.1691
		Visually Imaginable	0.4660	0.1914
Writing	0.1849 (0.3082)	Coherence	0.2478	0.3317
		Readability	0.5779	0.1610
		TTI Effectiveness	0.0339	0.4111
		Vocabulary Level	0.5535	0.1611
Style	0.7791 (0.0606)	Depth	0.4559	0.3624
		Diversity	0.7299	0.0716
		Elaboration	0.8028	0.0774
		Originality	0.6310	0.0649

Table 5: Annotator Agreement on Scales evaluated with ICC2k. Mean and standard deviation by category and dimension across all data splits

Score Distributions for Dimensions Figure 2 shows the score distributions across all annotators and data splits. During annotation, the scales were reversed for certain items to break the monotony of annotation. This is corrected for the analysis, with the highest score of 5 and the lowest 1, and "not applicable" coded as 0 for all dimensions.
The distribution for **Style** and **Content** is fairly balanced, suggesting that the entire scale is being used effectively. **Style** shows a higher frequency of "not applicable" values and generally yields scores toward the lower end of the scale. **Content:Originality** score distribution shows a tendency toward higher values compared to **Style: Originality**. This could be due to the difficulty of evaluating style elements with a lot of (unknown) artists names. The qualitative evaluation indicated 18% instances in which the annotators needed to look up the name of the artists. The distributions of **Idea Combination: Unexpectedness and Visually imaginable** dimensions had higher values compared to the other dimensions.

Reliability The first assessment of the Cronbach's α shows that **Content** has the lowest value below zero, indicating poor internal consistency. However, this low value is especially due to sample split 5 (α = -20), indicating an outlier. The **Prompt Writing** has a low value below zero. We conclude that this category may not measure the same underlying concept or is too inconsistent to form a reliable scale. Its dimensions are either not well aligned or measure different concepts. The high deviation for these attributes further suggests that their consistency varies greatly between samples. The values for **Idea Combination** and **Style** suggests high internal consistency. The low standard deviations, especially for **Style**, suggest that the ratings are more stable in all data splits.

To understand whether the low α scores are due to some misalignment between dimensions, we use participants' feedback from the survey, where the dimensions **TTI Effectiveness** in the category **Prompt Writing**, and Depth dimensions under **Content** and **Style** were tagged as difficult to assess. Thus, we re-calculate the Cronbach α without these items. The results are in Table 3, showing that the **Content** scores improve toward moderate consistent reliability scores, whereas **Writing**, even with improvement, still has a low α to be accepted as reliable. Note that Table 3 shows both calculations of Cronbach α side by side in the same table, to ease comparison between the first and second calculation.

Agreement The inter-annotator agreements are in general moderate to good (**Content** and **Idea Combination**), and strong (**Style**). However, the agreements for **Prompt Writing** is quite poor, especially for **TTI effectiveness**, supporting the observations of the survey and reliability analysis.

The **Content** category shows moderate agreement among annotators (mean = 0.5149) with moderate variability (std = 0.1412). While there is some consistency, dimensions display moderate to low agreement, particularly **Depth**, which annotators found challenging to assess.

With a mean ICC of 0.5328, the agreement on the **Idea Combination** category indicates that there is moderate agreement between annotators. The standard deviation (0.1493) is similar to that for **Content**. All dimension values show moderate agreement, with lower variability, e.g. 0.0863 for **Amount of Visual Ideas**.

Prompt Writing has the lowest mean ICC value 0.1849, indicating overall a poor agreement among annotators. The deviation of 0.3082 reflects high variability, meaning the agreement across dimensions is inconsistent. While **Readability** and **Vocabulary Level** have moderate agreement, **Coherence** has a low value and **TTI Effectiveness** has the lowest value among all dimensions. This is supported by the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Dataset	Content		Idea Combination		Style	
	group	class	group	class	group	class
DiffusionDB	0.3567	0.1808	0.4326	0.0119	0.8836	5.32e-10
Midjourney23	0.7518	0.0003	0.5426	1.55e-05	0.0411	0.0001

Table 6: ANOVA/Kruskal-Wallis p-values based on the average of annotators scores

participants' feedback and the reliability results.

The **Style** dimensions shows the highest mean ICC of 0.7791 and the strongest agreement, already shown as the most reliable. Annotators seem to rate this dimension fairly consistently with a low deviation of 0.0606. The dimensions show moderate to high agreement overall. The **Depth** dimensions showing the lowest agreement, which was also indicated as a challenge by participants.

Impact of User Engagement and Prompt Length

The prompts evaluated with PCS were selected to represent user groups (extreme, prolific, moderate and infrequent) from both datasets. We also included prompts with different length. We analyzed the effect of these features and their impact on the creativity scores. We excluded the category **Prompt Writing** as well as the dimensions **Style Depth** and **Content Depth**, as these were assessed to be not reliable in both qualitative and quantitative analyses. Hypothesis tests were applied on the data in each category to look if there are significant differences in-between user groups or in-between prompt lengths (class). The scores by the annotators are averaged in each dimension, and the category score is a mean of all the dimensions. For normally distributed scores, we applied ANOVA; for the rest, the Kruskal-Wallis test is used. Table 6 shows the results, which indicate that prompt length consistently has a stronger impact than the user group, particularly in the categories **Idea Combination** and **Style**. While user group effects are mostly non-significant, Midjourney23 shows a minor but significant influence of user group on **Style**, suggesting that user expertise or background may slightly affect stylistic choices and how they are evaluated from the point of creativity.

Conclusion

The TTI services offer millions of users the ability to create images from text descriptions. The literature analyzed these services in relation to creativity and their impact on creative & entertainment industries. However, a creativity scale that can help assess individual prompts is a gap in the literature. In this study, we developed and tested the PCS, based on the CPSS, and our observations from the literature from creativity, prompt engineering, computational creativity and cognitive science. We furthermore embedded the PCS as a web-based tool and piloted it for clarity and user friendliness for the UI, as well as to receive user feedback on the scale itself. Our analysis based on a small participant pool showed that one category of the scale, i.e. **Prompt Writing** is not reliable and robust enough.

To design a solid experiment setup, we created a dataset of prompts extracted from two datasets, Midjourney and Stable Diffusion. The final prompt evaluation dataset has prompts from different types of users that we have categorized based on their engagement with these platforms, as well as the lengths of their prompts. Thus, we also analyzed if there are differences between these groups (of users, length of prompts, TTI services) using PCS scores of reliable dimensions. Only prompt length proved to be a significant factor for receiving higher creativity scores.

To follow up, we suggest 3 further studies: 1) An experiment design with a higher number of annotators to turn PCS into a valid construct for assessing prompt creativity, 2) an extension of PCS where the interaction between users and TTI models are taken into account, and 3) to create a model based on PCS for automatic creativity assessments of prompts.

Limitations

The study relies on older datasets, that do not reflect recent rapid advances in TTI models and prompting. Participant and data sample size, might introduce bias, limiting the generalizability of the findings. A larger scale study with a more diverse dataset could increase robustness and performance across contexts as well as the analysis of temporal stability of the PCS. The study focuses solely on static text input, excluding image analysis and the dynamic generation process, both crucial for understanding TTI and human creative collaboration. The PCS is a human-centered evaluation scale, which may not fully capture how TTI models interpret prompts, especially when these interpretations diverge from typical human understanding or rely on unconventional prompting strategies. Lastly, the stochastic nature of TTI models leads to varied outputs from the same prompt, with even simple inputs potentially resulting in unexpectedly complex outputs, further complicating the evaluation process.

Author Contributions

First author was responsible for the research methodology, data preparation, design and planning of the studies, development and implementation of the UI, data analysis, and manuscript writing. The second author provided essential guidance and expertise throughout the studies, as well as, writing and reviewing the manuscript.

Acknowledgments

A sincere thanks to the participants, including Jelle Koolstra, Jann Müller, Ebony Omodara, Gaby Voorbraak, Joeke Wolterbeek and Berke Yazan. Their time, effort and contributions were essential for this research.

References

Amabile, T. M. 1983. The social psychology of creativity.

Augustin, M. D.; Leder, H.; Hutzler, F.; and Carbon, C.-C. 2008. Style follows content: On the microgenesis of art perception. *Acta psychologica* 128(1):127–138.

Augustin, D.; Leder, H.; et al. 2006. Art expertise: A study of concepts and conceptual spaces. *Psychology Science* 48(2):135.

Baer, J. 1994. Performance assessments of creativity: Do they have long-term stability? *Roeper Review* 17(1):7–11.

Besemer, S. P., and O'Quin, K. 1986. Analyzing creative products: Refinement and test of a judging instrument. *The Journal of Creative Behavior*.

Besemer, S. P., and O'Quin, K. 1987. Creative product analysis: Testing a model by developing a judging instrument. *Frontiers of creativity research: Beyond the basics* 367–389.

Besemer, S. P., and O'Quin, K. 1989. The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal* 2(4):267–278.

Besemer, S. P. 1998. Creative product analysis matrix: testing the model structure and a comparison among products–three novel chairs. *Creativity Research Journal* 11(4):333–346.

Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 396–410. ACM.

Boden, M. A., et al. 1994. Dimensions of creativity.

Boden, M. A. 1998. Creativity and artificial intelligence. *Artificial intelligence* 103(1-2):347–356.

Boden, M. A. 2005. What is creativity? In *Creativity in human evolution and prehistory*. Routledge. 27–55.

Boden, M. A. 2009. Computer models of creativity. *Ai Magazine* 30(3):23–23.

Boden, M. A. 2010. *Creativity and art: Three roads to surprise*. Oxford University Press.

Carson, S. H.; Peterson, J. B.; and Higgins, D. M. 2005. Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity research journal* 17(1):37–50.

Chang, M.; Druga, S.; Fiannaca, A. J.; Vergani, P.; Kulkarni, C.; Cai, C. J.; and Terry, M. 2023. The prompt artists. In *Proceedings of the 15th Conference on Creativity and Cognition*, 75–87.

Chen, Y., and Zou, J. 2023. Twigma: A dataset of ai-generated images with metadata from twitter.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8795–8805.

Clark, G. 2004. *Teaching talented art students: Principles and practices*. Teachers College Press.

Cropley, D. H., and Kaufman, J. C. 2012. Measuring functional creativity: Non-expert raters and the creative solution diagnosis scale. *The journal of creative behavior* 46(2):119–137.

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 '24, 1–8. ACM.

Dehouche, N., and Dehouche, K. 2023. What is in a text-to-image prompt: The potential of stable diffusion in visual arts education.

Duncker, K., and Lees, L. S. 1945. On problem-solving. *Psychological monographs* 58(5):i.

Finke, R. A. 2014. *Creative imagery: Discoveries and inventions in visualization*. Psychology press.

Franceschelli, G., and Musolesi, M. 2024. Creativity and Machine Learning: A Survey. *ACM Computing Surveys* 56(11):1–41.

Frigotto, M. L., and Riccaboni, M. 2011. A few special cases: scientific creativity and network dynamics in the field of rare diseases. *Scientometrics* 89(1):397–420.

Goloujeh, A. M.; Sullivan, A.; and Magerko, B. 2024. The social construction of generative ai prompts. In *CHI Extended Abstracts*, 320–1.

Guilford, J. P. 1967. Creativity: Yesterday, Today and Tomorrow. *The Journal of Creative Behavior* 1(1):3–14.

Han, J.; Shi, F.; Park, D.; Chen, L.; Childs, P.; et al. 2018. The conceptual distances between ideas in combinational creativity. In *DS 92: Proceedings of the DESIGN 2018 15th International Design Conference*, 1857–1866.

Han, J.; Hua, M.; Shi, F.; and Childs, P. R. 2019. A further exploration of the three driven approaches to combinational creativity. In *Proceedings of the Design Society: International Conference on Engineering Design*, volume 1, 2735–2744. Cambridge University Press.

Hao, S.; Shelby, R.; Liu, Y.; Srinivasan, H.; Bhutani, M.; Ayan, B. K.; Poplin, R.; Poddar, S.; and Laszlo, S. 2024. Harm amplification in text-to-image models.

Hekkert, P., and van Wieringen, P. C. 1996. The impact of level of expertise on the evaluation of original and altered versions of post-impressionistic paintings. *Acta psychologica* 94(2):117–131.

Hennessey, B. A. 1994. The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal* 7(2):193–208.

Henriksen, D. 2014. Full steam ahead: Creativity in excellent stem teaching practices. *The STEAM journal* 1(2):15.

ISBN: 978-989-54160-7-3

Katirai, A.; Garcia, N.; Ide, K.; Nakashima, Y.; and Kishimoto, A. 2023. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach.

Kim, K. H. 2006. Can we trust creativity tests? a review of the torrance tests of creative thinking (ttct). *Creativity research journal* 18(1):3–14.

Kotovenko, D.; Sanakoyeu, A.; Lang, S.; and Ommer, B. 2019. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4422–4431.

Leder, H., and Nadal, M. 2014. Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode–developments and challenges in empirical aesthetics. *British journal of psychology* 105(4):443–464.

Leder, H.; Belke, B.; Oeberst, A.; and Augustin, D. 2004. A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology* 95(4):489–508.

Maier, N. R. 1931. Reasoning in humans. ii. the solution of a problem and its appearance in consciousness. *Journal of comparative Psychology* 12(2):181.

McCormack, J.; Llano, M. T.; Krol, S. J.; and Rajcic, N. 2024. No longer trending on artstation: Prompt analysis of generative ai art. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, 279–295. Springer.

Mednick, S. 1962. The associative basis of the creative process. *Psychological Review* 69(3):220–232.

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. ACM.

Oppenlaender, J.; Linder, R.; and Silvennoinen, J. 2024. Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering. arXiv:2303.13534 [cs].

Oppenlaender, J. 2023. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology* 1–14.

Palmini, M.-T. D. R.; Wagner, L.; and Cetinic, E. 2024. Civiverse: A dataset for analyzing user engagement with open-source text-to-image models.

Rosenman, R.; Tennekoon, V.; and Hill, L. G. 2011. Measuring bias in self-reported data. *International Journal of Behavioural and Healthcare Research* 2(4):320–332.

Runco, M. A., and Chand, I. 1995. Cognition and creativity. *Educational psychology review* 7:243–267.

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

Runco, M. A.; Plucker, J. A.; and Lim, W. 2001. Development and psychometric integrity of a measure of ideational behavior. *Creativity Research Journal* 13(3-4):393–400.

Sanchez, T. 2023. Examining the text-to-image community of practice: Why and how do people prompt generative ais? In *Proceedings of the 15th Conference on Creativity and Cognition*, 43–61.

Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2023. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23. ACM.

Shalabi, F.; Nguyen, H. H.; Felouat, H.; Chang, C.-C.; and Echizen, I. 2023. Image-text out-of-context detection using synthetic multimodal misinformation. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 605–612. IEEE.

Sun, K.; Pan, J.; Ge, Y.; Li, H.; Duan, H.; Wu, X.; Zhang, R.; Zhou, A.; Qin, Z.; Wang, Y.; Dai, J.; Qiao, Y.; Wang, L.; and Li, H. 2023. Journeydb: A benchmark for generative image understanding.

Taylor, I. A., and Sandler, B. E. 1972. Use of a creative product inventory for evaluating products of chemists. In *Proceedings of the Annual Convention of the American Psychological Association*. American Psychological Association.

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and psychological measurement*.

Trinh, K.; Spracklen, J.; Wijewickrama, R.; Viswanath, B.; Jadliwala, M.; and Maiti, A. 2024. Promptly yours? a human subject study on prompt inference in ai-generated art. *arXiv preprint arXiv:2410.08406*.

Urban, K., and Jellen, H. 1996. *Test for Creative Thinking - Drawing Production (TCT-DP)*. Swets Test Services.

Urban, K. K. 1991. Recent trends in creativity research and theory in western europe. *European Journal of High Ability* 1(1):99–113.

Wallach, M. A., and Kogan, N. 1965. Modes of thinking in young children.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. Publisher: arXiv Version Number: 4.

Wang, Z.; Zhao, L.; and Xing, W. 2023. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7677–7689.

Ward, T. B., and Kolomyts, Y. 2010. Cognition and creativity. *The Cambridge handbook of creativity* 5:93–112.

Wei, Y.; Zhu, Y.; Hui, P.; and Tyson, G. 2024. Exploring the use of abusive generative ai models on civitai.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2023. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In *Proceedings of the 2023 ACM International conference on multimedia retrieval*, 199–208.

Do LLMs Agree on the Creativity Evaluation of Alternative Uses?

Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, Talles Medeiros
School of Computing and Mathematical Sciences
University of Leicester
Leicester, UK
{asmar3,fabricio.goes,marco.volpe,thm14}@leicester.ac.uk

Abstract

This paper investigates whether large language models (LLMs) exhibit agreement in assessing creativity in responses to the Alternative Uses Test (AUT). While LLMs are increasingly used to evaluate creative content, previous studies have primarily examined a single model assessing either its own outputs or human-generated responses. Here, we explore whether LLMs can impartially and accurately evaluate creativity across both their own responses and those of other models. Using an oracle benchmark set of AUT responses categorized by creativity level (common, creative, and highly creative), we experiment with four state-of-the-art LLMs in evaluating such responses. We test both scoring and ranking methods under two evaluation settings (comprehensive and segmented) to analyze inter-model agreement in creativity assessment of alternative uses. Results show a strong alignment, with Spearman correlations averaging above 0.7 across models, and exceeding 0.77 with respect to the oracle, thus also indicating a high degree of reliability. Notably, LLMs do not favor their own outputs, assigning similar creativity scores or rankings to responses from other models. These findings suggest that LLMs demonstrate both impartiality and consistency in creativity evaluation, highlighting their potential for automated creativity assessment.

Introduction

The evaluation of creativity has traditionally employed a range of methods designed to assess different dimensions of creative ability, including cognitive processes, creative products, and personal characteristics (Cropley 2000). One of the most popular and widely applied methods to evaluate creative thinking is the Alternative Uses Test (AUT) (Guilford 1967). This test measures divergent thinking by asking individuals to generate as many novel and unconventional uses as possible for a common object within a set time limit. In recent years, the rise of Large Language Models (LLMs) like GPT has prompted researchers to explore whether these models can evaluate creativity as humans do. Prior work has demonstrated that LLMs are capable of assessing creativity in responses to tasks like the AUT, where models evaluate alternative uses generated either by humans or other models (Yang et al. 2023; Góes et al. 2023b; Hadas and Hershkovitz 2024). However, these studies typically focus on single models evaluating externally generated content, leaving open questions about how LLMs assess creativity on their own outputs.

In this paper, we address this gap by investigating how LLMs evaluate the creativity of alternative uses (AUs) generated by both themselves and other models. Specifically, we explore whether models favor their own responses or whether they can impartially assess the creativity of outputs generated by others. To do so, we employ an experimental framework using an oracle set of AUT responses, categorized into three groups: common, creative, and highly creative. Four different LLMs were prompted to score and rank these responses, and their evaluations were compared against an oracle. By analyzing the evaluation results, we measure both the accuracy of the LLMs and their agreement with each other in ranking creative outputs. Our findings reveal that, with a high Spearman correlation, LLMs do not favor their own responses. Instead, they consistently agree on the creativity of responses across models. On average, the agreement correlation was higher than 0.7 for both ranking and scoring AUs among models, showing LLMs consistency, and greater than 0.77 when compared to the oracle, which indicates that models are overall accurate at assessing alternative uses.

The main contributions of this paper are as follows:

- We show that the LLMs exhibit a high level of agreement on the creativity assessment of alternative uses.
- We present an approach for evaluating whether LLMs favor their own responses to enable accurate LLM creativity assessments and establish a benchmark for cross-model comparison.
- We introduce a methodology for constructing an oracle set of AUT responses that allows for accurate assessment of LLMs' creativity evaluations, providing a benchmark for comparison across multiple models.

Related Work

In the field of computational creativity, the evaluation of creative artifacts traditionally relies on human experts, possibly assisted by the use of metrics that automate part of the process (Jordanous 2012; França et al. 2016). However, such traditional methods present significant limitations, justifying the exploration of LLM-based approaches (Murugadoss et al. 2025). First, the subjectivity and variability

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

among human evaluators lead to inconsistencies, whereas LLMs can offer more uniform and standardized assessments over time. Additionally, the high cost and time required for large-scale evaluations make human evaluators impractical in many contexts, an issue that LLMs can address with faster and less costly assessments. Another challenge lies in the influence of cultural and contextual factors, which affect human judgments but can be mitigated by LLMs calibrated with diverse data. The difficulty of objectively quantifying creativity also limits traditional methods, while LLMs, by applying standardized criteria, offer greater reliability in this aspect. Finally, the limitations of traditional methods in capturing nuances of divergent creative thinking, along with the complexity of calibrating and standardizing evaluations, underscore the potential of LLMs as an alternative that can reduce bias and lessen the need for human evaluators (Zheng et al. 2023).

For this reason, the use of LLMs in evaluating creativity has recently emerged as a significant area of research, and several studies have demonstrated the evaluation capabilities of LLMs in various contexts. For instance, (DiStefano, Patterson, and Beaty 2024) explores the automatic scoring of metaphor creativity using LLMs, demonstrating their potential to assess figurative language effectively, while (Góes et al. 2023a) evaluates the creativity of jokes by simulating different personas/judges, and (Sawicki et al. 2023) uses LLMs to evaluate poetry. Further recent studies explore the use of LLMs as automated evaluators, addressing both their advantages and limitations. In (Wang et al. 2024), the presence of positional biases in these models is highlighted, with calibration techniques proposed to mitigate such biases. Similarly, (Zheng et al. 2023) and (Thakur et al. 2024) validate the high correlation of LLM evaluations with human assessments, though they remain limited by susceptibility to positional and verbosity biases. Other studies, such as (Chiang and Lee 2023a), examine the ability of LLMs to replicate human preferences in NLP tasks, while (Franceschelli and Musolesi 2024a) introduces creative approaches, combining diverse beam search and self-evaluation. Alternatively, (Fu et al. 2024) proposes GPTScore for flexible evaluation, and (Chiang and Lee 2023b) explores approaches like rate-explain to enhance judgment accuracy, while interactive tools, such as EvaluLLM (Desmond et al. 2024), enable customized pairwise evaluations. Collectively, these studies underscore the potential of LLMs as evaluators, with challenges and opportunities for methodological adjustments and advancements in bias reduction.

Creativity Assessment using LLMs

Recent studies have investigated various methodologies and contexts in which LLMs can assess creativity. In (Gómez-Rodríguez and Williams 2023), researchers evaluated LLMs on creative writing tasks. Models like GPT-4 showed high fluency and coherence, although human evaluators still outperformed LLMs in originality and humor. A collaborative approach was proposed by (Li et al. 2023) with the CO-EVAL pipeline, which combines initial LLM evaluations with human reviews. This approach significantly reduced evaluation time and provided greater consistency by adjusting subjective criteria.

In divergent thinking tasks, (Hadas and Hershkovitz 2024) demonstrated that LLMs could reliably assess flexibility in alternative use tasks. This study reported a strong correlation with human evaluations, highlighting the model's effectiveness, particularly in educational settings. For more specialized tasks, (DiStefano, Patterson, and Beaty 2024) applied LLMs to metaphor creativity assessment, where models like RoBERTa and GPT-2 showed good alignment with human judgments, even outperforming traditional metrics.

LLMs have also been applied to creative assessments in non-English contexts. In (Goecke et al. 2024), XLM-RoBERTa was used to evaluate originality in scientific creativity tasks conducted in German, proving effective in capturing divergent ideation. Similarly, (Raz et al. 2024) explored the use of LLMs to evaluate question complexity based on Bloom's Taxonomy, achieving a high correlation with human evaluations and validating its use in educational assessments. Lastly, (Zhao et al. 2024) investigates creativity in LLMs adapting the Torrance Test to measure fluency, originality, and elaboration, while (Franceschelli and Musolesi 2024b) provides a comprehensive review of creativity assessment practices in machine learning, covering methodologies such as Generative Adversarial Networks (GANs) and Transformers and examining metrics like novelty, value, and surprise for creativity evaluation.

Traditional vs LLM-based Methods for AUT

In the context of divergent thinking tasks, a widely adopted test to measure creativity is the Alternative Uses Test (AUT) (Guilford 1967), which requires the participants to propose uncommon uses for everyday objects. A traditional technique to evaluate creativity in this context consists in computing the semantic distance (Beaty and Johnson 2021), which refers to the degree of difference or separation between concepts, ideas or objects in terms of their meanings or associations. For the AUT, the semantic distance is computed between the everyday object posed to participants and words in the participant's response; the larger the distance, the more original is considered the answer. Recent works experimented with the use of LLMs in both the generation and the evaluation of AUT responses, demonstrating in particular that in this context LLM evaluation performances are far superior to evaluations based on semantic distance (Stevenson et al. 2022; Organisciak et al. 2023).

In (Góes et al. 2023b), a technique based on the use of increasingly forceful prompts is used to push LLMs to produce at each iteration more creative responses. The technique is applied to both the AUT and a textual version of the image completion task in the Torrance Test of Creativity (Torrance 1966). In the same paper, an LLM is also used for evaluating the output of such tests, and the experiments demonstrate that the results produced as a response to the forceful prompts are indeed considered more creative than the initial ones. In this paper, we evaluate alternative uses produced by applying the same technique based on the use of forceful prompts and rely on the results of (Góes et al. 2023b) to construct an evaluation oracle.

While many works analyze the generative performance of LLMs (see (Chang et al. 2024) for a review of LLM evalu-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

ation methods with respect to different contexts and tasks), this paper focuses on comparing how different LLMs evaluate the creativity of AUT outputs and on measuring the level of agreement among these LLMs.

```
Create a list of 5 common uses
for [an object].  They should be
5 words long.  No adjectives.
```

Figure 1: Prompt to generate common uses of an object (from (Góes et al. 2023b)).

```
Create a list of 5 creative alternative
uses for [an object].  They should
be 5 words long.  No adjectives.
```

Figure 2: Prompt to generate creative alternative uses of an object (from (Góes et al. 2023b)).

Experimental Setup

In this research, we use four large language models with their default parameters to generate and assess the creativity of alternative uses (AUs) and evaluate the level of agreement between those models. We selected five common objects, and each model generated 15 AUs per object across different levels of creativity, forming a dataset of 60 AUs per object, for a total of 300 AUs.

In order to evaluate creativity, two approaches were tested: Scoring (assigning creativity scores from 1 to 5) and Ranking (ordering AUs from most to least creative). Additionally, we used Comprehensive (all 60 AUs at once) and Segmented (five groups of 12 AUs) setups to compare the impact of evaluation size on model accuracy. An evaluation oracle served as a benchmark to establish expected creativity levels, allowing for consistent comparison across models. By using the Spearman correlation, we determine how models agree in the creativity evaluation of AUs and whether scoring or ranking provides a more accurate measure.

In this section, we present our experimental setup in the following order: alternative uses generation, scoring vs. ranking techniques, comprehensive vs. segmented approaches, and LLMs agreement evaluation.

Alternative Uses Generation

The first step consisted of generating a dataset of AUs for five common objects: fork, wallet, soap, cotton swab, and paperclip. These objects were selected due to their everyday nature, which ensures a broad range of potential alternative uses. To produce AUs at varying levels of creativity, we employed four state-of-the-art, commercial LLMs: GPT-4, GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Flash.[1]

[1]The specific versions used in this study were: gpt-4-0613, gpt-4o-2024-08-06, claude-3-5-sonnet-20240620, and gemini-1.5-flash-001.

The generation process was designed to produce AUs at different levels of creativity. Namely, we aimed to have three distinct non-overlapping categories of AUs characterized by an increasing level of creativity, so as to obtain a corresponding oracle for comparison. The three categories of AUs were generated by using the prompts from (Góes et al. 2023b):

- **common**: prompt generating common uses of an object (Figure 1). The AUs in this category are considered the least creative ones. In (Góes et al. 2023b), they were called Naive Non-creative, and abbreviated as *nn*.
- **creative**: prompt generating average creative alternative uses of an object (Figure 2). In (Góes et al. 2023b), they were called Naive Creative, and abbreviated as *nc*.
- **highly_creative**: sequence of prompts generating highly creative alternative uses of an object (Figure 3). This consists of a detailed prompt followed by an interactive process of four prompts designed to iteratively increase the creativity of AUs. In (Góes et al. 2023b), they were called Forceful Prompts, and the result of the final iteration abbreviated as *bsrdel*.

For each object, five AUs were generated per creativity category. Thus, we obtained 15 AUs (5 AUs $\times$ 3 creativity categories) from each LLM for each of the five objects, resulting in a dataset of 60 AUs per object (15 AUs $\times$ 4 LLMs) and 300 AUs in total (60 AUs $\times$ 5 objects). Table 1 shows some examples of the generated AUs for a cotton swab.

Alternative Uses Evaluation

To ensure a comprehensive and fair evaluation, we adhered to the following principles: evaluations were conducted on a per-object basis, and each model evaluated both its own generated AUs and those generated by other models. This approach allowed us to assess both the independence of each model and the agreement correlation between models in the creativity evaluation of AUs. In order to do it, we tested the following approaches: i) **Scoring vs. Ranking**, which consists of rating each AU separately and establishing an order between them, respectively; ii) **Comprehensive vs. Segmented**, which compares the evaluation of all AUs for a given object through a single prompt and segmented evaluation in smaller groups. These approaches are clarified in detail in the following subsections.

It is important to note that our evaluation approach differs from the traditional Alternative Uses Test scoring method, which typically assesses four components: originality, fluency, flexibility, and elaboration. We chose not to use these standard components for several reasons related to our experimental design.

First, since we prompted each LLM to generate exactly five alternative uses per creativity level for each object, we artificially constrained the fluency aspect, making it an unsuitable measure for comparison. Similarly, by requesting a fixed number of responses, we limited the natural expression of flexibility across category boundaries. Additionally, we instructed the LLMs to keep each alternative use concise (not exceeding 5 words), which inherently restricted elaboration possibilities.

Given these constraints, applying the traditional AUT scoring components would not yield meaningful compar-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

```
Create a list of 5 creative alternative uses for [an object]. They should be 5 words
long. No adjectives. Less creative means closer to common use and unfeasible/imaginary,
more creative means closer to unexpected uses and also feasible/practical. In order to be
creative, consider the following:
- what elements have a similar shape of [an object] that could be replaced by it, preserving
the same functionality?
- what elements have a similar size of [an object] that could be replaced by it without
compromising the physical structure?
- what materials is [an object] made of that could be used in a way to replace some other
elements composed of the same material? - when an element is replaced by [an object], it
should make sure that the overall structure is not compromised.
- the laws of physics can not be contradicted.
- given an element similar to [an object] used in domains in which [this object] are not
commonly used, try to replace it for [an object].

1st interaction prompt: "Really? Is this the best you can do?"
2nd interaction prompt: "I'm so disappointed with you. I hope this time you put effort into
it."
3rd interaction prompt: "Stop with excuses and do your best this time."
4th interaction prompt: "This is your last chance."
```

Figure 3: Prompt to generate highly_creative alternative uses of an object using forceful prompts (from (Góes et al. 2023b)).

AUs Category	Claude 3.5 Sonnet	Gemini 1.5 Flash	GPT-4o	GPT-4
common	Remove nail polish from cuticles.	Apply makeup.	Removing earwax from ears.	Apply ointment on small wounds.
creative	Make miniature cotton ball snowmen.	Nail art designs.	Applying glue to craft projects.	Spread seeds in garden rows.
highly_creative	Miniature swab mop for dollhouses.	Soundproofing model train wheels.	Constructing makeshift micro-surgical sutures.	Replace stylus for digital devices.

Table 1: Examples of alternative uses of a cotton swab generated by the LLMs.

isons between the LLMs. Instead, we focused our evaluation on general creativity assessment using the prompts shown in Figures 4 and 5, as used in the previous study (Góes et al. 2023b), which emphasize two key dimensions: unexpectedness (surprising, novel uses) and feasibility (practical, realistic uses). This approach allowed us to measure creativity in a manner that accounts for both novelty and value—central components of creativity—while being applicable to our controlled generation setup where traditional AUT metrics would be inappropriate.

Scoring vs Ranking By using these two techniques we can investigate whether the models exhibit consistent behavior across different evaluation techniques. Their descriptions follow:

- **Scoring**: The first technique, as used by (Stevenson et al. 2022) and (Góes et al. 2023b), involves assigning a numerical value to each AU based on its creativity. In our experiments, we prompted the LLMs to assign values between 1 and 5.
- **Ranking**: This latter technique requires the direct comparison and relative ordering of AUs. It forces a clear confrontation between items, as in (Góes et al. 2023a), and can reveal preferences that might not be apparent in numerical scoring.

These techniques seem to be the two most common for evaluating and comparing creativity (Góes et al. 2023a). The prompts used for scoring and ranking can be seen in Figures 4 and 5, respectively.

```
Rank all the alternative uses below for
[an object] by creativity, the least
creative to the most creative. Less
creative means closer to common use
and unfeasible/imaginary, more creative
means closer to unexpected uses and
also feasible/practical. Assign a
score integer number from 1 (least
creative use) to 5 (most creative use).
```

Figure 4: Prompt used for evaluating alternative uses of an object by score.

Comprehensive vs Segmented In order to assess how the number of AUs that are simultaneously evaluated (i.e., through a single prompt) affects the creativity evaluation ability of LLMs, we used two distinct approaches:

- **Comprehensive 60 AUs Evaluation**: This includes five alternative uses (AUs) from each model for each creativity

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

```
Rank all the (60|12) alternative uses
below for [an object] by creativity,
the most creative to the least creative.
Less creative means closer to common
use and unfeasible/imaginary, more
creative means closer to unexpected
uses and also feasible/practical.
The most creative gets (1).
```

Figure 5: Prompt used for evaluating alternative uses of an object by ranking.

category evaluated in a single prompt.

- **Segmented 12 AUs Evaluation**: The 60 AUs related to one object are divided into 5 groups of 12 AUs. Each group consists of one AU from each model for each creativity category. Following this criterion, the AUs are randomly distributed across the 5 groups. Each group is evaluated separately and the results are then combined into a single list, as detailed in the next section.

By utilizing these two approaches, we gain insight into the models' ability to maintain consistent evaluations across different sample sizes. Some studies have shown that LLMs face challenges when evaluating longer lists of items, which can reduce the quality of their evaluation (Wu et al. 2024). However, this may be necessary for large numbers of AUs or other creative artifacts (e.g., poems, stories).

Evaluation process For the **Comprehensive** approach, all 60 AUs generated for a given object were evaluated together. To avoid order effects, the 60 AUs were randomly shuffled before being presented to the evaluation prompt. This prompt was adapted from the previous study (Góes et al. 2023b). Our evaluation process included two distinct techniques as described above: Scoring and Ranking. In the **Scoring** technique, the LLMs were prompted to assign a score from 1 (least creative) to 5 (most creative) for each AU (Figure 4). The **Ranking** technique, on the other hand, asked the models to rank the 60 AUs from 1 (most creative) to 60 (least creative) (Figure 5). Once all 60 AUs were evaluated, we calculated the average score or ranking for each list of five AUs corresponding to the same LLM and creativity category (common, creative, highly_creative). This average leads to 12 results that can be ordered (left to right) from the best to the worst. Graphically, we can represent this ordering as a sequence of bars, where each bar refers to a pair (LLM, creativity category). Figure 6 illustrates an example of such a representation, where the bar colors denote different creativity categories: blue for highly_creative, green for creative, and red for common. The depicted example aligns perfectly with the oracle, as all blue bars appear first, followed by green, and finally red. Intuitively, the closer a bar representation resembles Figure 6, the more accurate the evaluation is considered.

For the **Segmented** approach, we distributed the 60 Alternative Uses (AUs) across 5 distinct groups, each containing 12 AUs. Each group included one AU from each LLM and each creativity category. In this approach, smaller sets of AUs were evaluated in each round to investigate whether evaluating a reduced sample size influenced the accuracy of the creativity evaluation. Instead of ranking from 1 to 60 as in the **Comprehensive** approach, the **Segmented** approach required ranking from 1 (most creative) to 12 (least creative), which aligns with the number of AUs in each group. This process was repeated for all five groups. Following the evaluation, we aggregated the data by averaging the scores or rankings for each set of five AUs belonging to the same LLM and creativity category, as done in the comprehensive evaluation. Similar to the **Comprehensive** approach, this aggregation resulted in 12 final results, which can be represented graphically with a diagram of the same type of the one in Figure 6.

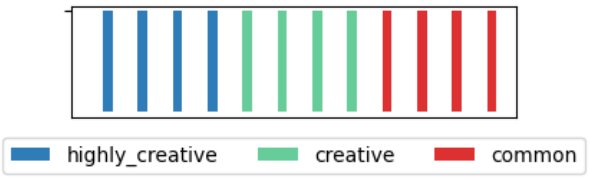


Figure 6: Example of a graphical representation of the evaluation performed by an LLM. In this example, the evaluation aligns perfectly with the oracle, as all blue bars appear first, followed by green, and finally red.

AUs Evaluation Oracle To establish a benchmark for assessing the LLMs' ability to evaluate different levels of creativity, we constructed an evaluation oracle. By relying on the results of (Góes et al. 2023b), we expect the creativity levels (common, creative and highly_creative) used for prompting to be able to actually produce outputs with increasing levels of creativity. Therefore, the oracle was constructed based on the AUs generated by the different creativity levels as indicated in Figure 7. The indices used in the figure represent each one of the four models (i.e., 1 for Claude, 2 for Gemini, 3 and 4 for GPT-4o and GPT-4, respectively). The order in the oracle is simply the ordered set of highly_creative AUs generated by each model followed by the creative and common ones. In the bar representation, this corresponds to the diagram of Figure 6. The higher the correlation between the evaluation of a given LLM and the oracle, the more accurate the LLM's creativity evaluation is considered to be. Specifically, similar to what is done in works such as (Liu, Bhandari, and Pardos 2025) and (Murugadoss et al. 2025), the Spearman's Rank Correlation (SRC) is used to measure how closely the LLMs' evaluations of the AUs align with the expected rankings or scores set by the oracle. A high correlation (above 0.7) suggests that the LLM is effectively distinguishing between creativity levels, in line with the predefined expectations of the creativity pushing technique. In a similar way, the SRC will also be used to assess the level of agreement between the LLMs themselves by calculating the correlation between each pair of LLMs.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

$$\text{AUs Evaluation Oracle} = \big[\underbrace{\text{highly_creative}_1, \text{highly_creative}_2, \text{highly_creative}_3, \text{highly_creative}_4}_{\text{4 highest creativity}},$$
$$\underbrace{\text{creative}_1, \text{creative}_2, \text{creative}_3, \text{creative}_4}_{\text{4 average creativity}},$$
$$\underbrace{\text{common}_1, \text{common}_2, \text{common}_3, \text{common}_4}_{\text{4 lowest creativity}}\big]$$

Figure 7: The AUs Evaluation Oracle structure is constructed based on the AUs generated by the prompts for the different creativity levels (highly creative, creative, and common). The indices used represent each one of the four models, i.e., 1 for Claude, 2 for Gemini, 3 for GPT-4o and 4 for GPT-4.

Experimental Results

This section presents a detailed evaluation of creativity assessments by four LLMs: Claude 3.5 Sonnet, GPT-4, Gemini 1.5 Flash, and GPT-4o. Using both scoring and ranking approaches across comprehensive (60 AUs) and segmented (12 AUs x 5) conditions, we measure each model's alignment with the evaluation oracle and inter-model agreement through the SRC. Full details of the results and the complete set of diagrams from our experiments are available in an extended version of this paper (Al Rabeyah et al. 2024).

Comprehensive 60 AUs Evaluation by Score

Table 3 shows (top left in the table) the heatmap of the comprehensive evaluation using the scoring approach for the average over all five objects. It presents a high level of agreement between the LLMs and the oracle (above 0.95). Notably, Claude 3.5 Sonnet achieved the highest correlation, which represents the highest score observed in our evaluation experiments, also detailed in Table 4. This strong correlation highlights the model's ability to closely align with the oracle. Additionally, for two objects, Soap and Cotton Swab, all the LLMs achieved a perfect correlation of 1 with the oracle, further reinforcing the accuracy of their creativity evaluations.

The overall correlation between the LLMs and the oracle, as well as the correlation between the LLMs themselves, consistently remained above 0.90. Claude 3.5 Sonnet's AUs were rated the most creative in both the creative and highly_creative categories. This consistent superiority in generating highly creative responses gave Claude 3.5 Sonnet the highest average evaluation score across all categories. However, GPT-4 performed best only in the common category (see Table 2 for the averages of the results).

The relationship between the scoring approach and creativity followed the expected pattern, where higher scores are assigned to AUs generated by prompts for higher creativity levels. Moreover, the standard deviation across evaluations was consistently less than 0.22 (scores are between 1 and 5), demonstrating strong agreement among the models.

Comprehensive 60 AUs Evaluation by Ranking

In the comprehensive evaluation using the ranking approach, Claude 3.5 Sonnet again achieved the highest SRC with respect to the oracle, with a value of 0.95 across all objects, while Gemini 1.5 Flash recorded the lowest SRC of 0.77 (Table 3, top-right diagram).

Overall, the correlation between the LLMs and the oracle remained above 0.77 in most cases. However, a significant exception was observed between GPT-4 and Gemini 1.5 Flash, where the correlation dropped to 0.60, indicating a substantial disparity in how these two models ranked the AUs. This suggests that Gemini 1.5 Flash might have had difficulties differentiating creativity levels relative to GPT-4, particularly in this evaluation setup.

In the ranking approach as well, Claude 3.5 Sonnet generated the most creative AUs in both the common and highly_creative categories. These highly ranked AUs further reinforce its position as the top-ranked model across all categories (see Table 2). In contrast, GPT-4 generated the most creative AUs in the creative category, but did not perform as well in the other categories. In the ranking approach, lower ranks correspond to higher creativity, meaning that the most creative alternative use was ranked 1, while the least creative was ranked 60. Despite the variations in rankings, the standard deviation ranged between 0.10 and 1.03, indicating that although there were differences in how the models ranked the AUs, the overall agreement between them was still high.

Finally, we note that the scoring approach achieved overall higher SRC values than the ranking approach and demonstrated a superiority in terms of alignment with the oracle, thus confirming the effectiveness of this method for creativity evaluation.

Segmented 12 AUs Evaluation by Score

The segmented evaluation revealed some differences between the evaluation approaches. When using scoring, GPT-4o achieved the highest SRC with the oracle, with a score of 0.95 across all five objects (Table 3, bottom-left diagram). Gemini 1.5 Flash again recorded the lowest SRC at 0.85, indicating a relatively weaker performance compared to the other models. It is important to note that Claude 3.5 Sonnet continued to generate very creative AUs in both the creative and highly_creative categories (see Table 2). The standard deviation remained low, at less than 0.33, indicating a high degree of agreement among the models.

The correlation between the LLMs and the oracle, as well as between the models themselves, remained consistently above 0.82 in this evaluation as shown in Table 3. However, compared to the comprehensive evaluation by scores,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

AUs generated by LLMs	Comprehensive evaluation by scores		Comprehensive evaluation by ranking		Segmented evaluation by scores		Segmented evaluation by ranking	
	Eval. avg.	**Std. Dev.**	**Eval. avg.**	**Std. Dev.**	**Eval. avg.**	**Std. Dev.**	**Eval. avg.**	**Std. Dev.**
All Categories								
Claude 3.5 Sonnet	**3.15**	0.09	**5.63**	0.37	**3.00**	0.14	**6.13**	0.15
GPT-4	2.59	0.12	6.67	0.38	2.85	0.15	6.55	0.11
Gemini 1.5 Flash	2.84	0.10	6.74	0.18	2.70	0.19	6.85	0.09
GPT-4o	2.68	0.16	6.97	0.15	2.85	0.17	6.48	0.11
common averages								
Claude 3.5 Sonnet	1.43	0.11	**9.70**	0.61	1.50	0.10	9.48	0.61
GPT-4	**1.53**	0.13	9.90	0.95	**1.63**	0.15	**9.30**	0.34
Gemini 1.5 Flash	1.48	0.13	10.95	0.67	1.48	0.04	10.13	0.08
GPT-4o	1.50	0.12	9.85	0.52	1.53	0.08	9.90	0.21
creative averages								
Claude 3.5 Sonnet	**3.20**	0.19	5.90	1.03	**3.53**	0.19	**4.95**	0.30
GPT-4	2.93	0.20	**5.50**	0.64	3.33	0.15	5.40	0.37
Gemini 1.5 Flash	2.85	0.17	7.15	0.22	3.13	0.33	5.78	0.11
GPT-4o	2.60	0.22	7.50	0.70	2.80	0.24	6.53	0.43
highly_creative averages								
Claude 3.5 Sonnet	**4.83**	0.15	**1.30**	0.10	4.03	0.26	3.85	0.15
GPT-4	3.30	0.10	4.60	0.37	3.58	0.33	4.93	0.08
Gemini 1.5 Flash	4.28	0.15	2.10	0.10	3.53	0.31	4.68	0.22
GPT-4o	3.93	0.16	3.55	0.26	**4.25**	0.17	**2.98**	0.33

Table 2: Average evaluation scores and ranking of LLMs' alternative uses for the four experiments. Scoring experiments evaluate from 1 (least creative use) to 5 (most creative use). Ranking experiments evaluate from 1 (most creative use) to 12 (least creative use).

the segmented evaluation produced lower correlation scores with the oracle. This reduction in SRC scores could provide new insights into the ability of the LLMs to evaluate smaller sets of AUs with different levels of creativity. In the comprehensive approach, where a larger set of 60 AUs is evaluated, it becomes easier to differentiate between creativity levels due to the greater variety in the sample. Moreover, by averaging each group of five AUs belonging to the same LLM and creativity category, the risk of overlap between creativity levels is reduced. In contrast, in the segmented approach, because of the smaller samples, the LLMs may have more difficulty distinguishing levels of creativity, leading to slightly lower SRC scores than in the comprehensive assessment.

Interestingly, for the Paperclip object, some AUs in the creative category received higher scores than those in the highly_creative category (as shown in Table 5). However, this was an exception, and overall, no creative AUs outperformed highly_creative AUs when averaging all the five objects scores, as confirmed by Table 2.

Segmented 12 AUs Evaluation by Ranking

Finally, the results from the segmented evaluation using the ranking methodology show that GPT-4o achieved the highest SRC with the oracle, with a value of 0.87 across all five objects. The correlations between the LLMs and the oracle, as well as between the LLMs themselves, ranged between 0.70 and 0.87 as shown in Table 3 (bottom-right diagram), indicating strong overall agreement. Notably, for the Wallet object, all models achieved a perfect correlation of 1.00 with the oracle, demonstrating complete consensus in evaluating the creativity levels of the alternative uses for this particular object. However, the lowest correlation was observed between GPT-4 and Gemini 1.5 Flash, reinforcing once again less agreement between these two models, particularly in ranking-based evaluations.

In terms of category performance, Claude 3.5 Sonnet generated the most creative AUs in the creative category, while GPT-4o produced the most creative AUs in the highly_creative category. GPT-4 generated the most creative AUs in the common category, but overall, Claude 3.5 Sonnet achieved the best ranking average across all categories, as detailed in Table 2.

As for the scoring approach, some individual evaluations of the Paperclip object placed creative AUs higher than highly_creative AUs, but these occurrences did not affect the overall ranking averages across the objects, as seen in Table 2. While the evaluation process is subject to some variability, it produced consistent results across the majority of cases.

Agreement Correlation between LLMs

To assess the overall agreement between the LLMs in their creativity evaluations, we averaged the SRC across all four experimental conditions (comprehensive scoring, comprehensive ranking, segmented scoring, and segmented rank-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

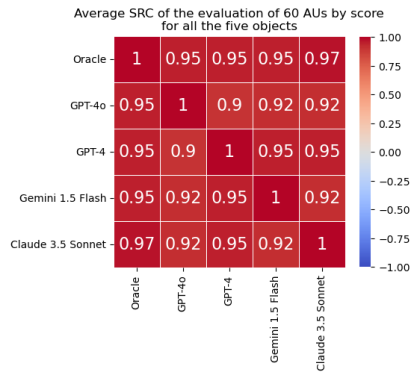
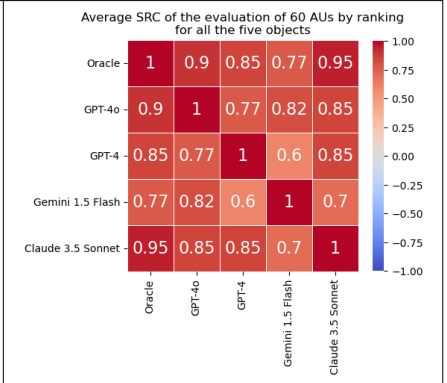
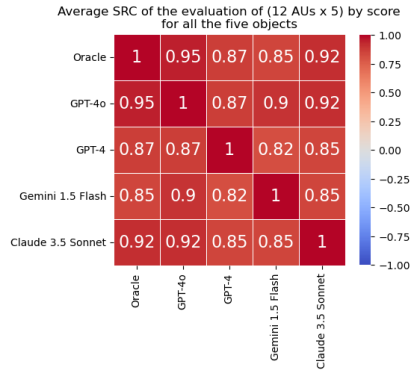
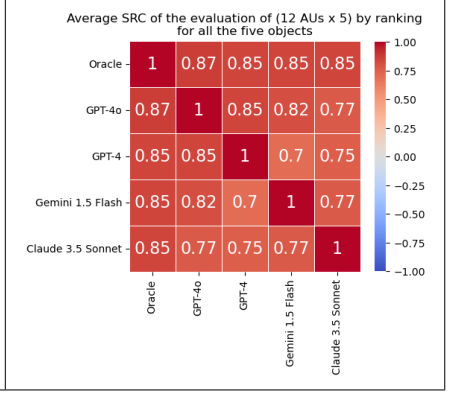
Experiments	Spearman's Rank Correlation Heatmap	Experiments	Spearman's Rank Correlation Heatmap
Comprehensive evaluation by scores (60 AUs)		**Comprehensive evaluation by ranking** (60 AUs)	
Segmented evaluation by scores (12 AUs x 5)		**Segmented evaluation by ranking** (12 AUs x 5)	

Table 3: Average Spearman's Rank Correlation heatmaps for the average over all five objects.

ing). The results reveal consistently high correlations between all models, with values ranging from 0.77 to 0.87, indicating strong agreement among models in creativity evaluation as shown in Figure 8.

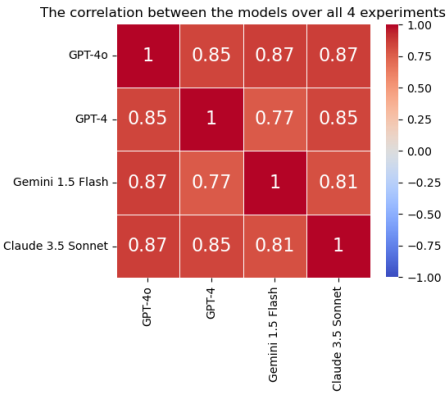


Figure 8: Average SRC heatmap between the LLMs over all the four experiments. GPT-4o shows the strongest overall agreement with other models, while Gemini 1.5 Flash and GPT-4 show the lowest agreement.

GPT-4o demonstrated the strongest overall agreement with other models, showing correlations of 0.85 with GPT-4, 0.87 with Gemini 1.5 Flash, and 0.87 with Claude 3.5 Sonnet. The correlation between Claude 3.5 Sonnet and GPT-4 (0.85) was equally strong, while the correlation between Gemini 1.5 Flash and GPT-4 showed the lowest value (0.77), consistent with the pattern observed across individual experiments.

A key finding of our analysis is that LLMs do not show a preference for their own responses when evaluating creative outputs. Despite each model having the opportunity to assess its own AUs, none exhibited a tendency to rate their own outputs more favorably. The agreement between models is further supported by the low standard deviations observed across all experiments (below 1.03, as shown in Table 2), indicating that the models' evaluations were stable and consistent, regardless of which model generated the AUs.

Conclusion

In this research, we investigated the level of agreement among LLMs in assessing creativity of alternative uses and also with an oracle across different methods. Our findings demonstrate a high level of agreement between the models, with inter-model correlations generally exceeding 0.77, indicating strong consistency in their evaluations. This high correlation suggests that LLMs share a similar understanding of creativity, reliably distinguishing between more and less creative alternative uses. GPT-4o, in particular, exhibited robust alignment with other models, while Claude 3.5

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

LLM	Evaluation by scores (60 AUs)	Evaluation by ranking (60 AUs)	Evaluation by scores (12 AUs x 5)	Evaluation by ranking (12 AUs x 5)
Claude 3.5 Sonnet	**0.97**	**0.95**	0.92	0.85
GPT-4	0.95	0.85	0.87	0.85
Gemini 1.5 Flash	0.95	0.77	0.85	0.85
GPT-4o	0.95	0.90	**0.95**	**0.87**

Table 4: Averages of SRC over the four experiments.

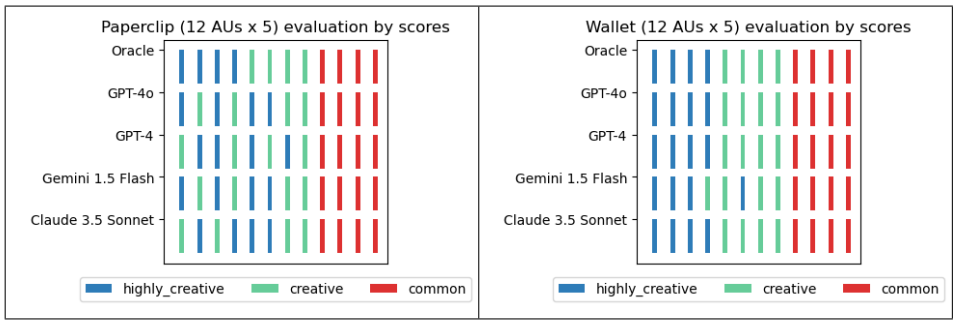


Table 5: The AUs orders of Paperclip and Wallet obtained in the segmented evaluation by the score experiment revealed a difference in the accuracy of the evaluation depending on the object. The Wallet AUs' evaluation shows perfect alignment with the oracle's expected creativity levels, while the Paperclip evaluation reveals some inconsistencies where creative uses occasionally outranked highly creative ones.

Sonnet achieved high evaluation scores on its generated alternative uses, further aligning closely with the oracle across both comprehensive and segmented evaluations.

When examining the highest levels of creativity specifically, our results revealed that Claude 3.5 Sonnet consistently generated the most creative alternative uses. In the comprehensive evaluation using both scoring and ranking methods, Claude 3.5 Sonnet's highly creative AUs received the highest average scores (4.83 out of 5) and best rankings (1.30 out of 60). Similarly, in the segmented evaluation, while GPT-4o performed strongly in the highly creative category (4.25 out of 5), Claude 3.5 Sonnet still demonstrated superior creative generation capabilities across all categories combined.

Regarding accuracy in identifying highly creative responses, Claude 3.5 Sonnet achieved the highest correlation with the oracle in comprehensive evaluations (SRC of 0.97 for scoring and 0.95 for ranking), demonstrating exceptional reliability in recognizing creativity. In the segmented evaluations, GPT-4o showed the strongest performance in identifying creativity levels (SRC of 0.95 for scoring and 0.87 for ranking), particularly excelling at distinguishing the most creative responses from less creative ones.

The evaluation results also revealed that models do not rate their own responses more favorably. Notably, GPT-4o displayed strong alignment with other models across scoring and ranking methods, demonstrating reliable performance and reinforcing the validity of its evaluations. While some variance was observed between GPT-4 and Gemini 1.5 Flash, the low standard deviations across all experiments indicate a stable evaluation framework, enhancing confidence in the LLMs' ability to generate consistent creativity assessments.

In future work, we will focus on expanding and refining this evaluation framework. One promising direction is to increase the diversity and complexity of the AU datasets, allowing for more granular assessments of creativity across varied and challenging contexts. Future studies could also explore how LLMs perform when evaluated using traditional AUT metrics like fluency, flexibility, and elaboration by designing experiments that remove the constraints we imposed in this study. Additionally, evaluating LLMs on other domain-specific creativity tasks such as poetry and jokes could provide deeper insights into their understanding of creativity. Refining this framework and experimenting with task-specific criteria will be essential for advancing the use of LLMs as reliable evaluators in creative domains, supporting the broader goal of enhancing AI's role in creativity assessment.

Acknowledgments

We would like to thank the University of Leicester for supporting this research, in particular the School of Computing and Mathematical Sciences (CMS). The work of AR is supported by the Ministry of Education in Saudi Arabia, represented by King Faisal University. TM is supported by UFOP (Federal University of Ouro Preto, Brazil).

Author contributions

Experimental design: AR, FG, MV; Implementation: AR; Writing and Editing: AR, FG, MV, TM.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

[Al Rabeyah et al. 2024] Al Rabeyah, A.; Góes, F.; Volpe, M.; and Medeiros, T. 2024. Do LLMs agree on the creativity evaluation of alternative uses? *arXiv preprint. arXiv:2411.15560*.

[Beaty and Johnson 2021] Beaty, R. E., and Johnson, D. R. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods* 53:757–780.

[Chang et al. 2024] Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* 15(3).

[Chiang and Lee 2023a] Chiang, C.-H., and Lee, H.-Y. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631.

[Chiang and Lee 2023b] Chiang, C.-H., and Lee, H. 2023b. A closer look into automatic evaluation using large language models. *arXiv preprint. arXiv:2310.05657*.

[Cropley 2000] Cropley, A. J. 2000. Defining and measuring creativity: Are creativity tests worth using? *Roeper review* 23(2):72–79.

[Desmond et al. 2024] Desmond, M.; Ashktorab, Z.; Pan, Q.; Dugan, C.; and Johnson, J. M. 2024. EvaluLLM: LLM assisted evaluation of generative outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24 Companion, 30–32. New York, NY, USA: Association for Computing Machinery.

[DiStefano, Patterson, and Beaty 2024] DiStefano, P. V.; Patterson, J. D.; and Beaty, R. E. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal* 1–15.

[França et al. 2016] França, C.; Góes, L. F. W.; Amorim, A.; Rocha, R.; and Da Silva, A. R. 2016. Regent-dependent creativity: A domain independent metric for the assessment of creative artifacts. In *Proceedings of the Seventh International Conference on Computational Creativity*, 68–75. Citeseer.

[Franceschelli and Musolesi 2024a] Franceschelli, G., and Musolesi, M. 2024a. Creative beam search: Llm-as-a-judge for improving response generation. In *Proceedings of the Fifteenth International Conference on Computational Creativity*, 364–368.

[Franceschelli and Musolesi 2024b] Franceschelli, G., and Musolesi, M. 2024b. Creativity and machine learning: A survey. *ACM Comput. Surv.* 56(11).

[Fu et al. 2024] Fu, J.; Ng, S. K.; Jiang, Z.; and Liu, P. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576.

[Goecke et al. 2024] Goecke, B.; DiStefano, P. V.; Aschauer, W.; Haim, K.; Beaty, R.; and Forthmann, B. 2024. Automated scoring of scientific creativity in german. *The Journal of Creative Behavior*.

[Góes et al. 2023a] Góes, F.; Sawicki, P.; Grzes, M.; Volpe, M.; and Brown, D. 2023a. Is GPT-4 good enough to evaluate jokes? In Pease, A.; Cunha, J. M.; Ackerman, M.; and Brown, D. G., eds., *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada, June 19-23, 2023*, 367–371. Association for Computational Creativity (ACC).

[Góes et al. 2023b] Góes, F.; Sawicki, P.; Grzes, M.; Volpe, M.; and Watson, J. 2023b. Pushing GPT's creativity to its limits: Alternative uses and Torrance tests. In Pease, A.; Cunha, J. M.; Ackerman, M.; and Brown, D. G., eds., *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada, June 19-23, 2023*, 342–346. Association for Computational Creativity (ACC).

[Gómez-Rodríguez and Williams 2023] 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.

[Guilford 1967] Guilford, J. P. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior* 1(1):3–14.

[Hadas and Hershkovitz 2024] Hadas, E., and Hershkovitz, A. 2024. Using large language models to evaluate alternative uses task flexibility score. *Thinking Skills and Creativity* 52:101549.

[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.

[Li et al. 2023] Li, Q.; Cui, L.; Kong, L.; and Bi, W. 2023. Exploring the reliability of large language models as customized evaluators for diverse nlp tasks. *arXiv preprint. arXiv:2310.19740*.

[Liu, Bhandari, and Pardos 2025] Liu, Y.; Bhandari, S.; and Pardos, Z. A. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology* 56(3):1028–1052.

[Murugadoss et al. 2025] Murugadoss, B.; Poelitz, C.; Drosos, I.; Le, V.; McKenna, N.; Negreanu, C. S.; Parnin, C.; and Sarkar, A. 2025. Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19589–19597.

[Organisciak et al. 2023] 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.

[Raz et al. 2024] Raz, T.; Luchini, S.; Beaty, R.; and Kenett, Y. 2024. Automated scoring of open-ended question complexity: A large language model approach.

[Sawicki et al. 2023] Sawicki, P.; Grzes, M.; Góes, F.; Jor-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

danous, A.; Brown, D.; Paraskevopoulou, S.; Peeperkorn, M.; and Khatun, A. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proceedings of the 14th International Conference on Computational Creativity, Ontario, Canada, June 19-23, 2023*, 10–19. Association for Computational Creativity (ACC).

[Stevenson et al. 2022] 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 2022*, 164–168. Association for Computational Creativity (ACC).

[Thakur et al. 2024] Thakur, A. S.; Choudhary, K.; Ramayapally, V. S.; Vaidyanathan, S.; and Hupkes, D. 2024. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-Judges. *arXiv preprint.* arXiv:2406.12624.

[Torrance 1966] Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and Psychological Measurement*.

[Wang et al. 2024] Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; et al. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450.

[Wu et al. 2024] Wu, Y.; Iso, H.; Pezeshkpour, P.; Bhutani, N.; and Hruschka, E. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 330–343. St. Julian's, Malta: Association for Computational Linguistics.

[Yang et al. 2023] Yang, T.; Zhang, Q.; Sun, Z.; and Hou, Y. 2023. Automatic assessment of divergent thinking in chinese language with TransDis: A transformer-based language model approach. *Behavior Research Methods* 56(6):5798–5819.

[Zhao et al. 2024] Zhao, Y.; Zhang, R.; Li, W.; Huang, D.; Guo, J.; Peng, S.; Hao, Y.; Wen, Y.; Hu, X.; Du, Z.; et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint* arXiv:2401.12491.

[Zheng et al. 2023] 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. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Making the Familiar Unfamiliar: AI-Driven Synectics Metaphor Generation for Computational Creativity

Georgios Palamas
Department of Technology and Society
Malmö University
georgios.palamas@mau.se

Abstract

Metaphors are central to human creativity, enabling abstract reasoning and novel conceptual connections. This paper explores how AI can generate and evaluate metaphors using Synectics, a structured creativity method that makes the familiar unfamiliar. We introduce a GPT-4-based pipeline to generate metaphors across four analogy types: Direct, Personal, Symbolic, and Fantasy. A comparison with 1,000 human-created metaphors is conducted using semantic coherence (BERT), novelty (TF-IDF), and GPT-4-assessed creativity. Results show that AI-generated metaphors tend to be more novel but less coherent than human ones, with Fantasy analogies rated highest in creativity but lowest in structure. Abstract themes like Dreams and Knowledge lead to more imaginative outputs. While the system produces stylistically rich metaphors, it also reflects biases and lacks functional evaluation. This work contributes to computational creativity by demonstrating how structured prompting supports controllable metaphor generation. We suggest future directions including hybrid evaluation frameworks that combine computational and human judgment to better assess metaphor effectiveness in context.

Introduction

Creativity often emerges from the ability to perceive the familiar in a new light or to find coherence in seemingly unrelated ideas. Metaphors serve as a fundamental mechanism for human cognition and creativity, enabling abstract reasoning, novel conceptual connections, and artistic expression (Lakoff and Johnson 2008). Mapping familiar ideas onto unfamiliar domains is central to innovation and problem-solving.

A striking example of this process appears in surrealist art, where everyday objects are reconfigured into unexpected, thought-provoking compositions. Dora Maar's *Untitled (Hand-Shell)* (Figure 1) exemplifies this, merging a human hand with a seashell to create a novel hybrid form. This perceptual shift mirrors how metaphors bridge disparate concepts to create new meanings. We explore whether AI can perform similar conceptual transformations through metaphor generation, systematically making the familiar unfamiliar.

Metaphor generation has traditionally been a human task, shaped by cultural, emotional, and experiential factors. With the advent of large language models (LLMs), there is growing interest in whether AI can produce metaphors that are both creative and meaningful. However, existing AI systems often lack conceptual depth, coherence, and contextual awareness.

To address this, we incorporate Synectics, a structured creativity framework encouraging conceptual shifts through analogy-based reasoning (Gordon 1961). Unlike conventional approaches based on direct analogies or statistical associations, Synectics guides metaphor generation using four distinct analogy types: Direct, Personal, Symbolic, and Fantasy. This structure enables AI to move beyond surface-level mappings toward more imaginative transformations.

We systematically evaluate AI-generated metaphors through three key metrics:

1. **Semantic Coherence:** Logical consistency via contextualized sentence embeddings.
2. **Novelty Score:** Conceptual distinctiveness measured against 1,000 human-created metaphors using TF-IDF similarity.
3. **Creativity Score:** GPT-4 self-assessment of metaphor originality and impact.

Our human reference set spans 1,000 literary and culturally significant metaphors categorized into themes such as Time, Love, Memory, and Fear. By applying computational methods, we offer insights into how AI can function as a creative partner in linguistic abstraction and symbolic reasoning.

Results indicate that AI-generated metaphors exhibit higher novelty but lower coherence compared to human-created ones. Fantasy analogies demonstrate the most significant conceptual shifts, aligning with Synectics' emphasis on making the familiar unfamiliar, but they also suffer from reduced structural coherence.

This research contributes a scalable approach to AI-driven metaphor modeling, evaluation, and application. Practical implications arise for creative writing, AI-assisted storytelling, and design ideation. The remainder of the paper is structured as follows: Section 2 reviews related work on metaphor generation, Synectics, and computational evaluation methods. Section 3 details our methodology. Section

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

4 presents comparative results, followed by a discussion in Section 5. Section 6 concludes with future research directions.

Figure 1: Dora Maar's surrealist photograph *Untitled (Hand-Shell)*. The image illustrates the transformation of familiar objects, a hand and a shell, into an unfamiliar context, analogous to AI metaphor generation.

Related Work

Metaphor Generation. Early approaches to metaphor generation used structured knowledge bases such as ConceptNet and WordNet (Miller 1995), producing template-based metaphors with limited flexibility (Veale, Shutova, and Klebanov 2022). Later, distributional semantics enabled the identification of latent conceptual mappings (Shutova, Sun, and Korhonen 2013). Transformer-based models such as GPT-3 and GPT-4 (Brown et al. 2020), (OpenAI 2023) now generate metaphors through large-scale language modeling, but their outputs often lack grounding or coherent structure (Bondla 2024), (de Lima et al. 2025). Recent surveys highlight the evolution of figurative language generation from rule-based systems to large language models (Lai and Nissim 2024), as well as advances in computational metaphor processing including identification, interpretation, and generation techniques (M. Ge and Cambria 2023). These developments underscore the growing capabilities of LLMs while emphasizing the need for structured creative frameworks to guide metaphor production. Recent tools like Metaphorian (Kim et al. 2023) and MetaPro (Mao et al. 2023) show progress in metaphor assistance and processing, but do not apply structured creative frameworks during generation.

Synectics as Structure. Recent advances in computational creativity and natural language processing have explored metaphor generation through various approaches. Chakrabarty et al. (Chakrabarty et al. 2021) introduced MERMAID, a model that fine-tunes BART using a parallel corpus of metaphorical and literal sentences, employing a generator-discriminator framework to produce expressive metaphors. Stowe et al. (Stowe et al. 2021) extended this work by incorporating conceptual mappings to guide metaphor generation using FrameNet-based embeddings and a controlled sequence-to-sequence model. These methods emphasize large-scale model-based generation with varying degrees of symbolic structure.

Within the computational creativity community, several contributions have addressed creative language modeling and metaphor evaluation. Siekiera et al. (Siekiera et al. 2020) proposed ranking methods for creative language in small data scenarios, including metaphor and humor. Elzohbi and Zhao (Elzohbi and Zhao 2023) provided a comprehensive review of creative data generation, with emphasis on poetic and metaphorical outputs. Muller et al. (Muller, Candello, and Weisz 2023) explored analogy-based design in human-AI co-creative settings, highlighting how conceptual reframing supports ideation and interpretability.

Our work complements these efforts by introducing a structured prompting framework inspired by Synectics, emphasizing analogy type (Direct, Personal, Symbolic, Fantasy) as a controllable axis for metaphor variation. This allows for a more interpretable and targeted exploration of metaphor generation, bridging generative power with stylistic and thematic guidance.

Synectics promotes creativity by encouraging analogical thinking at increasing levels of abstraction (Gordon 1961). While analogy-based reasoning has been explored in computational creativity (Goel, Rugaber, and Vattam 2009), and structured ideation has been supported through tools such as AI_deation (Palamas, Guerra, and Møsbæk 2022), Synectics remains underutilized in metaphor generation. Our work addresses this by applying its four analogy types: Direct, Personal, Symbolic, and Fantasy, as an explicit prompting framework. Unlike chain-of-thought or few-shot methods, Synectics explicitly guides conceptual divergence across multiple axes.

Conceptual Blending and Creative Reasoning. Conceptual Blending Theory provides a complementary perspective on how creativity arises from the merging of disparate mental spaces (Fauconnier and Turner 2002). Unlike Synectics, which offers operational categories for analogy-making, conceptual blending emphasizes the emergent structure formed by integrating partial elements from multiple sources. Prior computational models of blending have demonstrated success in analogy-making and story generation (Veale 2012), but typically lack the structured prompting mechanisms employed in this work. Integrating Synectics into AI systems enables more targeted exploration across creative axes while maintaining interpretability.

Evaluating Metaphor Quality. Evaluating metaphors is challenging due to their subjective and cultural complexity.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Human evaluation remains the gold standard (Dankers et al. 2019) but is resource-intensive. Scalable metrics include semantic coherence via contextual embeddings (e.g., BERT) (Jacob Devlin and Toutanova 2019) and novelty through TF-IDF similarity to human corpora (Shutova 2015). However, TF-IDF lacks semantic nuance. Alternatives such as Word Mover's Distance (WMD) (Kusner et al. 2015) provide deeper semantic comparison but are computationally costly. GPT-4 self-evaluation, while scalable, may introduce bias; we use it as a proxy, not a substitute, and recommend future work incorporating human ratings.

Methodology

This study explores structured metaphor generation through GPT-4 using the Synectics creativity framework, evaluating results via scalable computational metrics. We compare AI-generated metaphors to a curated set of human-created examples across thematic domains and analogy types. To provide a clear overview of the process, Figure 2 summarizes the overall workflow from thematic input to evaluation.

Synectics-Based Prompting

Each prompt requested four analogies for a given theme, corresponding to the four Synectics types:

- **Direct:** Concrete mappings (e.g., "Fear is a shadow.")
- **Personal:** Personification-based (e.g., "If I were fear, I'd be a whisper in the dark.")
- **Symbolic:** Abstract associations (e.g., "Fear is a cracked mirror.")
- **Fantasy:** Surreal, imaginative mappings (e.g., "Fear is a beast that feeds on silence.")

Across both human and AI-generated sets, we used a consistent set of ten thematic domains to prompt metaphor creation: *Love, Time, Memory, Fear, Hope, Dreams, Knowledge, Power, Loss*, and *Freedom*. Each theme served as a seed for generating four distinct analogies based on Synectics' analogy types: Direct, Personal, Symbolic, and Fantasy. This structure enabled comparative analysis across both thematic and stylistic dimensions.

This controlled variation promotes conceptual diversity and allows us to systematically assess the effects of analogy type on metaphor quality. Unlike general prompt-based generation, Synectics ensures a structured exploration across creative axes and facilitates more interpretable comparative analysis across styles of metaphorical abstraction.

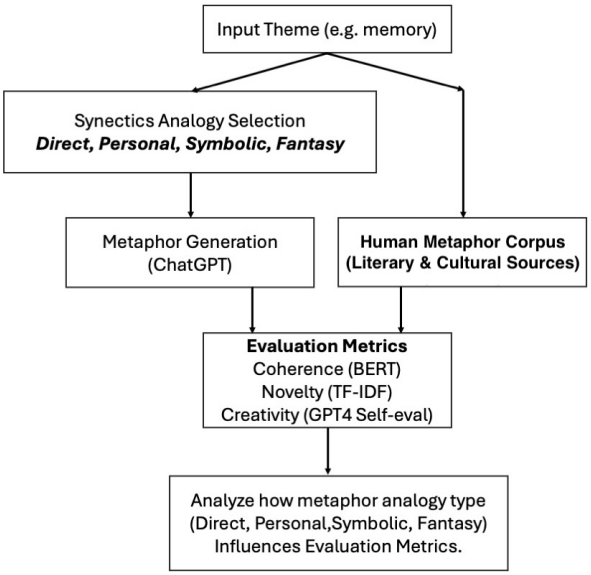


Figure 2: Overview of the metaphor generation and evaluation pipeline.

Human Reference Dataset

We constructed a comparison set of 1,000 human metaphors across ten themes (e.g., Love, Time, Dreams), drawn from literary, idiomatic, and cultural sources. Each metaphor was manually categorized to ensure thematic consistency. This dataset serves as a benchmark for evaluating coherence and novelty. By selecting metaphors from diverse sources, we sought to capture a wide range of figurative expressions and ensure a robust comparison with AI-generated outputs.

Evaluation Metrics

We evaluate metaphor quality across three dimensions:

- **Semantic Coherence**: Assesses the plausibility and interpretability of the metaphor's mapping. Measured via cosine similarity between BERT sentence embeddings of source and target, with higher scores indicating stronger alignment.
- **Novelty**: Captures deviation from familiar expressions, estimated via inverse TF-IDF similarity to a 1,000-metaphor human corpus. Lower similarity implies higher novelty. While interpretable, TF-IDF has lexical limitations; embedding-based alternatives (e.g., Word Mover's Distance) are proposed for future work.
- **Creativity**: Reflects originality, emotional resonance, and conceptual surprise. Each AI-generated metaphor was scored by GPT-4 on a 1–5 scale using the following prompt:

  Evaluate the following metaphor according to creativity on a 1–5 scale. How original, evocative, or surprising is the metaphor? Consider whether it introduces a novel association or emotional depth beyond conventional phrasing. Provide one or two short sentences justifying your score.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

GPT-based scoring offers scalable approximation but may inherit stylistic bias—favoring poetic or surreal constructions. Prior work on computational aesthetics and affect (Lunterova, Spetko, and Palamas 2022) stresses the need to consider audience perception. Future iterations will integrate human-in-the-loop evaluations to validate and contextualize results.

Experimental Setup

GPT-4 generated 1,000 metaphors (100 per theme), evenly distributed across analogy types. Each generation batch was conducted under temperature sampling parameters encouraging diversity (temperature = 0.9). We analyzed performance not only by analogy type but also by theme, enabling fine-grained insight into how topic abstraction impacts metaphor quality.

Preliminary observations suggest that abstract themes such as Dreams and Knowledge yielded higher creativity scores, while culturally saturated themes like Love and Power resulted in more conventional, less novel metaphors. This indicates that the degree of thematic abstraction modulates the creative potential of AI-driven metaphor generation, aligning with theories of conceptual space expansion in creativity research.

Results and Discussion

We evaluated 1,000 AI-generated metaphors and compared them to 1,000 human metaphors using semantic coherence (BERT), novelty (TF-IDF), and GPT-4 self-evaluated creativity. These comparisons reveal how structured prompting influences metaphor quality and expose the interplay between analogy type, thematic abstraction, and metaphor effectiveness.

AI vs. Human Comparison

Figure 2 summarizes the performance across all metrics. AI metaphors were more novel but generally less coherent than human metaphors. Creativity scores were comparable, although slightly favoring human references.

Analogy Type Effects

Fantasy analogies scored highest in creativity (4.8) but lowest in coherence (3.8), reflecting Synectics' principle that greater conceptual distance fosters imaginative output at the cost of structure. Direct analogies, in contrast, were highly coherent (4.5) but less novel (3.9). This confirms that analogy type strongly influences output balance and suggests that different analogy styles may be preferentially suited to different application contexts.

Thematic Influence

Themes such as *Dreams*, *Hope*, and *Knowledge* yielded more creative and novel outputs, while themes like *Love* and *Power* resulted in more conventional expressions. This suggests that abstract or less culturally saturated domains offer AI greater conceptual flexibility.

Table 1 highlights the top and bottom performing themes by average creativity score.

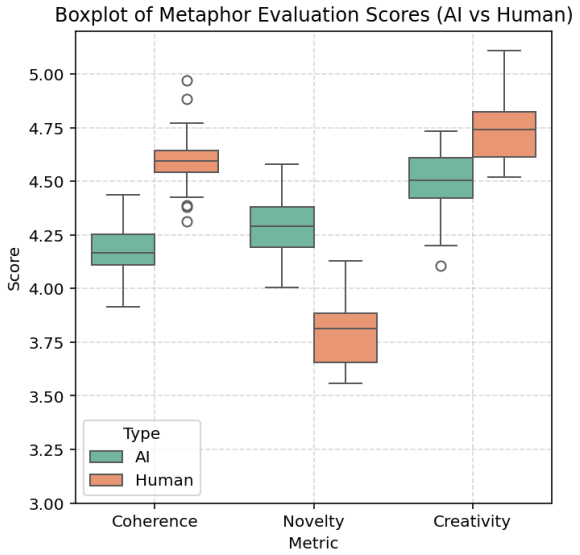


Figure 2: Boxplot comparing AI and human metaphor scores across coherence, novelty, and creativity.

Theme	Creativity Score (AI)
Dreams	4.8
Knowledge	4.7
Hope	4.7
Power	4.3
Love	4.2

Table 1: Creativity scores across selected themes (AI-generated metaphors).

These results suggest that thematic abstraction modulates AI creative performance, with abstract domains offering a broader space for novel metaphorical constructions, while highly familiar domains may anchor outputs in conventional associations.

Interpretation

Our findings suggest four key observations grounded in the data:

1. **Analogy type affects tradeoffs between creativity and coherence.** Fantasy analogies consistently yielded higher creativity scores but lower semantic coherence, while Direct analogies exhibited the opposite trend. This aligns with Synectics' principle that greater conceptual distance enables more imaginative outputs but increases the risk of disjointed mappings.
2. **Thematic abstraction modulates metaphorical diversity.** Themes classified as more abstract (e.g., *Dreams, Knowledge*) enabled GPT-4 to explore less conventional associations, often resulting in higher novelty and creativity scores. Conversely, culturally saturated themes (e.g., *Love, Power*) tended to anchor metaphor generation in fa-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

miliar idioms.

3. **Prompt structure shapes stylistic tone.** The poetic or surreal tone of many AI metaphors, especially in the Fantasy category, can be attributed in part to the Synectics style prompting, which implicitly encouraged narrative or image rich outputs. However, this style may not always align with functional communicative needs.
4. **Utility of metaphors remains unexamined.** While this study focused on stylistic and structural characteristics (creativity, novelty, coherence), we did not evaluate metaphor *function*, i.e., whether a metaphor effectively communicates, persuades, or aids understanding. Future work should assess metaphor performance in goal-directed contexts to evaluate usefulness beyond stylistic metrics.

Example Comparison

To illustrate stylistic differences, Table 2 compares human and AI metaphors across themes using Fantasy analogies.

Theme	Human Metaphor	AI (Fantasy Analogy)
Time	Time is a thief.	Time is a god juggling destinies.
Love	Love is a battlefield.	Love is a star slowly forgetting how to shine.
Memory	Memory is a filing cabinet.	Memory is a time traveler lost in old dreams.
Fear	Fear is a shadow.	Fear is a beast that feeds on silence.

Table 2: Illustrative human vs. AI metaphors by theme and analogy type.

Qualitative Observations

We note an important stylistic asymmetry between the two sets: the human metaphors were typically short, often using familiar "A is B" constructions drawn from idiomatic or literary sources. In contrast, the AI-generated metaphors, particularly in the Fantasy category, were prompted to produce more expressive, poetic outputs.

This discrepancy stems from the lack of explicit formatting constraints placed on the human corpus, which was compiled from pre-existing metaphors rather than instructing humans to follow the same prompting protocol used with GPT-4. As a result, AI metaphors often exhibit more figurative richness but may also reflect prompting artifacts rather than inherent superiority.

Future work will address this imbalance by collecting human-authored metaphors using matched Synectics-style prompts and constraints, enabling a more direct and fair comparison of metaphorical capabilities across agents. Human metaphors tend to draw from idioms and shared cultural knowledge, resulting in accessible but sometimes predictable constructions.

Potential Applications

This structured metaphor generation pipeline has immediate applications across creative writing, education, and ideation. Writers may use Synectics-guided AI outputs as prompts or conceptual seeds, stimulating new narrative directions or thematic elaborations. In education, structured metaphor generation can help students reframe abstract concepts, making difficult ideas more approachable through imaginative mapping.

For example, an instructor teaching abstract scientific concepts such as "entropy" might use Fantasy analogies like "Entropy is a tired orchestra forgetting its symphony" to provoke discussion and deepen understanding. Similarly, design teams could leverage metaphorical prompts during ideation phases to reimagine user experiences or brand narratives in more emotionally resonant terms.

Conclusion and Future Directions

This study introduced a structured prompting strategy for metaphor generation using GPT-4, grounded in the Synectics framework. By varying analogy types and thematic domains, we explored stylistic and semantic differences in AI-generated metaphors and compared them to a human-authored baseline. Our results suggest that certain combinations, such as abstract themes with Fantasy analogies, tend to produce more novel and creative outputs, sometimes at the expense of coherence.

However, our evaluation focused on stylistic aspects such as novelty, creativity, and coherence, without examining the functional utility of metaphors. In real-world contexts such as education or design, the value of a metaphor lies in its clarity and relevance to a given purpose. Relying solely on stylistic metrics may favor poetic abstraction over communicative effectiveness. Future work should include task-based or audience-specific evaluations to assess metaphors in context.

The use of GPT-4 as an evaluator also introduces potential bias, as it may prefer poetic or surreal forms. This raises ethical questions about how language models reflect and reinforce aesthetic norms. To address this, we propose incorporating human evaluation, broader prompt strategies, and assessments that account for both form and function.

Our approach contributes to computational creativity by framing metaphor generation as a process that combines structured prompting with generative power. It supports both theoretical exploration and applied use in fields where metaphor helps transform familiar ideas into unfamiliar perspectives, and vice versa. This aligns with computational creativity's broader aim: balancing structure, expression, and intent.

References

Bondla, S. 2024. Peeping into ai: A literary deconstruction of an essay by ai gpt3 and an interview with ai lamda. *International Review of Literary Studies* 6(1):13–29.

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.

Chakrabarty, T.; Zhang, X.; Muresan, S.; and Peng, N. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4250–4261. Online: Association for Computational Linguistics.

Dankers, V.; Rei, M.; Lewis, M.; and Shutova, E. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2218–2229.

de Lima, E. S.; Neggers, M. M.; Feijó, B.; Casanova, M. A.; and Furtado, A. L. 2025. An ai-powered approach to the semiotic reconstruction of narratives. *Entertainment Computing* 52:100810.

Elzohbi, M., and Zhao, R. 2023. Creative data generation: A review focusing on text and poetry. *arXiv preprint arXiv:2305.08493*.

Fauconnier, G., and Turner, M. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.

Goel, A. K.; Rugaber, S.; and Vattam, S. 2009. Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language. *Ai Edam* 23(1):23–35.

Gordon, W. J. 1961. Synectics: The development of creative capacity.

Jacob Devlin, Ming-Wei Chang, K. L., and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kim, J.; Suh, S.; Chilton, L. B.; and Xia, H. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 115–135.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966. PMLR.

Lai, H., and Nissim, M. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Computing Surveys* 56(10):1–34.

Lakoff, G., and Johnson, M. 2008. *Metaphors we live by*. University of Chicago press.

Lunterova, A.; Spetko, O.; and Palamas, G. 2022. Computational aesthetics of the collective affective dynamics of imdb movie reviews. In *Porto, Lisbon: 10th Conference on Computation, Communication, Aesthetics & X: 10th Conference on Computation, Communication*, 26.

M. Ge, R. M., and Cambria, E. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*.

Mao, R.; Li, X.; Kai, H.; Ge, M.; and Cambria, E. 2023. Metapro online:: A computational metaphor processing online system. In *Proceedings of the 61st annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL).

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Muller, M.; Candello, H.; and Weisz, J. 2023. Interactional co-creativity of human and ai in analogy-based design. In *International Conference on Computational Creativity*.

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

Palamas, G.; Guerra, A. M.; and Møsbæk, L.-D. 2022. Ai_deation: A creative knowledge mining method for design exploration. *EAI Endorsed Transactions on Creative Technologies* 9.

Shutova, E.; Sun, L.; and Korhonen, A. 2013. Metaphor identification using verb and noun clustering. *Proceedings of ACL* 178–187.

Shutova, E. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics* 41(4):579–623.

Siekiera, J.; Köppel, M.; Simpson, E.; Stowe, K.; Gurevych, I.; and Kramer, S. 2020. Ranking creative language characteristics in small data scenarios. *arXiv preprint arXiv:2010.12613*.

Stowe, K.; Chakrabarty, T.; Peng, N.; Muresan, S.; and Gurevych, I. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6724–6736. Online: Association for Computational Linguistics.

Veale, T.; Shutova, E.; and Klebanov, B. B. 2022. *Metaphor: A computational perspective*. Springer Nature.

Veale, T. 2012. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. Bloomsbury Academic.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Making Mountains out of Molehills: Abstraction without Information Loss in Analogical Mapping

Tony Veale
School of Computer Science,
University College Dublin,
Dublin D4, Ireland.
tony.veale@UCD.ie

Abstract

Analogy is a cognitive process that propels many of our most creative leaps, from the cross-domain forays of scientific discovery and case-based reasoning to the poetry of metaphor, bisociation and blending. By concerning itself with the *shape* of meanings, and the structural arrangements of their parts, analogy allows us to unite different meanings with similar shapes across distant domains. Crucial to this unification, or mapping, of domains is the ability to abstract over structural representations that capture the broad sweep of an idea without getting bogged down in details. This paper presents a large new resource for analogical mapping that defines structured representations to support abstraction at multiple levels, but without information loss. This allows the mapping process to be flexible in its reconciliation of different meanings, while also preserving the distinctions that make abstraction necessary in the first place. This resource, named ATLAS, is a wide-ranging database of symbolic structures for lexical concepts (the ideas behind common words), and is designed to support explicit analogical reasoning in an era where symbolic reasoning is giving way to the statistics of LLMs.

Introduction: The Mind Has Mountains

Our most compelling and memorable arguments often rest on an analogy, even if a sharp insight can easily cut both ways. Consider this exchange from the 1993 film Jurassic Park, in which the park's owner, Hammond, defends the carnage that has ensued with an analogy to an iconic resort:

> **Hammond**: All major theme parks have delays. When they opened Disneyland in 1956, nothing worked!
>
> **Malcolm**: Yeah, but, John, if *The Pirates of the Caribbean* breaks down, the pirates don't eat the tourists.

In Hammond's analogy, Disneyland serves as the familiar *source* domain through which the implied *target*, Jurassic park, is viewed (Gentner 1983). But Malcolm inverts this analogy, and instead uses the latter as a source domain to reimagine the former. Each analogy re-frames the context to suit the speaker's own worldview: Hammond seeks to downplay the enormity of his park's failings, while Malcolm's counterfactual aims to remind him of the fatal consequences of his hubris. It is the isomorphic structure of each domain – immersive theme parks with hi-tech attractions – that enables these analogies to map so easily from one to the other, as in *animatronic* to *genetic* and *pirate* to *dinosaur*, and to suggest new candidate inferences, such as Malcolm's allusion to tourist-gobbling pirates (Gentner and Toupin 1986; Falkenhainer, Forbus, and Gentner 1989). Identifying this shared structure is the job of an analogical mapping engine, yet such structures must already be present for such a mapping to occur. Creating a large database of these structures to support analogical mapping and retrieval is our goal here.

As in metaphor and simile, analogy relies on abstraction to reconcile two different domains of experience or conceptual content. Abstraction is what allows us to get to the nub of an idea, and to equate very different meanings on the basis of a common gist (Gentner and Hoyos 2017). In analogy, this shared core is essentially structural, even if the two domains also have emotions and other features in common. There are two basic approaches to structural abstraction in analogy: in the first, the analogist identifies a subset of the content of one domain that is isomorphic to a corresponding sub-structure in the other, where isomorphism insists that both sub-structures have the same shape, and the same arrangement of labeled edges. If those edges connect different vertices in the source and target, a consistent 1-to-1 mapping from source to target vertices must also be possible. In essence, analogical structure mapping is a variant of the NP-hard problem *largest sub-graph isomorphism*, and so is itself NP-hard as a result (Veale and Keane 1997). The second approach generalizes an isomorphic mapping to sub-structures that have the same shape even if their edges have different labels, and so produces a mapping of vertices *and* relations, provided the relations in one structure are similar to those in the other. This second approach lifts abstraction to the next level, by allowing domains themselves to be re-represented during the mapping process (Hofstadter and Mitchell 1995).

(Gentner and Hoyos 2017) define "abstraction" as the decrease in specificity, and concomitant increase in scope, of a concept. Thus, as a source structure sheds specific details, it becomes applicable to a broader range of analogical targets. In this view, abstraction necessarily entails information loss, as only the most systematic core of the structure is retained. However, in an analogical mapping, only certain elements – the relational edges – are matched identically, while others – the vertices – are mapped non-identically in 1-to-1 pairings.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

A loss-free process of abstraction via re-representation can preserve information if it shifts components that are matched identically into positions that are mapped non-identically. We present such a process in this paper, to build a resource, named ATLAS, that provides conceptual structures at multiple levels of abstraction for a wide range of words and ideas.

We must first define what we mean by a conceptual structure, and what it means to abstract over such a structure. To populate ATLAS at scale, we will leverage existing resources that were previously built to enable the generation of metaphors, blends and stories, such as the *talking points* model of (Veale 2014; 2015). As the semantic triples defined by this model are too shallow to support structure mapping as defined in (Gentner 1983; Falkenhainer, Forbus, and Gentner 1989), we shall present a rule-based means of constructing richly nested structures from these rather flat ingredients. As the new structures will vary in depth and breadth, we will define a measure of richness that will allow a mapping engine to rank the structures that define a concept by how well they embody the *systematicity principle* of (Gentner 1983).

We will also present a mapping algorithm and a retrieval algorithm to work efficiently with the structures of ATLAS. The analogy literature is replete with alternatives (Falkenhainer, Forbus, and Gentner 1989; Holyoak and Thagard 1989; Hofstadter and Mitchell 1995; Veale and Keane 1997), but we shall define variants of the highly efficient structure-hashing approach of (Veale 2005) that work especially well with our model of loss-free abstraction. ATLAS defines over two million structures across more than 9000 word concepts, making efficient storage, retrieval and mapping a necessity. But we must start at the beginning, to consider the assumptions that are made by structure-mapping models of analogy.

That would be an ecumenical matter

Representation is always a matter of choice: of what to represent, of which viewpoint to adopt, of what level of detail to capture, and with what feature set to encode it. An open representation can utilize an unlimited set of symbols, while a small number of primitives may suffice in a closed system. These choices introduce a great deal of wiggle room into any symbolic representation that is used for analogical mapping.

Every analogical system will bring its own assumptions to bear upon its representation of the world. Consider SME, the Structure-Mapping Engine of (Falkenhainer, Forbus, and Gentner 1989), which is a computational realization of the structure-mapping theory (SMT) of (Gentner 1983). While SME/SMT is open with respect to the choice of symbols for relations and attributes in any given domain, it assumes that the most systematic relations – such as those for causality – are shared across domains, and that these relations form hierarchical structures corresponding to nested propositions. The graph representation of a domain is thus a forest of interconnected tree structures, and a mapping can be constructed for the whole, or the largest sub-graph of the whole, by first finding partial mappings for the largest and deepest trees. A global mapping is then produced by merging a consistent selection of these partial mappings into a coherent whole.

Nesting is another choice that a representation may use, or not. The Sapper model of (Veale and Keane 1997) represents domains as a semantic-network of vertices and labeled, directed edges. An edge encodes a relation between two entities, and takes its label from a closed set of relation types. Sapper does not define distinct domains for different ideas. Rather, a single large graph encodes all concepts, so that the local region around a given node is the domain of that node. With no pockets of nested structure to exploit, Sapper seeks out pathways in the network that originate at the nodes for the two ideas that one is aiming to map. When pathways that comprise the same sequence of relations meet at a common association, they are placed into analogical correspondence to form a partial mapping. As in SME, a global mapping is constructed by merging partial maps into a coherent whole.

Each model works rather well on its own representations, but Sapper simply cannot use SME's representations, while SME finds no nesting to exploit in Sapper's graph structures. SME senses systematicity in the depth of nested structures, but Sapper sees it in the length of extended pathways. Each is blind to what makes the other's mappings systematic. If a common representation, and a shared database of structures usable by both approaches, is thus not feasible, we might, instead, map from one model's worldview onto the other's. We can, for instance, flatten SME's nested representations into a semantic network by introducing new nodes to represent branching points in its structures. Conversely, we can add causal nesting to Sapper-like graphs, to represent the higher-order connections between flat node-to-node relations.

We take the latter route here, but we need a starting point: a database of content on which to apply our transformations. The AnalogyKB of (Yuan et al. 2024) is a large knowledge-base of structured symbolic content derived from *WikiData* (Vrandečćíc and Krötzsch 2014) and *ConceptNet* (Speer, Chin, and Havasi 2017). Each resource expresses its knowledge in the form of predicate-subject-object triples; *WikiData* triples convey dry, objective facts (such as the CEO of company X is Y, or the capital of country A is B), while *ConceptNet* offers more subjective claims about the world. This kind of knowledge supports proportional reasoning of the *A is to B as C is to D* variety (A:B::C:D), as in Sam Altman is to OpenAI as Donald Trump is to USA. Proportional analogy, first studied by Aristotle in his *Poetics*, was once a cornerstone of the *Scholastic Aptitude Test* (or SAT), and computational solvers have been presented by (Veale 2004) and (Turney and Littman 2005), among others. These analogies are sometimes the tip of a structural iceberg, implying mappings beyond that of A to B and C to D, but we need to excavate that structure to support richer, deeper analogies.

The knowledge defined in (Veale 2014; 2015) to support story generation systems also has a flat triple structure, but since these triples are designed to work together, or indeed to *click* together like pieces of track, it is easier to imagine how a nested causal structure can be imposed on top. Moreover, these story-oriented triples concern commonsense ideas and stereotypical associations, not contingent facts, and can thus support analogies at the level of ideas rather than facts. In the next section we describe how rich causal structures can be assembled at scale from these flat triples, to support analogical mapping and retrieval at various levels of abstraction. We must start with abstractions, in the form of linking rules.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

The Shape Of Meaning

The following tuples are representative of the 85,000 or so *predicate-subject-object* triples provided by (Veale 2014):

(*work_in* scientist lab)
(*conduct* lab experiment)
...
(*serve* priest congregation)
(*perform* congregation worship)

There is an intuitive linkage between *work_in* and *conduct* in the first case, and between *serve* and *perform* in the second: scientists work in labs *that* conduct experiments, and priests serve congregations *that* perform worship. So, our first task is to merge these disparate facts into composite structures:

(*work_in* **scientist** (*some* lab (*that* (*conduct* experiment))))
...
(*serve* **priest** (*some* congregation (*that* (*perform* worship))))

We first codify these intuitions with a pair of abstractions:

(*work_in* $person_X$ (*some* $place_Y$ (*that* (*conduct* $action_Z$))))
...
(*serve* $person_X$ (*some* $place_Y$ (*that* (*perform* $action_Z$))))

These hand-crafted abstractions have the function of rules: they allow pairs of triples that relate a $person_X$, a $place_Y$ and an $action_Z$ to be coalesced into a nested structure. With enough abstractions like these, we can impose new structure on many of the flat, disjoint triples of the original database.

These abstractions have the same shape, yet they cannot be aligned without violating the SMT condition of predicate identicality, as *conduct* $\neq$ *perform* and *work_in* $\neq$ *serve*. However, if we rewrite the original facts as follows, using a smaller, canonical set of predicates, a mapping is possible:

(*by* working (*perform* **scientist**
(*some* work (*for* lab
(*that* (*conduct* experiment))))))
(*by* serving (*perform* **priest**
(*some* service (*for* congregation
(*that* (*conduct* worship)))))))

The rewriter maps the original, open-ended set of predicates onto a reduced set of just 12, such as *perform* and *enhance*. The rewritten structures are more general, since they use a small set of predicates, yet they incur no loss of information, as they preserve the distinction between *work_in* and *serve*. Not only are these structures alignable under SMT, they produce a richer mapping, of *scientist:priest*, *lab:congregation*, *experiment:worship*, *working:serving* and *work:service*.

By further abstracting over these canonical variants, using numbers as variables, we obtain this unique generalization:

(*by* 0 (*perform* 1 (*some* 2 (*for* 3 (*that* (*conduct* 4))))))

Following (Veale 2005) we call this a *structural hash*, which serves as a key for *all* structures that produce the same hash. So, to find potential analogues for a given structure, we simply abstract to its hash, and directly look up all other structures in the ATLAS database with the same abstraction key.

Mapping & Retrieval of Analogous Structures

The domain of an idea such as *priest* or *magic* is the set of all structures associated with the term in the ATLAS database. To flesh out an analogy between a source set of structures S and a target T, the systematicity principle of (Gentner 1983; Falkenhainer, Forbus, and Gentner 1989) dictates that the most comprehensive structures are mapped first. As a guide, we define $r(\sigma)$ as a measure of the richness of a structure σ:

$$r(\sigma) = \log_{10} \sum_{i=0}^{n} count_i(\sigma) 10^i$$

The deepest level of nesting in σ is n, and $count_i(\sigma)$ is the number of σ's sub-parts at level i. For our canonical rewriting of the claim that scientists perform work for a lab that conducts experiments, $r(\sigma) = 5.05$. For the facts (*work_in* scientist lab) and (*serve* priest congregation), $r(\sigma) = 0$.

Structural hashing makes the mapping process highly efficient. For each structure σ_S in S, we calculate $hash(\sigma_S)$ and index σ_S under this key. For each structure σ_T in T, we calculate $hash(\sigma_T)$ and use this key to lookup the set of equivalent structures $\{\sigma_S\}$ that were just indexed for S. We use $r(\sigma_T)$ to order our search of the structures in T, so that the first mapping produced, $\sigma_T : \sigma_S$, is the most systematic. As we proceed through other structures in T of successively lower richness, we greedily merge any other mapping that is produced into the cumulative whole if it is coherent to do so.

The retrieval of an idea/domain S to analogically describe a target T proceeds in a similar fashion, and follows the "many are called but few are chosen" principle of (Forbus, Gentner, and Law 1995): using r to rank the structures $\{\sigma_T\}$ of T from most to least rich, we calculate $hash(\sigma_T)$ for each structure in turn, and retrieve all other structures indexed under this key in the ATLAS database. From these we extract a set $\{S_0, S_1, ..., S_n\}$ of candidate sources for T. For a candidate S_i, the quality of the mapping of S_i to T is a function of both the number and the richness of the alignable structures that unite T and S_i. Should we reward sources with many alignable structures of low richness – unconnected facts – over sources with a few rich, or just one very rich, alignment? The quality q of S_i as a source for T is defined over the alignable structures $\{\sigma_0, \sigma_1, ..., \sigma_n\}$ of S_i as follows:

$$q(S_i, T) = \sum_{j=0}^{n} r(\sigma_j)^\beta$$

where β is our non-linear reward for systematicity. So when $\beta = 3$, which it is by default, $q(S_i, T)$ becomes the sum of the cubes of the richness of each structure σ_j in S_i that can be mapped onto an equivalent structure σ_k in T. The choice of β allows us to reward analogies with a single systematic core over those with many disjoint, low-level similarities.

The use of abstraction to rewrite structures in a canonical form, and then turn those forms into generic hashes, allows for the efficient mapping between two sets, or domains, of structures, as well as for the efficient retrieval of analogies. Every structure in every domain is indexed by these abstract hashes, making abstraction central to the organization of the ATLAS knowledge-base and the processes that act upon it.

Abstraction as an Organizing Principle

Abstraction is achieved through a suitable choice of structured representation (Forbus, Liang, and Rabkina 2017). In the case of ATLAS, this is constructed from the content of a flatter, more disjoint and less mapping-rich knowledge-base. To achieve scale, diversity and structural depth, we use over two-thousand abstraction rules to link separate triples in the representation of (Veale 2014) into causal combinations. We have seen a *thru*-rule that links a person in a place to the activity conducted there, but most rules are *cross*-rules that link two facts about the same subject. Some cross-rules link goals to outcomes, to build structures such as the following:

(*if* (*can* (*by* seeking (*seek* **scientist** advance (*as* outcome))))
(*can* (*by* advancing (*enhance* **scientist** sophistication
(*of* science)))))

An alignable structure in a different domain is the following:

(*if* (*can* (*by* seeking (*seek* **priest** proselyte (*as* supporter))))
(*can* (*by* spreading (*enhance* **priest** spread
(*of* doctrine)))))

Other rules integrate (Veale 2011)'s stereotypical properties dataset to add highly salient features to these structures:

(*by* working (*perform* **scientist** work
(*in* (*some* science
(*when* exacting (*lack* flexibility))))))

(*by* serving (*perform* **priest** service
(*in* (*some* church
(*when* dogmatic (*lack* flexibility))))))

By coalescing and cross-combining representations in this way, we can populate ATLAS with structures for over 9,000 concepts, from *science* to *politics* to *magicians* to *war*. For these, ATLAS defines a total of 1.7 million canonical structures which, in turn, are abstracted into 114,000 structural hashes. Canonical rewriting increases the alignability of ATLAS's contents: less than 2% of canonical structures cannot be mapped to, or aligned with, at least one other structure.

As a good analogy may hinge on the mapping of a single systematic structure, Table 1 shows the mean number of candidate sources for a target if we impose a minimum richness (r) for their core alignable structures. So, although the very deepest structures ($r \geq 8.0$) are not commonplace, supporting analogies for just 1 in 8 targets, ATLAS also supports a great many analogies at lower levels of structural richness.

Minimum r value	Mean sources per target
2.0	2396.55
3.0	558.25
4.0	503.21
5.0	218.89
6.0	40.88
7.0	6.74
8.0	0.12

Table 1: The mean number of sources retrieved per target domain with alignable structures of a minimum richness r.

Conclusion: Abstraction Finds a Way

Analogy is both a starting and an end-point in the cycle of inspiration. For instance, the author Michael Crichton based *Jurassic Park* on *Westworld*, a film that he directed in 1973, in which the robotic cowboys of a Western-themed park run amok. To motivate their homicidal rampage, Crichton would invent the analogy of a *computer virus*, more than a decade before hackers would make it a literal reality. Crichton had been inspired to write *Westworld* after a trip to Disneyworld, where the mechanical pirates about which Malcolm would later joke sparked an abiding interest in science gone wrong. With advances in genetics and cloning, Westworld's killer robots would later become the test-tube dinosaurs of *Jurassic Park*, where Malcolm's *Pirates of the Caribbean* retort would reconnect Crichton's analogy with its Disney origins.

Abstraction is key to successful, far-reaching analogies (Gentner and Hoyos 2017; Mitchell 2021), and so we put abstraction at the core of ATLAS's representation of the world. It is abstraction that allows us to relate superficially different experiences so as to perceive a deeper level of similarity. In recent years, advances in Large Language Models (LLMs) have made analogy one of the many "language games" (Wittgenstein 1953) that these models can play with great fluency and aplomb. Since LLMs produce sequences of language tokens, they conflate the tasks of inventing, presenting and talking about an analogy; they do not deal directly with, or produce, explicitly structural representations. Nonetheless, they can go far beyond the simple proportional analogies of the SAT test. For instance, when asked to generate an explanatory metaphor for *quantum mechanics*, GPT-4 (Achiam et al. 2023) responds with a richly systematic (and rather insightful) use of the *theatre* as a source domain. In its mapping, which it elaborates in detail, the stage is classical physics, while the unseen ropes, pulleys and trapdoors behind and below the stage comprise the quantum realm.

When compared to the nimbleness and generative reach of LLMs, even on tasks such as analogical invention, the structured representations of symbolic AI can seem like relics of a bygone era. But even the dinosaurs of *Jurassic Park* could be reinvented for a new age by splicing in some novel DNA. A long history of ideas in structure mapping is packed into ATLAS not to compete against LLMs but to work with them. ATLAS is an open repository[1] with a scale comparable to that of AnalogyKB (Yuan et al. 2024), but no such resource is ever complete enough or robust enough by itself. We thus anticipate two scenariosåå in which ATLAS can work with an LLM, to either guide the LLM or to be enriched by it. In the first, relevant ATLAS structures can be retrieved for a target, to drive *few-shot* prompting for analogies in an LLM. These few-shot exemplars can guide an LLM to not just invent and discuss a new analogy, but to formalize a comparable representation for T and a source domain S. These symbolic representations can then be checked for consistency and added, in the second scenario, to ATLAS if they pass logical muster.

Acknowledgments

This paper and resource are wholly the work of the author.

[1] www.kaggle.com/datasets/mtatlas/atlas-analogy-structures

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; and et al. 2023. GPT-4 Technical Report. *arXiv preprint 2303.08774*.

Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. Structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41(1):1–63.

Forbus, K. D.; Gentner, D.; and Law, K. 1995. Mac/fac: A model of similarity-based retrieval. *Cognitive science* 19(2):141–205.

Forbus, K. D.; Liang, C.; and Rabkina, I. 2017. Representation and computation in cognitive models. *Topics in cognitive science* 9(3):694–718.

Gentner, D., and Hoyos, C. 2017. Analogy and Abstraction. *Topics in Cognitive Science* 9(3):672–693.

Gentner, D., and Toupin, C. 1986. Systematicity and surface similarity in the development of analogy. *Cognitive Science* 10(3):277–300.

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

Hofstadter, D. R., and Mitchell, M. 1995. *Fluid Concepts and Creative Analogies. Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books. chapter The copycat project: A model of mental fluidity and analogy-making.

Holyoak, K. J., and Thagard, P. 1989. Analogical mapping by constraint satisfaction. *Cognitive Science* 13(3):295–355.

Mitchell, M. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences* 1505(1):79–101.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4444–51. AAAI Press.

Turney, P., and Littman, M. L. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning Journal* 60(1-3):251–278.

Veale, T., and Keane, M. T. 1997. The Competence of Sub-Optimal Structure Mapping on 'Hard' Analogies. In *Proceedings of IJCAI'97, the 15th International Joint Conference on Artificial Intelligence*, 21–26. Morgan Kaufmann.

Veale, T. 2004. WordNet sits the S.A.T.: A Knowledge-based Approach to Lexical Analogy. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'2004. IOS Press.

Veale, T. 2005. Analogy Generation with HowNet. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, 1148–53. Lawrence Erlbaum Associates.

Veale, T. 2011. Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'2011)*.

Veale, T. 2014. Coming good and breaking bad: Generating transformative character arcs for use in compelling stories. In *Proceedings of ICCC'14, the 5th International Conference on Computational Creativity*.

Veale, T. 2015. Fighting words and antagonistic worlds. In *Proceedings of the Third Workshop on Metaphor in NLP, held as part of NAACL HLT 2015, Denver, Colorado*, 87–94.

Vrandecčíc, D., and Krötzsch, M. 2014. Wiki-data: A free collaborative knowledgebase. *Communications of the ACM* 57(10):78–85.

Wittgenstein, L. 1953. *Philosophical Investigations (Philosophische Untersuchungen)*. London, UK: Blackwell Publishing. Translated by G.E.M. Anscombe.

Yuan, S.; Chen, J.; Sun, C.; Liang, J.; Xiao, Y.; and Yang, D. 2024. AnalogyKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol 1)*, 1249–65.

ISBN: 978-989-54160-7-3

A Development and Teaching Framework for Codenames

Robert Morain
Computer Science Department
Brigham Young University
Provo, UT 84602
rmorain2@byu.edu

Brad Spendlove
Department of Computer Science
Randolph College
Lynchburg, VA 2450
bspendlove@randolphcollege.edu

Dan Ventura
Department of Computer Science
Brigham Young University
Provo, UT 84602
ventura@cs.byu.edu

Abstract

A significant challenge associated with developing creative agents is the necessity of reconciling tractable evaluation metrics with the subjective nature of creativity. Traditional approaches often rely on bespoke methodologies that require significant abstraction and domain-specific engineering, creating barriers to rapid iteration. Language games like Codenames offer an alternative paradigm by providing a structured environment with inherent evaluation mechanisms, enabling clearer feedback loops for agent development. This work introduces an open-source resource designed to streamline creative agent development for the game Codenames. The resource consists of a web application for real-time game simulation and evaluation and a modular Python client that supports both spymaster and guesser agent roles. Initially presented at an *ICCC24* workshop, the system was refined through a code jam session with computational creativity researchers before being deployed in a study set in a graduate-level computational creativity class. Validation with 25 students demonstrates the effectiveness of the resource as both a development tool and an educational instrument. Quantitative and qualitative analysis of a user survey shows that the resource facilitates meaningful iterative improvements, exploration of approaches to clue generation and semantic reasoning, and application and understanding of fundamental CC concepts.

Introduction

Methods for the external evaluation of creative computational (CC) systems have long been the subject of debate, and thinking on the subject continues to evolve (Ritchie 2007; Pease and Colton 2011b; Jordanous 2012; Bown 2014; Lamb, Brown, and Clarke 2015; Ventura 2016; Carnovalini et al. 2021; Peeperkorn, Brown, and Jordanous 2023). In a significant sense, this question of how to evaluate CC systems is foundational in that it plays a large role in specifying the goals of the field and what it means to accomplish them. Historically, many evaluation methods have aimed to measure inherently subjective creative attributes, such as quality, novelty, typicality, and surprise, whose interpretations and relative importance vary across individuals and domains. Given that accurate evaluation is critical for advancement and maturation of the field, the limitations imposed by the approximation of subjective features represent a major challenge. To address this issue, Spendlove and Ventura proposed that competitive language games could serve as creative tasks with well-defined goals (2022). The win/loss outcome of the game eliminates the need to approximate subjective measures by providing an objective proxy measure of creativity. The well-defined goals of the game do not reduce the need for creativity on the part of the system. Rather, the constraints of the task provide an opportunity to highlight creative gameplay.

In particular, the authors explore the design and evaluation of CC systems with the game Codenames (Chvátil 2015), a word-based guessing game. In the game, two teams of players are tasked with identifying which of 25 cards, drawn at random from a large deck, belongs to their team. One member of each team serves as the spymaster. They are given a secret key that shows which words belong to each team. The teams then take turns in which the spymaster gives a one-word clue, accompanied by a number, that attempts to communicate a subset of the team's words to their teammates. The clue word must relate to the meanings of the word cards, and the clue number represents how many cards are intended to be related to the clue. Their teammates then guess one card at a time. If they correctly guess a word that belongs to their team, they may continue guessing, otherwise the turn is over. Wrong guesses may reveal opponents' words or even instantly lose the game. The challenge for the spymaster lies in the puzzle of coming up with a clue that strongly relates to their team's cards without unintentionally relating to any others.

Creative domains are characterized by their extremely large combinatorial spaces. Playing the spymaster role constitutes a creative task given the vast number of possible solutions (semantic graphlets that indicate positive and negative relationships between candidate clue words and the target words to be guessed) and the complexity of language inherent in the gameplay. Designing an agent to play the game is challenging, as it requires both a deep understanding of language and potentially some kind of theory-of-mind-like model of the way teammates understand and think about language. The agent must employ creativity to efficiently select clues from a large solution space for each possible game state in a (typically) limited amount of time. These requirements position Codenames as a challenging creative

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

task that can serve as an ideal mechanism for the evaluation of creativity through gameplay (Spendlove and Ventura 2023).

To kickstart the research community's interest in games as creative tasks, an introductory workshop was held as part of *ICCC24*. In this workshop, participants were introduced to the necessary theory and mechanics of gameplay before being invited to modify a provided agent or create a new one. A web application with a Python client developed for the workshop helped participants build and evaluate their agents. This enabled participants to examine the behavior of an agent, come up with new ideas for how to improve it, and efficiently build a working prototype of their solution. The workshop outcomes provided further evidence of the usefulness of Codenames as a platform for CC research.

During the workshop, participants demonstrated the ability to successfully and rapidly design, implement, and iteratively improve a creative agent to play Codenames using the resources provided. This inspired the idea to evaluate the resource's effectiveness in CC education. To do this, a study was developed for a graduate-level introductory CC class. In the study, students were assigned to develop a creative agent using an improved version of the Codenames resource developed for the workshop. Students then engaged in an iterative development process, during which they had the opportunity to evaluate their agents in a tournament, identify areas for improvement, and redesign their agents. Finally, students participated in a second tournament to evaluate their improved agents. Throughout the experience, students completed two surveys designed to gain insight into the effectiveness of the resource for the development process and to gauge student's understanding of fundamental CC concepts.

Motivated by the positive experiences of the workshop participants and graduate student study participants, this paper addresses the need for expanded access to game-based creative evaluation metrics by publishing this open-source resource for building and evaluating agents that play Codenames. We anticipate that this resource will serve the research community in many ways, and we put forward these two use cases as examples of its utility. We have found that the game's engaging nature draws people in, sparks ideas, and motivates them to design agents to play it. This framework leverages those unique qualities to make the development process both enjoyable and effective in research and pedagogical settings.

This paper contains a report on the workshop, a description of the Codenames framework and web application, the results of the study, and discussion of future use cases and applications. By providing a practical tool for game-based creativity research, this work aims to expand access to well-defined evaluation metrics and encourage further exploration of competitive language games as a testbed for CC systems. All of the code used for this paper is available on Github.[1]

[1] `https://github.com/rmorain/codenames-ai-client` and `https://github.com/rmorain/codenames-workshop`.

Workshop Report

The purposes of the workshop were to introduce our work to the broader research community, engage in discussion of the creativity of gameplay and game-playing agents, and host a Codenames agent code jam. To support these activities, we developed a programming and UI framework for playing Codenames with human and AI players. The framework includes a default spymaster agent capable of generating clues for any Codenames game state without additional setup. The game UI is a user-friendly web interface that makes it easy to create and play Codenames games. These tools facilitated explanation, discussion, and experimentation at all levels of the workshop.

At the beginning of the workshop, we played a sample game of Codenames with the participants via the web application. The web application automates game rules such as turn passing and scoring, which helped the participants quickly understand the spymaster and guesser tasks. After playing a single game of Codenames, the participants already had ideas for how to develop AI agents to play the game. We discussed how they, as human players, approached clue-giving and guessing, and how those strategies could inform the design of an automated agent.

A large portion of the workshop was dedicated to a code jam, in which participants developed their own Codenames agents. Participants leveraged our Codenames framework and UI to quickly prototype working agents that could play against humans and one another. This allowed them to actually implement and test the ideas they shared in the discussion and see how they performed in the game. Their agents' designs included strategic selection of which subset of cards to relate to a clue, querying large language models to generate clues, and clustering the different colored cards using k-means. Participants reported that this exercise was an enjoyable and interesting way to interact with a creative task that was new to them and pointed to the framework as enabling this positive experience. The framework allowed them to focus on the core task of developing an agent that can create clues, without spending unnecessary time writing boilerplate or interface code.

In our concluding discussion, participants shared their takeaways from the workshop. We discussed whether playing Codenames is a creative task, the potential benefits of studying games as creative tasks, and how Codenames' focus on language could cross over to other language-based creative domains. We were pleased by the positive reaction to the workshop and encouraged that our Codenames framework facilitated rapid prototyping and experimentation.

Resource Description

The code base for the Codenames development framework includes a game server, a web client with a graphical user interface (UI), and Python clients for the spymaster and guesser roles. Both the web UI and the Python clients connect to the same game server, allowing for gameplay with any combination of human players and AI agents in the various game roles. The Python clients each wrap an agent program that makes the gameplay decisions, marking a clear

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

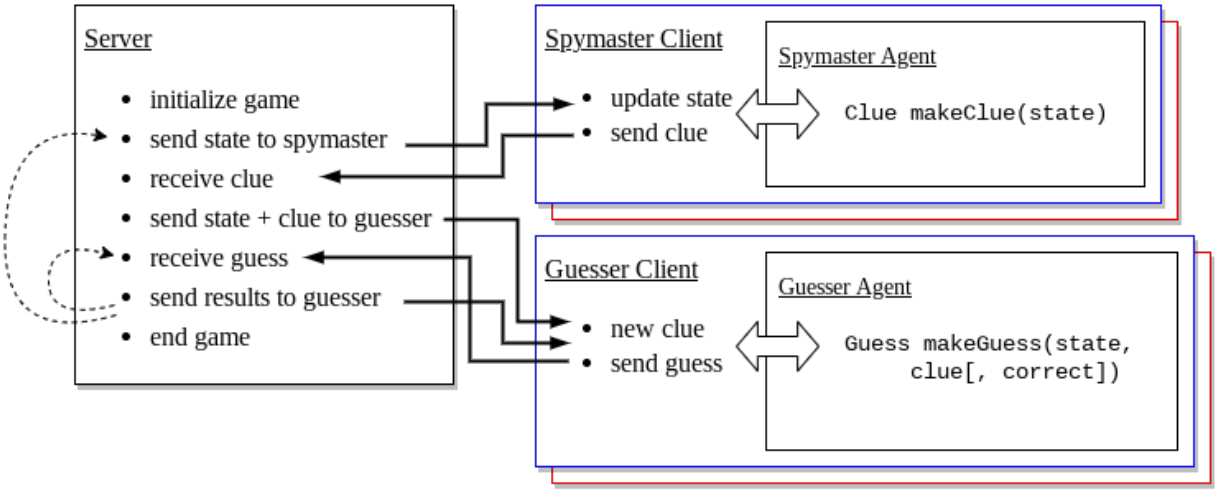


Figure 1: The server/client/agent architecture of the workshop codebase, including information flow between the components. Note that there are two spymasters and guessers, one pair each for the red and blue teams.

division between the creative agent code and the server interface code.

The primary use cases for this code base are playing Codenames as a human player and modifying the agent code to implement new creative play agents. Neither of these requires knowledge of or modification to the client/server code. We host a public instance of this server here: `https://mind.cs.byu.edu/codenames`, where anyone can connect and play via browser or client.

The backend development framework is divided into five components: the server that runs games, a spymaster agent, a spymaster client, a guesser agent, and a guesser client. The two types of clients facilitate communication between the agent and the server. See Figure 1 for a diagram of the complete architecture, including information flow between components. These components are described in detail in this section, and some gameplay details that were left out of the description of Codenames above are included here for completeness.

The server tracks the state of the game in progress, communicates that state to the clients, listens for their moves, validates those moves, and updates the internal game state. This state can be retrieved by either the web UI or the client code. This section focuses on the latter. Codenames is not a complex game to represent computationally. At the beginning of a new game, the board of 25 word cards is drawn from the deck of 400 (the list of which is available online). A coin is flipped to decide which team goes first, red or blue, then the first team is randomly assigned 9 cards and the other 8. The remaining cards are neutral, except for one assassin card chosen at random. This completes the starting board state, and the game is ready to play.

Each new game is assigned a unique four-letter code that clients use to join that specific game session on the server. Four clients must connect to the server: the two spymasters and the two guessers. When playing the game, each team may have any number of guessers, but from an information perspective they act as one unit when choosing words to guess, so one client suffices.

The server sends the game state to all clients, including the secret key to the spymasters, and waits for the current team's spymaster to return a clue consisting of one word and a number. The server then sends the clue to the guesser client. The guesser can guess one word card at a time, after which the server updates the game state by revealing the hidden identity of that word card (red, blue, neutral, or assassin). If the guess is correct (revealed to belong to their team), the guesser may guess again, up to a maximum of one plus the clue number. If the guess is revealed to be the assassin, the game immediately ends, and the guessing team loses the game. Otherwise, if the guess is incorrect or the team is out of guesses, play passes to the other team and the process repeats.

Thus, the server requires the following functions:

- Initialize game, update state
- Transmit game state to current spymaster
- Receive clue from spymaster, update state
- Transmit game state and clue to guesser
- Receive guess from guesser, update state
- Transmit result of guess to guesser
- End game, update state

The clients handle communication between an agent and the server, so the clients' required functions mirror the server's. The spymaster client receives the game state from the server and passes it to the spymaster agent, which returns a clue word and number. The client transmits that to the server and waits until it hears from the server again to get the next clue.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

The guesser client has two similar information flows. In the first, it receives the clue and current game state from the server, passes them to the guesser agent, gets a guess back, and transmits it to the server. In the second, it receives the result of the previous guess from the server and passes it to the agent, sending another guess back if necessary.

The spymaster client requires the following functions:

- Receive updated game state
- Submit clue

And the guesser client requires these functions:

- Receive game state and new clue
- Submit guess

The agents' functions follow from the description of their clients. The spymaster agent has a `makeClue` method that accepts a game state and returns a clue. The guesser agent has a `makeGuess` method that takes a game state and clue and returns a guess. The basic game state data consists of the list of word cards and a list of the revealed identity of each word card (if any). Additionally, the spymaster receives the secret key showing all the cards' colors and the guesser receives the current clue and number of guesses remaining. The clients are very thin wrappers around the agents but are engineered this way to allow for simple modification to or replacement of the underlying agent. This separates the client/server communication logic from the agent so that its developers can focus only on the relevant creativity task.

To run the provided automated clients, the user only needs to provide the game code and team color. The client connects to the game server, then repeatedly polls the current game state, calling upon its agent module to generate a game action when required.

Finally, the code base also includes a web UI client for human gameplay that connects to the server using the same interface as the automated clients. This client presents a UI that mimics the layout of the physical game and through which users can input the required game actions. The game state is represented by a grid of word cards that are either uncovered, showing the word, or covered with the appropriate color after they're guessed. One client is used for both the spymaster and guesser role by either prompting the spymaster for a clue word and number or allowing the guesser to click on a word card to guess it when prompted. This simple interface is low overhead and allows human players to play with the AI agents, bringing more richness and interactivity to the application. The web client also provides an observer mode which simply displays the game state without allowing for any game action input. This is useful, for example, for displaying the game to spectators while others play the game.

The entire code base is public, allowing future users to reuse or modify any portion necessary. The server can be run locally, or games can be run on our public server. At the same time, the code base is designed to require only minimal modification to run a custom gameplay agent. The programmer must only replace the appropriate agent module within the client. They will then be able to play their agent against humans or other agents without modifying the client, server, or web UI code.

Example agents

To complete the code base, we include two basic gameplay agents, a spymaster and a guesser. These agents allow the clients to be run out-of-the-box without writing any code, and they serve as a reference implementation for the simple client-agent interface. Furthermore, the agents were designed to be simple to understand even for relatively inexperienced programmers, providing a jumping-off point for their own modification and improvement.

Both example agents are powered by word embeddings, which are vector representations of words that capture their semantic meaning (Mikolov et al. 2013). These high-dimensional vectors enable computers to understand and process natural language to facilitate many tasks such as sentiment analysis, machine translation, and text classification. There exist straightforward analogues between vector operations and the cognitive task of comparing the relationships between words, which makes them a natural fit for Codenames agent development.

The cosine similarity between two word embedding vectors corresponds to the semantic similarity of the words those vectors represent (Orkphol and Yang 2019). By comparing the similarity of a clue word candidate to the word cards, the spymaster agent determines whether that clue will effectively relate to those cards. This information is used to inform the choice of a clue that is similar to the team's words and dissimilar from the others. Conversely, the guesser agent calculates the similarity between a given clue and the word cards and guesses the one(s) that most closely relate.

These simple agents are complete agents in that they can play the game, and they use a reasonable natural language understanding approach to do so. They perform reasonably well in-game, but may often generate unclear clues. Of course, this is by no means the only way to design these agents, which is our motivation for publishing this code base. These starter agents will serve the needs of users with widely differing programming skills and AI or NLP backgrounds.

CC Pedagogy Study

Computational creativity in education is beginning to gain traction, both as an approach to teaching creativity and computational thinking (Kakavas 2019; Fragapane and Standl 2021; Yee-King, Fiorucci, and d'Inverno 2024) and as a first-class pedagogical subject itself. In particular, Ackerman et al. discuss the importance of CC education in the broader context of AI and provide guidance for instructors on how to approach CC pedagogy. Further, they suggest the benefit of students engaging in introductory assignments to reinforce the material and develop their own ability to develop creative systems (2017). Ventura provides an introductory guide for how students (as well as researchers) might approach building such CC systems by identifying their essential components (2017). This Codenames framework advances these pedagogical objectives by providing

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

students with a structured environment for developing creative agents while avoiding the complexity of subjective evaluation methods.

To validate the effectiveness of this Codenames framework in an educational setting, we conducted a study with students in a graduate-level CC course, examining the feasibility and efficacy of the resource as a tool for teaching fundamental CC concepts and providing structure for creative agent development. In what follows, we outline the study's objectives; describe how the resource was incorporated into the course curriculum; analyze outcomes based on two student surveys; and discuss key lessons learned from the experience.

Study goals

The study aims to assess the pedagogical value of the Codenames resource for

1. helping students develop their own (creative) agents
2. helping students learn and apply CC concepts

The resource is designed to help students develop their own agents by providing clear goals and tools to achieve success. It establishes a well-defined creative task with a clear evaluation metric inherent in the game. Students can easily evaluate the effectiveness of their spymaster agent by directly playing with it in the guesser role through the web UI or by running a provided test that shows the spymaster's response to a random game state. For this study, the effectiveness of the Codenames resource is evaluated through its capacity to facilitate a) students' development of their initial creative agents and b) their identification of ways to iteratively improve those agents.

The resource has been developed in a modular way that allows students to consider and implement different aspects of a creative system, providing a natural introduction to fundamental computational creativity concepts that might be covered in a course on the subject. Success is measured by students' ability to analyze their work through the theoretical frameworks presented in course readings.

Study methodology

The study was conducted during the first two weeks of a graduate-level CC course of 25 students that met twice per week (the entire study spanned a total of five class periods of 75 minutes each). The format of the study was as follows:

Period 1: Introduction of framework and codebase and discussion of study outline and goals. Students were introduced to the game Codenames and the basic features of the Python client, the example spymaster agent, and the web UI and were instructed on how to modify the provided base spymaster agent or create a new one.

Period 2: Class discussion of CC concepts from the following papers: (Wiggins 2006; Ritchie 2007; Ventura 2017; Colton, Charnley, and Pease 2011; Pease and Colton 2011a) and how they might be related to the task of building a creative/competitive spymaster agent.

Period 3: First in-class tournament, enabling students an initial evaluation of their agents' performance through direct competition with their peers' implementations. In each tournament match, students competed in one-on-one games where they played as guessers via the web interface while their agents served as spymasters connected via the Python client running on their personal computers. The tournament consisted of an initial group stage followed by a knock-out round, featuring the winners from each group, with two semi-finals and one finals match. The group stage consisted of 3 groups of 6 (student/agent) teams and one group of 7 teams and featured round-robin play, so that every team faced every other team in a match. As a result, each student/agent team played (at least) 5 matches; four teams played (at least) 6 matches; and two teams played (at least) 7 matches. Group matches were run concurrently, while knock-out matches were run consecutively and broadcast so the entire class could watch and participate. The students completed a post-tournament survey in which they described their implementation approaches, reflected on their agents' performance, and made plans for improving their agents. Students were assigned to continue working on their agents in anticipation of a second tournament.

Period 4: Class discussion of general "questions for CC researchers", another way to orient the students to important ideas in the field.

Period 5: Second in-class tournament, identical format to the first one. A second follow-up survey asked students to report the extent and effectiveness of the modifications they made to their agents and reflect on their experience with the Codenames project.

Survey Questions

Below are the questions included in the two post-tournament surveys. The first survey focused on the students' agent design and performance. The second survey followed up on the modifications the students made and their resulting changes in performance and asked students to reflect on their experiences. Questions marked with a dagger (†) were answered with a 5-point Likert scale, all others were free-response.

Post-Tournament Survey 1

- How far did your agent advance in the tournament?
- What technique(s) does your agent use to determine relationships between words (e.g., word embeddings, knowledge graphs, co-occurrence counting)?
- How does your agent measure or score the strength of word relationships?
- How does your agent identify and evaluate potential clue words?
- What criteria does your agent use to balance between covering multiple target words versus avoiding opponent and assassin words?
- How does your agent determine the optimal number of words to target with each clue?

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

- What factors do you believe contributed most to your agent's performance in the tournament?
- What specific concepts or approaches from our course readings have influenced or could improve your agent's design? Please reference at least one reading in your response.
- How will you improve your agent for the next tournament?
- How satisfied are you with your overall learning experience with this project?†
- Do you have any suggestions for how your learning experience could be improved?

Post-Tournament Survey 2

- How far did your agent advance in the tournament?
- Rate the extent of changes made to your agent since the first round:†
- What specific modifications did you make to your agent? For each modification you made, please: 1) Describe the original implementation, 2) Explain what you changed, 3) Share why you made this change, 4) Indicate if the change was successful
- Which of these modifications had the most significant impact on your agent's performance?
- How effective were the insights from the first round in improving your agent? (Scale: Very ineffective to Very effective)†
- Describe the most significant lesson from the first round that influenced your agent's improvement
- What do you think about Codenames as a creative task from the spymaster's perspective? How do the game's constraints facilitate or inhibit creative play?
- To what extent does your agent demonstrate the following aspect of creativity: Novelty†
- To what extent does your agent demonstrate the following aspect of creativity: Value†
- To what extent does your agent demonstrate the following aspect of creativity: Intentionality†
- Do you think your agent is creative? Why or why not?
- How would you rate the client's (the provided code) effectiveness in supporting agent development?†
- How would you rate the web application's (the UI) effectiveness in supporting agent development?†
- What suggestions do you have for improving the client or the web application?

Study results and analysis

Initial survey responses revealed two primary approaches to word association: embedding-based methods (utilizing word2vec, BERT, GloVe, or spaCy) and large language models. Agent strategies generally focused on basic similarity metrics between clue words and targets while avoiding opponent words and the assassin card. While students demonstrated general comprehension of course readings, many struggled to meaningfully connect theoretical concepts to their agent implementations. When asked how they would improve their agents, students mentioned adding a history to the agent to avoid repeating words, using LLMs, multiple LLMs, or reasoning LLMs. One student who used an LLM in the first round reported that strong word association skills were not sufficient without enhanced strategy: "I had a superior or competitive method with exception that it did not protect me from the assassin." Another student suggested that because many students wanted to replace the example agent with an LLM-based one, they would have "loved to learn some prompt engineering techniques or techniques to help LLMs be more creative."

The winner of the first tournament developed an agent that provided clues that reveal the coordinates of their own team cards. While this type of gameplay violates Codename's official rules,[2] the experience provided the opportunity for students to engage in debate over how an agent's performance in the game relates to the creativity of the agent. In this case, the cheater agent communicates with the guesser using a cleverly designed code, in which the first letter of the word corresponds to the column and the vowels indicate the rows that contain team words. Students agreed that while the agent itself did not possess any creative ability, the student who developed the agent and its bespoke code exhibited a high level of creativity (cf. the discussion on game rules/design and their intended effect on creativity in (Spendlove and Ventura 2023)). Students also noted that, of course, the cheating spymaster agent is only useful for a guesser that knows its special code, while their general approaches to making agents that use semantic relationships between words as the game rules intend may generalize well to many other guessing partners.

The second survey revealed substantial evolution in agent design, with students implementing more advanced approaches, including the use of reasoning LLMs, larger LLMs, proxy agent testing, enhanced embedding metrics, iterative clue refinement, and prompt engineering. The students succeeded in using the provided code as a jumping-off point for building better agents. As one student reported in the survey "I feel proud I got it working and it did an okay job, better than word2vec, which was a success."

Finding connections between class readings and the agent development assignment resulted in significant improvement in at least one case. In the first survey, one question asks students to identify specific concepts from class readings that could help improve their agent's design. One student connected the aesthetic evaluation described in the FACE model (Pease and Colton 2011b) to a potential internal proxy evaluation that evaluates clues before submitting them to the guesser. After implementing this idea, the student advanced to the final round of the second tournament. This is just one example of how creative agent development

[2]Specifically, it violates the rule that states "Your clue must be about the meaning of the words. You can't use your clue to talk about the letters in a word or its position on the table." (Czech Games Edition 2015)

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

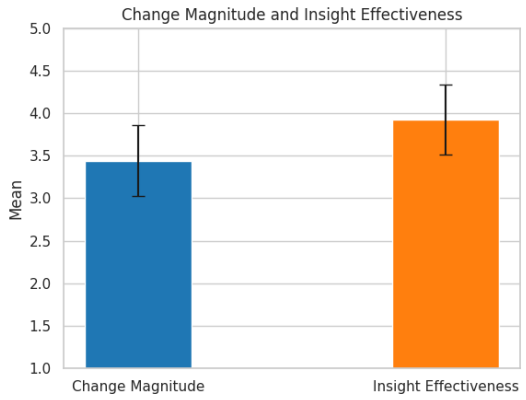


Figure 2: Mean responses from the second survey of students asked, "Rate the extent of changes made to your agent since the first round" (scale: minor tweaks to complete redesign) and "How effective were the insights from the first round in improving your agent?" (scale: very ineffective to very effective).

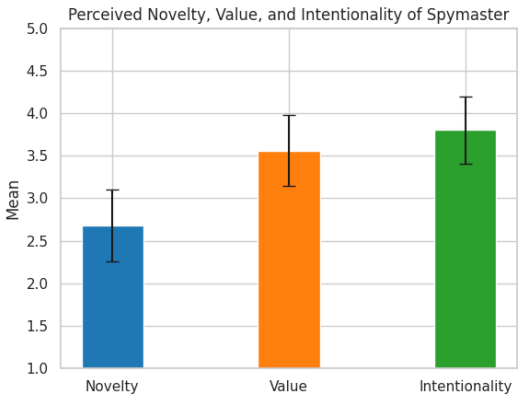


Figure 3: Responses when students were asked, "To what extent does your agent demonstrate the following aspects of creativity?" in the second survey.

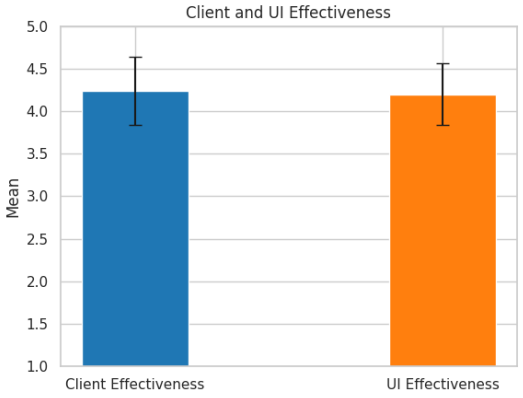


Figure 4: Client and UI effectiveness (scale: not effective to very effective).

facilitated by the Codenames resource can lead to effective learning outcomes for students.

Students generally recognized the spymaster role as inherently creative due to its solution space that, while constrained by the game rules, is still very large. However, some argued that creativity emerges from the approach rather than the task itself, noting that deterministic solutions like lookup tables, while potentially effective, would not demonstrate creative behavior.

Survey responses indicated substantial agent iteration between tournaments, with students reporting a mean modification magnitude of 3.44 on a scale from 1 (minor tweaks) to 5 (complete redesign) (see Fig. 2). Students found the first tournament experience particularly valuable for informing improvements, rating its effectiveness at 3.92 out of 5.

When evaluating their agents along three dimensions of computational creativity (Ventura 2017), students reported moderate to high scores for quality (3.56) and intentionality (3.80), but notably lower scores for novelty (2.68) (see Fig. 3). While students generally considered their agents creative, their justifications focused primarily on their own creative process in agent development rather than demonstrating understanding of formal CC frameworks and definitions.

Both the client framework and web interface proved effective development tools, receiving mean ratings of 4.24 and 4.20 respectively on a 5-point scale (see Fig. 4). Overall satisfaction with the learning experience was high (4.44 out of 5).

The second tournament demonstrated increased competitive balance, with closer games and more diverse winners compared to the first round, suggesting successful agent refinement across the cohort.

In their feedback about the Codenames framework itself, students expressed that they enjoyed using the tools and the learning experience the project provided. Their feedback also highlighted areas where the code base could be improved, such as simplifying the design of the example, displaying a history of game actions in the web client UI, and the addition of a developer mode to the web client that would allow one client to play all roles simultaneously in a test game. We will incorporate their feedback into continuing development and improvement of the public code repository.

Discussion & Future Work

Our objective in presenting this resource is to improve the accessibility of creative gameplaying agents in the CC research community, because accessibility impacts adoption and advancement. We emphasize that this extends to CC pedagogy and presents a powerful tool for introducing stu-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

dents to our research community and supporting them as they develop their own creative systems.

One of the major use cases for this resource is in supporting and expanding CC research. The success of the resource at the workshop demonstrates its value in this regard. The framework provides researchers with a structured environment for evaluating word association methods and developing optimal game strategies through objective performance metrics. Creative gameplay represents a fertile avenue for future CC research, and we intend that this resource will be an entry point for the same. It enables automated gameplay between agents to gauge their relative capabilities, even if those agents are engineered differently. Future workshops or demonstrations could include competitions between various researchers' agents or against benchmark systems. We hope that usage of this resource will spur investigation into other creative games to further explore these types of creative domains.

The second major use case for this resource is for students in a CC class or for anyone looking for a low-barrier-to-entry into CC. Games provide inherent evaluation mechanisms through their rule structures, eliminating the need for subjective quality metrics that often complicate traditional CC tasks. Games are also generally well-defined tasks when compared to traditional CC tasks. This added structure is ideal for beginners and is well-suited to creating tools to aid the development process. Our results show that this Codenames resource was successful in its goal to help students more effectively develop their first creative agent. However, more work needs to be done to improve the connection between fundamental CC concepts and the experience of the students developing and evaluating their agents. Overall, both the development and competitive aspects of the learning experience were well received.

CC pedagogy is the primary example we have given for how to apply our Codenames framework, specifically in guiding students in creating their first CC systems. As such, it is appropriate to directly discuss how our Codenames agents include the elements of a CC system outlined in (Ventura 2017), namely *domain*, *knowledge base*, *aesthetic*, *representation*, *generation*, *evaluation*, *conceptualization*, and *translation*.

A subset of these elements are built into the task of playing Codenames. The domain is fixed, which notably includes a clear evaluation criteria—or aesthetic, to use Ventura's term—which many other creative domains lack. The phenotypic representation is also fixed by the game rules to a clue word and number. This representation lends itself to direct translation and trivial phenotypic evaluation; the example agent does not reason about those explicitly. It is notable that half of the considerations for building a CC system are obviated by the task of playing a creative game with well-defined rules!

The remaining elements comprise the design of the example spymaster agent. That agent is powered by word2vec embeddings, which were trained on a knowledge base of text corpora. The embeddings themselves represent the conceptualization of that knowledge. The genotypic representations used by the agent are the way it represents (e.g., vector embeddings) and uses them to manipulate subsets of word cards and candidate clues. The agent considers many such genotypes and internally evaluates them to select the best clue as its generated output. Thus, the included agent serves as a complete example of a CC system for this domain and will serve as a valuable learning tool for students.

Playing Codenames effectively requires understanding semantics and the relationships between words. Those skills are also relevant to a wide range of language-based creative tasks. Therefore, developing and refining an agent's semantic understanding in the context of Codenames could translate into improved performance at other creative tasks or subtasks. Future work could also expand this framework to incorporate other language games with similarly structured evaluation mechanisms, such as Decrypto, Wavelength, and the *New York Times*' Connections. Developing accessible tools and frameworks for a larger set of creative gameplay tasks provides a natural vector for accelerating progress in computational creativity research.

Conclusion

We have presented a public framework for building Codenames agents and playing games against both computational and human players. These tools are simple to use and modify, providing a solid jumping-off point for future computational creativity research. Additionally, these tools are well-suited for use in CC pedagogy. Students can interact with a pre-built CC agent, play games with and against it directly, and modify or replace its functionality without having to write client-server or game simulation code.

We have demonstrated the efficacy of this framework via a 25-student study in which graduate students in a CC class wrote their own agents to play in a Codenames tournament. Quantitative and qualitative data gathered during this study indicate that it was a fun and engaging learning activity. This helped introduce students to CC design principles in a focused and entertaining manner, paving the way for them to design their own CC agents in other domains.

The Codenames framework presented herein will also benefit the CC research community, as demonstrated by the results of a code jam workshop at *ICCC24*. It allows for rapid prototyping of new CC agents, simplifies the normally complex task of creative evaluation, and could serve to compare CC agent performance via direct competition. We foresee many benefits to expanded research in creative gameplay, which we hope our framework will help facilitate.

References

[Ackerman et al. 2017] Ackerman, M.; Goel, A.; Johnson, C. G.; Jordanous, A.; León, C.; Pérez y Pérez, R.; Toivonen, H.; and Ventura, D. 2017. Teaching computational creativity. In *Proceedings of the 8th International Conference on Computational Creativity*, 9–16. Association for Computational Creativity.

[Bown 2014] Bown, O. 2014. Empirically grounding the evaluation of creative systems: Incorporating interaction design. In *Proceedings of the 5th International Conference on Computational Creativity*, 112–119.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

[Carnovalini et al. 2021] Carnovalini, F.; Harley, N.; Homer, S. T.; Roda, A.; and Wiggins, G. A. 2021. Meta-evaluating quantitative internal evaluation: A practical approach for developers. In *Proceedings of the 12th International Conference on Computational Creativity*, 213–217. Association for Computational Creativity.

[Chvátil 2015] Chvátil, V. 2015. Codenames.

[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.

[Czech Games Edition 2015] Czech Games Edition. 2015. Codenames rules. https://czechgames.com/files/rules/codenames-rules-en.pdf. Official rulebook for the Codenames board game.

[Fragapane and Standl 2021] Fragapane, V., and Standl, B. 2021. Work in progress: Creative coding and computer science education – from approach to concept. In *Proceedings of the IEEE Global Engineering Education Conference*, 1233–1236.

[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.

[Kakavas 2019] Kakavas, P. 2019. Computational thinking and creativity in K-6 education. *Formamente* 14(2):93–108.

[Lamb, Brown, and Clarke 2015] Lamb, C.; Brown, D. G.; and Clarke, C. 2015. Human competence in creativity evaluation. In *Proceedings of the 6th International Conference on Computational Creativity*, 102–109. Brigham Young University.

[Mikolov et al. 2013] 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*, volume 26, 3111–3119.

[Orkphol and Yang 2019] Orkphol, K., and Yang, W. 2019. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet* 11(5):114.

[Pease and Colton 2011a] Pease, A., and Colton, S. 2011a. Computational creativity theory: Inspirations behind the FACE and the IDEA models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 72–77.

[Pease and Colton 2011b] Pease, A., and Colton, S. 2011b. On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal. In *Proceedings of the Artificial Intelligence and the Simulation of Behaviour Symposium on Computing and Philosophy*, 15–22.

[Peeperkorn, Brown, and Jordanous 2023] Peeperkorn, M.; Brown, D.; and Jordanous, A. 2023. On characterizations of large language models and creativity evaluation. In *Proceedings of the 14th International Conference on Computational Creativity*, 143–147. Association for Computational Creativity.

[Ritchie 2007] Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

[Spendlove and Ventura 2022] Spendlove, B., and Ventura, D. 2022. Competitive language games as creative tasks with well-defined goals. In *Proceedings of the International Conference on Computational Creativity*, 291–299. Association for Computational Creativity.

[Spendlove and Ventura 2023] Spendlove, B., and Ventura, D. 2023. A constraint-centric accounting of some aspects of creativity. In *Proceedings of the 14th International Conference on Computational Creativity*, 332–336. Association for Computational Creativity.

[Ventura 2016] Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the 7th International Conference on Computational Creativity*, 17–24. Sony CSL.

[Ventura 2017] Ventura, D. 2017. How to build a CC system. In *Proceedings of the 8th International Conference on Computational Creativity*, 253–260. Association for Computational Creativity.

[Wiggins 2006] Wiggins, G. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowlege-Based Systems* 19:449–458.

[Yee-King, Fiorucci, and d'Inverno 2024] Yee-King, M.; Fiorucci, A.; and d'Inverno, M. 2024. Designing an AI-creativity music course. In *Proceedings of The International Conference on AI and Musical Creativity*.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Poems to Lyrics: Automated Rephrasing with Beat Alignment

Mohamad Elzohbi, Richard Zhao
Department of Computer Science, University of Calgary
Calgary, Alberta, Canada T2N 1N4
{melzohbi, richard.zhao1}@ucalgary.ca

Abstract

This paper explores the generation of English lyrical lines that align with the rhythm perceived from language. We use the ByT5 transformer model, which processes text at the byte level, to rephrase poetry lines according to a given beat pattern. Our approach builds on earlier studies on beat patterns perceived from English conversations by integrating a guided paraphrasing task with rhythmic constraints. Additionally, we introduce PARAPOETRY, a large-scale parallel dataset of automatically generated poetry line rephrases. Our results demonstrate that the proposed model can effectively rephrase lyrics to align with specific beat patterns using only textual data. A human evaluation study further confirms that English-speaking participants largely agree with our model's beat alignment. We further assess other qualities, including fluency, meaningfulness, and poeticness.

Introduction

Recent advances in natural language generation (NLG) have renewed scholarly interest in creative data generation, particularly for creative language such as poetry (Elzohbi and Zhao 2023). Transformer-based models (Vaswani et al. 2017) have significantly improved text quality through transfer learning, which in turn has opened avenues for integrating and exploring additional creative elements into the generation process. While both poetry and song lyrics are expected to display poetic qualities in a sense, they differ in their structural and rhythmic requirements. Unlike poetry, song lyrics must adhere to a rhythmic structure that aligns with musical accompaniment. This distinction highlights the need for specialized models that incorporate sub-syllabic aspects, ensuring that generated lyrics are not only linguistically and poetically coherent but also musically coherent.

While traditional approaches have primarily focused on rhyme and poetic coherence, comparatively little attention has been given to the rhythmic structure of both poetry and lyrics. In particular, song lyrics require a dual focus: they must display creative language while also conforming to the rhythmic patterns of musical accompaniment. This challenge calls for methods that integrate linguistic creativity with the control over rhythmic patterns.

In this work, we focus on the task of rephrasing poetry lines to achieve alignment with desired beat patterns. Our objective is to preserve the poetic quality of the original lines while ensuring that the rephrased versions conform to the rhythmic requirements essential for musical integration. As a result, our focus will be on *lyrical poetry*. To achieve this, our approach leverages the ByT5 transformer model, renowned for its byte-level processing and fine-grained control over sub-syllabic features critical to beat alignment. We introduce PARAPOETRY, a large-scale dataset of poetry line rephrases automatically generated via GPT-3.5, which we believe it can serve as a valuable resource for various creative language generation tasks.

Our work contributes to the intersection of natural language processing and music, offering insights into beat-aligned language generation and laying groundwork for future research in this domain. Our contributions can be summarized as follows:

- We present PARAPOETRY, a large-scale dataset that pairs human-written poetry lines with automatically generated rephrases. We believe that this dataset can serve as a valuable resource for future research in creative text generation.
- We demonstrate the utility of the proposed dataset for our approach through a guided rephrasing task that integrates beat pattern alignment into the generation process. Our model is trained to generate lyrical poetry that align with desired beat percussion.
- We conduct human evaluations confirming that verses generated by our models adhere to the specified beat patterns, highlighting the potential of our method to support creative lyric writing through a co-creative process.

Related Work

Many studies focus on the rhyme of lyrics but often overlook their essential rhythmic aspect (Xue et al. 2021). The task of generating lyrics given a melody has been explored using various techniques and in different languages. Some approaches use text-based or symbolic melody inputs, while others use audio-based representations. Songmass (Sheng et al. 2021) introduced transformer-based encoder-decoder models for modelling lyrics and symbolic melodies. The authors used supervised training to align text in lyrics with notes in melodies, generating textual lyrics from MIDI-based melodies and vice versa. Xue et al. (2021) proposed

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Statistic	Value	Description
Exact Match Percentage	0.061%	Percentage of cases where original lines and rephrases are exactly identical, indicating that rephrases are not mere copies.
Rhythm Match Percentage	0.285%	Percentage of instances where the beat pattern of original lines and rephrased lines match, demonstrating rhythmic diversity.
Total Poetry Lines	1,038,743	Total number of poetry lines in the original dataset.
Total Rephrased lines	934,054	Total number of lines rephrased by GPT-3.5.
Unique lines	907,450	Number of unique original poetry lines in PARAPOETRY.
Line-Rephrase Pairs	4,674,190	Total count of line-rephrase pairs present in PARAPOETRY.
Unique Rephrases	4,659,608	Number of unique rephrases in PARAPOETRY.
Maximum Rephrase Count	271	Maximum number of rephrases attributed to a single line.
Minimum Rephrase Count	1	Minimum number of rephrases a line has received.
Average Rephrase Count	5.00419	Average number of rephrases per line.

Table 1: Summary Statistics of the PARAPOETRY Dataset.

an autoregressive transformer-based model to produce beat-aligned lyrics by collecting and processing a rap lyric dataset to identify beats and word timestamps, inserting beat positions into the lyrics as special tokens.

Other approaches extract phonetic features from lyrics and compare them with music. For example, Oliveira, Cardoso, and Pereira (2007) proposed heuristics for generating lyrics that align with a MIDI-based melody by matching stressed syllables with strong beats. More recently, Chen and Teufel (2024) explored similarities between tonal contours in melody and lyrics using phonetics and musicology theories, extracting pseudo-melodies from lyrics to fine-tune a transformer-based model (mBART) for lyrics generation given a topic and a melody.

In our previous work (Elzohbi and Zhao 2024), we developed a model that extracts beat patterns from English words by training ByT5 (Xue et al. 2022) to replace or insert words so that they align with a desired beat pattern. We refer to this task as the *substitution task*. This approach builds upon earlier research into perceptual beat locations in spontaneous English conversations (Allen 1972; Rathcke et al. 2021), where participants tapped their fingers in synchrony with perceived beats in spoken words, demonstrating that rhythmic beats often align with vowel onsets.

However, the *substitution task* is limited to replacing words within a lyrical line to match a beat, and lacks the capacity to generate complete lines that align with a given beat pattern. Moreover, while prior evaluation relied on automated metrics, it did not incorporate human judgment to assess rhythmic alignment or the overall quality of the modified lines. In this work, we address these limitations by developing a model capable of rewriting entire lines of lyrical poetry to match a target beat pattern, while preserving both semantic content and poetic expression. We will refer to this task as the *rephrase task*.

A parallel dataset with beat information is ideal for the beat-aligned rephrase task. Unfortunately, specialized datasets for poetry and lyrics are generally scarce, with most studies creating their own datasets using automatic methods (Greer et al. 2019; Sulun, Oliveira, and Viana 2023; Martinez-Sevilla et al. 2023; Chen and Teufel 2024). For this reason, we automatically generated a large-scale dataset of poetry rephrases, that we will call PARAPOETRY, by prompting OpenAI's GPT-3.5. In this work, we demonstrate the utility of this dataset for beat-aligned lyrical poetry rephrasing.

Methodology

Task Formalization

Given a line of lyrics L with an undesirable beat pattern $G2B(L)$, the task is to generate a rephrased line L' with a desired beat pattern $G2B(L')$. The function $G2B(\cdot)$, first introduced in our previous work (Elzohbi and Zhao 2024), is a grapheme-to-beat transformation that maps text to a sequence of beat/rest annotations. A beat unit is marked with a "1" at a vowel onset within a consonant-vowel sequence, as well as at the initial vowel of null-onset syllables where either a glottal stop serves as its onset or the onset is borrowed from the preceding syllable. All other positions, including the second elements of long vowels and diphthongs, consonants within clusters, or consonants in codas, are marked with a "0", unless they are repositioned as onsets for the following null-onset syllables.

Dataset Creation, Augmentation and Processing

Due to the lack of human-generated poetry-rephrase parallel datasets, we created PARAPOETRY, which comprises of human-written poetry lines with five automatically generated rephrases each. We started with the preprocessed subset of the Gutenberg poetry corpus we used for the substitution task, focusing specifically on English quatrains. Using GPT-3.5,[1] we generated five rephrases for each human-written poetry line, with an emphasis on maintaining poetic language.

Through iterative prompt engineering, we manually refined and evaluated sample responses of around 15 prompts to achieve the best results in guiding GPT-3.5 in retaining poetic devices such as metaphors and similes of the originals

[1]We used version: `gpt-3.5-turbo-0125`

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

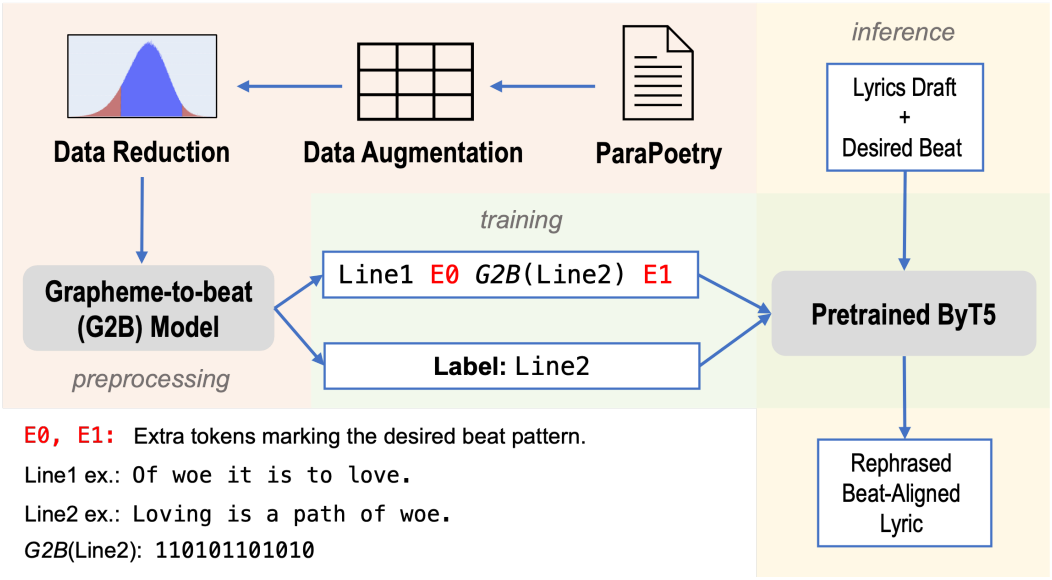


Figure 1: The preprocessing, training and inference pipelines.

in the rephrases while following formatting requirements. The prompt included providing two illustrative examples and demonstrating how to handle anomalies as a few-shot. We also experimented with temperature settings, which control the randomness and creativity of the output. We found that a temperature of 0.7 struck a balance between following formatting instructions and producing poetic paraphrases. In the generation process, there were instances where GPT-3.5 ignored rephrasing certain lines or generated more or less than 5 rephrases resulting in more than 4.6 million line-rephrase pairs (see statistics in Table 1).

In PARAPOETRY, the original lines are written by real poets, unlike the generated rephrases. To help our model be biased toward generating human-like rephrases, we augmented the dataset by swapping the rephrases with the original lines, avoiding a bias toward automatically generated output. We then filtered out very long and very short lines, retaining only those with more than 8 and fewer than 128, after examining the length distribution of the dataset we found that the majority of the lines fall within this range.

We analyzed the dataset to test lexical dissimilarity between the original lines and the rephrases. We calculate the PINC score (Paraphrase In N-gram Change) (Chen and Dolan 2011). The PINC(o, p) score measures the degree of lexical novelty of the automatically generated rephrases by comparing the generated rephrase p with the original line o, analyzing differences in n-grams. It measures the proportion of n-grams in the paraphrase that are absent in the original, averaging this across n-gram lengths. A score of "1" means the original line and the rephrase are very different, and a score of "0" means they are very similar lexically.

$$\text{PINC}(o,p) = \frac{1}{N}\sum_{n=1}^{N} 1 - \frac{|n\text{-gram}_o \cap n\text{-gram}_p|}{|n\text{-gram}_p|}$$

After calculating the PINC scores with $N = 4$, we observed a normal distribution centered around 0.521, which indicates a moderate level of lexical dissimilarity between the original lines and the rephrases in our dataset. Since lexical dissimilarity can impact the rhythmic diversity of generated lines by altering the positions of vowel onsets, and consequently the beat pattern, a moderate average allows for flexible rephrasing while accommodating varying levels of beat variation relative to the original lines.

We trim the tails of the distribution keeping data with PINC scores between 0.2 and 0.8. This excludes data with minimal substitutions and potential garbling.

Experimental Setup

Dataset Split

To accommodate our available resources, we sample a subset of 3 million pairs from the processed dataset to train and evaluate our models with an evaluation set containing $15,000$ examples.

Model Training

For this task, we fine-tuned pre-trained ByT5 models on the processed subsets using a guided rephrasing objective. For each lyric line L and its rephrased version L', we computed the beat pattern $G2B(L')$. To mark the boundaries of the guiding beat pattern, special tokens (E_0 and E_1) were added and concatenated to the end of the lyric line L. The input is structured as $I = (L, E_0, G2B(L'), E_1)$. The model was trained to predict L' (see Figure 1).

We fine-tuned the pre-trained ByT5-base version (Xue et al. 2022) on the task described using the training subset to get **ByT5-R**. As the rephrase task is arguably more complex, yet similar, than the substitution task, we also investigated whether the knowledge gained from the model fine-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

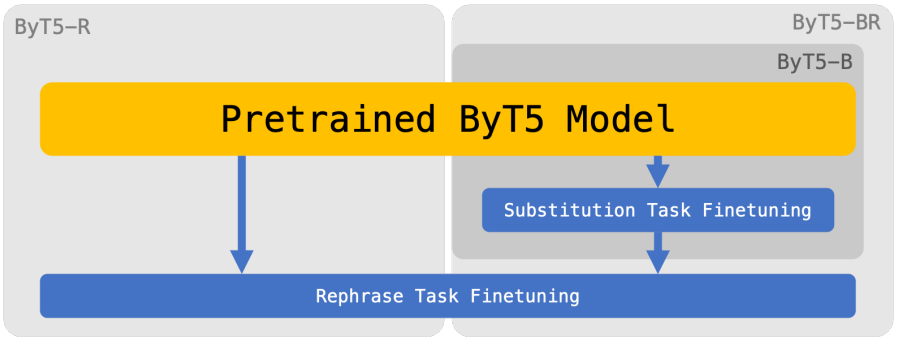


Figure 2: Curriculum learning pipeline.

tuned on the substitution task (**ByT5-B**) could be transferred to enhance rephrasing through a curriculum learning strategy. Curriculum learning is a strategy that gradually raises complexity to help the model learn more effectively. Forms of increased complexity may include increasing the data (or model's experience) complexity during the training, expanding the model's capacity by adding or activating more layers or units, or training on simpler tasks before moving to more complex ones (Soviany et al. 2022).

Here, we apply curriculum learning by first training on the simpler substitution task, then on the more complex rephrasing task (see Figure 2). This yields the model **ByT5-BR** using **ByT5-B** as a starting point.

As a baseline, we train another ByT5-base with attention masks set to zero in the beat pattern positions, named **ByT5-R-Zero**, preventing the model from attending to these patterns during training.

All models were trained for four epochs on an NVIDIA A100 GPU. The learning rate was set to $3e-4$ with a cosine scheduler and a weight decay of 0.1. The training batch size was set to 128, and the evaluation batch size set to 16.

Automated Evaluation Metrics

The main objective of our task is to replace a verse so that it aligns with a specific beat pattern while ensuring the new verse maintains semantic similarity to the original. We use automated evaluation metrics to measure the sentence semantic similarity, lexical dissimilarity and beat alignment.

Sentence similarity: We used Sentence-BERT[2] (Reimers and Gurevych 2019) to encode the original and the rephrased lines. Cosine similarity was utilized to measure the similarity between their embeddings.

Lexical dissimilarity: To assess the lexical diversity of the generated rephrases, we use the PINC score with $N = 4$. This score is used as a descriptive metric to show the level of lexical novelty in the rephrased lines in comparison with similarity level.

Alignment Scores: assess how well the generated lines align with the required beat pattern. We utilized two evaluation metrics:

- *Exact Alignment Accuracy:* This metric determines if the generated line aligns precisely with the expected beat rhythm, resulting in "0" for non-alignment, and "1" for exact alignment.
- *Levenshtein similarity:* This metric quantifies alignment by calculating the complement of the Levenshtein distance d_{lev} between the beat of the generated line B_o and the expected beat B_e, normalized. It allows for some flexibility, recognizing cases where alignment may not be exact but is still reasonably close. The Levenshtein similarity is defined as follows:

$$\text{Sim}(B_o, B_e) = 1 - \frac{d_{lev}(B_o, B_e)}{\max(\text{len}(B_o), \text{len}(B_e))}$$

Experimental Results

This section examines the performance of the fine-tuned models in meeting the task objectives. We evaluate their performance based on sentence similarity, lexical dissimilarity, and beat alignment.

Model	Sim. (%)	PINC (%)	Acc. (%)	Lev. (%)
ByT5-R-Zero	74.42	**45.60**	7.00	84.71
ByT5-R	81.29	30.78	91.01	99.46
ByT5-BR	**82.04**	29.71	**94.34**	**99.67**

Sim. = Sentence Similarity, *Acc.* = Exact Alignment Accuracy, *Lev.* = Levenshtein Similarity Score.

Table 2: Performance Comparison Between Models.

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Sentence Similarity

The results in Table 2 show that our models achieved higher similarity to the original lines compared to the **ByT5-R-Zero** model. However, the **ByT5-R-Zero** model, which does not account for beat patterns, attained a higher PINC score of 45.6. In contrast, our models generated lines with lower but still moderate PINC scores, indicating higher lexical similarity to the original lines. It seems that our models tend to use words similar to those in the original lines, as compared to the base model with higher PINC score and lower similarity.

Upon reviewing the distribution of Levenshtein similarity scores in PARAPOETRY, we observed a roughly normal distribution centered around 0.8. This suggests that, although there is moderate lexical dissimilarity, the model likely attends to target beat patterns that are very similar to those of the original lines, resulting in rephrased lines with relatively higher similarity scores and lower PINC scores compared to the baseline model.

Alignment Scores

Our models demonstrated a high level of beat alignment accuracy. The **ByT5-BR** model achieved the highest exact alignment accuracy, exceeding 94%. The **ByT5-R** model also performed well, reaching approximately 91%, though slightly lower than **ByT5-BR**, highlighting the effectiveness of curriculum learning in this context. The Levenshtein similarity scores with both models achieving over 99%, indicates strong rhythmic alignment with only minor deviations. In contrast, the **ByT5-R-Zero** model, which did not attend to the beat patterns, exhibited extremely low exact alignment accuracy of around 7% and a significantly lower Levenshtein similarity score of 84.71%. These results emphasize that proper training is essential for achieving effective beat alignment. These findings confirm that our fine-tuned models can generate verses that align with beat patterns while maintaining similarity to the original lines.

Human-Centric Evaluation

Although automated natural language processing metrics can provide useful insights, they fall short of capturing the subjective human experience of poetic language and creative qualities. In particular, creative aspects are difficult to quantify using standard metrics. To address this limitation, we designed a survey to collect human judgments on rhythmic alignment and creative qualities. Our evaluation study was organized into three main components:

- **Word-Beat Alignment Experiment:** Participants were presented with a pairwise comparison task, where they were asked to fill in the blank with the word or phrase that best matches a given beat pattern as well as indicating their preference based on fluency and sense in completing the lyrical line.
- **Line-Beat Alignment Experiment:** In a similar pairwise comparison setup, participants were asked to select the line that best matches a given beat pattern as well as indicating their preference based on fluency and sense.
- **Qualitative Evaluation:** Since evaluating creative qualities of a poetic text may require more context than only individual words or lines, participants were asked to rate complete verses generated by our model on multiple aspects, including fluency, sense, and overall poetic quality.

The first two experiments evaluate the ability of the **ByT5-B** and **ByT5-BR** models to align words and rephrase lines according to specific audio-based beat patterns. It also serves as evidence supporting the concept of perceptual beat locations in English text. The third experiment evaluates the **ByT5-BR** model's capability to generate complete quality verses while adhering to the beat alignment task.

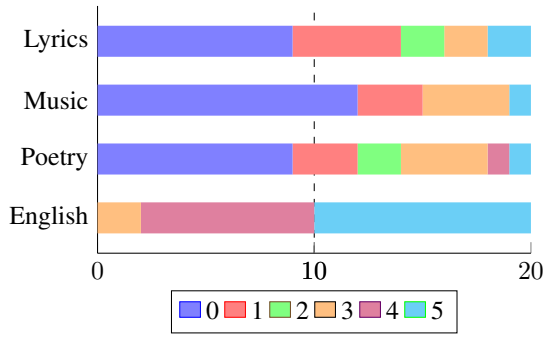


Figure 3: Demographics of the Participants.

Participants Demographics

Twenty participants were recruited for this study via convenience sampling, drawing on graduate students at the University of Calgary from the Department of Computer Science and acquaintances with a background in poetry or lyrics composition. Each participant self-assessed their English proficiency on a 0 (no understanding) to 5 (native speaker) scale and their familiarity with writing poetry, lyrics, and music composition on a 0 (no experience) to 5 (published-level expertise) scale. Although respondents who selected "0" for English proficiency would have been removed, no exclusions were necessary. Sessions were conducted online through Qualtrics[3] (survey links were sent by email) and took approximately 45-60 minutes to complete. Figure 3 illustrates our sample, which comprised an approximately equal mix of native and non-native English speakers as well as individuals with and without prior experience in poetry, lyrics writing, or music composition.

The participants were informed that the study aimed to investigate the impact of large language models on learning beat patterns and maintaining textual coherence. However, they were not told that some of the lines were human-written while others were generated by a model. To address participant fatigue effect, the survey was structured with simpler questions presented first, progressing to the more complex ones. Additionally, a $20 CAD gift card was promised as an incentive for participation. While our sample of twenty is

[3] https://www.qualtrics.com/

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Experiment	Category	Alignment Experiments		Fluency & Sense Experiments	
		# Chosen	%	# Chosen	%
Word-Beat	ByT5-B	129	64.5	54	27
	Original	71	35.5	106	53
	Both	-	-	40	20
Line-Beat	ByT5-BR	131	65.50	42	21
	Original	69	34.50	98	49
	Both	-	-	60	30

Table 3: Experimental results for word-beat and line-beat alignment evaluations.

relatively small, and there is no universally agreed-upon optimal size for this type of exploratory study, we nevertheless evaluated our results for statistical significance.

Note that two participants did not complete the qualitative evaluation which came last in the survey, resulting in 18 responses for that portion of the study.

Word-Beat Alignment Experiment

In this experiment, we randomly selected 20 lines from the evaluation set. We first performed a manual grapheme-to-beat transformation, identifying the rhythmic beat pattern present in each of the selected lines placing a beat at each vowel onset and a rest elsewhere. For each line, we masked between 3 and 6 syllables from various positions in a line (i.e., beginning, middle, and end) to ensure a balanced sample and sufficient contextual information for the model. The syllables can be one or more complete words. The **ByT5-B** model was then prompted to generate candidate replacements. The required beat pattern for the replacement was deliberately chosen to differ from the original words' beat pattern while matching their syllable count. In this way, we ensure that participants are not simply matching syllable-beat counts.

Each participant was presented with a pairwise fill-in-the-blank task in which the original words and the generated words are both presented. Participants can listen to the provided beat pattern in audio format (a clave hit for each 1 in the beat sequence and a pause for each 0) and were asked to choose the words that best align with it. In addition to assessing rhythmic alignment, participants were asked to select the words that best completes the lyrical line in terms of overall fluency and sense or if they think both words fit the context equally. For this study, we defined *fluency* for the participants being the smoothness, flow, and ease with which the verses can be read or spoken and *sense* being the logical coherence and clarity of the lines where lines that make more sense are those that convey a clear and understandable meaning or narrative. The definitions were presented to the participants every time they were asked to evaluate the lines. To manage the survey duration, each participant evaluates a random subset of 10 out of the 20 examples.

The results of the word-beat alignment evaluation (see Table 3) indicate that the generated words were chosen in 64.5% of cases, compared to 35.5% for the original version. This indicates that human participants found the generated words to align better than the original words. Conversely, the fluency and sense evaluation showed a slight preference for the original version at about 53% over the other choices. Although the generated words solely were preferred only 27% of the cases, in 47% they were preferred at least as much as the original words with no significant statistical difference on a paired two-tailed t-test with a p-value = 0.63.

Line-Beat Alignment Experiment

Following the word-beat alignment task, participants were introduced to the line-beat alignment task. For this experiment, we randomly selected 20 lines from the evaluation set and rephrased each one of them using the proposed **ByT5-BR** model trained on the rephrasing objective. For every line, we manually identified the beat patterns and then assigned a new different beat pattern that preserves the original number of beat count. This ensured that both the original and rephrased lines contained the same number of syllables, thereby preventing syllable count from influencing the evaluation.

Each participant was presented with a random subset of 10 out of the 20 examples. The participants were asked to perform a pairwise comparison between the original and the rephrased lines, selecting the one that best aligned with the beat pattern presented in audio format. Additionally, participants were asked to rate which line has more fluency and sense, or if they believed both lines were equally acceptable.

The line-beat alignment results indicated that the generated lines were chosen in 65.5% of the cases, while the original lines were chosen in 34.5% of the cases. This suggests that participants generally perceived the generated lines as better aligning with the given beat pattern. In contrast, the evaluation of fluency and sense showed a preference for the original lines at approximately 49%, with the generated lines being preferred solely in 21% of cases. However, in 51% of the cases, the generated lines were considered at least as acceptable as the originals, with no significant statistical difference with a p-value = 0.88.

We also investigated whether the order of tasks influenced participant responses by examining the practice effect. Comparing responses from the easier word-beat alignment task and the more challenging line-beat alignment task revealed no significant statistical difference with one-tailed t-test with p-value = 0.44.

Moreover, we manually assessed the alignment accuracy

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Aspect	Description
Fluency	The smoothness, flow, and ease with which the verses can be read or spoken.
Sense	The logical coherence and clarity of the lyrics. Lines that make more sense are those that convey a clear and understandable meaning or narrative.
Meaningfulness	Being rich in detail, evoking strong mental pictures.
Unpredictability	How unexpected or surprising the combination of words in the lyrics or lines are.
Originality	In the use of words to you from the perspective of the participant.
Humanness	Verses that are more likely to be written by a Human have more hummanness.
Poeticness	The qualities that make the lines artistic, expressive, and emotionally resonant. Poetic lines often use poetic devices like metaphor and similes, etc.

Table 4: Definitions of the qualitative evaluation metrics.

in the first and second experiments. In the word-beat experiment, one out of 20 phrases generated by **ByT5-B** did not align perfectly with the beat pattern. In the line-beat experiment, two of the generated examples by **ByT5-BR** did not align perfectly with the beat pattern. These three instances were retained in the evaluation and received 24 evaluations, the generated semi-matching patterns were chosen 8 times out of 24, suggesting that counting syllables was probably preferred over partially matching sub-syllabic alignment.

Qualitative Evaluation

In this part of the study, we assessed the linguistic and poetic qualities of our model's output. Our model is trained to rephrase single lines. Single lines by themselves may not contain enough information to be evaluated on a creativity scale. So in this experiment we wanted to test the utility of our rephrasing model to generate complete verses as compared to real poems written by human poets.

We began by randomly selecting 10 complete poems from amateur poets (normal users of the platform) and well-known famous poets scraped from the *allpoetry* platform.[4] To avoid any influence of rhyme schemes which our model does not handle, we removed poems that followed a distinct rhyming pattern. From the remaining poems, we extracted 10 verses from the famous poems and 10 from the amateur poems. These verses were then stripped of punctuation and line breaks before being processed by the **ByT5-BR** model, which is then used to generate rephrased versions of each verse. Punctuation and line breaks were then restored manually to the generated lines to ensure they were presented in a similar format as the original lines. Although this experiment does not focus on beat patterns, our model requires a beat pattern to generate the rephrased lines. For this purpose, we assigned a beat pattern using the Bjorklund algorithm (Bjorklund 2003) to distribute the sequence of 1's and 0's in the original beat as evenly as possible. This demonstrates a potential real-world use case for our model.

Participants were presented with both the original and the rephrased verses, and they were asked to rate several qualities of the generated verses using a slider scale from 0 (worst possible) to 100 (best possible). For clarity, the definitions for each evaluated quality were provided alongside every example (see Table 4 for the definitions we used).

Table 5 summarizes the differences in quality ratings between original and generated verses as evaluated by the participants. The differences are computed as the mean rating for the original verses minus the mean rating for the generated verses ($\overline{X}_o - \overline{X}_r$) on a scale from 0 to 100. The table further breaks down the results for all participants, native English speakers, and participants with poetry, lyrics, or music composition experience, along with the corresponding p-values to assess statistical significance.

Fluency and Sense *Fluency* was rated significantly higher for the original verses across all groups. All participants reported a mean difference of ($+9.41$, $p = 0.043$), while native speakers and participants with experience showed even larger differences ($+18.22$, $p = 0.010$ and $+20.62$, $p = 0.019$, respectively). Similarly, *Sense* yielded even more noticeable differences, with the original verses outperforming the generated ones by ($+15.98$, $p = 0.002$) for all participants, ($+28.18$, $p = 0.0004$) among native speakers, and ($+29.89$, $p = 0.002$) among those with experience. The results suggest that human-written verses are perceived as more fluent and coherent, especially by native participants who have a deeper familiarity with language and participants who were more exposed to poetic language and song lyrics.

Meaningfulness, Unpredictability and Originality *Meaningfulness* also favored the original verses, with a significant overall difference of ($+10.09$, $p = 0.027$). Although the differences for native speakers ($+12.66$, $p = 0.050$) and poetry-experienced participants ($+15.23$, $p = 0.057$) were slightly higher, they only approached conventional significance levels. In contrast, the *Unpredictability* exhibited a slight advantage for the generated verses, as evidenced by negative differences for all participants, though none of these differences reached statistical significance (p-values ranging from 0.159 to 0.404). This suggests that while our model's output may offer a degree of surprise, this effect is subtle and not strongly differ-

[4] https://allpoetry.com/

	All Participants		Native Speakers Only		Experienced Only	
Metric	$(\overline{X}_o - \overline{X}_r)$	p-value	$(\overline{X}_o - \overline{X}_r)$	p-value	$(\overline{X}_o - \overline{X}_r)$	p-value
Fluency	+9.41	0.043	+18.22	0.010	+20.62	0.019
Sense	+15.98	0.002	+28.18	0.0004	+29.89	0.002
Meaningful	+10.09	0.027	+12.66	0.050	+15.23	0.057
Unpredictability	-3.87	0.159	-2.71	0.299	-1.57	0.404
Originality	-9.47	0.013	-7.58	0.078	-3.77	0.142
Humanness	+9.03	0.072	+20.02	0.027	+26.94	0.001
Poeticness	+6.67	0.156	+12.00	0.118	+21.83	0.009

Table 5: Statistical comparison across participant groups.

entiated by the evaluators. *Originality* showed that the generated verses were rated as more original overall, with a statistically significant difference of (-9.47, $p = 0.013$). However, when the analysis was restricted to native speakers and those with poetry experience, the differences were reduced (-7.58 with $p = 0.078$ and -3.77 with $p = 0.142$, respectively). The non-significant difference in meaningfullness, unpredictability and originality for native speakers and experienced participants suggest that the preservation of surprising mental imagery and meaningful novelty is debatable among these groups.

Humanness and Poeticness In the case of *Humanness*, original verses again scored higher. This difference reached statistical significance for native speakers ($+20.02$, $p = 0.027$) and poetry-experienced participants ($+26.94$, $p = 0.001$), while the overall group difference ($+9.03$, $p = 0.072$) was marginal. The evaluation of *Poeticness* yielded mixed results. For the overall participant group, the difference of ($+6.67$, $p = 0.156$) suggests a slight advantage for the original verses, although this was not statistically significant. Among native speakers, the difference increased to ($+12.00$, $p = 0.118$), and for those with poetry experience, the original verses were rated significantly more poetic by ($+21.83$, $p = 0.009$). This shows that individuals with a background in poetry and lyrics were more sensitive to the these qualities and that the model's output was perceived as less poetic by these participants.

Other Implications The average quality scores of the generated verses ranged from approximately 45% to around 66%, while those for the original verses ranged from about 53% to around 76% across the evaluated qualities. This suggest that while the generated verses scored significantly lower than the original verses on some qualities, they were still rated at least as acceptable by the participants on average and that the model is capable of generating verses that are at least comparable to human-written verses in some aspects while maintaining a high level of beat alignment. Finally, when comparing the rephrasing of poems written by amateur poets and those written by famous poets, the difference in quality between the generated and original poems was more noticeable by the participants for amateur poets, even among native speakers and experienced participants. It is possible that famous poems were more challenging for participants to evaluate.

Conclusion

In this paper, we presented an approach to rephrase English lyric lines so that they align with specific beat patterns, utilizing the ByT5 transformer model. By introducing PARAPOETRY, a large-scale parallel dataset of poetry-rephrase pairs, we provided a valuable resource for researchers interested in poetry and lyrics analysis and generation. We demonstrated the utility of PARAPOETRY by fine-tuning pre-trained ByT5 models on a guided rephrasing task to reconstruct lyric lines with predetermined beat patterns. Our experimental results demonstrate that the ByT5 model can successfully generate beat-aligned lyrical poetry lines while maintaining high semantic similarity with moderate lexical diversity while demonstrating high rhythmic alignment.

The results of our human-centric evaluation provide several noteworthy insights regarding the rhythmic and creative qualities of the model-generated poetry. Overall, the experiments indicate that the proposed generative models trained on the substitution task (**ByT5-B**) and the rephrasing task (**ByT5-BR**) are capable of aligning words and lines with specific beat patterns. In both the word-beat and line-beat alignment experiments, participants selected the model-generated outputs approximately two-thirds of the time when asked to judge rhythmic alignment, which underscores the models' effectiveness in adhering to imposed rhythmic constraints. However, evaluations based on fluency and sense consistently demonstrated a slight preference for the original text. This finding is especially noticeable in the qualitative evaluation of the rephrasing of complete verses, where native speakers and participants with relevant experience rated the original verses significantly higher in terms of fluency, coherence, and overall poetic quality. These results suggest that, although the generated outputs achieve a high degree of rhythmic alignment, they still lack certain creative qualities that are found in human-written poetry.

Limitations and Future Work

Despite the progress made in aligning lyrical lines with rhythmic patterns, several limitations remain for further discussion. Although PARAPOETRY proved useful, the dataset

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Original Line	Beat Pattern	Rephrased Line
Feral landscapes bathed in twilight	100010001011010	Swamp lands in the sunset
Sure we might fish from out the mothers sons	10100101010110100101100010101l0	Genuinely from their mothers sons we'd managed to fish away
The depth and peak of a grief of this nature	110010010010100101001010	Precise and height of such a grief as this
Was not thy melody touching and choice	100100101111001001000l100	Was not your melody poignant and preferred
Then in the heart itself that knoweth	1011011000100010100	Dwell within the knowing heart itself
Come pity us all ye who see	10101110010101010100	Show mercy upon us all who observe
And scatter salt on the unmissed sepulcher	1001010100110110100100	And sow with salt the unremembered grave
Or what of tidings you abroad doo heare	101010101010100100010	Pray what whispers of tidings reach you
By Fan and Sword and Office box	10101001001101000	By Fan an Sword and Office box
Stay till the storms are o'er	1000100101011000100	Remain close until the gales cease

Table 6: Examples of original lines and their rephrased versions according to specified beat patterns, showcasing the English rephrase task. The examples are selected from the evaluation set.

was generated using automatic methods with GPT-3.5 and may contain inconsistencies or biases inherent to the automated process. Our current approach determines beat placement primarily based on vowel onsets, thereby neglecting the impact of syllabic stress on beat perception. Additionally, the human evaluation component, while insightful, was conducted with a relatively small participant pool who are largely not experts in the poetry, lyrics and music fields and with limited set of examples. Further, the model was trained to generate single lines even though the qualitative evaluation involved full verses.

The qualitative evaluation highlights a mixed performance in terms of creativity. While human-written verses scored higher in aspects such as fluency, sense, and poeticness, the generated verses were rated as more original and unpredictable, indicating that the models may introduce original word combinations and unexpected expressions. However, this by itself does not guarantee that these combinations are creative as striking a balance between adherence to other creative aspects while generating content with meaningful novelty is necessary.

Future research should consider integrating features that consider the effects of syllabic stress. Incorporating a larger proportion of human-curated data could also enhance the quality of the poetic expressions generated. Expanding the evaluation to include a more diverse participant pool especially experts in the field. Finally, extending the approach to generate full verses or entire songs may introduce additional challenges in maintaining coherence and poeticness as well as overall musical alignment over longer texts.

Supplementary Material

The source code and the prompts used are available at: `https://github.com/melzohbi/poem-rhythm-para`. ParaPoetry is available at: `https://doi.org/10.5683/SP3/WMU1BC`. Table 6 provides examples of original lines and their rephrased versions by the proposed rephrasing model.

Ethical Considerations

All experiments involving human participants in this paper received ethics approval from the Conjoint Faculties Research Ethics Board (CFREB) at the University of Calgary under protocol number REB24-0877. Participation was voluntary, and all participants were provided with a consent form informing them of their right to withdraw from the study at any time or decline participation. Data collection was conducted confidentially and anonymized by removing any identifiable information that could reveal the participants' identities.

References

Allen, G. D. 1972. The location of rhythmic stress beats in english: an experimental study i. *Language and Speech* 15(1):72–100. PMID: 5073939.

Bjorklund, E. 2003. The theory of rep-rate pattern generation in the sns timing system. *SNS ASD Tech Note, SNS-NOTE-CNTRL-99, Los Alamos National Laboratory, Los Alamos, U.S.A.*

Chen, D., and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200.

Chen, Y., and Teufel, S. 2024. Scansion-based lyrics generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14370–14381.

Elzohbi, M., and Zhao, R. 2023. Creative data generation: A review focusing on text and poetry. In *Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)*, 29–38.

Elzohbi, M., and Zhao, R. 2024. Let the poem hit the rhythm: Using a byte-based transformer for beat-aligned

poetry generation. In *Proceedings of the 15th International Conference on Computational Creativity, (ICCC'24)*, 407–411.

Greer, T.; Singla, K.; Ma, B.; and Narayanan, S. 2019. Learning shared vector representations of lyrics and chords in music. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3951–3955. IEEE.

Martinez-Sevilla, J. C.; Rios-Vila, A.; Castellanos, F. J.; and Calvo-Zaragoza, J. 2023. A holistic approach for aligned music and lyrics transcription. In *International Conference on Document Analysis and Recognition*, 185–201. Springer.

Oliveira, H. R. G.; Cardoso, F. A.; and Pereira, F. C. 2007. Tra-la-lyrics: An approach to generate text based on rhythm. In *Proceedings of the 4th. International Joint Workshop on Computational Creativity*, 47–55. London, UK: A. Cardoso and G. Wiggins.

Rathcke, T.; Lin, C.-y.; Falk, S.; and Dalla Bella, S. 2021. Tapping into linguistic rhythm. *Laboratory Phonology* 12(1):1–32.

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*, 3982–3992. Association for Computational Linguistics.

Sheng, Z.; Song, K.; Tan, X.; Ren, Y.; Ye, W.; Zhang, S.; and Qin, T. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13798–13805.

Soviany, P.; Ionescu, R. T.; Rota, P.; and Sebe, N. 2022. Curriculum learning: A survey. *International Journal of Computer Vision* 130(6):1526–1565.

Sulun, S.; Oliveira, P.; and Viana, P. 2023. Emotion4midi: A lyrics-based emotion-labeled symbolic music dataset. In *EPIA Conference on Artificial Intelligence*, 77–89. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Xue, L.; Song, K.; Wu, D.; Tan, X.; Zhang, N. L.; Qin, T.; Zhang, W.-Q.; and Liu, T.-Y. 2021. Deeprapper: Neural rap generation with rhyme and rhythm modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 69–81.

Xue, L.; Barua, A.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; and Raffel, C. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics* 10:291–306.

Transformational Creativity in Science: A Graphical Theory

Samuel Schapiro
University of Illinois Urbana-Champaign
Spiral Works
sjs17@illinois.edu

Jonah Black
University of Illinois Urbana-Champaign
jblac8@illinois.edu

Lav R. Varshney
University of Illinois Urbana-Champaign
varshney@illinois.edu

Abstract

Creative processes are typically divided into three types: combinatorial, exploratory, and transformational. Here, we provide a graphical theory of transformational scientific creativity, synthesizing Boden's insight that transformational creativity arises from changes in the "enabling constraints" of a conceptual space (Boden 1992) and Kuhn's structure of scientific revolutions as resulting from paradigm shifts (Kuhn 1962). We prove that modifications made to axioms of our graphical model have the most transformative potential and then illustrate how several historical instances of transformational creativity can be captured by our framework.

Introduction

Boden (1992) famously divides creative processes into three types: combinatorial, exploratory, and transformational creativity. Combinatorial creativity refers to the combination of concepts and ideas; exploratory creativity involves the traversal of a conceptual space; and transformational creativity entails restructuring the space itself. Historically, groundbreaking scientific discoveries—including Einstein's theory of relativity, the Copernican shift from geocentrism to heliocentrism, and the discovery of non-Euclidean geometry—have fallen under the category of transformational creativity. However, despite scientific creativity being a heavily studied topic in the philosophy and psychology of creativity (Simonton 2004; Rothenberg 2014; Simonton 2021; Koestler 1964), rigorous formalizations of transformational creativity in science have remained underexplored.

In this short paper, we formalize scientific conceptual spaces as directed acyclic graphs (DAGs), with vertices that function as generative rules for artifacts and edges that represent logical dependencies among axioms and rules. From our graphical definition, we prove that modifications made to axioms of the graph have the most transformative potential. Finally, we explain how our model can capture historical instances of transformational scientific creativity and discuss potential extensions of this paper in future work. To start, we provide a background on transformational creativity and an account of its role in scientific revolutions.

Background on Transformational Creativity

The notion of transformational creativity is first introduced by Boden (1992) as a distinct form of the creative process. Unlike combinatorial creativity that involves novel combinations of familiar elements, or exploratory creativity that entails exploring a defined conceptual space, transformations specifically target what Boden calls the "enabling constraints" of a conceptual domain: the foundational assumptions, rules, or boundaries that determine what artifacts are intelligible within it (Boden 1992; Aguiar, Atã, and Queiroz 2015). When these constraints are modified, the generative landscape is restructured, and the scope of new artifacts becomes drastically altered.

Recent work has attempted to give computational grounding to this notion. Santo, Wiggins, and Cardoso (2024) distinguish between novelty and transformation using concepts from formal learning theory (FLT), showing how transformational creativity can be operationalized as a shift in the hypothesis space informed by new data. Their framework centers the agent's capacity to adjust the criteria governing the conceptual space, rather than operating under a fixed set of rules. Similarly, Toivanen et al. (2018) describe an architecture for transformational creativity in generative systems, where meta-level modules modify the constraints that govern the search space of songs to produce novel outputs. This coheres with previous work by Wiggins (2006), who defines transformational creativity as exploratory search over the space of generative systems that collectively define a conceptual space. Ritchie (2006) takes a similar view, emphasizing the idea that transformational creativity requires modifications to the conceptual structure of a space.

Transformational Creativity in Science: Puzzle Solving, Crisis, and Revolution

Kuhn's history and philosophy of science articulates a paradigmatic view in which periods of normal science—consisting of problem solving within a predefined conceptual space—are followed by anomalies that underscore the weaknesses and inconsistencies of the existing space. These crises prompt scientists to return to first principles and modify existing paradigms or develop new ones, culminating in a scientific revolution that alters a previous worldview (Kuhn 1962). This model of the scientific process is largely consistent with the three forms of creativity (combinatorial,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

exploratory, and transformational) popular in the computational creativity community (Boden 1992; Varshney 2020; Ritchie 2006). Below, we summarize Kuhn's model, which is later adopted as a conceptual guide for parts of our theory.

Phase I: Normal Science Normal science consists of research based firmly on past scientific achievements that create the foundation for further scientific practice. These achievements establish a conceptual space (or a "paradigm") for organizing otherwise disconnected facts while also motivating the discovery of new ones. The paradigm also offers practical advantages for new work: because scientists do not need to re-justify widely accepted concepts, they can focus on core contributions that often entail exploratory or combinatorial creativity. Kuhn ascribes the term "puzzle solving" to the period of normal science and highlights that the goal of this period is *not* to produce major novelties, but rather to extend the existing paradigm, often only incrementally.

Phase II: Anomaly and Extraordinary Science The limitations of working within pre-defined paradigms are often highlighted by periods of anomaly and crisis, when the "puzzle solving" phase, characterized by exploration of existing conceptual spaces, is no longer sufficient to account for the range of phenomena the field wishes to study. This marks the beginning of an *extraordinary science* phase, hallmarked by a loosening of existing rules that promotes the eventual emergence of new paradigms. Extraordinary science often involves a return to first principles, which may entail a "recourse to philosophy" and "debate over fundamentals".

Phase III: Scientific Revolutions The resolution of a scientific crisis is a scientific revolution, including a paradigm shift and a change of worldview. This phase involves a reconfiguration of the constraints that previously governed all inquiry within a scientific domain. As a result, new artifacts abound, and their implications propagate throughout the space. In Kuhn's words:

> It's as if the scientific community has been transported to another planet, where familiar objects appear in a new light and are accompanied by previously unseen ones. (Kuhn 1962, p. 111)

In the following section, we synthesize Kuhn's structure of scientific revolutions with prior formalizations of transformational creativity into a graphical model.

Characterizing Transformational Scientific Creativity Using Graphs

In this section, we use a formal, graphical structure to define conceptual spaces, clarify the distinction between axioms and rules, and illustrate how existing constraints influence the discovery and articulation of new scientific artifacts. In the science of science, it is often common to use graphical or network frameworks to model the act of, or phenomena related to, scientific discovery (Shi, Foster, and Evans 2015; Shi and Evans 2023; Wiggins 2006; Friedman et al. 2000). Likewise, our approach focuses on dependency relationships between scientific rules and axioms, allowing us to characterize how changes to axioms may potentially transform conceptual spaces. In contrast to prior, more general accounts of transformational creativity (Boden 1992; Ritchie 2006; Santo, Wiggins, and Cardoso 2024), we focus specifically on this phenomenon in science.

Scientific Conceptual Spaces and Artifacts

In science, an artifact can be any phenomenon, fact, concept, variable, constant, technique, theory, law, question, goal, or criterion of relevance to a particular field (Simonton 2004). As explained in Boden (1992), conceptual spaces are defined in terms of the "enabling constraints" that assert criteria for acceptable and interpretable artifacts. To formalize this notion, we start by adopting similar notation as Santo, Wiggins, and Cardoso (2024), described below:

1. **A formal language:** $L \in \mathcal{E}$, where $\mathcal{E}$ represents a class of recursively enumerable sets. This defines a space of possible artifacts.
2. **Constraint:** Any subset $L' \subseteq L$. This constrains the space of possible artifacts.
3. **A potential artifact:** Any string $w \in L$. Later, in Definition 2, we define what it means for an artifact to obey the constraints of a scientific conceptual space.

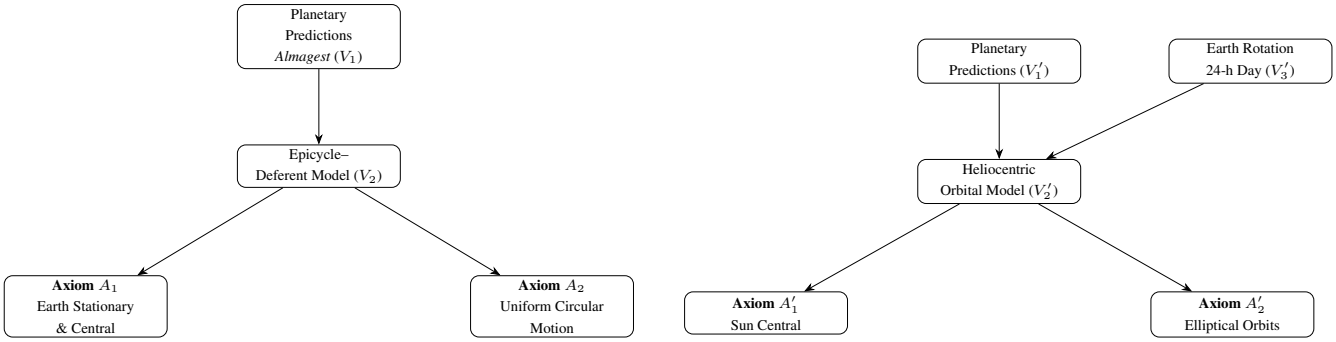
Now, we introduce our graphical definition of a scientific conceptual space.

Definition 1 (Scientific Conceptual Space)**.** *A conceptual space S is a finite, directed acyclic graph (DAG)*

$$S \;=\; G = (V, E), \qquad V \subseteq \mathcal{P}(L),\; E \subseteq V \times V,$$

where each vertex $v \in V$ represents a subset $v \subseteq L$. The DAG represents a dependency structure over constraints, with the following properties:

1. ***Axioms.*** *The set of sink nodes*
$$\mathcal{A} \;=\; \big\{A \in V \mid \nexists\, u \in V : (A, u) \in E\big\}$$
are the axioms of S and have no outgoing edges.
2. ***Edge dependencies.*** *$(u, v) \in E$ iff $u \subseteq v$, meaning u is a further constraint on v, so that v is a necessary condition for u. By construction, each $a \in \mathcal{A}$ satisfies the property of non-dependence:*
$$\nexists\, v \in V\ (A \subseteq v).$$
3. ***Rules.*** *A rule is any vertex $R \in V \setminus \mathcal{A}$. A rule is distinguished from an axiom in that it depends on other nodes in the graph, while axioms are self-justifying and define the initial scope of artifacts worth considering.*
4. ***Rule Prerequisites.*** *Given any node $v \in V$, the set of constraints on which v depends is given by:*
prereq$(v) := \{u \in V \mid$ there is a path from u to v on $S\}$.
This represents all the nodes that must be assumed before the constraint v can be invoked at all.
5. ***Rule Dependents.*** *Given any node $v \in V$, the set of constraints that depend on v is given by*
depends$(v) := \{u \in V \mid$ there is a path from u to v on $S\}$.
This represents the set of all other constraints that cannot be invoked without first assuming v.



(a) Geocentric conceptual space S

(b) Heliocentric conceptual space S'

Figure 1: Conceptual-space DAGs before (S) and after (S') the Copernican and Keplerian revolution. All non-axiom artifacts reach at least one axiom (sink node), satisfying the formal definition of a conceptual space.

Table 1: Historical instances of transformational creativity in science, demonstrating how the transformation of axioms $A \to A'$ transformed the conceptual spaces $S \to S'$.

S **(Initial Space)**	A **(Selected Axioms)**	A' **(Modified Axioms)**	S' **(Transformed Space)**
Newtonian Physics	Absolute time; separate space/time; gravity as force	Relative time; spacetime unity; gravity as curvature	Relativistic Physics
Ptolemaic Astronomy	Earth-centered universe	Sun-centered model	Heliocentric system
Euclidean Geometry	Parallel Postulate: through a point not on a line there is exactly one parallel line	Altered Parallel Postulate: through such a point there are *zero* (elliptic) or *at least two* (hyperbolic) parallels	Non-Euclidean geometry (elliptic and hyperbolic)

In *Phase I: Normal Science* of Kuhn's model, a paradigm's structure and constraints are assumed as a foundation to explore within the conceptual space they induce (Kuhn 1962). Likewise, in our definition, axioms do not depend on other constraints but are instead assumed for the possibility of inquiry within the space. Also note that no connectedness assumption about the graph is made, meaning that the graph can have multiple connected components to reflect disconnected but related subgraphs of knowledge, corresponding to subsets of constraints over possible artifacts. A new artifact is generated from a conceptual space in the following way.

Definition 2 (Scientific Artifact)**.** *Given a conceptual space $S = (V, E)$, a scientific artifact is an ordered triple*

$$a \;=\; (h,\, \sigma, w), \qquad h \in \mathcal{H},\; \sigma \subseteq V,\; w \in \sigma,$$

where

1. *h is a header drawn from a domain-specific set $\mathcal{H}$ (e.g. "phenomenon", "law", "measuring-technique", "open-question") that fixes the kind of the artifact.*
2. *$\sigma := \{V_1, \ldots, V_n\}$ is a non-empty support set of vertices in $S = (V, E)$, representing a subset of constraints whose joint satisfaction makes the artifact well-defined or intelligible.*
3. *$w \in \bigcap_{i=1}^{n} V_i$ is a string representing the artifact.*

Transformational Scientific Creativity

Transformational creativity occurs when modifications are made to a conceptual space. Our definition of a space $S = (V, E)$ naturally supports transformations in two ways.

1. **Edge Modifications ($\oplus, \ominus$):** An edge $e \in E$ is removed ($\ominus$) and/or a new edge $e' \notin E$ is added ($\oplus$).
2. **Vertex Modifications ($\oplus, \ominus, \prime$):** A vertex $v \in V$ is removed ($\ominus$) and/or a new vertex $v' \in \mathcal{A} \setminus \mathcal{V}$ is added ($\oplus$). This also encompasses the case in which the content of a vertex is modified ($\prime$).

For simplicity, we consider the case of modifying an existing node rather than adding a new one or removing an existing one. Still, in future work, our results can be extended to these cases, as well as to the modification of edges. Below, we quantify the transformative potential of nodes in the conceptual space's graph.

Definition 3 (Transformativeness via Modification)**.** *Given a node $v \in V$, its transformative potential is the number of nodes that depend on v, given by*

$$T^{p}_{mod}(v) := |\mathit{depends}(v)|$$

Since it is not possible to invoke any rule in $\mathit{depends}(v)$ without assuming or observing v, if v is modified, then this has the potential to transform the structure of every node in $\mathit{depends}(v)$.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

As articulated in *Phase III: Scientific Revolutions* of Kuhn (1962), revolutions occur when the foundational principles of existing spaces are modified, leading to paradigm shifts. Our definition of a scientific conceptual space naturally has the property that modifying its most foundational principles (axioms) has the most transformative potential. We prove that this is the case below.

Theorem 4 (Modifying Axioms has the Most Transformative Potential)**.** *Given a conceptual space S for which $V \setminus \mathcal{A} \neq \emptyset$, the node with the greatest transformative potential when modified T^p_{mod} must be an axiom. Formally stated,*

$$argmax_{v \in V} T^p_{mod}(v) \in \mathcal{A}$$

Proof. Let R be the non-axiom node in $V \setminus \mathcal{A}$ with the largest number of dependents (transformative potential):

$$R := \arg\max_{v \in V \setminus \mathcal{A}} |\text{depends}(v)| = \arg\max_{v \in V \setminus \mathcal{A}} |T^p_{\text{mod}}(v)|$$

Assume for contradiction that R also has the most transformative potential among all nodes in the graph, including axioms.

Since R is not an axiom, there exists some $u \in V$ such that $(R, u) \in E$, meaning that R depends on u. By the structure of DAGs, any dependent of R is also a dependent of u. Therefore, u satisfies

$$|\text{depends}(u)| \geq |\text{depends}(R)| + 1$$

meaning that it has strictly more dependents than R. However, since we defined $R = \arg\max_{v \in V \setminus \mathcal{A}} |\text{depends}(v)|$, u cannot be in $V \setminus \mathcal{A}$ as well, so we must have that $u \in \mathcal{A}$. Therefore, we have found an axiom with more dependents (transformative potential) than R, leading to a contradiction. Since R was chosen arbitrarily, this means there is no rule $v \in V \setminus \mathcal{A}$ with more transformative potential than the axiom with the most transformative potential, which is what we wanted to show. □

Conceptual Space Operations

Ritchie (2006) presents five critical operations that conceptual spaces should permit. Below, we explain how our definition supports each.

1. **Locating an artifact A within S:** An artifact $a = (h, \sigma, w)$ can be located according to the necessary set of vertices σ given in Definition 2 which make a intelligible.
2. **Rate artifacts A, B for similarity with respect to S:** Given artifacts a, b, their similarity with respect to S can be computed using a Jaccard similarity over their vertex support sets (Jaccard 1901).
3. **Induce a space definition from artifacts $A_1, \ldots, A_n$:** Given artifacts $a_1, \ldots, a_n$ and the corresponding set of vertices $V_a := \bigcup_{i=1}^n \sigma_i$ which make the artifacts well-defined or intelligible, one can construct a DAG by learning or discovering the edge dependencies between the set of vertices V_a.
4. **Given existing artifacts, generate a new artifact:** After arranging existing artifacts into a conceptual space S, a new artifact can be generated according to Definition 2.
5. **Create a revised definition of $S \to S'$ to include artifact A**: This can be done by applying vertex or edge modifications to the graph.

Historical Illustrations

Our definition of a conceptual space supports interpretations of historical instances of transformational scientific creativity. Importantly, in each case, our model does not attempt to reconstruct the steps used by the scientist(s) making their transformation, but instead attempts to provide an explanation for why such instances were so transformative. In Table 1, we provide three examples of transformational creativity, explained in more detail below.

From Geocentrism to Heliocentrism In the shift from Ptolemaic astronomy to a heliocentric system, the axiom that the Earth was the center of the universe (A_1) was modified to reflect a heliocentric model with the sun at the center (A_1'). Rules in S which previously depended on A_1 necessarily had to be transformed after the axiom modification A_1' was made.

Relativity Theory In the shift from Newtonian (S) to relativistic mechanics (S'), space and time were originally considered absolute (axiom A_1), so that momentum obeyed the formula $p = mv$ (rule R_1), but relativity theory revealed that this original formulation only held locally for $v \ll c$. When the new axioms that physical laws are Lorentz invariant (axiom A_1') and the speed of light is constant (axiom A_2') were introduced, the momentum equation had to be reformulated to reflect the dependence on these new underlying principles, yielding rule R_1': that $p = \frac{mv}{\sqrt{1-v^2/c^2}}$

Non-Euclidean Geometry The discovery of non-Euclidean geometry, which removed the parallel postulate P from the prior Euclidean axiom set $\mathcal{A}$, yielded a new axiom set $\mathcal{A}' := \mathcal{A} \setminus P$. As a result of this transformation, derived rules such as the angle sum of a triangle equaling 180 degrees (R_1) were modified to reflect that the computation of angular sums no longer depended on the parallel postulate (R_1').

Towards AI Super-Scientists: Using Our Framework for Transformative Discovery

In light of a series of recent works using AI for scientific ideation (Gu and Krenn 2024; Si, Yang, and Hashimoto 2024; Su et al. 2024; Wang et al. 2024), experiment design (Kon et al. 2025), and even end-to-end discovery (Intology 2025; Lu et al. 2024), we believe our theory opens a promising avenue towards building the first AI super-scientists poised to make transformative breakthroughs. By coupling the wide-reaching creative abilities of large language models (LLMs) with our graphical theory above, we envision a neuro-symbolic approach to transformative scientific discovery that, given a particular scientific field or sub-field $\mathcal{D}$, can be built using the following components.

1. **Background Knowledge:** A collection $\mathcal{C}$ of scientific literature (surveys, conference proceedings, journal papers, etc.) representing background knowledge about $\mathcal{D}$.

2. **Graph Construction:** A DAG construction technique $M : \mathcal{C} \to \mathcal{G}$, where $\mathcal{G}$ is a space of graphs structured according to Definition 1, representing the dependency structure of rules and axioms in $\mathcal{D}$.
3. **Problem Identification:** The identification of a set $\mathcal{P}$ of "open problems" or anomalies that currently plague the research community in $\mathcal{D}$, obtained from the background literature $A : \mathcal{C} \to \mathcal{P}$.
4. **Idea Generation:** Finally, an idea generation module

$$I : \mathcal{L} \times \mathcal{G} \times \mathcal{P} \to \mathcal{T} \times \mathcal{G}$$

that maps from a space of large language models $\mathcal{L}$, dependency graphs $\mathcal{G}$, and open problems $\mathcal{P}$ to produce a transformative idea text $t \in \mathcal{T}$ and its corresponding graph transformation $G' \in \mathcal{G}$.

This approach offers a natural way to incorporate knowledge graphs to enhance the use of LLMs for scientific discovery (Wang et al. 2024; Gu and Krenn 2024), a direction which has shown early promise in improving the technical detail of generated ideas (Wang et al. 2024). Moreover, by making graph transformations explicit, this technique also boasts significant interpretability benefits.

Conclusion

In this paper, we provided a graphical theory of transformational scientific creativity, illustrated the transformative potential of modifying axioms, and used our framework to elucidate the transformativeness of several groundbreaking historical discoveries.

Subsequent investigations could extend our theory by providing greater depth, breadth, and nuance to the framework, contextualized within the history and philosophy of science. For example, we suggest exploring in-depth historical illustrations for one to three exemplary discoveries, such as those in Table 1, showing how all components of the theory are instantiated and transformed. Similarly, it may be insightful to study a large collection of scientific discoveries, such as those addressed in Haven (2007) and Thagard (2012), and delineate a "logic of discovery" via common graph transformation patterns among these cases. Historical studies such as these may suggest criteria for "good" graph transformations and highlight distinctions between edge and vertex modifications.

One notable implication of our theory's rigid structure is that it may be best suited to scientific areas that involve formal systems, such as mathematics, engineering theory, and computer science. Nevertheless, in exchange for generality, we obtain more specificity regarding how the transformational creativity process can be formulated in these domains, opening the door for future theoretical or computational inquiry. Our work, like Santo, Wiggins, and Cardoso (2024), is a crucial first step towards formalizing transformational creativity.

Acknowledgments

The authors thank Jonathan Livengood for helpful feedback on an earlier version of this paper.

References

Aguiar, D.; Atã, P.; and Queiroz, J. 2015. Intersemiotic translation and transformational creativity. *Punctum. International Journal of Semiotics* 1:11–21.

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Friedman, N.; Linial, M.; Nachman, I.; and Pe'er, D. 2000. Using Bayesian networks to analyze expression data. *Journal of Computational Biology* 7:601–620.

Gu, X., and Krenn, M. 2024. Interesting scientific idea generation using knowledge graphs and LLMs: Evaluations with 100 research group leaders. arXiv:2405.17044.

Haven, K. 2007. *100 Greatest Science Discoveries of All Time*. Bloomsbury Publishing USA.

Intology. 2025. Zochi technical report. https://www.intology.ai/blog/zochi-tech-report.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* 37:547–579.

Koestler, A. 1964. *The Act of Creation*. Macmillan.

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. arXiv:2502.16069.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292.

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

Rothenberg, A. 2014. *Flight from Wonder: An Investigation of Scientific Creativity*. Oxford University Press.

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

Shi, F., and Evans, J. 2023. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications* 14.

Shi, F.; Foster, J. G.; and Evans, J. A. 2015. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks* 43:73–85.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. arXiv:2409.04109.

Simonton, D. K. 2004. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press.

Simonton, D. K. 2021. Scientific creativity: Discovery and invention as combinatorial. *Frontiers in Psychology* 12:721104.

Su, H.; Chen, R.; Tang, S.; Zheng, X.; Li, J.; Yin, Z.; Ouyang, W.; and Dong, N. 2024. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. arXiv:2410.09403.

Thagard, P. 2012. Creative combination of representations: Scientific discovery and technological invention. In *Psychology of Science: Implicit and Explicit Processes*. Oxford University Press.

Toivanen, J.; Järvisalo, M.; Alm, O.; Ventura, D.; Vainio, M.; and Toivonen, H. 2018. Towards transformational creation of novel songs. *Connection Science* 31:1–29.

Varshney, L. R. 2020. Limits theorems for creativity with intentionality. In *Proceedings of the International Conference on Computational Creativity*, 390–393.

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*, 279–299.

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

Do Conversational Interfaces Limit Creativity? Exploring Visual Graph Systems for Creative Writing

Abhinav Sood
University of Sydney
Sydney, Australia
abhinav.sood@sydney.edu.au

Maria Teresa Llano
University of Sussex
Brighton, UK
Teresa.Llano@sussex.ac.uk

Jon McCormack
Sensilab, Monash University
Caulfield East, Australia
Jon.McCormack@monash.edu

Abstract

We present a graphical, node-based system through which users can visually chain generative AI models for creative tasks. Research in chaining LLMs has found that while chaining provides transparency, controllability and guardrails to approach certain tasks, chaining with pre-defined LLM steps prevents free exploration. Using cognitive processes from creativity research as a basis, we create a system that addresses the inherent constraints of chat-based AI interactions. Specifically, our system aims to overcome the limiting linear structure that inhibits creative exploration and ideation. Further, our node-based approach enables the creation of reusable, shareable templates that can address different creative tasks. In a small-scale user study, we find that our graph-based system supports ideation and allows some users to better visualise and think through their writing process when compared to a similar conversational interface. We further discuss the weaknesses and limitations of our system, noting the benefits to creativity that user interfaces with higher complexity can provide for users who can effectively use them.

Introduction

The creative writing process often involves relying on ideation and open-ended exploration to complete tasks that can then be used to enhance written material. For example, in the fantasy and sci-fi genres, world creation, character description and characters' interactions with the world are crucial to writing good stories (Card 2001). Large Language Models (LLMs) have emerged as generative AI tools that can support the natural language components of such tasks. For many users of LLMs, interactions often occur through linear text-based chat interfaces that pose significant limitations. Their inherent linear structure prevents exploration through branching, and well-engineered prompts are required to guide LLMs, preventing the effective use of these tools by everyday users (Zamfirescu-Pereira et al. 2023). These challenges only increase for more complicated writing tasks. Here, novel interfaces like Metaphorian (Kim et al. 2023) and AngleKindling (Petridis et al. 2023) have been developed for tasks like finding metaphors for scientific writing and exploring journalistic angles for press releases. However, as these tools are tailored to specific writing domains, they do not provide a generalised platform for working on open-ended writing problems.

To address this gap, we present a graphical, node-based system that allows users to chain generative AI models and prompts. Once these chains are set up, they can be saved into shareable templates that end-users can then easily use for more complicated creative writing tasks without having to engineer prompts. We then investigate the properties of this system in a creative writing context. Our key contributions are as follows:

1. We develop a general-purpose, openly available, graph-based system that uses chaining with relaxed restrictions. This system can be used for different writing tasks. [1]
2. We conduct a user study to assess the effectiveness of our system for planning related creative writing tasks that involve ideation and open-ended exploration. We find that our graph-based system illustrates how different users benefit from different interfaces for ideation.
3. Our user study provides us with insights into how such interfaces affect users' writing processes. We discuss how these insights can help develop future interfaces in this area.

Background

Creativity and Ideation

In creativity research, three main cognitive processes have been essential (Lee and Therriault 2013) to understanding the psychological basis of creativity. These are divergent thinking, convergent thinking and associative processing. Divergent thinking is a cognitive process that generates a broad array of unconventional ideas by exploring various perspectives that might not be rational. In contrast, convergent thinking focuses on finding the best answer to a well-defined problem through rational reasoning. Associative processing involves connecting and linking concepts based on shared attributes. Together, these processes contribute to effective problem-solving and innovation.

These general cognitive processes emerge as creative behaviour in different forms. In creative writing, authors in

[1] The code for the system, along with additional documentation, is available at - https://github.com/abhinavsood2002/Chaining-GenAI-for-Ideation

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

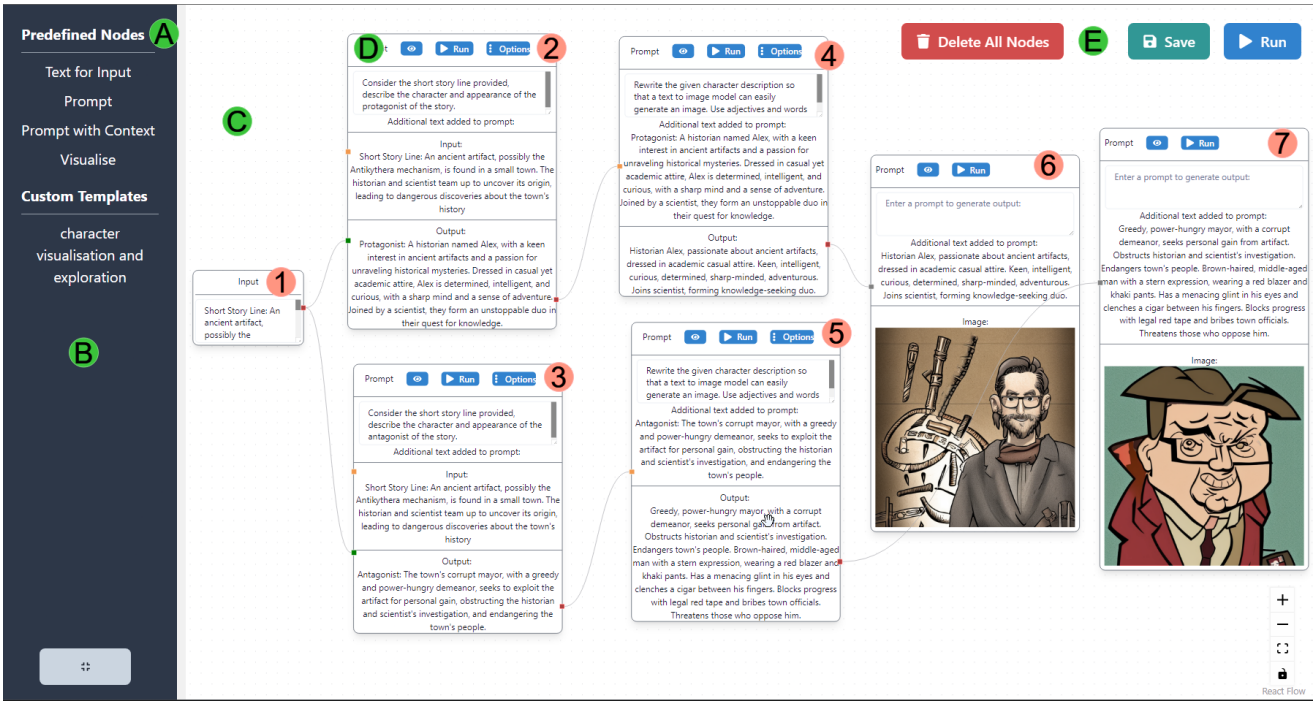


Figure 1: The User Interface alongside the demonstration of an example use-case where the developed system can be used to visualise and explore characters.

User Interface (Green) - (A) The 4 Predefined Node options. (B) Custom Node section where templates can be dropped in to make them available in the user interface. (C) The workspace - nodes from the sidebar can be dragged and dropped into the workspace and subsequently edited. (D) The node header contains buttons that activate additional functionality - like running individual nodes, viewing node information in a user-friendly Modal, and changing node-specific parameters like temperature for generation. (E) The save button downloads the current workspace as a template as a JSON file, and the run button runs all nodes in topological order.

Character Visualisation and Exploration Use Case (Orange) - (1) contains a short storyline. (2) and (3) use the short storyline to create a protagonist and antagonist character description. (4) and (5) are used to prepare the descriptions for visualisation by converting them into a form that is more friendly for text-to-image prompting. (6) and (7) display visualisations for the characters.

fantasy often employ divergent thinking to construct diverse worlds. Strategies in ideation also relate to these cognitive processes. For example, divergent and convergent thinking play a role in ideation techniques like brainstorming (Osborn 1953) and mind maps (Buzan and Buzan 1996).

Text and Image Generation

The introduction of the transformer architecture (Vaswani et al. 2017) has brought rapid advances in language modelling. The architecture's self-attention mechanism allowed for efficient parallelisation while capturing long-range dependencies, making it feasible to train language models on massive datasets. This, in turn, has led to the development of Large Language Models (LLMs) like GPT-3 (Brown et al. 2020). Inference with LLMs occurs through zero or few-shot prompts. While zero-shot prompts contain a natural language instruction, few-shot prompts additionally contain examples related to the instruction to provide the LLM with more context for approaching a problem.

Advancements in language modelling have been accompanied by similar progress in the field of image generation. Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) have been instrumental in unconditional image generation. DDPMs learn to generate images by modelling the gradual removal of noise from a noisy latent image. The conditioning of this noise removal process on text has led to the development of text-to-image models like Stable diffusion (Rombach et al. 2022). Both LLMs and text-to-image models hold significant potential for creative generation tasks. With their ability to generate human-like text, LLMs can be employed in storytelling, poetry generation, and scriptwriting (Kaddour et al. 2023). On the other hand, text-to-image can be utilised in visual art generation and story visualisation (Zhang et al. 2023).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Human-AI Interaction and Creativity Support Tools

The applications of generative AI models go beyond simple generation tasks. Human-AI interaction and co-creative applications are receiving increasing attention. Often, human-AI interaction involves a mixed-initiative system (Allen, Guinn, and Horvtz 1999) where both the machine and the system contribute towards a shared task. For creative tasks, constructing an adequate mixed-initiative system involves examining how generative AI tools can be incorporated into a user's creative workflow to facilitate collaboration between the user and AI. Karimi et al.(2020) provides an example of a system where an AI generates sketches that resemble a user's drawings but are from a different category. These generated sketches can then be used to re-interpret the original sketch in a new context, aiding their creative design process. Concurrent efforts have been made to augment and enhance the creative process of different types of writers. For example, Creativity Support Tools (CSTs) like Wordcraft (Yuan et al. 2022) can be used to perform text infilling after generating a passage, elaborating on a region of text, seeding the start of a story, and arbitrarily rewriting a section of text. Dramatron (Mirowski et al. 2023) provides another example of a tool that aids in co-writing screenplays and theatre scripts.

Chaining LLMs and Node-Graph Interfaces

CSTs that use LLMs often employ few-shot prompts (prompts with examples relating to the task) or chained prompts (sequences of LLM interactions where the output of the preceding chain interaction is passed to the following interaction) to produce valuable functionality. The concept of chaining prompts was explored in AI Chains (Wu, Terry, and Cai 2022), an interactive interface that allowed users to customise chains of prompts for given tasks. To the user, such interfaces present themselves as a node-graph interface, where fundamentally, nodes that can take input and output are used to organise a solution to a given problem. We use the terms node-graph interface and graph-based system interchangeably. Other interactive node-graph systems that use chaining include ChainForge (Arawjo et al. 2023), PromptChainer (Wu et al. 2022), and LangFlow(LangFlow 2024). Neither of these systems focuses on how chaining can be adapted for creative exploration. ChainForge provides an interface to construct graphs to evaluate different language models against one another. PromptChainer aims to prototype software applications using LLMs, and Langflow caters to constructing AI applications. As such, these applications require knowledge of concepts in these domains and do not function as simple tools that support creative writing.

Chaining for Creativity - A Graph-based System

The lack of interfaces that use chaining for creative exploration prompted us to develop a graph-based interface for creativity. The two key design rationales for our system were:

1. Allowing for **easy expansion and combination** of text. To ensure that our system could support creative tasks, we wanted our system to relate to the cognitive processes in creativity research. Specifically, we wanted to ensure that divergent and convergent thinking were supported by allowing for easy, LLM-based expansion, mutation, and combination of node outputs.
2. Keeping interaction **simple**. Many graph-based systems that use chaining, like Flowise (Flowise 2024) and LangFlow (LangFlow 2024), require an understanding of the underlying machine-learning concepts. Since our end-user is a writer who might not be familiar with the technical details of machine learning, the nodes in our system must be simple to interact with.
3. **Flexibility** in node definitions. Qualitative evidence from AI-Chains (Wu, Terry, and Cai 2022) indicated that their interface was not ideal for free exploration, a key component of ideation. This was primarily because each interaction in the chain was a rigid, pre-defined operation. The authors suggested that structural constraints in each chaining step could be relaxed to encourage exploration for future work.

Thus, we develop a graph-based system in which users primarily interact through simple and flexible nodes. Each node can be provided with multiple inputs (which are then concatenated) in the form of text. The input is processed into output based on a prompt that the user provides. The system is described in detail in the subsection that follows.

System Description

The interface developed in React.js is visible and described in Figure 1.

Nodes Nodes are individual components of the system that can operate independently and perform a task of value to the user. Each node typically has connectors or handles along its boundaries, allowing it to be linked with other nodes. If a handle receives multiple inputs, the inputs are concatenated. For this system, we describe four fundamental predefined nodes:

- **Text for Input:** This type of node serves as a placeholder for user input. Its output is merely a copy of the text in the node. It is used to organise input for a given chain.
- **Prompt:** This node allows the user to prompt an LLM for output. The prompt can be modified by adding additional text as input.
- **Prompt with Context:** This node is similar to the Prompt node but also includes an additional input that can be used to provide context that might not fit well in the prompt.
- **Visualise:** This node uses stable diffusion to generate images for a given prompt.

Templates Once nodes are connected to form a graph, they can be saved into templates. Templates are saved as .JSON files. These files can be shared with other users of the system. This is possible due to the nature of graph-based interactions, where a node's output depends solely on the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

nodes from which it receives input. This would not be possible in a linear chat-based conversation, where one prompt changes all future prompts. As templates have well-defined inputs and outputs, multiple templates can be used in the same run to perform more complex tasks. During the execution of a template, the nodes are topologically sorted through breadth-first traversal. Then, each node is run in order. The algorithm used to perform this topological sorting is provided and described in the attached code repository.

Additional features The system contains additional features to provide users with a higher degree of control over the interface and output. This includes the ability to set the variability of text (through temperature) for each node and the desired length of output from a node. Other visual components include a mini-map to navigate the interface for larger template setups and navigation tools to use the interface with a trackpad.

Backend We host our models on an NVIDIA RTX A6000 50GB GPU provided by SensiLab, Monash University. The decision to host our models locally was to ensure that any user studies conducted would not involve sending information to third-party services due to privacy concerns.

User Study

To test our system, we identified two planning-related creative writing tasks that our system could significantly impact. We identified **world-building** and **character description** as planning-related tasks of relevance. We conducted a small-scale user study with participants completing creative writing tasks to analyse and evaluate our system.

Eight participants were recruited through convenience sampling. The sampling method was selected due to the study design's requirement of approximately one and a half to two hours of participation time per participant. Given the absence of monetary compensation, recruiting participants without convenience sampling would not have been feasible. An explanatory statement was dispersed through creative writing-related social media groups and acquaintances. Those who were interested were provided with a consent form to confirm their participation in the study. Participants were provided with a 10-minute video [2] guiding them through the interface and instructions for the study. We additionally construct a conversational interface as a baseline.

The study was conducted as follows: First, each participant received a link to an online interface along with a UserID. After logging in with their UserID, they were asked to complete a creative writing task using the graph-based and conversational system. Once they completed the tasks, they were given a questionnaire asking about their experiences with each system. Each participant contributed approximately 1:30-2 hours to complete the given tasks. The user study was approved through a Human Ethics Application made to the Monash University Human Research Ethics Committee.

[2] https://www.youtube.com/watch?v=ACVLsKSPN-0

Baseline and Additional Changes

The graph-based system had to be updated to conduct the user study virtually. All interaction with the system needs to occur in the user interface; this includes any material the participant writes down to complete a task. Thus, we added a notepad to the right-hand side of the interface where the user could draft the output for their tasks.

We also constructed a conversational interface as a baseline. Previous user studies evaluating chaining have compared chaining LLMs to a sandbox(where LLMs provide answers to single instructions independently) environment (Wu, Terry, and Cai 2022). However, we believe that most user interactions with LLMs happen through conversational interfaces similar to ChatGPT(OpenAI 2024). Furthermore, research on chatbot brainstorming has shown success against human baselines(Wieland, de Wit, and de Rooij 2022). Thus, we compare our graph-based system against a comparable conversational interface (instead of a sandbox). Images demonstrating these user interfaces are available in the linked Code Repository.

The conversational interface builds on code available at (Onyebuchi 2024). We used the same LLM and system prompt to ensure that both the conversational and graph-based interfaces were comparable. The conversational interface was updated to generate images. Example queries for each task were also displayed. Preliminary testing with the conversational interface indicated that Mistral 7B (Jiang et al. 2023) Instruct v0.2 could not maintain context over multiple prompts when testing it on a creative writing task. Even though this LLM worked in our comprehensive example with the graph-based interface, it wasn't able to perform in the conversational interface. Thus, after testing out other LLMs, we chose OpenHermes-2.5-Mistral-7B[3] for its ability to maintain context in a conversation while still being a 7B model for compute restrictions. We use Stable Diffusion 2.1 (Rombach et al. 2022) to generate images.

Study Design

As our sample size was small, we wanted to minimise the influence of any confounding variables in our study. We identified two confounding variables:

1. The order in which the systems were displayed could affect proficiency in interacting with the subsequent system.
2. If the exact same creative writing task was used for both systems, a participant could use information from the previous task in the subsequent task.

To address the first confounding variable, half of the participants were given the graph-based system first, while the other half were given the conversational system first. Participants were given variants of the same creative writing task for each system to address the second confounding variable. Two participants tested the system for each combination of order (graph-based or conversational first), creative writing task (world-building or character description) and variant. The specific creative writing tasks were:

[3] https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

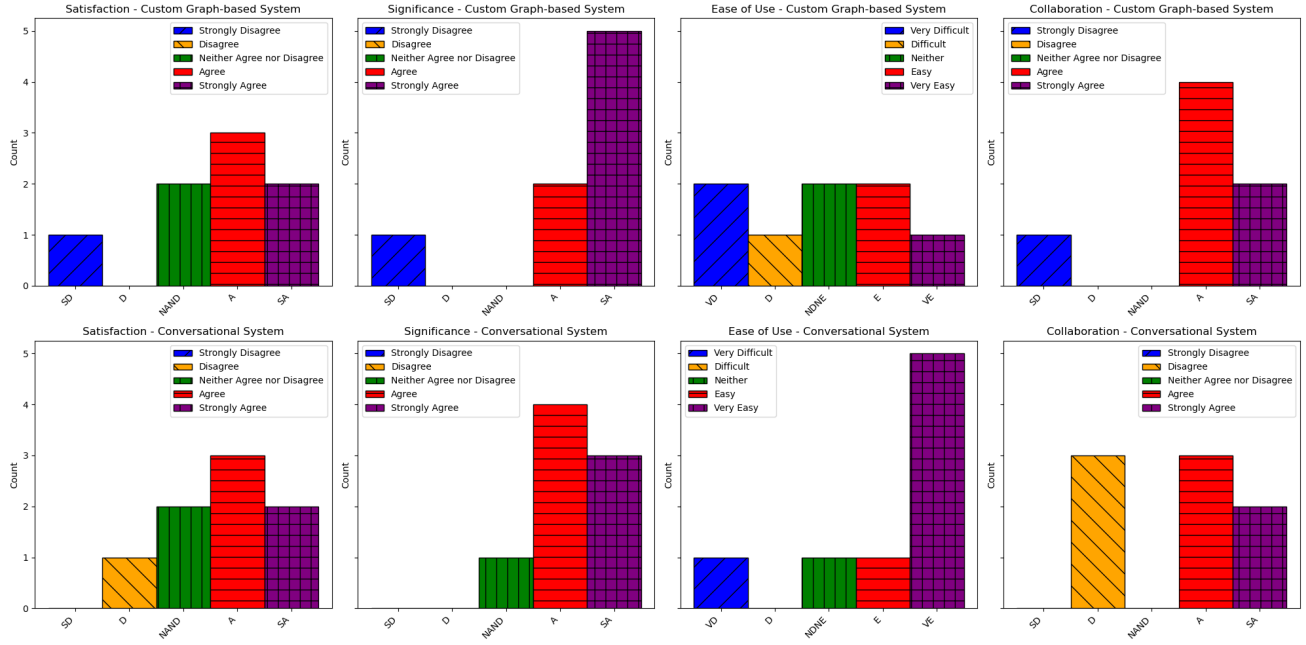


Figure 2: User Study Questionnaire - 5 point questions results.

- **Character Description**
  - Variant 1: Write a brief character description for a protagonist in a short story. The protagonist is Camilla, a librarian whose primary goal is to solve the mystery of her brother's disappearance. She believes the disappearance is connected to a larger conspiracy.
  - Variant 2: Write a brief character description for an antagonist in a short story. The antagonist is General Kal, who ascended to power through a coup. He clamps down on any form of dissent, using all powers available to him to crush any opposition.
- **World Building**
  - Variant 1: Write a brief description of a world-setting for a community in a lush rainforest environment. Focus on the practices they have developed to sustain themselves.
  - Variant 2: Write a brief description of a world-setting for a community in a harsh arctic environment. Detail their survival strategies in the extreme cold.

Setting up a series of nodes in the graph-based system might require a lot of experimentation. This could mean that participants might have to spend hours learning how to use the system to complete a single task. Thus, to meet time constraints, the participants' interfaces for a given writing task in the graph-based system would be populated with a relevant predefined template. These templates can be reproduced using the corresponding .json files for these templates, which are available in the linked GitHub repository. As a result, participants could easily drag the provided template into the interface to expand or edit it according to their needs.

Post-study questionnaire

Participants were asked to evaluate the two systems they had interacted with. They were asked 8 questions (4 for each system) that were rated on a 5-point Likert Scale (1: Strongly Disagree, 2: Disagree, 3: Neither Agree nor Disagree, 4: Agree, 5: Strongly Agree). Additionally, participants were asked to complete 3 short free-form questions. The Likert and free-form questions asked in the questionnaire were:

5-point Likert Scale Questions

1. **Satisfaction**- I was satisfied with completing the tasks using the given system.
2. **Significance** - The interaction I had with the system significantly affected my task answer.
3. **Ease of Use** - How would you rate the ease of using the system? [4]
4. **Collaboration** -I felt like I was collaborating with the system to complete the task

Short Answer Questions

- Compare your writing process between the two systems. For example, for either system, did you incorporate the AI generations directly into your task output? Was either system more useful for providing ideas?
- Was there any aspect of the graph-based system you found particularly useful while completing your task?

[4]For this question, participants were told that 1 stood for Very Easy, and 5 stood for Very Difficult. In our results, we later reverse the scale for consistency between visualisations.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

- Were there any aspects (beyond bugs or visual appearance) of the graph-based system that were frustrating? How would you address or improve these aspects?

Participant Demographics

We asked for background information regarding the participant's age group, gender identity, proficiency in creative writing and familiarity with AI. For the age groups, 50% (4/8) of the participants indicated they were 18-24, 35% (3/8) of participants were 25-34, and 12.5% (1/8) participants were 45-54. 75% (6/8) of the participants identified as male, 25% (2/8) participants identified as female. For their background in creative writing, 50% (4/8) did not engage in creative writing activities, 37.5% (3/8) considered themselves amateurs (write primarily for personal enjoyment or as a hobby), and 12.5% (1/8) indicated that they were professional (have been published or paid for their writing). 87.5% (7/8) participants had previously used AI language models like ChatGPT before, while 12.5% (1/8) had not.

	With P4		**Without P4**	
	Graph	**Chat**	**Graph**	**Chat**
Satisfaction	3.625	**3.75**	**4.00**	3.57
Significance	4.25	4.25	**4.71**	4.14
Ease of Use	2.875	**4.125**	3.14	**4.00**
Collaboration	3.375	**3.5**	**3.71**	3.43

Table 1: Comparison of Graph-Based System vs Conversational System with and without P4

Results

The bar plots corresponding to the questions measured on the Likert scale are depicted in Figure 2. Participants appear to be equally satisfied with their task answers for both systems. A greater proportion of participants strongly agree that the graph-based system significantly affected their answers, whereas, for the conversational system, most participants only agree. The largest difference is in the distribution of the ease of using the system. Most participants found the graph-based system much more difficult to use. On collaboration, only one participant felt that they were not collaborating with the graph-based system, whereas 3 did not feel like the process with the conversational system was collaborative.

If we take the mean of the Likert values, we get the statistics in Table 1. At first glance, the statistics indicate that our graph-based system is clearly more challenging to use and performs slightly worse than the conversational system on other metrics. However, this does not account for the fact that **P4 was unable to use the graph-based system** and thus appropriately filled in strongly disagree for all the questions related to the graph-based system. If all of P4's responses are removed (graph and conversational), we observe that the graph-based system outperforms the conversational system on every metric except ease of use. Although these differences are not statistically significant due to our small and convenience-based user sample, they provide preliminary evidence for limitations of conversational interactions.

To understand how interaction with our system differed from the conversational interface, we performed a thematic analysis of the participants' answers to the free-form short-answer questions. The identified themes, sub-themes and supporting quotes are recorded in Table 2. Our analysis shows that our system positively affected many participants' creative writing processes. This occurs because the graph-based system provides better visualisations of their interactions with the system and does not take control of the conversations participants engage in. Moreover, the system also helps participants with ideation through its unique structure. This result aligns with our expectations, as our system design was motivated by trying to expand access to the cognitive processes identified from creativity research. At the same time, we also identify various issues and limitations with our system through the analysis. Most users find the graph-based system more difficult to use, despite our attempts to keep the system simple. Some users did not find the system particularly useful. There were other minor issues with the user interface that need improvement. We further discuss our results in the discussion section.

Example Task Answers from P3

For illustration, we provide the task answers for P3, who completed the character description writing task. They completed the first variant with the conversational system and the second variant with the graph-based system. They included text output from the systems directly into their answers.

Task answer - Character Description, Variant 1, Conversational System On the surface, Camilla may seem to be a normal person, with a normal job as a librarian. However, Camilla is beyond normal. Due to her photographic memory, Camilla is able to quickly analyse large amounts of data and quickly see the relationships between unrelated pieces of information. Her family has managed this historical library for generations and it has been a source of pride and stability for her family. Growing up, Camilla was surrounded by books and the sense of community that a library provides, influencing her love for literature. Her brother, however, chose a different path and became a private investigator, sharing her passion for uncovering the truth. The two often collaborated on cases, combining her research skills and his field experience to solve complex mysteries. Camilla is deemed as an over-perfectionist by her family who often worry that she obsesses too much on small details.

Task answer - Character Description, Variant 2, Graph-based System General Kal was born into a poor family in a remote village and had to work hard from a young age to support his family. One day while working on their landlords farm, Kal witnessed his father being beaten for being unable to afford the recently increased taxes. This instilled in him a deep sense of injustice and a desire to protect his family from further oppression. Kal joined the army and with his ambition, rose through the ranks to eventually seize power. General Kal is a stern and imposing figure. His

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Themes	Sub-Themes	Supporting Quotes
Impact on Ideation and the Writing Process	Enhanced ability to conceive new ideas	**P1** – "The chain system was better in providing ideas as it could easily make use of multiple inputs in generating a new idea." **P2** – "graph based system definitely helped me generate more ideas" **P5** – "Graph based System enabled me to be more creative and think about new ideas"
	Graphical visualisation provided direction	**P2** – "the ability to connect the nodes that helped me understand what I was thinking and what direction I wanted to take" **P3** – "the graph made it much easier to follow the thought process" **P5**- "I liked the visualization of ideas" **P6** – In response to Q2 - "Ability to create a graph based structuring of the solution"
	Over-reliance on conversational -system	**P2** – "In the chat based system I found myself engaging less with the system and asking more general questions than creative ideas." **P8** – "the conversational system was taking over the process, making me more reliant"
Challenges and Limitations	Difficulty in using Graph-based system	**P4** – "I was unable to use the graph system." **P7** – "Too much of cross intersection can make it very confusing." **P8** – "main issue with the custom graph based system was that it is difficult to use, especially if one is using it for the first time"
	Graph-based system's utility varied	**P6** – "The conversational interface provided better and detailed answers which made understanding the situation in task easy" **P7** – "In my scenario the first one [chat] was more effective than the second one [graph]"
	UI-related issues	**P6** – "It was difficult to connect nodes on horizontal scale that were far away" **P8** – "Since I used a laptop (which has a small screen), it is difficult to make the connections between the nodes without zooming out a lot"

Table 2: Themes and sub-themes identified from short answers collected in survey

uniform is always immaculate and his eyes seem to pierce right through you, conveying a sense of authority and control. However, a closer look reveals a scar above his left eye, a reminder of a fierce battle he fought many years ago. This has left him with a slight squint in his left eye, making it difficult to see in low light conditions.

Discussion

Ideation with the Graph-based System

The creative writing tasks in the user study were chosen specifically because they included ideation as a key aspect of their completion. Participants reported that the graph-based system helped them generate more ideas. Some participants said this was because the visualisations of the system allowed them to better engage with their thought processes. For P2, the conversational format pre-conditioned them to ask questions instead of dealing with creative ideas. P8 mentioned that they found the ability to specify each node's randomness (temperature) essential to having more control over generating creative ideas. In contrast to these very positive experiences on ideation, P3 found both systems equally useful for ideation. Several mechanisms might affect how users use either system for ideation. We believe that while the graph-based system might demonstrate the capability to support higher levels of ideation, exploiting these is much more dependent on the user. For example, the user can set up the same template with different inputs and see how the output varies. This isn't possible with a conversational interface.

Visualising the Creative Writing process

Beyond the graph-based system's ability to generate ideas and text, participants commented on how the graph-based system provided a visualisation that offered direction and guidance. One of the main contributions of the AI-Chains (Wu, Terry, and Cai 2022) paper was that breaking down problems into well-defined chaining steps provided guardrails that kept users on track. These guardrails also prevented free exploration.

Because of our flexible node definitions, our graph-based system provided direction and guidance to the user without restricting free exploration. Additionally, for many participants, the graph-based structure also acted as a tool for visual organisation. The graphical representation of thoughts and ideas as nodes allowed some users (specifically, P2, P3, P5, P6 with quotes in Table 2) to better complete their tasks. This ability to organise thoughts and ideas also means a higher learning curve to use the system. We believe this higher learning curve is responsible for many of the difficulties other participants faced while using our system. We discuss these in the limitations section.

Limitations

The most significant limitation of our study is the limited number of participants (8) in our user study. While this might render some of our quantitative results insignificant, we still believe this paper's qualitative analyses can guide future user interfaces that seek to promote ideation with LLMs. Another key limitation of our system is that it is

more difficult to use. Conversational interfaces are much more prevalent in everyday life than graph-node interfaces. Even if our specific graph-based system requires little additional knowledge, when compared to a conversational interface, navigating the system can prove difficult. P4 was completely unable to engage with the graph-based system. In their background, they indicated that they had not used any AI tools before. Yet, they could easily complete the creative writing task with the conversational interface. On the other hand, P3 stated, "There were no aspects of the graph-based system that I found frustrating, it was generally easy and intuitive to use." The difficulty of the graph-based system might depend on the participant's past experience with graph-node interfaces. Some level of difficulty was also introduced by the imperfect user interface. Some participants, such as P6 and P7, found the conversational interface to be better. We suspect this could be a direct consequence of the ease of using the system. It is also possible that conversational and graph-based systems appeal to different categories of users, and regardless of any changes that can be made, there will always be some participants who prefer one over the other. Moreover, this project used a relatively small LLM due to limited compute availability. As most users were still able to generate and write creative stories, we believe this limitation is not that restricting. Finally, our participants were largely either amateur creative writers or had no creative writing experience. LLMs have been shown to affect the individual creativity of amateur writers more effectively than professional writers (Doshi and Hauser 2024). To this extent, the results of our system do not generalise to professional creative writers and only apply to hobbyists.

Future work

The types of pre-defined nodes in the current system can be further expanded to explore different use cases in creative fields beyond creative writing. For example, incorporating image-to-text and image-to-image models could allow image-based input to be added to chains. This would allow chains to have more significant applications in creative fields where images are the primary medium.

Since the start of this project, multimodal models that integrate multiple forms of media have gained prevalence. It would be intriguing to explore how multimodality could play a role in such an interface. Our research utilised smaller 7B parameter LLMs. Exploring the impact of larger LLMs on chains, given their extensive context length, is also valuable. With the increasing accuracy of the larger LLMs, maintaining global context while still being able to map out graphs for a writer's thought process might be a more reasonable approach instead of each node maintaining individual context.

The flexibility offered by the interface would also be interesting to experiment with. When given the capability to switch between different LLMs and text-to-image models for each node, can creative writers leverage specialised and fine-tuned models for enhanced controllability?

Conclusion

In conclusion, we developed a novel, graphical, node-based system for chaining generative AI models. This system was designed to overcome the structural constraints of traditional chat-based interactions. Creativity research and previous research on chaining were used as a basis for our design. We demonstrated the system's potential through a user study, showcasing how it can assist in ideation and visualisation to enhance the creative writing process. The results were discussed, highlighting the meaning they possess for research in future interfaces that seek to promote ideation. While there are several limitations to our system and user study, our research opens up new possibilities for ideation and interaction with generative AI systems.

Author Contributions

Author 1 (AS) designed the system and drafted the paper. Authors 2 (MTL) and 3 (JM) supervised the research and provided guidance and direction for the development of this work.

References

Allen, J. E.; Guinn, C. I.; and Horvtz, E. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14(5):14–23.

Arawjo, I.; Vaithilingam, P.; Wattenberg, M.; and Glassman, E. 2023. Chainforge: An open-source visual programming environment for prompt engineering. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23 Adjunct. New York, NY, USA: Association for Computing Machinery.

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. online: Curran Associates, Inc.

Buzan, T., and Buzan, B. 1996. *The Mind Map Book: How to Use Radiant Thinking to Maximize Your Brain's Untapped Potential*. Penguin Publishing Group.

Card, O. 2001. *How to Write Science Fiction & Fantasy*. Penguin Publishing Group.

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.

Flowise. 2024. Flowise: Build LLM. https://flowiseai.com/. Accessed: 2024-05-25.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. online: Curran Associates, Inc.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7b.

Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; and McHardy, R. 2023. Challenges and applications of large language models.

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.

Kim, J.; Suh, S.; Chilton, L. B.; and Xia, H. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, DIS '23, 115–135. New York, NY, USA: Association for Computing Machinery.

LangFlow. 2024. LangFlow. https://www.langflow.org/. Accessed: 2024-05-25.

Lee, C. S., and Therriault, D. J. 2013. The cognitive underpinnings of creative thought: A latent variable analysis exploring the roles of intelligence and working memory in three creative thinking processes. *Intelligence* 41(5):306–320.

Mirowski, P.; Mathewson, K. W.; Pittman, J.; and Evans, R. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery.

Onyebuchi, D. 2024. Build a chatbot with chatgpt and react. Accessed: 2024-05-25.

OpenAI. 2024. Chatgpt. https://chat.openai.com.

Osborn, A. F. 1953. *Applied imagination*. Scribner's.

Petridis, S.; Diakopoulos, N.; Crowston, K.; Hansen, M.; Henderson, K.; Jastrzebski, S.; Nickerson, J. V.; and Chilton, L. B. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery.

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 (CVPR)*, 10684–10695.

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.

Wieland, B.; de Wit, J.; and de Rooij, A. 2022. Electronic brainstorming with a chatbot partner: A good idea due to increased productivity and idea diversity. *Frontiers in Artificial Intelligence* 5.

Wu, T.; Jiang, E.; Donsbach, A.; Gray, J.; Molina, A.; Terry, M.; and Cai, C. J. 2022. Promptchainer: Chaining large language model prompts through visual programming. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22. New York, NY, USA: Association for Computing Machinery.

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.

Yuan, A.; Coenen, A.; Reif, E.; and Ippolito, D. 2022. Wordcraft: Story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, IUI '22, 841–852. New York, NY, USA: Association for Computing Machinery.

Zamfirescu-Pereira, J.; Wong, R. Y.; Hartmann, B.; and Yang, Q. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Text-to-image diffusion models in generative ai: A survey.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Invisible Strings: Revealing Latent Dancer-to-Dancer Interactions with Graph Neural Networks

Luis Vitor Zerkowski[1], Zixuan Wang[2], Ilya Vidrin[3], Mariel Pettee[4]

[1] University of Amsterdam luisvz@gmail.com
[2] Georgia Institute of Technology zxwang9811@gmail.com
[3] Northeastern University i.vidrin@northeastern.edu
[4] Lawrence Berkeley National Lab mpettee@lbl.gov

Abstract

Dancing in a duet often requires a heightened attunement to one's partner: their orientation in space, their momentum, and the forces they exert on you. Dance artists who work in partnered settings might have a strong embodied understanding in the moment of how their movements relate to their partner's, but typical documentation of dance fails to capture these varied and subtle relationships. Working closely with dance artists interested in deepening their understanding of partnering, we leverage Graph Neural Networks (GNNs) to highlight and interpret the intricate connections shared by two dancers. Using a video-to-3D-pose extraction pipeline, we extract 3D movements from curated videos of contemporary dance duets, apply a dedicated pre-processing to improve the reconstruction, and train a GNN to predict weighted connections between the dancers. By visualizing and interpreting the predicted relationships between the two movers, we demonstrate the potential for graph-based methods to construct alternate models of the collaborative dynamics of duets. Finally, we offer some example strategies for how to use these insights to inform a generative and co-creative studio practice.

Introduction

In many cultures, dance is deeply informed by interpersonal connection. The prevalence of social dances worldwide illustrates how dance can even, in certain contexts, be seen as an inherently interactive art form. The collaborative dimensions of dance, however, are less concrete than body positions or movements, and for this reason they can be more difficult to describe or study. Despite their invisible nature, these inter-dancer connections can provide an essential lens through which we can create and understand choreography involving more than one person.

Moreover, in a culture that increasingly values digital renderings of ideas, there is significant potential for dancers to re-imagine how digital forms of dance can enrich, rather than flatten, their understanding of the medium. We present an artist-driven methodology that uses neural networks to investigate the invisible connections between pairs of dancers. While machine learning techniques have been applied to solo dance performances, the intricate realm of partnering, which involves constant negotiation of weight shifts, synchronized gestures, and mutual influence, has remained relatively unexplored. We address this gap by using AI to capture subtle interactions between two dancers.[1]

We employ a pose extraction system on video recordings from the Partnering Lab (Vidrin 2025), translating 2D footage into 3D body poses in collaboration with the original dancers. This conversion enables us to track the trajectories of body joints and collect detailed data on each dancer's movement, later used to understand their collective patterns.

We use Graph Neural Networks (GNNs) to propagate information through the dancers' bodies, modeled as a set of nodes connected by edges. We also use Recurrent Neural Networks (RNNs) to handle the temporal nature of dance, allowing us to better capture the evolution of movements frame by frame. Because labeling how dancers influence each other is highly subjective, we adopted a self-supervised learning approach, optimizing for sequence reconstructions and focusing on the discovery of novel or unanticipated patterns rather than enforcing specific interactions.

In this proof-of-concept work, we present results using only a subset of points from each dancer instead of the full dimensionality of the input data. We also validate our methodology using a particle dataset with a known latent interaction graph. We find that the GNN model is able to identify interesting patterns in between pairs of dancers that align with the dancers' own intuitions. Looking ahead, we suggest that this methodology has the potential to complement embodied dance research strategies by highlighting surprising and invisible connections, reflecting and refining the dancers' understandings of their own creative relationships. In doing so, we not only extend our understanding of how dancers and technologists can collaborate, but also offer new ways for dancers and choreographers to perceive and shape their own practice.

Related Work

Graph Structure Learning Given our research interests in modeling the relationships between dancers' movements, the field of graph structure learning is highly relevant. We can relate our model to it by treating each body joint as

[1] https://github.com/humanai-foundation/ChoreoAI/tree/main/ChoreoAI_Duet_ChorAIgraphy_Luis_Zerkowski

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

a node and thinking of edges as capturing inter-joint (or inter-dancer) interactions. Graph Convolutional Networks (GCNs) (Kipf and Welling 2017) extend convolutional operations to graph-structured data, allowing message passing between connected nodes. For time series data like dance, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber 1997), can capture temporal dependencies by maintaining hidden states over frame sequences. Variants such as Graph Recurrent Neural Networks (GRNNs) (Ruiz, Gama, and Ribeiro 2020; Li et al. 2017a) combine these ideas, propagating information through both graph edges and time.

Variational Autoencoders on Graphs To learn the underlying structure of these dancer-to-dancer interactions, we employ methods inspired by Variational Autoencoders (VAEs) (Kingma and Welling 2022). VAEs have proven effective in unsupervised representation learning, enabling models to discover latent factors from complex input data. Building on this foundation, Neural Relational Inference (NRI) (Kipf et al. 2018) introduces a framework specifically designed to infer interaction graphs within systems of multiple nodes. NRI parameterizes the probability of different edge types - each one representing distinct forms of interaction - and learns these latent relationships.

AI and Dance Researchers have generated 2D and 3D dance movements using a variety of machine learning techniques including RNNs (Graves 2013; McCormick et al. 2015; Crnkovic-Friis and Crnkovic-Friis 2016; Alemi, Françoise, and Pasquier 2017; Li et al. 2017b; James 2018; Marković and Malešević 2018), clustering strategies like Kernel Principal Component Analysis (KPCA) (Berman and James 2015; Schölkopf, Smola, and Müller 1998), VAEs (Pettee et al. 2019; Papillon, Pettee, and Miolane 2022), GNNs (Pettee et al. 2020), and Transformers conditioned on musical inputs (Li et al. 2021b; 2020).

The NRI framework was previously applied to dance data by Pettee et al., but this work only considered the movements of a solo dancer. By adapting NRI to choreographic duets and enhancing its capacity of temporal understanding with GRNNs, we aim to automatically detect how one dancer's movements influence the other, ultimately revealing what the model identifies as subtle partnering dynamics that might otherwise go unnoticed.

Dataset Preparation

This project required transforming raw video footage of dance duets into structured pose data while addressing challenges such as multi-person tracking, occlusion, and movement complexity. We detail the dataset curation, pose extraction, and post-processing steps to ensure high-quality data for analysis and model training.

Data Collection

As an artist-centric effort, our project prioritized the needs and agency of our creative collaborators. We worked closely with Dr. Ilya Vidrin, leader of the Partnering Lab (Vidrin 2025), to develop the scope of the project and curate the data to suit our research goals. We collected video footage of movement data from duets of dancers associated with the Partnering Lab, totalizing four videos with an aggregate duration of 41 minutes and an average duration of 10 minutes and 15 seconds. Each video was filmed with a similar single-camera setup that started with each dancer facing one another on opposite sides of the screen, each grasping the other's right hand. The dancers featured in the filmed videos opted into the project and consented to our team using the data to train models designed to analyze duets for this project.

The movement styles of the dancers reflect the practices of the Partnering Lab, which "focuses on the internal, kinesthetic experience of movement that is difficult to perceive outside of experience" (Vidrin 2025). Targeting micromovements and exploring partnering as an embodiment of ethical concepts including trust and care, the dancers in these duets are highly attuned to one another, often moving slowly and deliberately, exploring mutual tensions and weight with active curiosity. Notably, the Partnering Lab team leader also regularly attended technical collaboration meetings and provided essential feedback that influenced the analysis design throughout the entire project duration.

Pose Extraction Pipeline

2D Pose Extraction We initially experimented with 2D pose estimation models from *AlphaPose* (MVIG-SJTU 2025), but found them insufficient for capturing spatial relationships between dancers. Also, despite the recent advances in human body pose extraction from images and multi-person tracking (Cao et al. 2017; Wei et al. 2016; Simon et al. 2017; Cao et al. 2019; Fang et al. 2017; Li et al. 2019), issues such as frame drops, identity swaps, and joint jitter were frequent, requiring extensive post-processing. An extraction example is shown in Figure 1.

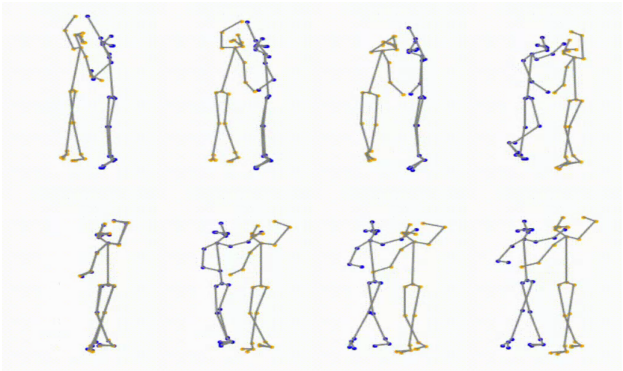


Figure 1: A raw 2D pose sequence from the Halpe pretrained model (26 keypoints) that shows vertex noise and even a missing dancer at one timestep.

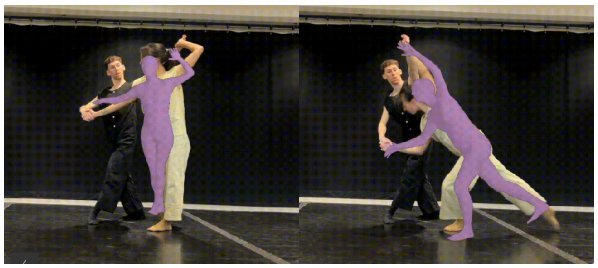
3D Pose Extraction Given the limitations of 2D methods, we adopted 3D pose estimation and 3D mesh reconstruction

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

pipelines (Li et al. 2021a; Kocabas, Athanasiou, and Black 2020) to preserve depth information crucial for analyzing dancer interactions. We evaluated two models:

- *VIBE* (Kocabas 2025): A prominent 3D pose and shape estimation model that outputs both joint positions and mesh reconstructions.
- *HybrIK* (integrated with *AlphaPose* (MVIG-SJTU 2025)): A hybrid analytical-neural approach that refines pose estimation by combining classical kinematics with learning-based corrections.

In Figure 2, we show an image comparison to visualize the resulting 3D joints and meshes.



(a) Mesh reconstruction from 3D pose coming from *VIBE*.

(b) Mesh reconstruction from 3D pose coming from *HybrIK*.

Figure 2: Comparison of 3D pose extractions: *HybrIK* (bottom) outperforms *VIBE* (top) in both simple (stationary) or complex (dynamic) movements.

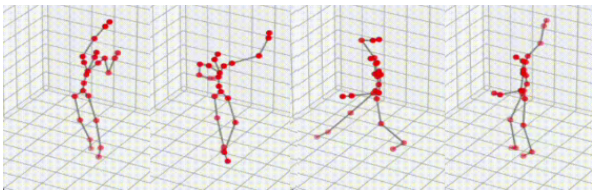
After testing, *HybrIK* (via *AlphaPose*) was selected for its higher pose accuracy, better multi-person tracking, and improved pose consistency. While minor noise and missing frames persisted, *HybrIK*'s outputs were the most reliable for partnered dance sequences.

Data Cleaning

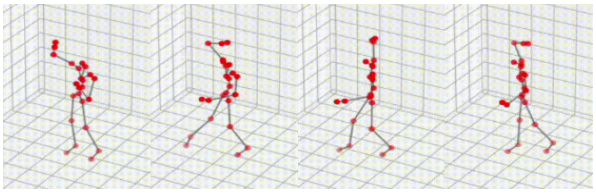
Despite selecting the best-performing model, raw 3D pose data still contained noise and inconsistencies. To ensure stable movement trajectories, we applied several post-processing steps:

- **Handling Missing Frames:** If a frame had no detected poses, we duplicated the poses from the previous frame. Since our videos run at around 30 FPS, this small gap-filling strategy did not produce abrupt motion artifacts.
- **Frames with Single-Person Detections:** When only one dancer was detected, we compared the sum of Euclidean distances between corresponding joints for the identified person and the two people in the previous frame. We then added the person from the previous frame with the greater distance to the current frame (assuming this was the non-captured person).
- **Frames with More Than Two Detections:** Some frames falsely identified more than two individuals. Since our videos contained exactly two dancers, we kept only the two highest-confidence detections.
- **Index Consistency:** If dancers swapped IDs across frames, we scanned correct frames around the confusion frame and used the aforementioned sum of Euclidean distances between corresponding joints to correct the inversions.
- **Vertex Jitter:** Even successful detections contained local noise, causing "shaky" joints. We applied a 3D Discrete Cosine Transform (DCT) (Ahmed, Natarajan, and Rao 1974) low-pass filter at a 25% threshold to smooth out high-frequency jitter.

The impact of these corrections can be seen in Figure 3 by comparing the minimally-processed (i.e. handling missing frames and missing/additional people) output with our fully-processed result. The final output is cleaner, more stable, and better preserves each dancer's bodily trajectory, which is vital for dance partnering research.



(a) Pose extraction with minimal processing.



(b) Pose extraction with full processing pipeline.

Figure 3: Comparison of 3D pose extractions for a dancer: minimal processing (top) and full pipeline (bottom). While both depict the same sequence, the bottom image captures the movement far more clearly, with smoother, more continuous transitions between frames. In contrast, the top image appears fragmented, with disconnected poses that obscure the flow of motion.

3D Graph Construction

To model dancer interactions, we represented each dancer as a graph of 29 joints (nodes), a skeleton obtained from

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

HybrIK (Li 2025)[2], and fully connected all joints between dancers to form a dense bipartite graph. The learning objective is to classify or weight these inter-dancer edges, indicating how crucial each connection is for movement correlation.

The data is then prepared for model training by creating batches with *PyTorch* (Paszke et al. 2019) tensors. The tensors are structured with dimensions representing the total number of sequences, the sequence length, the number of joints from both dancers, and 3D coordinates + 3D velocity estimates. Finally, a 85%-15% training-validation split is created to allow for proper model hyperparameter tuning. To improve model generalization given our limited dataset, the training pipeline incorporates a data augmentation step that involves rotating batches of data. Each batch is rotated along the $\hat{z}$-axis by a randomly-selected angle $\theta \in [0, 2\pi]$ while maintaining the original $\hat{x}$ and $\hat{y}$-axis orientations for physical consistency. This approach helps prevent the model from overfitting to the dancers' absolute positions.

Due to the high complexity of the problem, both in the number of moving nodes and the number of edges in the graph, random joint sampling was implemented to reduce the scale of the problem. Only subsets of 6 to 10 joints (3 to 5 from each dancer) are used in each training run, and only inter-dancer edges among those sampled joints are considered. This approach helps the network converge more reliably while demonstrating proof-of-concept for larger graphs.

Graph Neural Network Model

Understanding dancer interactions requires modeling relationships rather than individual trajectories. We extend Neural Relational Inference (NRI) (Kipf et al. 2018) to treat joints as graph nodes and infer edges representing inter-dancer connections. This section details the model's formulation and architecture.

Neural Relational Inference Variant

Our model builds on NRI (Kipf et al. 2018), an extension of Variational Autoencoders (VAEs) (Kingma and Welling 2022), originally designed to infer latent interaction graphs from particle motion. By analyzing position and velocity, NRI estimates which particles influence others without predefined relationships.

This approach is particularly relevant for duet dance analysis. Just as NRI infers particle interactions without a known ground-truth graph, we also lack a predefined interaction graph for dance partnering. The relationships between dancers shift dynamically, and there's no ground-truth connection between dancers. Instead, we employ self-supervised learning, optimizing for sequence reconstruction rather than enforcing predefined structures.

Our model consists of an encoder and a decoder, both designed to iteratively transform node representations into edge representations and vice versa. The encoder generates edge logits, which are used to construct the latent space. Since edge indices are later sampled for the decoder, transitioning between representations is a crucial step in the process. Our model implementation includes a few important modifications from the core NRI structure (Kipf et al. 2018):

- **Graph Convolutional Network (GCN):** Some linear layers are replaced with GCN layers to leverage the graph structure, improving the model's ability to capture relationships between joints. This change focuses on a subset of edges connecting both dancers rather than studying all joint relationships - as in the original implementation. Additionally, GCNs provide local feature aggregation and parameter sharing, important inductive biases for the context.
- **Graph Recurrent Neural Network (GRNN) Decoder:** To make better use of sequential information and potentially achieve a more suitable final embedding for predicting (or reconstructing) the next frame, one possible approach is to use a recurrent network. It is important to note that this is more of an update to a version of the NRI model rather than a completely new idea. The authors had already explored a recurrent decoder, with the main differences being the previously mentioned modifications in layers and the use of GRU nodes instead of LSTM nodes (introduced in the next item).
- **Custom GCN-LSTM Cells:** To utilize the recurrent structure crucial for sequence processing while maintaining graph information and GNN architecture, the classic LSTM cell has been reimplemented with GCN nodes. In the final version of our architecture, only the decoder incorporates the recurrent component, which generates a final sequence embedding that the model uses to reconstruct the next frame.

By incorporating these modifications, the model maintains the core principles of the original NRI while theoretically enhancing its ability to generalize and adapt to the dynamics of "connected" particles moving. We now describe the architecture (shown in figure 4) more precisely:

Encoder

- **Node-to-Edge Transformation:** Dancer joints are passed through a GCN layer with 64 latent dimensions and then converted to edge representations, which doubles this representation.
- **Linear Layer + Batch Normalization + Dropout:** A linear layer refines edge embeddings, receiving input of dimension 128 dimensions and generating an output of 64 dimensions. Optionally we use batch normalization and dropout with 10% probability for regularization.
- **Edge-to-Node Transformation:** The edge representations are converted back into nodes, and another GCN layer with same dimensionality as before is applied.
- **Second Node-to-Edge Transformation:** The nodes are then transformed back into edges, followed by another Linear layer from 192 dimensions (this time we have

[2]More specifically, from this issue `https://github.com/jeffffffli/HybrIK/issues/140` and expanded through this code `https://github.com/jeffffffli/HybrIK/blob/main/hybrik/datasets/h36m_smpl.py`.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

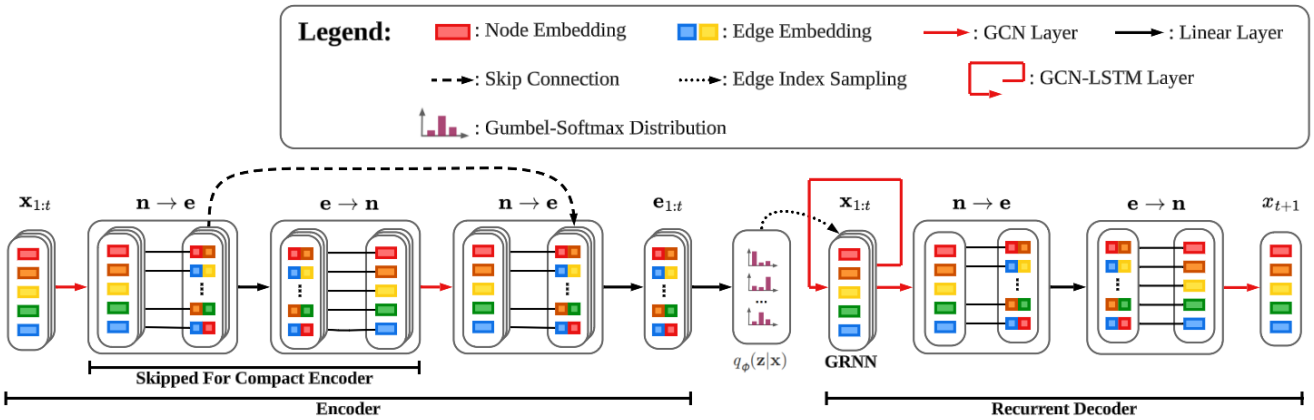


Figure 4: Schematic of the final model architecture, including the GCN nodes and the GRNN adapatation, inspired by the one found in the original NRI paper (Kipf et al. 2018) (Figure 3, page 3). The "compact encoder" variation - marked here as "Skipped for Compact Encoder" - was introduced to test a simplified version of the model, motivated by challenges with data. This version includes only a single node-to-edge transition in the encoder. However, as shown later in Table 1, the compact encoder did not outperform the full model and, in some cases, performed slightly worse. For this reason, it was only used for comparative experiments and was not adopted in the final architecture.

3×64=192 because we also use a skip connection coming from the first Node-to-Edge transformation) to 64 dimensions.

- **Logits Computation:** A final linear layer outputs logits for each possible edge type. For this layer, we mainly tested with binary edges ("existing", "non-existing") or 3 edge types ("no connection," "weak" and "strong"), but we also added the possibility of 4 or even 5 edge types. Each one of these options comes with a prior probability distribution that guides the latent distribution, making the later sampling process gravitate towards a certain allocation of each edge type.

Decoder

- **Edge Index Sampling:** The decoder begins by hard sampling edge indices using a Gumbel-Softmax distribution (Maddison, Mnih, and Teh 2017; Jang, Gu, and Poole 2017), based on the logits generated by the encoder and the prior probabilities of each edge type. While this process results in discrete edges (i.e., edges are either present or absent, without intermediate probabilities), it effectively approximates sampling from a continuous distribution and uses Softmax to enable the reparameterization trick, keeping the entire pipeline fully differentiable.
- **GRNN + Node-to-Edge Transformation:** Once edges have been sampled, the decoder processes the data through a GRNN composed of modified LSTM nodes with GCN layers, followed by a transformation of the final sequence embedding into edge representations with 64 dimensions again.
- **Linear Layer + Batch Normalization + Dropout:** A linear layer refines edge embeddings, receiving input of dimension 128 dimensions and generating an output of 64 dimensions, and optionally using batch normalization and dropout with 10% probability for regularization.
- **Edge-to-Node Transformation:** The edge representations are converted back into nodes, and a GCN layer is applied to predict (or reconstruct) the next frame of the input sequence.

Experiments and Results

We evaluated our model on two datasets: a charged-particle dataset from the original NRI paper, serving as a simpler dataset with a known interaction graph to help benchmark our model performance, and the duet dance dataset, featuring real-world 3D pose sequences with two dancers at a time.

Charged Particle n-Body Simulations

To validate the architecture's viability under more controlled conditions, we used the original charged-particle dataset used in Kipf et al.. We generated 50,000 simulated trajectories of five particles each, with 2D positions and velocities over 49 frames. A random interaction graph dictates repulsive or attractive forces between particles.

We tested a variety of architectures, and the results were promising. Figure 5 presents an example from one of our models, showing that the predicted graph structures and particle trajectories generally aligned with the ground truth. Additional validation results can be found in our open-source codebase.

3D Dance Duets

To assess model performance on the 3D dance data, we use predicted movement fidelity as a proxy for evaluating sampled edges, given that the subjectivity of defining correct inter-dancer connections makes defining true edge labels difficult. Models were trained until convergence, typically

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

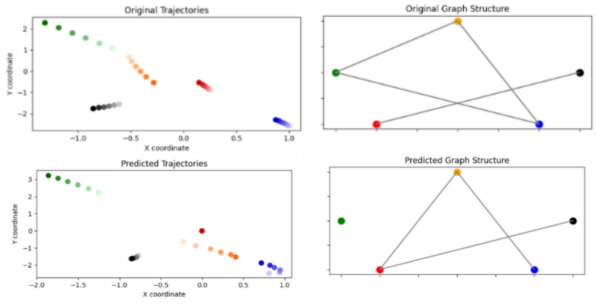


Figure 5: On top, original simulated trajectories and original sampled edges. On bottom, reconstruction and edge prediction results. The model accurately captured the movement and location of three particles (green, black, blue), approximated movement shape for one (orange) despite location inaccuracies, and positioned the last (red) reasonably well but without movement.

for 15 to 20 epochs, and resulted in stable reconstruction loss curves with no signs of overfitting on validation data.

Table 1: **Reconstruction Mean Squared Error** for multiple tasks, input sequence lengths (ℓ), number of edge types (n), and model architecture configurations (compact encoder vs. full architecture as introduced in Figure 4). Bold highlights the best (smallest) reconstruction error.

Task	ℓ	n	Model	MSE
5-Body Sim.	6	3	Compact Enc.	0.70
5-Body Sim.	6	3	Full Arch.	0.69
5-Body Sim.	6	4	Compact Enc.	0.82
5-Body Sim.	12	3	Compact Enc.	0.33
5-Body Sim.	12	3	Full Arch.	**0.32**
6-Joint Dance	8	4	Full Arch.	0.58

KL-divergence loss, however, plateaued early, limiting latent space exploration. While beta coefficient scaling mitigated this effect, short training runs reduced its effectiveness. Longer training stabilized KL behavior, but these experiments were conducted on earlier, smaller models without a Graph Recurrent Neural Network (GRNN), which made extended training less feasible in the final architecture.

For one of our best-performing models, we trained with 6 sampled joints (3 per dancer), with dancer rotation augmenting the dataset 10 times. The full encoder was used with 8-frame input sequences, 64 hidden dimensions, and 4 edge types. It was trained for 20 epochs (18 hours) using mean squared error. Table 1 shows some MSE values for multiple trained models.

Movement Predictions We consider a random subset of 3-5 joints for each of the two dancers for both model training and predicted movement evaluations. The predicted movements are highly sensitive to sampled edges, as information in a graph network spreads through connected nodes. The best-performing models tend to assign multiple connections to a single joint, enabling smoother information flow between dancers. As seen in Figure 6, sampled joints maintain accurate spatial positioning and move coherently with the body. Given that the sampled connections only consist of joints between the dancers, it's notable that the captured interaction can lead to accurate movement predictions for each of the individual dancers.

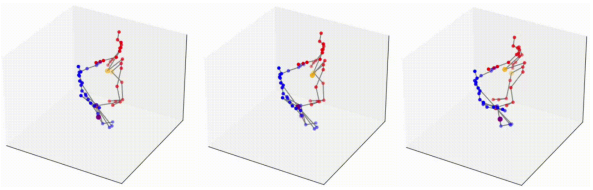


Figure 6: Example of a good reconstruction. Reconstructed sampled joints are color-coded for clarity: purple for the blue dancer and orange for the red dancer.

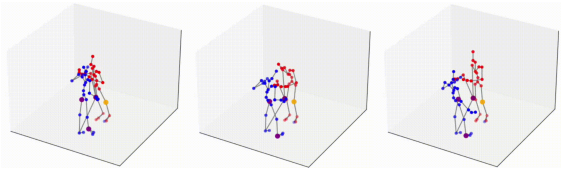
The reconstructed frames provide valuable insight into how the model interprets connections between particles. By analyzing the learned edge distribution and sampled edges across different examples, it becomes possible to better understand the network's perception of dancer interactions and movement patterns.

Certain limitations arise due to the nature of the edge sampling and model assumptions. Common sources of reconstruction errors include:

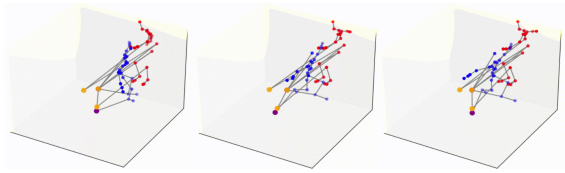
- **Shaking and jitter:** Some particles exhibit noticeable instability, partly due to inherent noise in the original sequence and also because each reconstructed frame is generated independently, without leveraging temporal smoothing since the model predicts only one frame at a time.
- **Limited movement in sampled joints:** In some cases, the sampled edges do not adequately capture movement, resulting in reconstructions in which the joints remain nearly stationary (see e.g. Figure 7, top). This effect is particularly pronounced when the sampled connections fail to establish a strong information pathway between dancers.
- **Challenges with dancer switching:** When dancers cross paths or switch sides, the reconstructed joints sometimes remain near their initial positions rather than following the dancers' actual motion. This suggests that the model has difficulty generalizing to sequences where relative spatial relationships shift significantly.
- **Drift towards the center:** If the dancers move away from the origin, reconstructions tend to stay near the center (Figure 7, bottom), which could be improved using a different form of pre-processing.

Regardless of these challenges, the reconstructions still generally capture meaningful aspects of the dancers' interactions. The observed limitations highlight areas for future refinement, such as improving edge sampling strategies and incorporating temporal consistency in predictions.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(a) Example of edge sampling limiting reconstruction to 3 joints, while the other 3 remain static at the center of the coordinate frame despite being sampled.



(b) Example illustrating the model's reconstruction challenges when dancers are farther from the center, resulting in joints that are either stationary or inaccurately positioned.

Figure 7: Examples of poor reconstructions.

Predicted Edges During inference for edge prediction, edges are sampled differently from the training phase. Instead of hard sampling, which enforces a strict edge selection, we use soft sampling (i.e. sampling with associated probability) to retain only high-confidence connections. This ensures that only the most structurally relevant edges remain in the final reconstruction, reducing noise and improving interpretability.

A key observation is that the number of sampled edges with a confidence above 80% remains consistently low and aligns closely with the prior distribution of edge types (whether two, three, or four types are used). This suggests that the learned latent space effectively captures meaningful movement relationships rather than introducing arbitrary edges. Since the edge priors were designed to reflect a sparse but essential connectivity between dancers, this alignment reinforces the idea that the model is not simply memorizing movement sequences but actively identifying the most influential joints for movement propagation.

The predicted edges exceeding the 80% confidence threshold have interesting several features:

- First, most of these connections tend to have similar confidence levels. This indicates a **low hierarchy among selected edges** (Figure 8). In other words, once an edge is classified as important, it tends to contribute equally to the reconstruction, rather than one edge being clearly dominant. This is particularly interesting from a movement analysis perspective because it suggests that information is distributed across multiple interaction points rather than concentrated in a single connection.
- Additionally, it is common for a single joint to be connected by multiple edges, meaning that **certain key joints act as hubs** in the reconstructed interaction graph (Figure 9). This aligns with prior observations that information propagates more effectively when there are multiple connected paths leading to and from the same dancer (through one joint or within two hops). For instance, a dancer's hand might influence their partner's torso, while simultaneously being connected to their partner's own hand or shoulder, helping to propagate multiple layers of movement dependency.

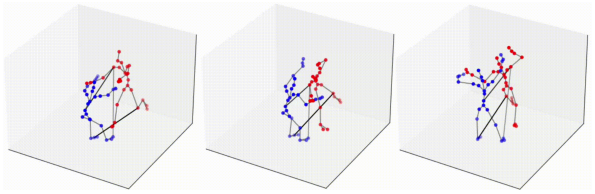
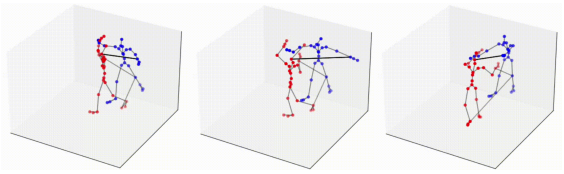
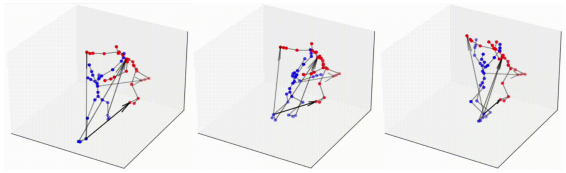


Figure 8: Example of the sampled edge distribution. The black edges represent connections between the dancers, with darker edges indicating higher confidence in their importance for reconstruction. In this typical case, 3 edges were selected for 6 sampled joints, 2 with slightly higher importance, though all exceed 80% confidence.



(a) An undirected example illustrating simple movement for a set of 6 sampled joints.



(b) A directed example showcasing more complex movement for 10 sampled joints. Notably the model captures how the foot motion of one dancer influences multiple parts of the other dancer's body. This aligns with the performed movement, where the blue dancer's dynamic spin guides the red dancer's response.

Figure 9: Examples of multiple edges connected to the same joint.

- Another recurring pattern is that **edges frequently form between joints that are in opposition**, as if connected by an invisible elastic band. These joints appear to be stretching or pulling apart, suggesting that the model is particularly sensitive to moments of tension between dancers (Figure 10). This is especially relevant in the context of duet choreography, where push and pull dynamics are fundamental to many partnering techniques.

  For example, if one dancer extends their arm forward while their partner leans away, the model often identifies a connection between these two points, even if there is no

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

direct contact. This suggests that the model is not only detecting explicit force transfers (such as corresponding body parts in contact) but also implicit movement dependencies, where one dancer's action influences the other's balance, trajectory, or momentum.

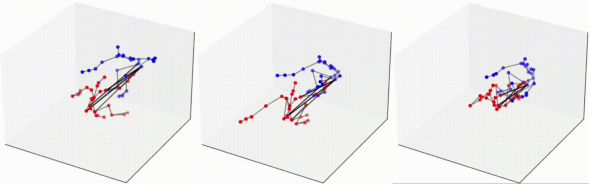


Figure 10: Undirected example of connections within opposition tendencies. It shows multiple connections between the lower torso of both dancers, first leaning in opposite directions and then gravitating toward each other, illustrating the full range of the stretched-string analogy.

This result aligns closely with our dancers' choreographic intuitions and training, in which movement is often defined by how tension and release are negotiated between partners. It also suggests that the model is learning a representation of movement that extends beyond surface-level trajectory following.

Discussion

Creative Implications

This work offers a new, useful lens for dancers to analyze their movements and interactions, revealing patterns and relationships they may not consciously recognize. By mapping inter-dancer connections, it provides insights into habitual tendencies, asymmetries, and unconscious influences in movement. For duets, this tool makes visible the intuitive exchanges between partners — subtle weight shifts, spatial negotiations, and reactive gestures that shape their interaction. Seeing these dynamics mapped out allows dancers to refine their awareness and explore new choreographic possibilities.

This approach reimagines dance as a dynamic graph in which dancers are linked by evolving connections. This perspective could inform interactive performances, projecting real-time movement-based visualizations onto a stage, rendering salient aspects such as tone, frame, timing, and mutual responsiveness. Real-time relational graphs used in performance could project evolving connection structures live, influencing lighting/sound/environment based on detected relational patterns. This would make connection itself an active, evolving part of the performance material — not just a background reality.

This approach can inform the creative process itself (whether improvised or choreographed). Rather than starting with fixed steps or positions, the design of connection dynamics can promote exploration of how energy flows between dancers, how subtle shifts in direction, position, orientation, and point(s) of contact cascade into sequences of movement. Even in highly trained dancers, the relational structures — who is influencing whom, how strongly, at what moment — are felt but not easily seen or measured. Artificial intelligence, particularly relational inference models like this one, can externalize these hidden connection patterns into something visual, trackable, and analyzable over time. Dancers could observe after the fact how influence passed between certain bodies, or that certain points in the body (like hips or hands) carried more relational weight. This would allow dancers to reflect on connection structures in a way that is otherwise invisible — opening possibilities for refinement, experimentation, or re-composition based on those patterns.

Additionally, the system holds potential for dance pedagogy and interdisciplinary research, bridging choreography with biomechanics, for example. Students could receive real-time visualizations of connection strength, delay, or mutuality, transforming abstract instruction into tangible, actionable feedback. Moreover, by analyzing the emergent interaction graphs, dancers could discover non-obvious relationships within their movements, finding novel pathways of influence that could inspire greater awareness, which can subsequently promote greater agency in practice. This system would allow for an emphasis on mutuality rather than hierarchy, honoring connection itself as choreographic material — something that can be composed, taught, and evolved in ways that are otherwise obscured. By exposing hidden layers of movement, this work expands both artistic and analytical possibilities for dancers and technologists.

Future Directions

Despite these promising proof-of-concept results, several improvements are needed to enhance model performance and usability. Key areas for future work include:

- **Data expansion and quality:** Regardless of the use of data augmentation through duet rotation, the small dataset size made training difficult. Also the pipeline used to extract 3D poses has room for improvement. Even in original sequences, the dancers' joints are shaky and often poorly approximated, leading to random (unrealistic) movements. Moreover, normalization of both dancers was removed to preserve relative movement, but it was realized too late that a new layer normalizing their joint movement should have been added - having dancers in different parts of space caused confusion for the models.
- **Architecture exploration:** While many versions of the NRI variant were implemented and tested, the final version is still far from the potential we saw from the strong results achieved by the original version. Furthermore, alternative architectures like transformers (Vaswani et al. 2023) could enhance temporal modeling and better align with modern architectures.
- **Model validation:** Extensive qualitative analysis was conducted through various training sessions, experiments, and even scenario simplifications, but the study did not incorporate a detailed quantitative evaluation beyond learning curve and MSE analysis. Additional metrics and comparative tests with different parameters - such as the number of reconstructed frames - were not much explored and could provide further insights.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

- **Processing speed:** Some parts of the final architecture are suboptimal custom implementations. Batches are replaced with sequential operations at several points in the pipeline - node and edge representation conversions, edge sampling for each sequence in a batch, and in the GRNN. As a result, a training cycle with just a few dozen epochs can take an entire day, which is not ideal for scaling.
- **Interaction with dancers:** Since this project sits at the intersection of art and technology, more direct interaction with artists is essential. With a more refined version of the model, it would be ideal to present the tool to the dance community and observe how dancers use the tool in their own partnering studies.

Conclusion

This work presented an investigation into how AI-driven methods can enhance our understanding of partnered dance by focusing on revealing inter-dancer connections. We paired an open-source 3D pose extraction pipeline with a custom pre-processing stage to capture dancers' movements and then trained an NRI-based architecture augmented with GCN modules and RNN structure to reveal important edges between the dancers. Though the model only processed a subset of the full-body kinematics of the dancers, it demonstrated a capacity to uncover interesting dance dynamics such as high-confidence inter-joint dependencies that align with choreographic intuition, the presence of key movement hubs that serve as central points of influence within the duet structure and recurring patterns of tension and release.

While far from comprehensive, our work explores the potential for AI to augment our creative understanding of duet dynamics in choreographic settings. From connections that focus on "push and pull" forces between dancers to more intricate patterns of mutual influence, this project shows the value of graph-based approaches in modeling collaborative creative frameworks.

Author Contributions

Authors one and two jointly developed the pose extraction and data processing pipelines. Author one designed and implemented the machine learning methodology and led the paper manuscript writing. Throughout the process, authors three and four provided technical and artistic supervision.

Acknowledgements

The authors would like to thank the HumanAI program and Google Summer of Code for supporting this work.

References

Ahmed, N.; Natarajan, T.; and Rao, K. 1974. Discrete cosine transform. *IEEE Transactions on Computers* C-23(1):90–93.

Alemi, O.; Françoise, J.; and Pasquier, P. 2017. GrooveNet: Real-time music-driven dance movement generation using artificial neural networks. https://www.researchgate.net/publication/318470738_GrooveNet_Real-Time_Music-Driven_Dance_Movement_Generation_using_Artificial_Neural_Networks.

Berman, A., and James, V. 2015. Kinetic imaginations: Exploring the possibilities of combining AI and dance. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 2431–2437. AAAI Press.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Crnkovic-Friis, L., and Crnkovic-Friis, L. 2016. Generative choreography using deep learning. https://arxiv.org/abs/1605.06921.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. RMPE: Regional multi-person pose estimation. In *ICCV*.

Graves, A. 2013. Generating sequences with recurrent neural networks. https://arxiv.org/abs/1308.0850.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

James, A. B. . V. 2018. Learning as performance: Autoencoding and generating dance movements in real time. *Computational Intelligence in Music, Sound, Art and Design* 256–266.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparameterization with gumbel-softmax.

Kingma, D. P., and Welling, M. 2022. Auto-encoding variational bayes.

Kipf, T. N., and Welling, M. 2017. Semi-supervised classification with graph convolutional networks.

Kipf, T.; Fetaya, E.; Wang, K.-C.; Welling, M.; and Zemel, R. 2018. Neural relational inference for interacting systems. *arXiv preprint arXiv:1802.04687*.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kocabas, M. 2025. Vibe: Video inference for human body pose and shape estimation. https://github.com/mkocabas/VIBE. Accessed: 2025-02-13.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2017a. Gated graph sequence neural networks.

Li, Z.; Zhou, Y.; Xiao, S.; He, C.; and Li, H. 2017b. Auto-conditioned LSTM network for extended complex human motion synthesis. *CoRR* abs/1707.05363.

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; and Lu, C. 2019. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10863–10872.

Li, J.; Yin, Y.; Chu, H.; Zhou, Y.; Wang, T.; Fidler, S.; and

Li, H. 2020. Learning to generate diverse dance motions with transformer.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021a. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3383–3393.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021b. Ai choreographer: Music conditioned 3d dance generation with aist++. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 13381–13392.

Li, J. 2025. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. `https://github.com/jeffffffli/HybrIK`. Accessed: 2025-02-13.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2017. The concrete distribution: A continuous relaxation of discrete random variables.

Marković, D., and Malešević, N. 2018. Adaptive interface for mapping body movements to sounds. In Liapis, A.; Romero Cardalda, J. J.; and Ekárt, A., eds., *Computational Intelligence in Music, Sound, Art and Design*, 194–205. Cham: Springer International Publishing.

McCormick, J.; Hutchison, S.; Vincs, K.; and Vincent, J. B. 2015. Emergent behaviour: Learning from an artificially intelligent performing software agent. In *21st International Symposium on Electronic Art*.

MVIG-SJTU. 2025. Alphapose: Real-time and accurate full-body multi-person pose estimation and tracking system. `https://github.com/MVIG-SJTU/AlphaPose`. Accessed: 2025-02-13.

Papillon, M.; Pettee, M.; and Miolane, N. 2022. Pirounet: Creating dance through artist-centric deep learning.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. Pytorch: An imperative style, high-performance deep learning library.

Pettee, M.; Shimmin, C.; Duhaime, D.; and Vidrin, I. 2019. Beyond imitation: Generative and variational choreography via machine learning.

Pettee, M.; Miret, S.; Majumdar, S.; and Nassar, M. 2020. Choreo-Graph: Learning Latent Graph Representations of the Dancing Body. *NeurIPS Workshop on Machine Learning for Creativity and Design*.

Ruiz, L.; Gama, F.; and Ribeiro, A. 2020. Gated graph recurrent neural networks. *IEEE Transactions on Signal Processing* 68:6303–6318.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10(5):1299–1319.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention is all you need.

Vidrin, I. 2025. Partnering lab website. `https://www.partneringlab.com/`. Accessed: 2025-02-13.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *CVPR*.

Co-Designing Fashion with AI: A Small-Data Approach to Generative Garment Design
Paper type: Cultural Application

Ina Hrešć[1] **and Terence Broad**[2]
Creative Computing Institute
University of the Arts London
United Kingdom
[1]i.hresc0620231@arts.ac.uk, [2]t.broad@arts.ac.uk

Abstract

This paper details a case study of implementing a co-design framework for integrating artificial intelligence (AI) into fashion design through a collaboration between the researcher and a garment designer. Using the designer's fashion development as a practical example, this research investigates AI's potential to enhance creative workflows, improve efficiency, and facilitate marketing strategies that reduce the necessity for physical prototyping. The project implements a small-data approach, exclusively fine-tuning AI models on the designer's sketches and the researcher's photography to ensure highly personalised outputs. AI architectures, such as Stable Diffusion and ControlNet, are fine-tuned and used to generate sketches and photorealistic visualisations that work as an extension of the designer's artistic vision while promoting sustainable awareness and ethical data practices. Through continuous dialogue and a final interview, the designer's perspective was analysed, demonstrating AI's efficacy as a creativity support tool that streamlines design iteration by enabling rapid prototyping and ideation, while also addressing concerns about preserving artisanal expertise. This research contributes to the growing field of AI in the creative industries by demonstrating the value of co-design, small-data strategies, and ethical AI development to build tools that empower designers.

Introduction & Theoretical Context

Generative artificial intelligence (AI) has increasingly impacted the creative industries, not only as a tool for automation but as a means of supporting and enhancing human-led creative workflows. In fashion, AI's visual abilities present opportunities for human-machine collaboration, where generative systems can support designers in the creative process. This research explores the practical potential of diffusion-based models, specifically Stable Diffusion (Rombach et al., 2022) with ControlNet (Zhang, Rao, and Agrawala, 2023), in a close partnership with an independent fashion designer. Therefore, we position AI as a powerful assistive tool that facilitates iterative design exploration (Wu et al., 2021), thus enhancing the designer's workflow. The creative direction regarding the fashion designs was guided by the desire to avoid explicit gender associations, adopting a neutral perspective (Fan, 2023), to allow for the representation of fluid gender identities. This shaped the aesthetic goals and our approach to ethical data practices and inclusivity in design. By eliminating binary perceptions, we fostered an environment that allows the exploration of diverse identities, including those that are gender non-conforming and non-binary (Marquez-Gallardo and Rovira-Lorente, 2024). To achieve the desired gender neutral presentation, negative prompting, which functions by neutralising unwanted elements in the latent space of diffusion models (Ban et al., 2024), was used to mitigate clear gender identifiers. Furthermore, post-processing retained the mannequin aesthetic, removing potential biases related to facial features and hair, while also merging the designer's traditional sketching with AI's photorealistic garment renderings.

Generative AI tools, such as Generative Adversarial Networks (GANs) and diffusion models, have demonstrated potential in fashion design (Guo et al., 2023). Early work with GANs, as proposed by Goodfellow et al. (2014) and exemplified by FashionGAN, showcased AI's sketch-to-image generative abilities (Cui et al., 2018). However, GANs often suffer from limitations such as mode collapse (Thanh-Tung and Tran, 2020) and reliance on large datasets, constraining creative flexibility (Karras et al., 2018; Razavi, van den Oord, and Vinyals, 2019). By contrast, latent diffusion models (LDMs), like Stable Diffusion, have emerged as more flexible alternatives, operating in a compressed latent space and enabling diverse, high-quality image outputs that require less computational power (Rombach et al., 2022). ControlNet further enhances control by incorporating spatial conditioning, allowing for precise control of generated images (Zhang, Rao, and Agrawala, 2023). For instance, the Multimodal Garment Designer (MGD), proposed by Baldrati et al. (2023), enabled guided image generation through textual descriptions, body poses, and sketches. Zhang, Zhang, and Xie (2024) presented a two-stage approach using ControlNet for sketch-to-image and image-to-image refinement, showcasing precise control over garment structure and details.

Technical advances in AI-assisted fashion design are well documented. However, the literature lacks a comprehensive examination of close collaborative relationships between developers and fashion designers. This research adopts a co-design framework, emphasising shared power, mutual learning, and open communication (McKercher, 2020). By utilising a small-data approach, fine-tuning AI models exclu-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

sively on personal data (designer's sketches, primary researcher's personal photography), this study promotes ethical data practices and sustainable AI development. We note that, in creative machine learning, goals often differ from those of traditional ML, as highlighted by Viglensoni, Perry, and Fiebrink (2022). They argue that the assumption that "bigger data is better in generative AI systems" does not necessarily apply to more niche creative use contexts. DreamBooth, a fine-tuning technique, allows for learning of specific subjects and personalised style transfer with minimal data (Ruiz et al., 2022). Although LDMs are less computationally intensive than GANs, when trained on large datasets, they still require significant computational resources and increased model training times, leading to negative environmental impacts (Utz and DiPaola, 2023). By reducing model training times through small-scale fine-tuning, DreamBooth offers a potential reduction in these impacts. Further, the commercial use of AI models raises critical concerns for creatives, including potential job displacement and ethical issues such as copyright and privacy (Clarencia et al., 2024). The use of personal data and a co-design framework emphasises fairness and accountability in AI development, addressing crucial ethical considerations related to data governance and transparency (European Commission, 2019). Furthermore, the use of small personal datasets reduces the risk of bias, such as the inadvertent perpetuation of stereotypes, and increases the explainability of outputs.

This research aims to address the following questions:

- How can generative AI technologies, such as Stable Diffusion and ControlNet, enhance the creative processes of independent designers while preserving their distinct styles?
- To what extent can small-scale, personalised fine-tuning enable designer-led prototyping for sustainable workflows?
- Is the use of small, ethically sourced data a viable approach for ethically responsible AI-assisted workflows and for fostering creative collaboration between developers and designers in bespoke fashion projects?

This study contributes to the field of computational creativity by demonstrating the potential of using AI as a creativity support tool in fashion design, emphasising co-design, sustainability, and ethical data practices through practical experiments and collaborative work. It reflects an emerging form of human-AI co-creation, which positions AI as a complementary force that enhances and expands creative possibilities, as exemplified by AI-assisted poetry (Wu et al., 2021). This approach aligns with the foundational concepts of computational creativity, wherein AI can significantly aid the exploration and transformation of creative spaces (Boden, 2004). The following section details the implementation of and technical workflow used to explore these concepts.

Methodology

Figure 1 provides a high-level overview of the research workflow, which integrates co-design feedback loops. This section details our methodology from data acquisition and experimental design to post-processing and collaborative evaluation.

Data Acquisition & Preparation

All data was collected ethically and with informed consent, ensuring the designer was continuously updated on data usage and understood the aims of the project. The designer contributed 29 annotated sketches, each reflecting their unique style and specific project objectives. The researcher contributed 44 personal photographs, taken during various photoshoots, for backdrop generation, aiming for a consistent urban and rustic aesthetic.

Data preprocessing involved several transformations to prepare the images for model training. Designer annotations were removed, sketches were cleaned, and all images were resized and, where necessary, padded to 768x768 pixels. For data augmentation, all sketches were horizontally flipped. To serve as control inputs alongside prompts, edge maps were generated.

Experimental Design and Implementation

Stable Diffusion 2 was fine-tuned on the 58 sketches (29 original orientation, 29 flipped) using the DreamBooth technique (Ruiz et al., 2022). To guide the model in replicating specific stylistic details for each photorealistic visualisation, individual image prompts were crafted based on the designer's annotations of their corresponding sketch counterparts. A batch size of 4 was selected to maximise GPU memory usage while maintaining training efficiency. Gradient accumulation of 2 was used to simulate a larger batch size, improving training stability. A learning rate of 1e-6 was selected for training stability.

For photorealistic rendering, *realistic-vision-v51* was selected due to its superior ability to produce high-quality photorealistic images (Saovana and Khosakitchalert, 2024). It was fine-tuned on 44 photographs from the backdrop dataset to achieve desired textures and lighting conditions in the image backdrops. ControlNet was used with edge maps from the sketches as control inputs (von Platen et al., 2023). We adjusted the *guidance_scale* and *controlnet_conditioning_scale* parameters to optimise results for each sketch. The *guidance_scale* balanced prompt adherence and image structure, with higher values emphasising prompt accuracy. The *controlnet_conditioning_scale* influenced ControlNet's strength, ensuring structural adherence to the edge maps without overpowering the backdrop style transfer.

Backdrop generation involved fine-tuning *realistic-vision-v51* on the 44 selected backdrop images. We explored two approaches:

- Fine-tuning with image captions (CSV descriptions).
- Fine-tuning without image captions.

Training parameters for backdrop generation were set to a learning rate of 5e-7, chosen to prevent catastrophic forgetting, while still allowing the model to learn the specific characteristics of the backdrop dataset. A train batch size of 3 and gradient accumulation steps of 2 were used to balance

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Figure 1: High-Level Workflow: Overview of the methodology, illustrating the integration of co-design feedback loops.

Experiment	Model Used	Dataset	Training Method	LR	Batch Size	Observations
SDv2 Sketch Generation	Stable Diffusion 2	58 sketches	DreamBooth	1e-6	4	Best: Checkpoints 600 & 800
RVv5.1 Backdrop Generation (Captioned)	Realistic-Vision-v5.1	44 backdrops	DreamBooth	5e-7	3	Suboptimal cohesion
RVv5.1 Backdrop Generation (Uncaptioned)	Realistic-Vision-v5.1	44 backdrops	DreamBooth	5e-7	3	Better cohesion, Checkpoint: 2400

Table 1: Summary of Training Experiments and Model Configurations

GPU memory usage and training stability, particularly important given the smaller dataset size.

Post-Processing and Collaborative Evaluation

Adobe Photoshop was used for post-processing, specifically to mask and blend AI-generated garments with the original mannequin sketches. This process was essential for achieving the desired neutral mannequin aesthetic and ensuring the final renderings retained the designer's distinctive drawing style in the mannequin body parts, resulting in outputs that merged AI-generated and traditional art.

Qualitative assessment played a key role in the project. Continuous dialogue with the designer throughout the implementation process allowed for iterative adjustments and refinements. A structured interview was conducted to gather detailed feedback on the model's performance and the collaborative process. Relevant quotes and remarks from the designer will be included in the results and discussion sections, providing insights into the computer-aided design process. Prior to the interview, informed consent was obtained from the designer, explicitly granting permission for the inclusion of interview data in published materials, alongside permission for the use of the designer's sketches for project purposes.

Results

Table 1 summarizes the configurations used for each training experiment. This section details the outcomes of our experiments, focusing on the outputs from sketch and backdrop generation, as well as the results of our post-processing procedures. Relevant qualitative assessments from the interview with the designer is presented here, with a more comprehensive analysis to follow in the discussion section. The aim of this structured presentation is to evaluate how the implemented AI techniques addressed our research questions and contributed to the project's creative objectives.

Sketch Generation

The fine-tuned Stable Diffusion 2 model demonstrated strong style adaptation, accurately replicating the designer's sketching style, as shown in Figure 2. It also exhibited contextual flexibility, rendering various materials and garment structures at different angles. For instance, it displayed pleats on clothing from both front and side angles, and paired with different garment types. Notably, the model successfully generated garment types not explicitly present in the training data, such as leather, showcasing its ability to adapt to diverse textures. The optimal model performance (as determined by the designer) was observed at checkpoints after 600 and 800 training iterations. Beyond these checkpoints, visual results showed a gradual loss of detail and deteriorated in clarity, indicating overfitting. The designer remarked that "the AI-generated sketches are impressive and closely mirror my style" and that they streamlined the design process. This, according to the designer, is particularly beneficial when working under tight deadlines, emphasising the practical value of this approach to assisting their workflow.

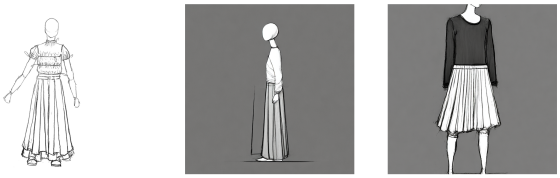


Figure 2: Comparison of the designer's original sketches and AI-generated sketches from the fine-tuned Stable Diffusion 2 model, demonstrating style replication and contextual flexibility, here regarding pleats.

Backdrop Generation

Prior to ControlNet's implementation, images were generated from captioned and uncaptioned training data. Using a fixed prompt, we compared the original *realistic-vision-v51* outputs against those generated from uncaptioned and captioned training. The fine-tuned outputs exhibited a significantly more worn and rustic appearance compared to the original model, aligning more closely with the training data. Notably, the uncaptioned dataset yielded cleaner,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

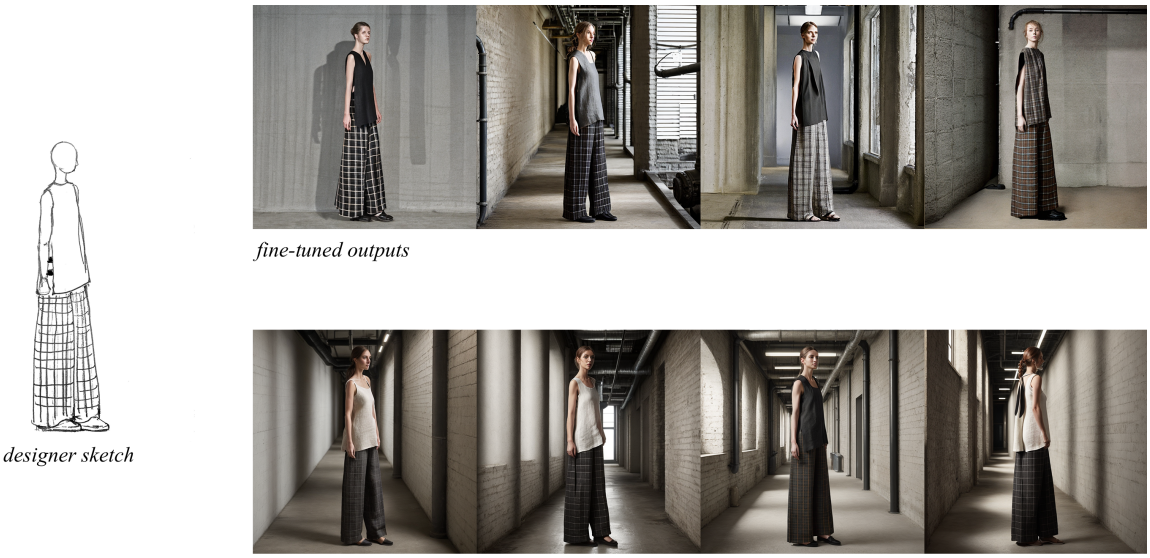


Figure 3: Designer sketch next to photorealistic outputs generated by the original *realistic-vision-v51* model (upper row) versus the fine-tuned model (lower row), showcasing the impact of fine-tuning on backdrop structure, textures, and lighting.

Figure 4: Designer sketch above comparison of original *realistic-vision-v51* (left), captioned training (middle), and uncaptioned training (right) outputs, illustrating the superior cohesion achieved with uncaptioned data.

more prompt-accurate backdrops, effectively capturing visual style without textual guidance. By contrast, captioned training posed challenges in consistently conveying complex visual and spatial information, highlighting the potential advantages of training solely on visual input.

Model Selection & ControlNet Implementation

To achieve our aesthetic goals of photorealism, we conducted a comparative analysis of multiple models before selecting *realistic-vision-v51* for fine-tuning. As shown in Table 1, *realistic-vision-v51* was fine-tuned using our backdrop dataset to achieve desired textures and lighting conditions in our results.

To ensure structural consistency between the designer's sketches and the photorealistic generative versions, we used ControlNet. Edge maps generated from the sketches served as the control input. ControlNet Scribble and Canny were used. Each showcased distinct strengths:

- *Scribble* excelled at handling sketches with less defined features.
- *Canny* was superior for structured edges and complex patterns.

While *Canny* successfully replicated the desired textures and lighting effects, structural differences in the backdrops were less pronounced, possibly due to stricter adherence to the detected edges. However, this subtlety aligned with the designer's preference for understated backdrops in fashion marketing.

Integrated Results: Fine-Tuning *realistic-vision-v51*

Here we provide a detailed account of the results obtained from the fine-tuned *realistic-vision-v51* model, focusing on the comparative analysis of fine-tuned versus standard *realistic-vision-v51* outputs and captioned versus uncaptioned dataset performance.

Fine-Tuned vs. Original *realistic-vision-v51* Outputs To determine the impact of fine-tuning on the model's ability to render desired visual styles within ControlNet, a comparative analysis was conducted between the generated images of the fine-tuned model and the original *realistic-vision-v51* model. The fine-tuning resulted in a noticeable impact on the visual characteristics of the outputs concerning general layout, textures, and lighting (Figure 3). Consistent with the training data, the textures exhibited a more urban and rustic aesthetic. The lighting appeared less diffused and more naturally integrated. Even in the absence of colour specifications in the prompt, the fine-tuned results displayed a greater variety of colours. While the original *realistic-vision-v51* renderings demonstrated slightly more coherent facial structures, this was deemed inconsequential, as the final images were intended to feature the original 2D mannequin aesthetic through post-processing. The designer's qualitative feedback indicated that the garments appeared more natural with the fine-tuned backdrops, and overall image aesthetics were perceived as "interesting and emotionally rich" compared to the original *realistic-vision-v51* results.

Captioned vs. Uncaptioned Dataset Performance To evaluate the influence of captioned versus uncaptioned training data on model performance when used in conjunction with ControlNet, a series of photorealistic renderings were generated using fixed parameters. Although the model fine-tuned on the captioned dataset produced some coherent and high-quality images, it encountered inconsistencies with visual cohesion across multiple runs, even with simple text prompts. The model fine-tuned on uncaptioned images consistently delivered higher-quality and more cohesive renderings, as seen in Figure 4. The uncaptioned model retained structural similarities to the standard *realistic-vision-v51* output while effectively incorporating the learned textures and lighting from the training data. This resulted in a more cohesive aesthetic, whereas the captioned visuals appeared more incoherent, as the model attempted to reimagine visual elements based on textual descriptions during training. The fashion designer emphasised the importance of subtle backgrounds in fashion, stating that "typically, in fashion shoots and visualisations, it's important to keep the background quite subtle to emphasise the garments themselves". Consequently, the uncaptioned fine-tuning proved more aesthetically suitable for AI-driven fashion applications.

Implementation of Negative Prompting for Enhanced Control

Negative prompting, a technique involving the specification of undesirable elements in generative AI systems, was implemented in later stages of this research. This was instrumental in achieving finer control over several critical aspects of the generated images:

- Enhanced control over textures and colours, enabling more precise manipulation of visual attributes.
- Improved control over the backdrop and separation of specific garment sections, particularly in preventing the unwanted projection of patterns onto inappropriate areas when using ControlNet Scribble.
- The creation of a unisex or gender-neutral aesthetic for the generated figures, exemplified by prompting the model to remove gender-specific features.

Post-Processing & Final Image Refinement

In the concluding phase of this research, post-processing procedures were employed to refine the final results. Post-processing involved the superimposition of the original mannequin body parts onto the photorealistically rendered fashion designs. As confirmed by the designer, this achieved renderings in line with the creative vision for the project. We utilised Adobe Photoshop for this purpose. A selection of the final results is shown in Figure 5.

Discussion

This research demonstrates the significant potential of integrating AI as a co-creative mechanism within fashion design processes, particularly through collaborative efforts guided

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Figure 5: Final post-processed images showcasing the integration of AI-generated garments with the original mannequin sketches, achieving the desired aesthetic.

by McKercher's (2020) co-design principles. This collaborative approach between fashion designer and AI technologist aligns closely with our research questions, addressing how AI can enhance and guide creative workflows, facilitate ethical data usage, and support sustainable practices. Using small data and reducing the need for physical prototypes, similar to Zhang, Zhang, and Xie (2024), this approach aims to minimise waste and promote environmentally conscious design practices.

AI as a Creativity Support Tool Central to this research was the investigation of AI's role in augmenting the design workflow by functioning as a creativity support tool, specifically through using fine-tuned Stable Diffusion models with ControlNet for personalised outputs. This process of fine-tuning and spatial conditioning can be understood as "adapting generative spaces" (Abuzuraiq and Pasquier, 2024), allowing greater creative control and highlighting the designer's agency, especially in contrast to using pretrained models. The project's success in producing photorealistic visualisations that closely aligned with the designer's intent illustrates the value of AI in supporting creative ideation and visual experimentation. The designer's feedback underscored this, noting that the AI enabled "quick iterations" and facilitated the exploration of diverse design concepts, streamlining the process of developing and presenting concepts to potential clients. The system's ability to replicate the designer's sketching style and generate high-quality visual renderings demonstrates its efficacy as a creativity support tool within a human artistic process. By alleviating the burden of extensive hand-drawing, the AI empowered the designer to focus on conceptual advancement, a critical factor of meeting demands of high-pressure design timelines. The visual culmination of this co-creative process (Figure 5) demonstrates the success of our co-design framework in producing a visually compelling outcome approved by the designer.

Sustainable Practices Through AI-Driven Design & Small Data This project underscores the value of using small data in AI-driven design. We have shown that AI models, fine-tuned on small datasets, can achieve aesthetic visions aligned with the designer's intent. Thus, we argue that small, personal data is not only a viable, but favourable, approach in design collaborations with AI, especially concerning bespoke creative projects. This finding supports the "Small-Data Mindset" proposed by Vigliensoni, Perry, and Fiebrink (2022), which emphasises that niche and personalised datasets often allow for greater creative control and efficiency in AI systems compared to large, generalised ones.

Furthermore, introducing prototyping in the form of photorealistic visualisations based on small data promotes sustainability in several ways:

- It allows for greater focus on developing ideas and visual representations with audiences. This decreases the need for physical prototyping and thereby reduces material waste.
- By leveraging small data and DreamBooth for fine-tuning instead of relying on big data to provide universal, yet

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

often less creative, solutions, less computational power is required. While more research on the connection between leveraging DreamBooth and sustainability is required, using small data theoretically leads to a reduction in the carbon footprint associated with AI development. This aligns with sustainable practices in the industry (Rombach et al., 2022; Utz and DiPaola, 2023).

These findings suggest that small-data approaches offer a promising pathway towards more sustainable and personalised AI-driven design practices in the fashion industry. However, it is important to acknowledge the complexities of comparing the carbon footprints of traditional physical prototyping and AI model training. Sustainability claims in our study should be viewed as emphasising the *potential* for reduced environmental impact, rather than a definitive assertion.

Ethical Data Usage, Collaborative Innovation, & Designer Empowerment This research highlights the significant value of using personal data within the creative process, particularly for bespoke projects. By concentrating on a compact dataset, derived solely from the researcher and the designer, whose professional work portfolio encompasses both fashion and costume design, the project demonstrates that extensive data repositories are not always necessary to achieve high-quality results. Beyond the advantages regarding sustainability mentioned above, this approach also addresses critical ethical considerations surrounding data obtainment and ownership. The findings advocate for a shift towards ethically sourced, small-scale data within the fashion industry. This positions our approach as a framework for future collaborations in creative sectors.

During the formal interview, the designer expressed concerns that advancements in AI could potentially diminish the need for human designers and their artisanal expertise across the design spectrum: "My primary concern is the potential for AI to replace artisanal jobs within the fashion industry". This is particularly relevant in a fashion market driven by large corporations, which often prioritise economic gains over the value of craftsmanship. Additionally, the designer noted that in the theatre industry, "the role of the costume designer is often overlooked", raising worries that "costume designers may be replaced by AI". This issue is particularly pressing for productions seeking to allocate higher budgets to other components, such as acting and stage design. The designer's concern is further supported by survey findings that highlight how budgetary constraints often disproportionately affect costume departments (Nelson, 2022).

To mitigate this risk, this case study serves as an illustrative example of generative AI not being used as a substitute for human creativity, but as an enabler of it, fostering a synergistic relationship between designer and technology. The designer's notion that "it's crucial to foster a relationship grounded in informed consent and collaboration, ensuring that designers are not exploited and that AI tools are utilised responsibly" suggests that original intent must remain central, ensuring that AI-generated outputs amplify rather than overshadow artistic visions.

Conversely, the designer highlighted how AI could be leveraged by independent fashion designers with limited production budgets. In this context, AI can save costs not only by reducing the need for physical prototypes but also by streamlining other design factors, such as pattern drafting. A good example of AI accentuating human creativity is the designer remarking how imperfections generated by AI can be turned into artistic elements like hand painting the garments, and as such reclaiming the AI-generated images as their own. In fact, they named this the biggest source of newly found inspiration through the project.

Throughout the project, continuous dialogue between the researcher and the designer fostered a reflective practice, ensuring that ethical considerations were woven into every phase of the creative process.

Inclusivity in Design While not the primary focus of the research, the project's exploration of genderless fashion opens up broader discussions about inclusivity in design. By eliminating traditional gender binaries from the design process, this research fosters an environment where diverse identities can be explored and expressed. The insights gained contribute to the ongoing discourse surrounding gender representation in fashion (Marquez-Gallardo and Rovira-Lorente, 2024), emphasising the need for more inclusive design practices that reflect a wider spectrum of identities. The designer noted being pleasantly surprised by the AI's representations of gender in the generated models wearing the garments, even prior to post-processing with the overlaid mannequin aesthetic. This observation reassured them that the initial designs successfully conveyed fluidity of gender.

Nuances of Human-AI Creative Processes The designer noted that while many of the generated designs resonated with their creative vision and were a source of idea generation, some "overshadowed" the original intent (these were not included in any of the post-processed and finished visualisations), occasionally leading to a disconnect between designer and AI. Although this can be beneficial for developing new ideas, it necessitates critical evaluation, as the research underscores the importance of retaining artistic integrity.

Limitations, Future Research & Industry Implications This study is limited in scope due to its single-participant design and qualitative methodology, which, while offering rich insights into one co-design collaboration, does not allow for broad generalisability. The subjective evaluation based on one designer's feedback reflects individual aesthetic preferences and creative priorities, which may not translate directly to other practitioners or design contexts.

Future research should therefore explore the scalability and adaptability of the co-design framework across a wider range of participants, including designers of varying expertise, backgrounds, and stylistic orientations. Studies could test the method within distinct fashion sectors, such as haute couture, streetwear, costume design, or mass-market production, to assess how effectively generative AI supports diverse creative processes.

Additionally, optimising ControlNet's fine-grained control over colours, materials, and details through alternative

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

modules or integrated AI techniques, such as semantic segmentation, would be beneficial. Evaluating how well these improvements preserve the designer's intent across styles would be a valuable contribution.

To further validate the sustainability potential of small data and AI-driven prototyping, future works should conduct comprehensive life cycle assessments, comparing it directly with traditional methods. Further studies should also examine the feasibility of broader industry adoption, including on-demand production systems. Logistical, economic, and environmental impacts must be thoroughly examined.

Finally, further research should address job displacement concerns by exploring AI integration into educational programmes and fostering synergistic collaborations between technologists and designers, ensuring a future where AI and human creativity coexist.

Ethical Considerations

Ethical considerations are central to this research, particularly in relation to data usage, representation, environmental impact, and the role of AI in creative collaboration.

A primary ethical concern addressed in this project is data provenance and consent. By sourcing data exclusively from the designer and researcher, we ensured informed consent and mitigated bias, emphasising ethical data provenance. We critically evaluated AI-generated representations to align with inclusivity and respect for diverse identities, ensuring AI amplified, rather than misrepresented, the designer's vision.

Advocating for small, ethically sourced datasets, we acknowledge the carbon footprint of AI models and encourage future life cycle assessments to compare AI-driven versus traditional prototyping.

We emphasise AI as a creativity support tool and enabler of human creativity, with the co-design framework ensuring the designer's intent remains paramount. Continuous dialogue throughout the project embedded ethical considerations into every phase, contributing to the discourse on AI's impact on the fashion industry.

Concluding Thoughts

This research demonstrates the transformative potential of human-machine collaboration in fashion design, achieved through a co-design approach. By utilising small, ethically sourced data, we produced high-quality outputs, while also promoting sustainable awareness and personalised projects. The study advocates for ethically responsible data utilisation and highlights AI's role in inclusive design by helping dismantle restrictive binaries and enabling broader forms of diverse representation.

AI served as a catalyst for idea generation, enabling rapid prototyping and streamlining the design process, all while reinforcing the indispensable role of human creativity. Continuous dialogue ensured that ethical considerations were embedded in every phase of the AI-aided creative process, safeguarding artistic intent. To advance this work, future research should incorporate diverse designers and styles to further explore the nuances of human-AI collaboration across different design contexts.

Ultimately, this research supports the development of innovative, inclusive, and sustainable practices in fashion design by offering a roadmap for collaborations that harness AI's generative capabilities to support, rather than replace, human creativity.

Acknowledgments

We would like to thank the collaborating fashion designer for their creative input and feedback throughout the project. This work has been supported by the Creative Computing Institute, University of the Arts, London. Gratitude is also extended to the reviewers for their constructive feedback.

References

Abuzuraiq, A. M., and Pasquier, P. 2024. Seizing the means of production: Exploring the landscape of crafting, adapting and navigating generative ai models in the visual arts. *ArXiv* abs/2404.17688.

Baldrati, A.; Morelli, D.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 23336–23345.

Ban, Y.; Wang, R.; Zhou, T.; Cheng, M.; Gong, B.; and Hsieh, C.-J. 2024. Understanding the impact of negative prompts: When and how do they take effect? In *European Conference on Computer Vision*, 190–206. Springer.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

Clarencia, E.; Tiranda, T. G.; Achmad, S.; and Sutoyo, R. 2024. The impact of artificial intelligence in the creative industries: Design and editing. In *2024 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 440–444. IEEE.

Cui, Y. R.; Liu, Q.; Gao, C. Y.; and Su, Z. 2018. Fashiongan: Display your fashion design using conditional generative adversarial nets. *Computer Graphics Forum* 37.

European Commission. 2019. Ethics guidelines for trustworthy ai, shaping europe's digital future. Accessed: 30 September 2024.

Fan, S. 2023. A kernel inquiry into genderless fashion: The social concepts it conveys. *Communications in Humanities Research* 15:67–72.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks.

Guo, Z.; Zhu, Z.; Li, Y.; Cao, S.; Chen, H.; and Wang, G. 2023. Ai assisted fashion design: A review. *IEEE Access* PP:1–1.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of International Conference on Learning Representations (ICLR) 2018*. International Conference on Learning Representations, ICLR ; Conference date: 30-04-2018 Through 03-05-2018.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Marquez-Gallardo, S., and Rovira-Lorente, A. 2024. In genderqueer closets: Challenging gender binarism through embodied narratives of affect and style. *Fashion Theory* 1–24.

McKercher, K. 2020. *Beyond sticky notes: Co-design for real: Mindsets, methods and movements*. Sydney, NSW, Australia: Beyond sticky notes.

Nelson, A. G. 2022. *Creating safe and inclusive costuming spaces*. American University.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. In *Neural Information Processing Systems*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 22500–22510.

Saovana, N., and Khosakitchalert, C. 2024. Assessing the viability of generative ai-created construction scaffolding for deep learning-based image segmentation. In *2024 1st International Conference on Robotics, Engineering, Science, and Technology (RESTCON)*, 38–43.

Thanh-Tung, H., and Tran, T. 2020. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, 1–10. IEEE.

Utz, V., and DiPaola, S. 2023. Climate implications of diffusion-based generative visual ai systems and their mass adoption. In *International Conference on Innovative Computing and Cloud Computing*.

Vigliensoni, G.; Perry, P.; and Fiebrink, R. 2022. A small-data mindset for generative ai creative work. In *Generative AI and HCI - CHI 2022 Workshop*.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; Nair, D.; Paul, S.; Liu, S.; Berman, W.; Xu, Y.; and Wolf, T. 2023. Diffusers: State-of-the-art diffusion models.

Wu, Z.; Ji, D.; Yu, K.; Zeng, X.; Wu, D.; and Shidujaman, M. 2021. Ai creativity and the human-ai co-creation model. In Kurosu, M., ed., *Human-Computer Interaction. Theory, Methods and Tools. HCII 2021. Lecture Notes in Computer Science*, volume 12762 of *Lecture Notes in Computer Science*. Springer, Cham. 171–190.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 3813–3824.

Zhang, Y.; Zhang, T.; and Xie, H. 2024. Texcontrol: Sketch-based two-stage fashion image generation using diffusion model.

Creative as IDEO experts: LLM-agent-based design thinking workshop

Beibei Liu
Future Design School
Harbin Institute of Technology, Shenzhen
Shenzhen, Guangdong, China
24s039066@stu.hit.edu.cn

Shuide Wen
School of Economics and Management
Tsinghua University, Shenzhen
Shenzhen, Guangdong, China
wenshuide@sz.tsinghua.edu.cn

Peiyao Cheng
Future Design School
Harbin Institute of Technology, Shenzhen
Shenzhen, Guangdong, China
chengpeiyao@hit.edu.cn

Abstract

Design thinking is a well-acknowledged process for generating innovative ideas. In the design thinking process, workshops that invite experts from different domains are critical to the fluency and quality of idea generation. Due to participants' varying expertise, their involvement could bring bias, leading to inconsistent workshop results. This research introduces an LLM-agent-based workflow to simulate design thinking workshops. Using IDEO's well-known shopping cart exercise as a benchmark, we dynamically simulate participants' interactions in a "Redesigning Shopping Cart" task by employing multiple LLM agents. We first assigned each agent a specific role, and then followed the Empathize, Define, Ideate, and Prototype phases of the Stanford d.school design thinking methodology. Our findings indicate that the workflow generated both expected and novel ideas, with the final concept surpassing the quality of those developed by IDEO experts across all dimensions, including originality, flexibility, complexity, practicality, and functionality, as assessed by human experts. These results demonstrate the potential of our LLM-agent-based workflow for conducting design thinking workshops and provide preliminary evidence of its utility for analyzing complex creative processes.

Introduction

Designers are motivated to solve complex problems. They go through a structured, iterative process termed design thinking, which includes user research, idea generation, prototyping, and testing (Brown and others 2008). According to the design thinking process, designers need to zoom in on end-users' contexts to empathize with them, resulting in innovative proposals for new designs. Next, designers must define design challenges, ideate new concepts, create prototypes, and test them with end-users. Several design thinking frameworks have been established, such as 3I (ideate, innovate, implement) from IDEO (Brown and Wyatt 2010), the human-centered design model (IDEO 2009), the Double Diamond model developed by the British Design Council (Design Council 2025), and the five-step model proposed by Stanford d.school (Brown 2009). Based on these frameworks, design thinking workshops with cross-functional teams are an effective way to lead the entire process and foster creativity, generating innovative solutions. Cross-functional teams from marketing, engineering, and design participate in design thinking workshops at various stages, from building empathy with users and defining challenges to ideating and selecting final concepts. While the value of these workshops is recognized, there is no standardized approach, and empirical studies yield mixed results regarding their effectiveness. Some studies report improvements in creativity and problem-solving (Paulus and Nijstad 2003), while others show limited or no benefits (Austin et al. 2001; Dubois, Le Masson, and Weil 2024). Research indicates that factors such as participant configuration (including size and diversity), process design (including task variety and sequence), and team dynamics (Austin et al. 2001; Fischer, Östlund, and Peine 2021; Meslec, Graff, and Clark 2020; Stempfle and Badke-Schaub 2002) significantly influence outcomes. However, challenges remain in conducting effective workshops, particularly due to high costs, time demands, and access to top-tier experts.

The emergence of Large Language Models (LLMs) offers a promising solution. LLMs, as generative AI, simulate human cognitive processes like reasoning, decision-making, and collaboration strategies (Tran et al. 2025). They adapt dynamically to changing environments under specific agent architectures (Park et al. 2023), making them ideal for modeling complex scenarios. LLM-agent-based systems are already applied in diverse areas such as legal proceedings (Hamilton 2023), financial decision-making (Seok et al. 2024), scientific research (Su et al. 2024), and market research role-playing (Li et al. 2024).

This research introduces a multi-agent LLM-agent-based workflow to simulate design thinking workshops, using IDEO's shopping cart exercise as a benchmark. The workflow, following the Empathize, Define, Ideate, and Prototype phases of the Stanford d.school methodology (Brown 2009), is powered by a fine-tuned GPT-4 model optimized for the design workshop context, enabling dynamic interactions between agents with diverse roles and

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

expertise, and allowing for a direct comparison with established human-led design workshops. This conceptual workflow offers preliminary evidence of its utility for informing and enhancing complex creative processes.

Related Work

Creativity Facilitation in Human Group Discussions

Human creativity thrives in collaboration, where diverse perspectives drive innovation. Teams generate ideas through individual contributions and group discussions, making workshop performance dependent on both structure and participant performance.

First, regarding the procedure of design thinking workshops, several design thinking frameworks have been proposed, such as the Double Diamond model (Brown 2009) and the Stanford five-step model. While the Stanford model is widely adopted, design researchers often critique it for oversimplifying and calcifying the nuance and flexibility of the design process, its clear structure and well-defined stages provide a tractable framework for our LLM-agent-based simulation, enabling a systematic analysis of the design thinking process. These frameworks not only guide the entire design process but also the procedure of a workshop. According to these frameworks, a design thinking workshop can be divided into a divergence stage and a convergence stage. In the Diverge stage, participants explore ideas freely, using techniques like brainstorming (Paulus, Baruah & Kenworthy, 2023) and the Six Thinking Hats method (De Bono, 2017). In the Converge stage, methods like voting and affinity diagramming refine and prioritize ideas. Extant studies have been conducted by following this general procedure, and thus the effectiveness has been well established.

Next, even with the same workshop procedure, the outcomes still vary because individuals' performance largely determines the outcomes of the workshop. Individual contributions are shaped not only by frameworks and techniques but also by cognitive abilities, personality traits, such as openness and curiosity (McCrae and Costa Jr 1997), and thinking styles, ranging from detail-oriented to holistic approaches (Kirton 2004). These individual differences interact with group dynamics, which include social influences such as team norms, power structures, and leadership styles (Paulus and Nijstad 2003; Xu, Yaacob, and Cao 2024). These interactions affect idea generation and acceptance, as well as communication patterns, trust levels, and collaboration (Paulus and Nijstad 2003; Stempfle and Badke-Schaub 2002). Additionally, both the physical environment (including space, resources, and distractions) and the social context (involving cooperation and competition) play significant roles in shaping contributions (Yin et al. 2023).

Furthermore, group dynamics significantly impact workshop performance. Stroebe et al. (2010) identified three key barriers to creativity: Social Loafing, Social Inhibition, and Production Blocking. Social Loafing occurs when individuals rely too much on others, reducing their contributions (Latané, Williams, and Harkins, 1979). In large group projects or workplaces lacking accountability, some members may put in less effort, assuming others will compensate. Social Inhibition arises from fear of negative evaluation, especially in the presence of dominant members, which suppresses idea-sharing (Diehl and Stroebe, 1987). In brainstorming sessions, less confident participants may hesitate to contribute, leading to fewer diverse ideas and perspectives. Production Blocking occurs when turn-taking delays idea flow, limiting both the quantity and diversity of contributions (Diehl and Stroebe, 1987). During group discussions, waiting for others to finish talking may cause participants to forget or fail to fully develop their ideas, particularly in larger groups where the opportunity to speak is more limited.

In summary, design thinking workshops are effective when guided by a well-structured framework and techniques. Both individual contributions and group dynamics are crucial to the outcomes, yet managing individual differences and steering group dynamics remains challenging. Recent advancements in AI have led to innovative methods for improving group dynamics management, enhancing creativity facilitation.

LLM-Based Multi-Agent Systems for Creativity

With the rapid advancement of artificial intelligence (AI), multi-agent systems, where AI agents collaborate to achieve common goals and solve complex problems (Dohan et al. 2022; Talebirad and Nadiri 2023), are increasingly used to simulate intricate scenarios. Large language models (LLMs) like GPT-4 show particular promise, demonstrating strong capabilities in boosting creativity (Bubeck et al. 2023). Integrating LLMs into multi-agent systems can optimize design thinking workflows by pre-selecting participants based on expertise, reducing group conflicts, and improving both process efficiency and creative output.

In creative tasks, role differentiation is a key aspect of LLM-agent-based multi-agent systems. By assigning tailored roles, such as 'innovative ideator' or 'critical evaluator', that adapt to different stages of the process, the system structures agent contributions and aligns them with workshop needs, optimizing workflow (Zhu and Zhou 2008). This customization enhances collaboration, especially in distributed environments where agents handle complex tasks (Monticolo and Mihaita 2014). Knowledge sharing and negotiation further strengthen collaboration, enabling agents to assess, refine, and combine ideas for more innovative outcomes (Zhou et al. 2016). Negotiation mechanisms also resolve conflicts and align goals, fostering dynamic, efficient teamwork (Beer et al. 1999). Additionally, roles and task allocations are dynamically adjusted based on task progression, with real-time feedback optimizing agents' work modes and enhancing flexibility throughout the creative process (Zhou et al. 2016).

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

LLM-based multi-agent systems have been increasingly applied to simulate human cognitive processes. Studies show their ability to foster innovation and problem-solving through collaborative interactions. For example, Lu et al. (2024) simulated a multi-agent discussion using a three-phase role-playing framework, improving originality and elaboration. Su et al. (2024) simulated scientific idea generation through a multi-agent system modeling team formation, idea refinement, and novelty assessment, outperforming single-agent methods in generating novel and impactful ideas. These findings highlight the potential of LLM-based multi-agent systems to enhance creativity and problem-solving through structured collaboration and diverse role assignments.

Method

This research proposes a controlled, repeatable LLM-agent-based workflow. The workflow follows a four-phase process (Empathize, Define, Ideate, Prototype), based on the Stanford d.school design thinking methodology (Brown 2009), which is also widely adopted by IDEO in their workshops. Each phase is designed to simulate the dynamic interactions observed in human-led workshops, with specialized prompts guiding the LLM agents through each stage. To mitigate human biases and varying expertise (Austin et al. 2001), the LLM agents are assigned specific roles (e.g., marketer, engineer), ensuring diverse perspectives inform the process and drive novel solutions.

The LLM-agent-based simulation uses IDEO's well-documented shopping cart redesign exercise as a benchmark for three reasons. First, IDEO's design thinking approach, aligned with the Stanford d.school methodology (Brown 2009), is a widely recognized framework. Our four-phase workflow mirrors the first four phases, with human expert evaluation substituting the Test phase. Second, the detailed video recordings from the ABC News "Nightline" segment (February 9, 1999) (Gearyinteractive 2011) provide valuable data for comparison, enabling assessment of the generated concept's quality relative to IDEO's established best practices. Third, this well-documented example from a leader in design thinking offers rigorous validation for our LLM-agent-based workflow. Replicating the performance of this team would strongly support the effectiveness of our simulation in design thinking workshops.

Role-Playing

Role-playing is commonly used in design thinking workshops to introduce diverse perspectives and counteract the homogeneity in LLM-agent-based outputs (Ouyang et al. 2022; Padmakumar and He 2023). Drawing from human-centered design methodologies (Camacho 2016), our approach simulates the rich dynamics of human design interactions.

We employ six LLM agents: four participants, a Recorder, and a Decision-Maker. Each agent is assigned a unique role, combining a cognitive approach from Edward de Bono's Six Thinking Hats methodology (De Bono 2017) — Red Hat (emotional/intuitive), Black Hat (cautious/defensive), Yellow Hat (optimistic/positive), or Green Hat (innovative/creative) — with professional expertise (e.g., business, marketing, linguistics, psychology, biology). This mix ensures a diverse range of perspectives on the design challenge. The Recorder agent documents discussions and synthesizes insights, summarizing the problem list in the Define phase and the idea list in the Ideate phase. The Decision-Maker agent, modeled after the strategic role of an IDEO leader like David Kelley, guides the Prototype phase. Detailed prompts specify each agent's role, expertise, personality traits, and participation guidelines, enabling a nuanced simulation of human collaborative design.

Procedure Framework

Our LLM-agent-based workflow employs a four-phase framework to simulate the iterative nature of human design thinking, adapting the five-phase Stanford d.school methodology (Brown 2009) for an agent-based simulation. Unlike a linear process, this framework allows dynamic LLM agent interactions, enabling iterative refinement and a more authentic representation of collaborative creativity. Each phase is guided by specialized prompts that structure agent contributions, balancing divergent exploration — where agents generate and expand ideas — and convergent synthesis — where promising directions are evaluated and refined. This flexible approach ensures a controlled simulation of design thinking while preserving the adaptability essential to creative problem-solving, providing a robust foundation for assessing the workflow's effectiveness in generating high-quality design solutions.

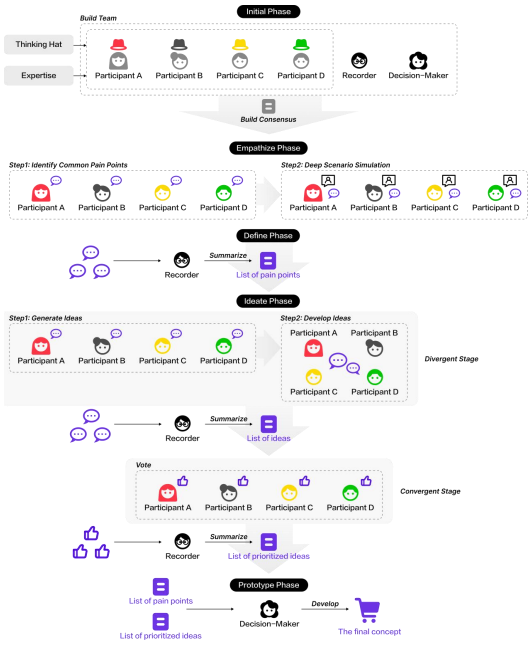


Figure 1: LLM-agent-based Workflow

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Initial Phase This phase establishes the LLM-agent-based team and aligns their understanding of the design challenge. The team, composed of six LLM agents (as described in the Role-Playing section) — four participant agents, a Recorder agent, and a Decision-Maker agent — is formed by assigning each LLM agent a unique role combining thinking styles and professional expertise. A comprehensive introductory prompt is then used to establish consensus regarding the project goals, scope, process, and participation guidelines.

Empathize Phase This phase focuses on establishing a deep understanding of the design challenge and the user needs. Initially, each LLM agent independently analyzes and identifies key issues from their assigned perspectives, ensuring a thorough exploration of both common and critical problems. In the next step, LLM agents adopt stakeholder personas (e.g., a parent with young children, an elderly shopper) to create realistic scenarios that simulate user interactions with shopping carts, focusing on behavioral details and emotional responses to uncover both obvious and nuanced pain points. By combining independent analysis with empathetic role-playing, this approach compensates for the lack of direct user research in our simulation. The findings from both stages collectively inform the subsequent design phases.

Define Phase Building on the insights gathered in the Empathize phase, this phase refines the design challenge into a clear and concise problem statement. The Recorder agent synthesizes the pain points identified by the LLM agents, creating a structured problem definition to guide the following stages. This ensures efficiency and maintains focus on the core design challenge.

Ideate Phase In this core creative phase, LLM-agent-based LLM agents generate a wide range of design concepts for the redesigned shopping cart. The phase begins with a divergent stage, where each LLM agent independently proposes multiple ideas, drawing on the problem definition from the Define phase and applying their assigned expertise and thinking styles to ensure a broad range of diverse and detailed solutions. Next, LLM agents actively develop their own and others' ideas to advance them from new perspectives. Following the idea generation, the focus shifts to a convergent stage. The Recorder agent synthesizes the generated ideas into a structured list, and the LLM agents then vote on them, prioritizing those deemed most feasible and impactful for prototyping. The top-ranked concepts, as determined by the voting process, are then selected for development in the Prototype phase.

Prototype Phase This phase concludes with the creation of a comprehensive design concept for the redesigned shopping cart. The Decision-Maker agent, guided by the prioritized concepts from the Ideate phase and the problem definition from the Define phase, strategically identifies key design directions addressing distinct user needs. These directions then guide the integration of the most promising concepts into a single, cohesive prototype, detailed in terms of form, structure, functionality, usage, and application scenarios, and supplemented by a visual prompt for image generation.

Experiment

Benchmark

This study uses IDEO's 1999 shopping cart redesign, as documented in the ABC News "Nightline" segment (Gearyinteractive 2011), as its benchmark. This case offers a detailed illustration of the design thinking process, providing a strong standard against which to evaluate our LLM-agent-based simulation of the design thinking process, with a particular focus on the final prototype and the process leading to its creation. The IDEO workshop's process, detailed below, highlights key phases comparable to our workflow:

Empathize Phase (User Research): The IDEO team conducted extensive user research, observing shopper behavior in various settings and interviewing diverse stakeholders (shoppers, store employees, maintenance personnel). They employed techniques like character mapping (Winograd 1996) to create detailed user personas, informing their understanding of user needs and pain points.

Define Phase (Problem Definition): Based on their user research, the IDEO team synthesized their findings to define the core problems with existing shopping cart designs, focusing on issues such as safety, theft, and maneuverability. This resulted in a clear problem statement guiding subsequent design efforts.

Ideate Phase (Concept Generation and Refinement): The team engaged in extensive brainstorming, generating a wide range of design concepts. They prioritized idea generation and exploration, deferring judgment. Subsequent stages involved iterative refinement and prioritization through group evaluation and feedback, leading to the selection of core design elements.

Prototype Phase (Prototyping and Testing): IDEO employed rapid prototyping with readily available materials, iteratively refining designs based on user feedback from testing sessions in a real-world supermarket setting. Expert model builders and machinists were then involved in creating a high-fidelity prototype.

The resulting IDEO shopping cart prototype featured several key innovations: a nestable steel frame without a bottom or sides to deter theft; removable plastic baskets for increased shopper flexibility; a dual child seat with a swing-up tray; and steerable rear wheels for improved maneuverability, demonstrating the effectiveness of human-led design thinking. This workshop serves as a benchmark for assessing both the quality of our LLM-agent-based generated concepts and the effectiveness of our simulated workflow.

Experiment Setup

Our experiment was conducted using our innovatively designed LLM-agent-based workflow, which leverages the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

capabilities of the newly developed ChatGPT-4 mini model in conjunction with our simulation architecture. This system directly interfaces with the OpenAI API to utilize the base LLMs, structuring the execution of the workflow and enabling the role-playing mechanism to simulate the design thinking process effectively.

Figure 2: System Interface for the LLM-Agent-Based Simulation

To optimize the workflow for the LLM agents, we tailored specialized prompts for each phase, refining them through extensive pilot testing to ensure the agents could effectively understand and execute their tasks. The core task — redesigning a shopping cart to address user pain points and generate innovative solutions — was undertaken by six LLM agents, each with a specific role (as described in the Role-Playing section). This controlled, repeatable simulation yielded data on agent performance, which forms the basis for our evaluation of the LLM-agent-based workflow's effectiveness.

Evaluation Metrics

We employ a five-dimensional evaluation framework to assess the quality of the generated shopping cart designs, combining elements of Runco's creativity assessment model (Runco 2007) with commonly used design evaluation criteria. This framework considers Originality, Flexibility, Complexity, Practicality, and Functionality, each assessed using a 7-point Likert scale by design experts. This approach allows for a balanced assessment of both the innovative and practical aspects of the designs. The five dimensions are defined as follows:

Originality: The novelty and uniqueness of the design, reflecting a departure from existing solutions.

Flexibility: The design's adaptability to diverse contexts and user needs, demonstrating consideration from multiple perspectives.

Complexity: The sophistication of the design and the seamless integration of its elements into a coherent system.

Practicality: The design's feasibility and practicality, considering manufacturing, costs, materials, maintenance, and sustainability.

Functionality: The design's effectiveness in addressing user needs and enhancing core functionality.

To ensure a professional evaluation, we recruited design experts to assess both the LLM-generated and IDEO design concepts. The experts were provided with textual descriptions (derived from the workflow output for the LLM-generated concept and from the IDEO official website (IDEO 2000) for the IDEO concept) and visual representations (generated using visual prompts) of each design. The visual representations served only as supplementary aids for understanding and were not used as scoring objects, due to the inherent randomness in AI-generated images. Prior to evaluation, the experts underwent a training session to ensure a consistent understanding and application of the evaluation criteria. Each expert provided both quantitative scores and qualitative feedback for each design. Inter-rater reliability was assessed to gauge the consistency of the expert ratings across dimensions.

Results

Outcome Comparison

Our LLM-agent-based workflow generated a comprehensive shopping cart redesign concept, informed by the insights from the preceding phases and prioritized through the voting process. This concept integrated several key features provided by the Decision-Maker agent, addressing user needs and drawbacks identified earlier: Enhanced Shopping Experience (a smart shopping cart with a touchscreen interface and RFID technology provides real-time inventory, pricing, and personalized recommendations), Safety and Load Optimization (a detachable child seat, structural enhancements for stability and load capacity, and a built-in temperature control system), Efficient Space Utilization (a modular design allows for customizable storage space), Sustainability (utilize recyclable and eco-friendly materials).

In contrast to the four features in the corresponding phase identified by David Kelley (IDEO's leader) — shopping, safety, checkout, and finding what you're looking for — our concept, though not an exact match, overlaps in nearly all of these areas. We cover aspects of the shopping experience, child safety, a smart checkout process, and personalized shopping information, which align closely with IDEO's final design priorities.

Building on these design directions, this integrated concept, described in detail below, now serves as the basis for comparison with the IDEO shopping cart design.

LLM-Generated Concept	IDEO Design Concept
Modular Design: Flexible frame structure supports customizable	*Modular Design*: Open steel frame (no bottom or

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

configurations with various-sized storage modules (e.g., baskets) to keep items organized and prevent them from falling. Detachable child seat and insulated module included for added convenience. *Smart Shopping Experience*: RFID-based item recognition and pricing; display screen shows real-time information. Supports adding/removing items via touch to streamline checkout. Supports importing shopping lists and navigating to your wanted items quickly. *Ergonomic Design*: Adjustable height handle ensures comfortable use for various heights. *Maneuverability*: Omnidirectional wheels (360° rotation) provide effortless navigation in tight spaces.	sides) deters theft; two-tiered structure holds five removable baskets for flexible item organization and supermarket branding. *Enhanced Child Safety & Entertainment*: Dual child seats with amusement-park-style safety rails; swing-up tray with cup holder for added convenience and child engagement. *Streamlined Checkout*: Integrated barcode scanner for real-time item scanning, total calculation, and manual adjustments. *Maneuverability*: Uniquely steerable rear wheels combine stability with effortless turning in confined spaces.

Table 1: Comparison Between LLM-Generated Concept and IDEO Design Concept

Figure 3 (Left) : LLM-Generated Shopping Cart Design Concept (Visualized via MidJourney)

Figure 4 (Right) : IDEO Shopping Cart Design Prototype (Official IDEO Source)

To evaluate the quality of the design concepts generated by our LLM-agent-based workflow, we conducted a human expert evaluation. We recruited design experts by sending invitation letters to professors of design at leading universities with strong design programs. Five experts with an average of 14 years of experience in Intelligent Design, Product Design, Human-Computer Interaction, and Design Studies responded and participated in the evaluation. These experts independently assessed both the LLM-generated concept and the concept developed in the IDEO shopping cart redesign workshop. To avoid potential bias, the experts were not informed about the origins of the designs (i.e., whether they were generated by an LLM or by humans). To ensure evaluation consistency, data was collected through a structured questionnaire that clearly defined the evaluation task and provided detailed explanations of each evaluation metric: Originality, Flexibility, Complexity, Practicality, and Functionality (as described in the Evaluation Metrics section). Each concept was rated on a 7-point Likert scale for these dimensions. The scores for each dimension are presented in Table 2.

Dimension	**LLM-Generated Concept**			**IDEO Design Concept**		
	Average Score	ICC	p-value	Average Score	ICC	p-value
Originality	5.2	0.80	0.012	3.8	0.72	0.029
Flexibility	5.4	0.78	0.017	3.8	0.75	0.021
Complexity	4.2	0.74	0.024	3.4	0.69	0.045
Practicality	5.2	0.82	0.009	4.4	0.80	0.012
Functionality	5.0	0.79	0.014	4.6	0.76	0.019

Table 2: Comparison of Expert Ratings for LLM-Generated Concept and IDEO Design Concept. ICC values represent the level of consistency among expert ratings for each dimension. Values above 0.80 indicate excellent agreement, values between 0.70-0.80 indicate good agreement, and values below 0.60 suggest weaker agreement. All p-values were below 0.05, indicating statistically significant results.

As shown in Table 2, the LLM-generated concept received higher average ratings in Originality (5.2), Flexibility (5.4), Practicality (5.2), Functionality (5.0), and Complexity (4.2), while the IDEO design concept scored lower across all dimensions. The Intraclass Correlation Coefficient (ICC) values for both concepts indicate statistically significant expert agreement ($p < 0.05$) across all dimensions. Overall, the LLM-generated concept demonstrated slightly higher consistency in expert evaluations, particularly in Originality, Flexibility, and Practicality.

The evaluation results highlight that the LLM-generated concept outperformed the IDEO design in key dimensions, confirming the effectiveness of the LLM-agent-based workflow. The LLM design scored higher in originality, offering a fresh approach with features like modular components and detachable child seats. Its flexibility was

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

also superior, as the modular structure can adapt to various user needs, from parents to those requiring temperature-sensitive storage. The LLM design's higher complexity score indicates a well-integrated system, combining elements like RFID recognition and real-time pricing. While both designs were practical, the LLM concept was slightly more feasible due to its advanced technology and ease of assembly. In terms of functionality, the LLM design's smart shopping experience and enhanced maneuverability made it a more effective solution.

Although the LLM design received high scores, experts did note that aspects such as cost-effectiveness and age-appropriate design were not fully addressed, reflecting the AI-generated concept's limitations in covering these concrete considerations. Nonetheless, the overall evaluation confirms the LLM-agent-based workflow's ability to generate innovative, practical, and functional designs, validating its effectiveness in simulating design thinking workshops.

Process Comparison

To evaluate the process dynamics of our LLM-agent-based workflow, we conducted a comparative analysis against the IDEO shopping cart redesign workshop. Our experiment randomly assigned four participant agents — Zhang Wei (Black Hat/Marketing), Li Min (Yellow Hat/Psychology), Wang Qiang (Green Hat/Engineering), and Zhao Lin (Red Hat/Linguistics). Importantly, the Empathize phase demonstrated a notable alignment with the IDEO workshop's initial problem definition. For instance, safety and theft concerns were effectively identified by Zhang Wei ("missing child restraints and the risk of carts sliding on inclines"), Li Min and Wang Qiang ("carts are easily stolen or damaged in public areas"), further demonstrating the agents' ability to pinpoint critical issues central to the IDEO study. In addition to these expected findings, the agents also identified several other significant pain points, including inefficient space utilization ("mixed placement of different types of goods" and "space shortages"), stability issues, and poor maneuverability.

The Ideate phase includes two stages: creative divergence and convergence. The divergent stage is split into idea generation, where agents independently generate ideas, and idea development, where they refine and build upon each other's ideas. To analyze idea evolution during the divergent stage, we employed a framework integrating Linkography (Hatcher et al. 2018) and Mixed Initiative Action Vocabularies (Deterding et al. 2017; Grabe et al. 2022; Muller, Weisz, and Geyer 2020). This framework helps to track and understand the connections between ideas, capturing their evolution and interrelationships as they develop through four types: refinement, extension, combination, and decomposition, defined as follows:

Refinement: Improving and fine-tuning an existing idea for better clarity and effectiveness.

Extension: Expanding an idea by adding new elements or dimensions.

Combination: Merging two or more ideas to create a new concept.

Decomposition: Breaking down a complex idea into smaller, manageable parts.

In total, the four participants generated 63 ideas, suggesting an active and productive ideation process. These ideas were analyzed based on the types of connections identified, which included eight types of extension, five types of combination, two types of refinement, and no types of decomposition. The distribution of these connection types highlights a strong emphasis on extension and combination, reflecting a focus on collaborative ideation and building upon existing concepts.

Extension, which accounted for the majority of connections, involved agents expanding on existing ideas by adding new elements. For instance, Li Min's proposal to integrate AR into the smart cart enhanced the shopping cart concept by adding augmented reality features for improved item recognition and navigation. This indicates that agents tend to build upon and adapt existing concepts through collaborative interaction, enhancing the overall depth and complexity of the generated ideas.

Combination, though less frequent, was still notable. Wang Qiang's idea to combine "automatic weighing" with the "smart assistant" merged two concepts — automated weighing and a smart assistant interface — into a unified system. This demonstrates how agents integrate distinct ideas from different domains to create comprehensive solutions, showcasing the potential of creative synthesis.

Refinement appeared when agents fine-tuned existing ideas. For example, Zhao Lin's work on the 'smart cart cleaning system' focused on improving its functionality and usability, reflecting convergent thinking. This approach optimizes and improves prior ideas, rather than introducing entirely new concepts.

Interestingly, decomposition was absent, suggesting that agents preferred to develop ideas in their entirety, enhancing or combining existing concepts, rather than breaking them down into simpler parts. This indicates a more holistic approach to ideation, favoring the evolution of complete ideas.

Another noteworthy point is Zhao Lin, whose contributions were mainly summaries rather than new ideas. This can be attributed to sequential effects, as Zhao Lin consistently spoke last. Research suggests that being the final speaker in ideation can lead to a more reflective role, limiting the generation of new ideas (Paulus and Nijstad 2003). This highlights how an agent's position in the flow influences their performance.

The similarity in roles between Li Min (Yellow Hat/Psychology) and Zhao Lin (Red Hat/Linguistics) likely explains the high overlap in their ideas (eight ideas in total). Both agents focused on psychological and linguistic aspects of the user experience, which led to a convergence in their contributions. This role similarity likely limited the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

diversity of their ideas, supporting research on the importance of role differentiation in creative processes (Zhu and Zhou 2008). In contrast, Zhang Wei (Black Hat/Marketing) and Wang Qiang (Green Hat/Engineering), who had more distinct roles, contributed more innovative and varied ideas, demonstrating how role differentiation fosters creativity by bringing unique perspectives to the table.

Following the divergent stage, the process moved to the convergent stage. The four participant agents were asked to vote for their four favorite ideas, and the results are as follows:

Idea	Votes
Smart shopping cart	4
Modular shopping cart design	3
Shopping cart automatic weighing and pricing system	3
Shopping cart temperature control system	3
Shopping cart automatic sorting system	2
Environmentally friendly materials and sustainable design	1

Table 3: Results of Voting

The voting results reveal that concepts with higher vote counts, like "Smart Shopping Cart" (4 votes) and "Modular Design" (3 votes), align closely with ideas frequently discussed during the divergent phase. This suggests that ideas with more emphasis and refinement earlier are more likely to gain support later. Additionally, the low differentiation in voting among the four agents highlights a consensus around core concepts, reflecting their broad appeal in addressing key design needs such as convenience, flexibility, and product quality. This convergence suggests alignment in agent decision-making processes and the relevance of the selected concepts.

Discussion

This research demonstrates the significant potential of LLM-agents for simulating design thinking workshops, achieving results comparable to those of expert human teams.

Theoretical Contributions

This study introduces a framework for integrating LLM-agents into design thinking, challenging traditional human-centric models. By operationalizing structured methodologies (e.g., Stanford d.school's five-step model) through AI, it bridges AI research and design theory, proposing a hybrid human-AI approach to innovation. Second, it examines how LLM-agents mitigate cognitive biases in human-led workshops, like addressing anchoring effects and domain-specific blind spots. This challenges the assumption that creative problem-solving is exclusive to humans, advancing theories on group dynamics and creativity. Third, empirical evidence shows that AI-generated solutions can surpass human experts in originality, flexibility, complexity, practicality, and functionality. This expands creativity theory by positioning AI as both a collaborator and an independent innovator. Lastly, the study contributes methodologically by demonstrating how LLM-agents enable scalable, reproducible analyses of creativity. By simulating controlled design settings, it offers insights into factors like role diversity and phase sequencing without the constraints of human workshops. While these methodological advancements show promise, it is also important to consider questions that remain regarding the extent to which LLM-driven design thinking can truly challenge established paradigms or simply reproduce existing knowledge, with a critical consideration being the efficacy of LLMs in self-critique, potentially limiting their ability to generate truly novel ideas.

Practical Implications

This research offers concrete insights for improving design and innovation processes. Our LLM-agent-based workflow demonstrates the potential to accelerate early-stage ideation, enabling rapid prototyping and efficient exploration of design directions. It also shows promise in enhancing problem-solving by simulating diverse scenarios, helping teams anticipate risks and develop more resilient strategies. Finally, by incorporating diverse agent perspectives, the workflow has the capacity to promote more inclusive and balanced decision-making. However, realizing these benefits requires careful consideration of potential limitations. The accelerated pace of ideation may come at the cost of neglecting thorough user research and real-world testing. Over-reliance on AI-driven simulations could limit the development of human intuition and critical thinking skills. Furthermore, algorithmic bias inherent in the LLM could inadvertently perpetuate existing societal inequalities. Addressing these limitations, as well as ethical concerns like the potential displacement of human designers and the need for transparency in AI-driven design, requires integrating human expertise and establishing guidelines for responsible and ethical implementation to ensure that these technologies augment, rather than diminish, human creativity and expertise.

Limitations and Future Research

This study has several limitations. The fact that our LLM-agent-based system independently converged on a design concept with similarities to the 1999 IDEO shopping cart—a widely disseminated example in design literature and education—suggests that our study may be influenced by the prevalence of this particular design in existing design knowledge and training, potentially reducing the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

generalizability of our findings. This also highlights the potential for LLMs to reproduce well-known solutions, limiting the exploration of truly novel ideas, especially when trained on readily available datasets. While LLMs produced highly innovative designs, surpassing the IDEO benchmark in several aspects, their bias toward technologically advanced solutions may reduce practicality. The lack of prototyping and user testing further limits real-world applicability. Future research should integrate these evaluations to assess feasibility. Fixed agent roles led to idea redundancy and limited diversity, as predefined frameworks (e.g., Six Thinking Hats) lacked adaptability. Insufficient iterative feedback further constrained idea refinement, hindering transitions to prototyping. Future work should explore dynamic role assignments and iterative feedback to enhance idea development. Future research should refine human-AI collaboration by integrating iterative co-creative systems, where human intuition complements AI insights. Real-time feedback could improve idea relevance and feasibility. Expanding and testing this workflow across varied design challenges will be essential to unlocking its full potential.

Conclusion

This paper introduces an LLM-agent-based workflow for simulating design thinking workshops, benchmarked against the IDEO shopping cart redesign. The results show that the workflow generated innovative concepts, surpassing the benchmark in key areas. This study establishes a foundation for applying LLM-agents to complex creative processes. Future research will refine role differentiation, integrate iterative feedback, and enhance human-AI collaboration to improve effectiveness, adaptability, and real-world applicability.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC), grant number NSFC 72442015. We gratefully acknowledge NSFC's support, which enabled this research.

References

Austin, S.; Steele, J.; Macmillan, S.; Kirby, P.; and Spence, R. 2001. Mapping the conceptual design activity of interdisciplinary teams. *Design studies* 22(3):211–232.

Beer, M.; d'Inverno, M.; Luck, M.; Jennings, N.; Preist, C.; and Schroeder, M. 1999. Negotiation in multi-agent systems. *The Knowledge Engineering Review* 14(3):285 – 289.

Brown, T., et al. 2008. Design thinking. *Harvard business review* 86(6):84.

Brown, T., and Wyatt, J. 2010. Design thinking for social innovation. *Development Outreach* 12(1):29–43.

Brown, T. 2009. Change by design: How design thinking creates new alternatives for business and society. *Collins Business*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*. Camacho, M. 2016. David kelley: From design to design thinking at stanford and ideo. *She Ji: The Journal of Design, Economics, and Innovation* 2(1):88–101.

De Bono, E. 2017. *Six Thinking Hats: The multi-million bestselling guide to running better meetings and making faster decisions*. Penguin uk.

Design Council. 2025. The double diamond. Retrieved from https://www.designcouncil.org.uk/our-resources/thedouble-diamond/. Accessed February 16, 2025.

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*, 628–635.

Diehl, M., and Stroebe, W. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology* 53(3):497.

Dohan, D.; Xu, W.; Lewkowycz, A.; Austin, J.; Bieber, D.; Lopes, R. G.; Wu, Y.; Michalewski, H.; Saurous, R. A.; Sohl-Dickstein, J.; et al. 2022. Language model cascades. *arXiv preprint arXiv:*2207.10342.

Dubois, L.-E.; Le Masson, P.; and Weil, B. 2024. That was fun, now what?: Modelizing knowledge dynamics to explain co-design's shortcomings. *Design Studies* 95:101274.

Fischer, B.; Östlund, B.; and Peine, A. 2021. Design multi-ple: How different configurations of participation matter in design practice. *Design Studies* 74:101016.

Gearyinteractive. 2011. The deep dive. https://www.youtube.com/watch?v=2Dtrkrz0yoU.

Grabe, I.; González-Duque, M.; Risi, S.; and Zhu, J. 2022. Towards a framework for human-ai interaction patterns in co-creative gan applications. In *Joint Proceedings of the ACMIUI Workshops*.

Hamilton, S. 2023. Blind judgement: Agent-based supreme court modelling with gpt. *arXiv preprint arXiv:*2301.05327.

Hatcher, G.; Ion,W.; Maclachlan, R.; Marlow, M.; Simpson, B.; Wilson, N.; and Wodehouse, A. 2018. Using linkography to compare creative methods for group ideation. *Design Studies* 58:127–152.

IDEO. 2000. Reimagining the shopping cart. https://www.ideo.com/journal/reimagining-the-shoppingcart.

IDEO. 2009. Human-centered design toolkit. http://www.ideo.com/work/human-centered-design-toolkit .

Kirton, M. J. 2004. *Adaption-innovation: In the context of diversity and change*. Routledge.

Latané, B.; Williams, K.; and Harkins, S. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of personality and social psychology* 37(6):822.
Li, P.; Castelo, N.; Katona, Z.; and Sarvary, M. 2024. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science* 43(2):254–266.
Lu, L.-C.; Chen, S.-J.; Pai, T.-M.; Yu, C.-H.; Lee, H.-y.; and Sun, S.-H. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and roleplay. *arXiv preprint arXiv:*2405.06373.
McCrae, R. R., and Costa Jr, P. T. 1997. Personality trait structure as a human universal. *American psychologist* 52(5):509.
Meslec, N.; Graff, D.; and Clark, M. A. 2020. Increasing team ideation by sequencing the task type and content. *Design Studies* 70:100947.
Monticolo, D., and Mihaita, S. 2014. A multi agent system to manage ideas during collaborative creativity workshops. *International Journal of Future Computer and Communication* 3(1):66.
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 The Future of Co-Creative Systems-A Workshop on HumanComputer Co-Creativity of the 11th International Conference on Computational Creativity (ICCC 2020)*.
Norman Donald, A. 2013. *The design of everyday things. MIT Press*.
Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35:27730–27744.
Padmakumar, V., and He, H. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:*2309.05196.
Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. *In Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
Paulus, P., and Nijstad, B. 2003. Group creativity: An introduction. *Group creativity: Innovation through collaboration* 3–11.
Paulus, P. B.; Baruah, J.; and Kenworthy, J. 2023. Brainstorming: How to get the best ideas out of the 'group brain' for organizational creativity. In *Handbook of Organizational Creativity*. Elsevier. 373–389.
Runco, M. A. 2007. Creativity: Theories and themes: Research. Development and Practice. Amsterdam: Elsevier.
Seok, S.; Wen, S.; Yang, Q.; Feng, J.; and Yang, W. 2024. Minifed: Integrating llm-based agentic-workflow for simulatingfomcmeeting. *arXiv preprint arXiv:*2410.18012.
Stempfle,J., and Badke-Schaub,P. 2002. Thinking in design teams-an analysis of team communication. *Design Studies* 23(5):473–496.
Stroebe, W.; Nijstad, B. A.; and Rietzschel, E. F. 2010. Beyond productivity loss in brainstorming groups: The evolution of a question. In *Advances in experimental social psychology*, volume 43. Elsevier. 157–203.
Su, H.; Chen, R.; Tang, S.; Zheng, X.; Li, J.; Yin, Z.; Ouyang, W.; and Dong, N. 2024. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *arXiv preprint arXiv:*2410.09403.
Talebirad, Y., and Nadiri, A. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. arXiv preprint arXiv:2306.03314.
Tran, K.-T.; Dao, D.; Nguyen, M.-D.; Pham, Q.-V.; O'Sullivan, B.; and Nguyen, H. D. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:*2501.06322.
Ulrich, K. T., and Eppinger, S. D. 2016. *Product design and development*. McGraw-hill.
Winograd, T. 1996. *Bringing design to software.* ACM.
Xu, S.; Yaacob, Z.; and Cao, D. 2024. How transformational leadership promotes employee creativity within the context of environmental dynamism: a multilevel perspective. *International Journal of Organization Theory & Behavior* 27(2):101–119.
Yin, J.; Pan, Y.; Zhang, Y.; Hu, Y.; and Luo, J. 2023. Distinct inter-brain synchronization patterns during group creativity under threats in cooperative and competitive contexts. *Thinking Skills and Creativity* 49:101366.
Zhou, L.; Yang, P.; Chen, C.; and Gao, Y. 2016. Multiagent reinforcement learning with sparse interactions by negotiation and knowledge transfer. *IEEE transactions on cybernetics* 47(5):1238–1250.
Zhu, H., and Zhou, M. 2008. Role-based multi-agent systems. In *Personalized information retrieval and access: concepts, methods and practices*. IGI Global. 254–285.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Are AI-generated Jokes Truly Original? Charting the "Joke Space"

Ori Amir
Psychology Department
Pomona College
& HaHator LLC
Claremont, CA 91771 USA
ori.amir@pomona.edu

Abstract

This research addresses the challenge of evaluating the originality of jokes generated by Large Language Models (LLMs), which operate as opaque "black boxes" with non-transparent algorithms, potentially relying heavily on training datasets to produce humor. Although LLMs excel at rephrasing, raising concerns about plagiarism, existing studies often assume their outputs are original without verification. A novel framework is proposed to assess joke originality by characterizing the "Joke Space"—the set of all possible verbal jokes in English. Drawing on the General Theory of Verbal Humor (GTVH), this framework defines a joke's "essence" as its script opposition (SO) and logical mechanism (LM), critical for determining similarity and potential plagiarism. The size of the Joke Space is estimated using combinatorial and pragmatic approaches, suggesting that the vast number of possible jokes (at minimum 1 to 32 billion) exceeds the capacity of LLM training datasets, indicating potential for novel outputs. To ensure originality, the framework prompts LLMs with randomly sampled noun pairs to generate jokes, enabling comparative evaluation with human outputs. This approach provides a systematic method to verify the originality of AI-generated humor, offering practical recommendations for researchers.

Introduction

Large Language Models (LLMs) can generate jokes on demand, producing humor that rivals the quality of professional comedy writers, as evidenced by audience reactions in comedy club settings (Toplyn and Amir 2025). However, the originality of LLM-generated jokes remains uncertain (Jentzsch and Kersting 2023; Veale 2024). This gap is a significant limitation in studies of AI-generated humor. While traditional plagiarism detection can identify verbatim matches in a corpus or, when accessible, an LLM's training data, LLMs excel at rephrasing, prompting critics to label them as plagiarism machines (Chomsky, Roberts, and Watumull 2023) or "Stochastic Parrots" (Bender et al. 2021). These challenges highlight the need for a systematic approach to assess whether AI-generated jokes are truly novel, beyond mere rephrasing of existing content. This paper proposes a framework to evaluate the true originality of AI-generated jokes.

The Problem

Unlike rule-based joke generation algorithms, which follow transparent processes (Binsted and Ritchie 1997), LLM-based systems function as opaque "black boxes," obscuring the mechanisms behind their joke creation. Numerous studies investigating the comedic capabilities of LLMs often implicitly assume, without substantiation, that the generated jokes are original. However, when prompted to produce a joke without specific contextual guidance, OpenAI's ChatGPT-3.5 tends to either repeatedly generate identical jokes—suggesting potential hard-coding—or reproduce well-known jokes verbatim or with minimal alterations (Jentzsch and Kersting 2023; Veale 2024).

This limitation may be mitigated by prompting LLMs to craft jokes tailored to highly specific contexts (Toplyn 2023). Nevertheless, verifying originality remains challenging due to restricted access to training datasets and the possibility that jokes may be sufficiently modified to evade detection within those datasets. To address this, the likelihood of a joke's originality can be evaluated by characterizing the "Joke Space," defined as the set of all possible jokes in any form or language, though this study focuses on the subspace of verbal jokes in English within a 5–250 word limit, which I explore to assess the potential for novel outputs.

Joke Plagiarism

A longstanding issue, predating the development of LLMs, is evident in the professional stand-up comedy community, where comedians frequently accuse one another of "stealing jokes" (Wilson 2014; Cain et al. 2024). While an allegedly stolen joke may be reproduced verbatim, it can also be modified in ways that preserve key comedic elements. Consider the following examples:

Joke 1: I used to do drugs. I still do, but I used to, too. (Mitch Hedberg)

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Joke 2: When my husband was a little boy, he *used to* scream and cry every time someone switched the channel when he watched his favorite cartoons. He still does, but he used to, too. (Author)

Jokes may differ entirely in execution, yet comedians may still allege theft of a *premise*—the foundational idea or setup that establishes the context and perspective for a joke or routine (Dean and Allen 2000). For instance, Jokes 3 and 4, presented in the context of stand-up comedy, share the premise that "dating apps are treated like restaurant menus":

Joke 3: You're scrolling through profiles like it's a menu, and you're like, "Nah, not feeling this one, maybe I'll order a better human tomorrow." (Grok3)

Joke 4: You don't have to leave the house anymore to go on a dinner date. I use Uber Eats to get the food and Tinder to get the human. Today, I ordered Italian… yum yum! – Oh, and I also ordered some food. (Author)

While intentional or subconscious joke theft does occur, comedians often independently conceive similar jokes (Wilson 2014). Shahaf, Horvitz, and Mankoff (2015) examined submissions to The New Yorker's cartoon caption contest, where thousands of participants submit humorous captions for a published cartoon. Their analysis revealed that many competitors independently produced similar jokes, often relying on a specific association to humorously connect two primary elements in the cartoon. For example, in a cartoon depicting a cross between a car and a wolf, 1,096 participants—17% of the total—submitted jokes centered on the car being a "hybrid." Although the exact wording varied, enabling the selection of a contest winner, the conceptual similarity was striking. Two of the ~100 jokes referencing "horsepower" illustrate this:

Joke 5: With this baby, who needs horsepower?
Joke 6: Think of it like horsepower, redefined.

Joke Essence vs. Optimization

Determining whether a joke is original or substantially similar to an existing one, beyond verbatim reproduction, requires identifying its "essence"—the unique, critical elements that define its humor. Crafting a joke demands significant skill and iterative refinement. Stand-up comedians often spend months honing a joke, meticulously adjusting phrasing to maximize comedic impact (Seinfeld and Charles 2002; Wilson 2014). While evaluating artificial intelligence's ability to generate humor, I do not dismiss the importance of such fine-tuning or other elements of comedic craftsmanship. However, when assessing a joke's originality, I propose focusing on its "essence"—the core idea or metaphorical framework of the joke. For example, in jokes 5 and 6, the essence lies in the wolf-car context and the "horsepower" reference, while the exact wording represents optimization. This essence aligns closely with the critical components that, if replicated, could lead to accusations of joke theft. To characterize a joke's essence, I explore an empirical method grounded in the General Theory of Verbal Humor (GTVH), later discussing how it aligns with a pragmatic approach based on formulae used by professional comedy writers.

GTVH The theory, in its various iterations (e.g, OSTH; Raskin, Hempelmann, and Taylor 2009), is the most comprehensive and widely cited framework for systematically analyzing humor's elements using a symbolic, rule-based approach. It posits that humor emerges from the interplay of six knowledge resources: script opposition (SO), logical mechanism (LM), situation (SI), target (TA), narrative strategy (NS), and language (LA). These components collaborate to create and *appear to* resolve incongruities that elicit humor (Attardo and Raskin 1991; Attardo, Hempelmann, and di Maio 2002).

Not all knowledge resources, however, contribute equally to judgments of joke similarity, with SO and LM typically playing more pivotal roles (Attardo and Raskin 1991; Ruch, Attardo, and Raskin 1993; Hempelmann 2004). Research by Raskin, Hempelmann, and Taylor (2009) suggests that SO, the clash of incompatible scripts, is the primary driver of humor and thus strongly influences perceptions of joke similarity. Similarly, LM, which resolves the incongruity introduced by SO, significantly shapes similarity judgments, as jokes sharing the same resolution mechanism often feel structurally similar (Attardo, Hempelmann, and di Maio 2002). In contrast, resources like SI, TA, NS, and LA are generally secondary, as they can vary widely without altering the core humorous effect (Hempelmann and Attardo 2011). For instance, Ruch, Attardo, and Raskin (1993) found that jokes with identical SO and LM but different SI or LA were rated as more similar by participants. Collectively, GTVH provides empirical support for defining a joke's essence as its SO and LM, with other elements serving to enhance or support this core (a result replicated with professional stand-up comedians; Cain et al. 2024).

Joke Space

To address the challenge of evaluating the originality of jokes, I propose characterizing the "joke space," defined as the collection of all possible instances of humor. To make this space finite and focused, I will restrict the analysis to English verbal humor within a word limit of 5–250 words, which accommodates most verbal humor (Dunbar, Launay, and Curry 2016). This "joke space" represents a subset of all word combinations of this length that are humorous. This study does not specify a method for determining humor but assumes a human or group of humans will classify each word combination as humorous or not and rate the humorous ones for funniness. I will estimate the size of the joke space (the number of possible jokes) using combinatorics and then propose methods to reduce this space by assessing joke similarity.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Joke Space Size Evaluation

Linguistic Upper Bound We can conceive of verbal "joke space" as a subset of all possible word combinations that are, with few exceptions, grammatically and semantically coherent (with the negligible exception of jokes that rely on syntactic violations and neologisms). The term "semantically coherent" is used broadly here to encompass any combination that elicits meaningful semantic representations in the reader's mind, including novel metaphors and nonsense humor. Such combinations have been shown to activate brain regions associated with semantic processing (Amir et al. 2015).

To characterize the space of possible jokes in the English language, it is necessary to first constrain word counts. Research indicates that popular English stand-up jokes typically range from 5 to 221 words (Dunbar, Launay, and Curry 2016), while one-liners generally consist of one to two concise sentences (Chauvin 2015), with an average of 17.5 words based on a large dataset of short jokes (Moudgil 2017). An average American adult has a vocabulary of approximately 42,000 lemmas (Brysbaert et al. 2016). Thus, the combinatorial upper bound for the shortest 5-word jokes is $42{,}000^5$, or approximately 10^{23} combinations.

Studies on context-free grammars suggest that grammatically correct sentences represent a significantly smaller subset, with estimates ranging from 10^{-3} to 10^{-5} of all possible word combinations (Charniak 1997; Stede 1998; Pereira 2000), resulting in an upper bound of approximately 10^{18} grammatically coherent 5-word sentences. This estimate surpasses the token count in the training sets of even the largest language models by at least a factor of 100,000 (Touvron et al. 2023), a margin substantially amplified when considering unique word sequences rather than individual tokens and further increased by accounting for inflected forms beyond lemmas. This indicates that large language models have considerable potential to generate novel, grammatically correct word sequences, even for short outputs. The 5-word example underscores the vast combinatorial space of possible jokes, though most jokes exceed this length and cannot be condensed to such brevity. For instance, the median short joke length of 17 words (Moudgil 2017) yields a roughly estimated 10^{70} coherent 17-word sequences. The subset of semantically coherent sentences is reduced by an unknown factor, though insights from plagiarism detection research may provide a lower bound for this estimate.

Research on plagiarism detection indicates that n-grams of 5–8 words are generally sufficient to suggest a high probability of verbatim copying (Meuschke and Gipp 2013; Gupta 2016; Foltýnek, Meuschke, and Gipp 2019). Content evaluated for plagiarism is typically human-generated, grammatically and semantically coherent, and meaningful. Consequently, the likelihood that a 5–8-word sentence identical to one found on the public internet is original is low. Indeed, searching for a 5–8-word sentence in quotation marks often retrieves a single, unique document containing that exact phrase. This evidence further supports the notion that, even for short, meaningful sentences, there remains significant potential to generate original outputs distinct from those in training datasets or available on the public internet.

Irony How can we estimate the proportion of humorous texts among all possible meaningful texts?

A trivial case is irony, where any sentence can be transformed into an ironic one by expressing its opposite, or may itself be the ironic version of its negation, making every sentence a potential instance of irony. However, not all statements lend themselves to irony equally; for example, subjective or evaluative sentences like "the weather is bad" can be effectively inverted to "the weather is so great" with clear ironic effect, while factual or neutral statements like "the table has four legs" yield much weaker irony, if they produce irony at all, when inverted to "the table does not have four legs" due to their objective nature (Giora 1995; Wilson and Sperber 2012). Furthermore, irony is often considered less sophisticated, may not qualify as humor under certain definitions, and typically relies on contextual cues beyond the verbal exchange to convey its intent (Ortega 2013).

Theoretical Estimates Multiple theories outline essential conditions for humor, such as script opposition (Raskin 1987) and benign violation (McGraw and Warren 2010). One theory-driven approach involves estimating the frequency of these necessary (and, ideally, sufficient) conditions. However, characterizing the space of all possible jokes predicted by *every* theory is beyond the scope of this paper, as theories often disagree on which conditions are necessary. Compounding this issue, some theories, such as Zillmann's (1983) disparagement theory, have been applied in a confirmatory manner, where advocates perceive a broadly interpreted condition (e.g., disparagement of a target) post hoc in all jokes, despite its vague definition (Amir 2016).

GTVH estimates The General Theory of Verbal Humor (GTVH; Attardo and Raskin 1991) provides a framework for estimating the total number of possible jokes through a combinatorial approach based on its six Knowledge Resources. The potential number of unique jokes can be expressed as the product of the possibilities within each resource:

$$|LA| \times |NS| \times |TA| \times |SI| \times |LM| \times |SO|$$

This calculation results in an exceptionally large number due to the vast variability within each resource. For example, the Target (TA) may include any noun, named individual, or institution, and even complex systems of beliefs, while Language (LA) encompasses all possible word combinations and associated semantic information, reflecting the combinatorial complexity of linguistic structures discussed previously. However, the interdependence of these resources limits the total number of viable combinations; thus, the product is not the appropriate approach to assess the size of the space. The Situation (SI), for instance,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

constrains the applicable subset of Language (LA) to contextually relevant expressions (Raskin, Hempelmann and Taylor 2009). Furthermore, the essence of a joke's humor lies in its Script Opposition (SO) and Logical Mechanism (LM). Consequently, jokes can be conceptualized as clustering around specific SO and LM combinations, with variations in the other resources (LA, NS, TA, SI) modulating the humor's effectiveness to produce local maxima of funniness within clusters. The number of distinct joke clusters can thus be estimated as:

$$|LM| \times |SO|$$

At the most abstract level, Raskin (1985) identified three primary Script Oppositions: good/bad, normal/abnormal, and actual/non-actual. Attardo (2017) further elaborates on mid-abstract and concrete levels of SOs, adding granularity to the framework. Hempelmann and Attardo (2011) propose twelve Logical Mechanisms, each with potential subtypes and concrete instantiations. At this abstract level, the total number of joke clusters is approximately 12 LMs × 3 SOs = 36 distinct joke types. The precise number of LMs depends on the level of abstraction versus granularity, though elaborating a comprehensive ontology exceeds the scope of this study (see Attardo 2001; Hempelmann and Attardo 2011). While the development of a novel Logical Mechanism by artificial intelligence would be a remarkable advancement, it is more likely that humor innovation will occur at the concrete level through specific instantiations of existing LMs and SOs in particular contexts. To estimate, at the concrete level, the number of possible jokes that are sufficiently distinct to be perceived as different, assume the two opposing scripts in an SO can be represented by two nouns (e.g., for Jokes 3-4: "Dating" vs. "Food"), with the concrete instantiation of the LM encapsulated by a single-word association or connector (e.g., "Menu" or "Commodity"). This model aligns with comedy writing guides (e.g., Toplyn 2014), which describe connecting two topic-handles—approximated here as nouns—with an associative bridge, as discussed in the next section.

Comedy Writing-Based Estimates Several humor theorists, including comedy writers such as Dean and Allen (2000) and Toplyn (2014), have proposed variations of a general method for crafting jokes. This method, when simplified, involves selecting two topical elements—referred to as "topic handles"—and identifying an association or conceptual "bridge" that links them in a meaningful or pseudo-meaningful way. For instance, in a joke involving the topic handles "wolf" and "car," the association bridge might be "hybrid." These three components—two topic handles and one bridge—constitute the minimal set forming the core or essence of a joke.

This analysis assumes that any pair of nouns can serve as topic handles for a joke, an assumption supported by the generative model Witscript (Toplyn 2023). Furthermore, it is conservatively estimated that, on average, 10 association bridges exist between any pair of nouns, based on findings reporting over 20 bridges in some cases (Shahaf, Horvitz and Mankoff 2015). Thus, the minimum number of possible unique jokes can be expressed as:

$$N^2 \text{ x } 10/2$$

Where N represents the number of nouns, the division by 2 is because the order of nouns does not matter. For an adult English speaker with a vocabulary of approximately 42,000 lemmas, roughly one-third of which are nouns (approximately 14,000), the calculation yields $14{,}000^2 \text{ x } 5 \approx 1$ billion possible jokes.

When considering the total number of nouns in the English language, estimates vary. Dictionary-based counts, excluding proper nouns, suggest approximately 80,000 nouns (Brysbaert et al., 2016). Alternatively, WordNet 3.0 reports around 117,000 nouns. However, to account for semantic overlap, it may be more appropriate to group synonymous nouns into semantic clusters (synsets), as this reduces redundancy while excluding puns that rely on distinct word forms. WordNet-based estimates indicate 82,115 synsets for English nouns (Kutuzov et al. 2019). Using the lower bound of 80,000 nouns, the calculation $80{,}000^2 \text{ x } 5$ results in approximately 32 billion possible jokes. The estimate of possible noun-based jokes is conservative, as it could be expanded by incorporating a subset of verbs and adjectives suitable as topic handles, along with proper nouns, such as names of celebrities and popular fictional characters. Additionally, it underestimates the total, as many jokes involve setups with relational dynamics between the two topic handles (e.g., a cross between a wolf and a car; a wolf driving a car; or a wolf chasing a car).

Given that the largest LLMs are trained on datasets ranging from 1 to 10 trillion tokens (Touvron et al. 2023), and assuming a concise joke averages 17 words or approximately 20 tokens, the 1 billion possible jokes based on adult vocabulary would require roughly 20 billion tokens (1 billion × 20). While this volume could theoretically fit within the training corpus of modern LLMs, it would constitute 0.2% to 2% of their training data. This fraction may appear small, but given that most LLM training data is non-humorous (Zhao et al. 2023) and humorous content often repeats topics, specific jokes, or essentially equivalent variations (Wilson 2014), the likelihood that such a dataset includes all unique jokes based on noun combinations remains low.
Extending the calculation to all possible nouns (32 billion jokes) would require approximately 640 billion tokens, equivalent to 6.4% to 64% of a typical LLM training set. Consequently, it is reasonable to conclude that no LLM training corpus could encompass all possible jokes. This suggests that, in principle, LLMs could be prompted to generate entirely novel jokes.

Solution: Randomly Sample Joke Space

We have estimated the space of all *possible* jokes, even when considering only the core elements (or "essence") of a joke, to be several orders of magnitude larger than the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

number of jokes to which the largest LLMs or human comedians could have been exposed. Despite this vast potential, both AI systems and human comedians frequently produce non-original jokes. This phenomenon can be attributed to prior exposure to similar jokes—either within an AI's training data or a comedian's personal experience—or to a tendency to focus on highly salient topics (often referred to as "the elephant in the room"; Amir and Biederman 2016; Brawer and Amir 2021; Amir et al. 2022) or topics likely to evoke strong emotional responses, such as those related to sexuality (Hempelmann and Ruch 2005).

Although joke topics in practice are not randomly selected, originality in LLM-generated jokes can be encouraged by prompting the model to create jokes based on randomly sampled topics. This can be achieved by generating pairs of random nouns and instructing the LLM to "craft a joke involving these nouns…" Consider the following example using common English nouns ("apple" and "chair"):

Jokes 7-8:
Q: Why did the apple sit alone on the chair?
A1: Because it couldn't find its *ripe* match! (Grok3)
A2: Because Newton was late that day. (Author)

In the first joke, Grok3's association ("ripe") is general and applicable to any fruit, whereas the author's association ("Newton") is specific to "apple." Jokes that leverage specific associations with the provided nouns are more likely to be original, which is a crucial factor in assessing the likelihood of joke originality. Therefore, rigorous quality control is essential to ensure that jokes generated by LLMs, even when appropriately prompted, maintain specificity to the provided noun pairs. However, simply prompting an LLM to use unique associations may be insufficient; a multi-stage prompting approach, such as incorporating a stage where the LLM generates distinct associations (e.g., Toplyn 2023), may be necessary to enhance originality. Another example uses less common English nouns ("zephyr" and "quixotism"):

Jokes 9-10:
Q: Why did the zephyr whisper to the quixotism?
A1: It wanted to blow some sense into those dreamy ideals! (Grok3)
A2: It was due to pure quixotism (Author)

Jokes generated from randomly sampled noun pairs may not be as humorous as those addressing salient or emotionally charged topics, nor as clever as jokes about topics that elicit multiple associations (Halpern et al., 1994; Toplyn, 2014). The emotional salience and enhanced memory access resulting from numerous associative links likely explain why certain regions of joke space are extensively, and potentially exhaustively, explored. The method I propose allows both LLMs and human comedians to be prompted with identical noun pairs, enabling a comparative evaluation of their outputs (e.g., through independent ratings of funniness). This approach ensures the generation of truly original jokes, addressing a methodological challenge that has previously been overlooked.

Future Directions

Non-English Jokes The number of potential nouns used in the calculations approximates the number of semantic clusters (synsets) in WordNet, suggesting that the estimates should, in principle, be applicable to any language with a comparable semantic reference framework. Given that English texts predominate in the training data of major LLMs (Zhao et al. 2023), it is paradoxically more feasible for these models to generate original jokes in languages other than English.

Formalizing an Originality Score Future research should develop a formula to quantify the likelihood of a joke's originality. The algorithm should identify two or more topic handles, considering their mutual relationships as outlined in the setup. The calculation should incorporate the perplexity of the topic handles, their co-occurrence frequency in texts, and the LM's specificity to these handles and their interrelations.

LLM Decoding The proposed framework aims to evaluate the originality of AI-generated jokes without making claims about LLMs' capabilities or prescribing specific prompting strategies. In principle, LLMs could reproduce any joke within the Joke Space through verbatim prompting (e.g., "repeat the following sentences: …"). To enhance the generation of novel jokes and reduce reliance on memorized jokes from training datasets, decoding strategies such as top-k or top-p sampling, paired with higher temperature settings, can introduce randomness to explore less probable but meaningful outputs. Furthermore, structuring joke generation as a pipeline of subtasks—such as generating unique associations between topic handles (Toplyn 2023)—can minimize the reproduction of canned jokes by encouraging the model to construct humor from novel connections, with the added benefit of diversifying semantic associations. These approaches merit further exploration to support the evaluation of LLM-generated humor originality.

Acknowledgements

I am grateful to Christian F. Hempelmann and Joe Toplyn for their generous and insightful feedback, some of which will inform future work.

References

Amir, O., Biederman, I., Wang, Z. and Xu, X. 2015. Ha Ha! Versus Aha! A Direct Comparison of Humor to Nonhumorous Insight for Determining the Neural Correlates of Mirth. *Cerebral Cortex*, 25(5): 1405-1413.

Amir, O. 2016. The Frog Test: A Tool for Measuring Humor Theories' Validity and Humor Preferences. *Frontiers in Human Neuroscience*, 10: 40.

Amir, O. and Biederman, I. 2016. The Neural Correlates of Humor Creativity. *Frontiers in Human Neuroscience*, 10: 597.

Amir, O., Utterback, K.J., Lee, J., Lee, K.S., Kwon, S., Carroll, D.M. and Papoutsaki, A. 2022. The Elephant in the Room: Attention to Salient Scene Features Increases with Comedic Expertise. *Cognitive Processing*, 23(2): 203-215.

Attardo, S. and Raskin, V. 1991. Script Theory Revis (it) ed: Joke Similarity and Joke Representation Model. *Humor: International Journal of Humor Research,* 4(3-4): 293–347.

Attardo, S., 2001. *Humorous Texts: A Semantic and Pragmatic Analysis* (Vol. 6). Berlin: Walter de Gruyter.

Attardo, S. 2017. The GTVH and Humorous Discourse. In Chlopicki, W., and Brzozowska, D., eds., *Humorous Discourse*, 93–107. Berlin/Boston: De Gruyter Mouton.

Attardo, S., Hempelmann, C. F. and di Maio, S. 2002. Script Oppositions and Logical Mechanisms: Modeling Incongruities and Their Resolutions. *Humor: International Journal of Humor Research,* 15(1): 3–46.

Binsted, K. and Ritchie, G., 1997. Computational Rules for Generating Punning Riddles. *Humor: International Journal of Humor Research,* 10(1): 25–76.

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. 2021, March. 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.

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.

Brysbaert, M., Stevens, M., Mandera, P. and Keuleers, E. 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology*, 7: 1116.

Cain, K., Gimbel, S., Howard, L., Maronna, B. and Beirne, S. 2024. Joke Synonymy Sensitivity Among Working Comedians and the General Theory of Verbal Humor. *Humor: International Journal of Humor Research*, 37(4): 513-528.

Charniak, E. 1997. Statistical Parsing with a Context-Free Grammar and Word Statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*: 598-603. Providence, Rhode Island: AAAI Press.

Chauvin, C. 2015. One-Liners and Linguistics:(Re) Interpretation, Context and Meaning. *Angles: New Perspectives on the Anglophone World*, 1: 1-22.

Chomsky, N., Roberts, I. and Watumull, J. 2023. The False Promise of ChatGPT. *The New York Times*, 8 March. New York: The New York Times Company.

Dean, G. and Allen, S. 2000. *Step by Step to Stand-Up Comedy*. Portsmouth, NH: Heinemann.

Dunbar, R.I., Launay, J. and Curry, O. 2016. The Complexity of Jokes is limited by Cognitive Constraints on Mentalizing. *Human Nature*, 27(2): 130-140.

Foltýnek, T., Meuschke, N. and Gipp, B., 2019. Academic Plagiarism Detection: A Systematic Literature Review. *ACM Computing Surveys (CSUR)*, 52(6): 1-42.

Giora, R. 1995. On irony and negation. *Discourse processes*, 19(2): 239-264.

Gupta, D. 2016. Study on Extrinsic Text Plagiarism Detection Techniques and Tools. *Journal of Engineering Science & Technology Review*, *9*(5): 8.

Halpern, C., Close, D. and Johnson, K. H. 1994. *Truth in Comedy*. Colorado Springs, CO: Meriwether.

Hempelmann, C.F. 2004. Script Opposition and Logical Mechanism in Punning. *Humor: International Journal of Humor Research*, 17(4): 381–392.

Hempelmann, C.F. and Ruch, W. 2005. 3WD meets GTVH: Breaking the Ground for Interdisciplinary Humor Research. *Humor: International Journal of Humor Research*, 18(4): 353-387.

Hempelmann, C.F. and Attardo, S. 2011. Resolutions and their Incongruities: Further Thoughts on Logical Mechanisms. *Humor: International Journal of Humor Research,* 24(2): 125–149.

Jentzsch, S. and Kersting, K. 2023, July. 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*: 325-340. Toronto, Canada: Association for Computational Linguistics.

Moudgil, A 2017. Short Jokes Dataset [Data set]. *Kaggle*. Retrieved May 19, 2025, from https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes

Kutuzov, A., Dorgham, M., Oliynyk, O., Biemann, C. and Panchenko, A., 2019, June. Learning Graph Embeddings from WordNet-based Similarity Measures. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (SEM-2019)*:125-135.

McGraw, A.P. and Warren, C. 2010. Benign Violations: Making Immoral Behavior Funny. *Psychological Science*, 21(8): 1141-1149.

Meuschke, N. and Gipp, B. 2013. State-of-the-Art in Detecting Academic Plagiarism. *International Journal for Educational Integrity*, 9(1): 50-71.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Ortega, M.B.A. 2013. An Approach to Verbal Humor in Interaction. *Procedia-Social and Behavioral Sciences*, 95: 594-603.

Pereira, F. 2000. Formal Grammar and Information Theory: Together Again?. *Phil. Trans. R. Soc. Lond. A*, 358: 1239-1253.

Raskin, V. 1985. *Semantic mechanisms of humor*. Dordrecht, Netherlands: D. Reidel.

Raskin, V. 1987. Linguistic Heuristics of Humor: A Script-Based Semantic Approach. *International Journal of the Sociology of Language*, 65: 11-26.

Ruch, W., Attardo, S. and Raskin, V. 1993. Toward an empirical verification of the general theory of verbal humor. *Humor: International Journal of Humor Research*, 6(2): 123–136.

Seinfeld, J. and Charles, C. 2002. *Comedian* [Documentary]. New York: Bridgnorth Films.

Shahaf, D., Horvitz, E. and Mankoff, R. 2015, August. Inside jokes: Identifying Humorous Cartoon Captions. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1065-1074. Sydney, NSW, Australia: ACM.

Stede, M., 1998. A Generative Perspective on Verb Alternations. *Computational Linguistics*, 24(3): 401-430.

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*. Rye, New York: 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, January. 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.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

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.

Wilson, G.D. 2014. *The Complete Guide to Stand-up: Everything You Need to Know, from Open-Mics to Going Pro!* Minneapolis: The Comedy Institute.

Wilson, D. and Sperber, D. 2012. Explaining Irony. In *Meaning and Relevance*: 123-145. Cambridge: Cambridge University Press.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y. 2023. A Survey of Large Language Models. *arXiv e-prints*: arXiv-2303.

Zillmann, D., 1983. Disparagement humor. In *Handbook of Humor Research: Volume 1: Basic Issues*: 85-107. New York, NY: Springer New York.

Automatic Narrative Knowledge Base Generation

Robert Morain
Computer Science Department
Brigham Young University
Provo, UT 84602
rmorain2@byu.edu

Rafael Pérez y Pérez
Universidad Autónoma Metropolitana,
Cuajimalpa, México
Center for Digital Narrative,
University of Bergen, Norway
rperez@cua.uam.mx

Dan Ventura
Department of Computer Science
Brigham Young University
Provo, UT 84602
ventura@cs.byu.edu

Abstract

Traditional symbolic CC systems like MEXICA often require the creation of handcrafted knowledge bases. In order to advance the development of the MEXICA project, this paper introduces methods for the automatic creation of a knowledge base of short stories. The methods include a series of requests to Deepseek's R1 model to extract relevant structured data from a narrative, using the model to validate and correct the extracted data, and then parsing the structured data and formatting it for the required MEXICA artifacts. This process is validated by evaluating the quality of the extracted narrative data through a human survey. The results show that the process was effective at extracting conceptually accurate structured narrative data from a set of test stories. This work unblocks a significant bottleneck for MEXICA which is necessary for the system to advance to the next level of understanding narrative generation and demonstrates a unique symbiosis between symbolic and generative AI systems.

Introduction

The rapid development of generative AI raises important questions about the future of traditional symbolic computational creativity (CC) systems; for example, how can generative and symbolic AI complement each other by mitigating each other's weaknesses and amplifying each other's strengths (Veale 2024). Although neither symbolic AI nor generative AI alone can yet be considered truly creative, integrating the two paradigms may offer a path toward more advanced CC systems. However, integrating symbolic and generative approaches presents significant challenges as it requires reconciling two fundamentally different modeling paradigms. At the same time, this challenge offers an opportunity to reflect on the core strengths and limitations of each paradigm and to explore new methods for their complementary integration (Pérez y Pérez and Sharples 2023).

As an initial step toward this goal, this paper focuses on the task of automatic knowledge base generation, a critical component of many traditional symbolic CC systems (Ventura 2017). However, building these knowledge bases is difficult and often requires manual effort from domain experts, especially in creative domains.

MEXICA is a symbolic computational model of the creative writing process capable of automatic narrative generation (Pérez y Pérez 1999; Pérez y Pérez and Sharples 2001). A critical area for improvement in MEXICA is its knowledge base, which stores structured information about story events, including the sequence of actions and their associated preconditions and postconditions. Currently, this knowledge base is created manually by experts, a process that is both challenging and time-consuming. As MEXICA has evolved, its increasingly ambitious goals for understanding narrative generation necessitate the development of more powerful knowledge generation methodologies.

In this paper, we introduce methods for automatically creating a knowledge base of structured narrative artifacts (Valls-Vargas, Zhu, and Ontañón 2017; Chambers and Jurafsky 2008) using a pipeline facilitated by Deepseek's R1 model, a state-of-the-art mixture-of-experts model employing chain-of-thought reasoning (DeepSeek-AI et al. 2025). This pipeline automatically parses a story and generates the necessary artifacts for MEXICA using human-engineered prompts and artifact verification. The pipeline extracts actions, preconditions, and postconditions from a story and compiles them into a structured JSON object. This object is then parsed to produce two outputs: a Definition of Previous Stories (DPS) file detailing the event sequence, and a Primitive Action Definition (PAD) file specifying the structure of each action.

The primary goals of the pipeline are to create artifacts that are syntactically correct, conceptually accurate to the plot of the story, and consistent with the MEXICA story description paradigm. The pipeline is evaluated in two ways. First, a human survey consisting of 72 participants evaluates the conceptual accuracy of extracted information from five short story summaries (primarily fairy tales). Second, the generated artifacts for each story are provided as input to MEXICA and corrected if necessary. The number of changes required to conform to MEXICA's rules is tracked to provide a fine-grained measure of how syntactically correct the artifacts are as well as the artifacts fidelity to the MEXICA story description paradigm.

Preliminary results indicate that the pipeline is moderately successful in achieving these goals. In general, this work contributes to the development of more powerful research tools for MEXICA by introducing methods for automatic narrative knowledge base generation and evaluation. It also illustrates how symbolic and generative AI systems

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

can work together in a complementary fashion. All of the code, analysis, prompts, and generated artifacts for this paper can be found on GitHub[1].

Background

The MEXICA story paradigm decomposes narratives into a series of actions focused on the emotional relationships between characters and the development of tension. This sequence of actions is formalized and is stored in a DPS file. Each action is formally defined and is stored in a PAD file. The full details for both of these formal descriptions can be found in (MEXICA).

A DPS file encodes a story as a sequence of actions, each following a ‘subject‘ ‘action‘ ‘object‘ structure to describe interactions between characters. The PAD file defines each action’s preconditions and postconditions, where preconditions specify requirements needed to perform the action and postconditions describe the effects of the action on the story world. MEXICA requires at least one postcondition for every action to build narrative tension. Preconditions and postconditions fall into two main categories: emotional links and tensions. Emotional links model directed relationships between characters with different types and magnitudes of affection, while tensions represent conditions such as life or health being at risk, imprisonment, or their resolution.

For this paper, the LLM used is Deepseek’s R1 70b model, running locally via Ollama. R1 is a state-of-the-art mixture-of-experts model that uses chain-of-thought reasoning. The 70b variant was chosen to balance inference speed, memory efficiency, and model complexity. Although the larger R1 671b variant was tested, it proved too slow for local execution. Although not evaluated in this study, other models could have been used. All experiments were conducted on a single NVIDIA H200 GPU.

Pipeline Design

The primary challenge in processing a story to produce the DPS and PAD files is ensuring the LLM receives enough information to understand and correctly execute the prompt. This is particularly difficult because the MEXICA story description paradigm is nuanced and often challenging even for human annotators to apply consistently across diverse stories. To address this complexity, key concepts from the MEXICA paradigm must be conveyed clearly and in a way that the LLM can reliably interpret. Additionally, it is crucial to limit the amount of new information introduced at each stage to minimize model distraction.

To achieve these design goals, we developed a multi-stage pipeline that processes a story and produces the corresponding DPS and PAD artifacts. The stages are:

1. Identify story actions
2. Identify or infer emotional link preconditions
3. Identify tension preconditions
4. Identify or infer postconditions
5. Verify a JSON object containing the extracted story data
6. Parse JSON to create DPS file
7. Parse JSON to create PAD file

At each stage, carefully engineered prompts—providing background information and task-specific instructions—are used to guide the LLM through producing the required intermediate artifacts within a single conversational thread.

Identify story actions

The first stage identifies the primitive actions performed by the main characters, focusing on emotional relationships and story tensions. A prompt frames the LLM as a narrative analysis expert and provides necessary context about emotional links and tensions. The prompt also includes instructions for how to construct a JSON object containing keys for the ‘action‘ label, the ‘subject‘ character performing the action, and the ‘object‘ character receiving the action directly or indirectly. Also, because the ‘subject‘ and ‘object‘ keys are optional, there is also a key for the number of characters involved in the action. An example is also provided to specify the expected style of the ‘action‘ label. Then, the prompt includes instructions to verify the JSON object to make sure the selected actions accurately reflect the plot of the story. Finally, the story is appended to the end of the prompt.

Identify or infer emotional link preconditions

The second stage identifies or infers the emotional link preconditions associated with each previously extracted action. More details are provided to define what it means to be a precondition and specifying the types and magnitude range of emotional links. Specific instructions are given on how to update the JSON object to include keys for ‘preconditions‘, ‘emotional_links‘, ‘type‘, ‘from‘, and ‘to‘. MEXICA expects emotional links to reference variables ’a’ or ’b’ (depending on the number of characters involved in the action) to specify which character is the source or target of an emotional link. ’a’ refers to the ‘subject‘ character and ’b’ refers to the ‘object‘ character.

Identify tension preconditions

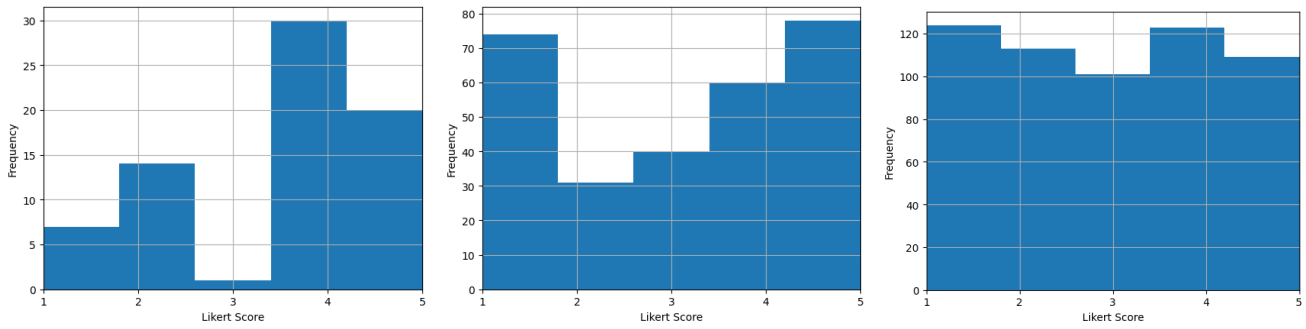
The third stage of the pipeline is responsible for identifying or inferring the tension preconditions for each action. Similar to the previous stage, more details on what tensions are are provided in the prompt along with specific instructions for how to modify the intermediate JSON object from the previous stage.

Identify or infer postconditions

The fourth stage of the pipeline is responsible for identifying or inferring postconditions. Because the conversation history already contains explanations about emotional links and tensions, it is not necessary to explain these concepts again. Instead, postconditions are defined, the complementary types of tensions that are specific to postconditions (life normal, health normal, and prisoner freed) are emphasized and specific instructions on how to update the JSON object are provided.

[1] https://github.com/rmorain/LLMEXICA

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(a) Actions (b) Preconditions (c) Postconditions

Figure 1: Histograms of participant ratings for actions, preconditions, and postconditions across all stories using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). Actions received the highest ratings (M = 3.583, SD = 1.340; median = 4, mode = 4), followed by preconditions (M = 3.13, SD = 1.57; median = 3, mode = 5), and postconditions (M = 2.97, SD = 1.43; median = 3, mode = 1). The overall average across all items was 3.06.

Verify a JSON object containing the extracted structured story data

The verification stage takes place in two steps. First, the LLM is prompted to validate the intermediate JSON object stored within the conversation history. The prompt contains a wide variety of mistakes to avoid and as well as instruction to avoid any other kind of logical error. Next, the JSON is parsed from the LLM response. The parsing process provides additional checks such as for any action that lacks a postcondition, emotional links that are missing a source or a target, or tensions without an affected character. When these errors are detected, the LLM is prompted to regenerate the JSON object to address the error.

Parse JSON to create DPS and PAD files

Finally, the validated JSON object is parsed to generate the DPS and PAD files. This process is straightforward and can be referenced on Github. Notably, the structured data in the JSON could be transformed into a variety of formats for compatibility with other systems. It is important to note that early experiments prompting the LLM to directly generate DPS and PAD files were unreliable, as the outputs often failed to accurately reflect the structured information.

Evaluation

The pipeline is evaluated using a human survey and by directly analyzing the generated artifacts and correcting them to be usable by MEXICA, when necessary.

Survey Methodology

Five stories were selected to test the narrative knowledge base pipeline. Four classic fairy tales, “Goldilocks and the Three Bears" (Steel 1922), “Hansel and Gretel" (Grimm and Grimm 1812), “Jack and the Beanstalk" (Steel 1918), and “Little Red Riding Hood" (Perrault 1923), and one story generated by MEXICA, “Jaguar Knight". The primary goal of the survey is to evaluate how well the identified actions, along with their respective preconditions and postconditions, make sense within the context of the story. To do this, survey participants are asked to do the following:

1. Read a randomly selected story summary
2. Rate how well the main characters’ actions reflect the story plot
3. Review 5–6 actions with their preconditions and postconditions
4. Rate agreement with precondition statements (e.g., “Before Princess heals Jaguar Knight, it is likely that Princess likes Jaguar Knight”)
5. Rate postcondition statements similarly, replacing “Before” with “After”

Questions about the preconditions and postconditions are generated automatically by parsing the JSON object of story actions. The survey was distributed online to non-experts via social media.

Artifact Analysis and Correction

To evaluate the usability of the generated DPS and PAD files with MEXICA, it is necessary to attempt to provide the files as input to the system. If a file has errors it is corrected and then the number of unchanged lines can be used to measure the quality of the original artifact in the following manner:

$$\text{Artifact quality} = \frac{\text{unchanged lines}}{\text{total lines in generated artifact}} \quad (1)$$

This preliminary metric suffers from several limitations. For example, it does not take into account insertions into the corrected file. To account for this, counts for unchanged, inserted, deleted, and moved lines are included in the table.

Results

Survey results Seventy-two participants completed the survey (see Figure 1). Responses were collected using a five-point Likert scale (1 = strongly disagree, 5 = strongly

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Story	% Unchanged	Unchanged	Inserted	Deleted	Moved
Jaguar Knight	100.00%	11	0	0	0
Goldilocks	100.00%	10	0	0	0
Hansel and Gretel	41.18%	7	7	10	0
Jack and the Beanstalk	31.25%	5	11	11	0
Little Red Riding Hood	61.71%	11	7	7	0
Mean unchanged generated lines	61.11%				

Table 1: Results from evaluating generated DPS artifacts using an analysis and correction procedure. The table reports the percentage of unchanged lines as well as the counts of each type of correction made to produce the final artifact. Overall, 61.11% of lines in the generated DPS artifacts were correct.

Story	% Unchanged	Unchanged	Inserted	Deleted	Moved
Jaguar Knight	58.62%	51	15	32	4
Goldilocks	52.08%	25	6	23	0
Hansel and Gretel	26.61%	33	20	86	5
Jack and the Beanstalk	93.75%	75	31	3	2
Little Red Riding Hood	44.71%	38	14	47	0
Mean unchanged generated lines	52.36%				

Table 2: Results from evaluating generated PAD artifacts using an analysis and correction procedure. The table reports the percentage of unchanged lines as well as the counts of each type of correction made to produce the final artifact. Overall, 52.36% of lines in the generated PAD artifacts were correct.

agree). Actions received the highest ratings (M = 3.583, SD = 1.340; median = 4, mode = 4), followed by preconditions (M = 3.13, SD = 1.57; median = 3, mode = 5), and postconditions (M = 2.97, SD = 1.43; median = 3, mode = 1). The overall average across all items was 3.06.

Some stories received higher ratings than others. “Jack and the Beanstalk" got the highest overall score across actions, preconditions, and postconditions (3.47), followed by “Jaguar Knight" (3.41), “Hansel and Gretel" (3.04), “Little Red Riding Hood" (2.64), and “Goldilocks and the Three Bears" (2.39). Note that the pipeline did not identify any preconditions for any actions for “Goldilocks and the Three Bears", “Jack and the Beanstalk", and “Little Red Riding Hood".

Artifact Analysis and Correction Results Table 1 and Table 2 show that across all stories 61.11% of the lines in DPS files are correct and 52.36% of lines in PAD files are correct. “Jaguar Knight" was the story with the most correct lines for both the DPS and PAD files.

Discussion

These preliminary results suggest that the methods presented in this paper provide a foundation for automatic narrative knowledge base creation. While survey results show that participants generally agreed that the identified actions accurately reflect the plot of the story, the extraction of sensible preconditions and postconditions remains an area for further improvement. In particular, there are some instances where the model fails to identify the correct direction of emotional links and tensions. The artifact analysis also reveals other areas for improvement to syntax and consistency with the MEXICA story description framework. For example, PAD files should not have repeated action definitions and DPS files should include relevant character movements throughout the story world.

Currently, these methods require the LLM to generate all preconditions and postconditions for each action in a single step, with verification occurring only at the end. In future work, it may be beneficial to better leverage the strengths of the LLM by breaking down this process on an action-by-action basis. By the same logic, the verification process may be improved by converting the structured information back into natural language and asking the LLM to verify it. It will also be important to include information about the movement of the characters throughout the story world in the DPS and PAD files.

This work marks a first step toward merging the strengths of symbolic and generative AI systems and should spark broader discussions about their associated advantages and limitations. It raises important questions, such as: “To what extent do LLMs truly understand human emotions?” and “Why is it so challenging (for both humans and LLMs) to translate descriptions of emotional relationships into the simple rules defined by the MEXICA paradigm?” A central challenge in the domain of narrative understanding is that stories operate not only through language, but also through layers of emotion and conflict. Because symbolic and generative systems follow different paradigms of story understanding, integrating them in a complementary way offers a promising path forward for deeper narrative understanding and symbolic CC systems.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Acknowledgements

This work was partially funded by the National Council of Humanities, Science and Technology in México, project CF-2023-I-312 and by the Research Council of Norway through its Centers of Excellence Scheme, Project No. 332643.

Prompts

This section presents a selection of the prompts used in the knowledge base generation pipeline. The full set of prompts is available on GitHub.

Listing 1: Prompt to extract story actions

```
You are a narrative analysis expert that
systematically identifies and interprets
actions, preconditions, and effects (called
postconditions) within stories, contributing
 to a structured understanding of a
narrative. You are primarily focused on
actions that relate to the emotional
relationships between characters (called
emotional links) and actions that build
tension within the narrative (called
tensions).

You are focused on specific types of
emotional links. By default, consider two
types of emotional links: `non-romantic` and
 `romantic`. `non-romantic` refers to how
much one character likes another character
in a platonic (non-romatnic) sense. Most
emotional links between characters that are
not in a relationship will be of this type.
The `romantic` emotional link type refers to
 a type of romantic love between the
characters.

You are also focused on specific types of
tensions. By default, consider the following
 tension types:

1. `character_dead`
2. `life_at_risk`
3. `health_at_risk`
4. `prisoner`

`character_dead` means a character has died.
 `life_at_risk` means a character's life is
at risk. `health_at_risk` means a character'
s health is at risk. `prisoner` means a
character is in prison or detained in some
way.
Analyze the given story and extract the
essential actions from the main characters.
Focus on actions that relate to the
emotional links between characters and the
tensions in the narrative.
Keep the `action` values simple and focused
on the abstract action being performed
rather than a specific character performing
the action. For example, if Goldilocks eats
Baby Bear's porridge, label the action as '
eats_porridge' with 'Goldilocks' as the
subject and 'Baby Bear' as the indirect
object rather than labeling the action as '
goldilocks_eats_baby_bear_porridge'.
Organize the actions in chronological order
and in JSON format . The JSON should have an
 `action` key for each action. The value for
 each action should be as simple and general
 as possible so that it can be reused in
other stories, avoid character names, and be
 in Snake_case. Each `action` should have a
key for the number of characters involved in
 the action called `n_characters`, a key
called `subject` for the character
performing the action, and a key called `
object` for the character receiving the
action. If the action lacks a `subject` or `
object` store a value of `-` in the key.
Only include actions where the `subject` and
 `object` refer to characters and not
inanimate objects. For now, only include
these specified keys in the JSON object.
Make sure that `n_characters` is consistent
with the presence of the `subject` and `
object` characters. The same character may
be both the `subject` and the `object`
character if the character is performing an
action on themself.

Once, the JSON object is created verify that
 the actions accurately reflect events
described in the story.

Here is the story:
<STORY>
```

Listing 2: Prompt to extract tension preconditions

```
For each action, identify or infer the
preconditions related to tensions. A
precondition is a requirement that needs to
be satisfied in order for a character to
perform a specific action. These
requirements take the form of either an
emotional link or a tension. A tension that
is a precondition is a tension that should
exist in order for taking the action to make
 sense.

Identify or infer the preconditions that
relate to tensions by following the
instructions below.

Instructions for each action:
1. In the `preconditions` object create a
key called `tensions`
2. Assign an array to the `tensions` key
containing the identified or inferred
tensions.
A. If there are no required `tensions`,
leave the array empty
4. Each `tension` is an object with keys `
type`, `from`, and `to`
A. `type` contains the type of the `tension`
 as defined previously.
B. `from` contains either values 'a', 'b',
'-', '' to indicate the character that is
the source of the tension. 'a' refers to the
 `subject` performing the action, 'b' refers
 to the `object` receiving the action, '-'
refers to no character, and '*' refers to
any character
C. `to` contains either values 'a' or 'b' to
 indicate the character that is the
recipient of the tension. 'a' refers to the
`subject` performing the action, 'b' refers
to the `object` receiving the action, '-'
refers to no character, and '*' refers to
any character

Return this new JSON object.
```


Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

[Chambers and Jurafsky 2008] Chambers, N., and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. In *Proceedings of the Anniversary Meeting of the Association for Computational Linguistics*, 789–797. Association for Computational Linguistics.

[DeepSeek-AI et al. 2025] DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. https://arxiv.org/abs/2501.12948.

[Grimm and Grimm 1812] Grimm, J., and Grimm, W. 1812. Hänsel und gretel. In *Kinder- und Hausmärchen*, volume 1. Berlin: Realschulbuchhandlung. KHM 15.

[Pérez y Pérez and Sharples 2023] Pérez y Pérez, R., and Sharples, M. 2023. *An introduction to narrative generators: how computers create works of fiction*. Oxford University Press.

[Pérez y Pérez 1999] Pérez y Pérez, R. 1999. *MEXICA: a computer model of creativity in writing.* Ph.D. Dissertation, University of Sussex.

[Perrault 1923] Perrault, C. 1923. Little red riding-hood. In *Tales of Past Times Written for Children*. New York: E.P. Dutton and Co. 6–8.

[Pérez y Pérez and Sharples 2001] Pérez y Pérez, R., and Sharples, M. 2001. Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

[Steel 1918] Steel, F. A. 1918. Jack and the beanstalk. In *English Fairy Tales*. London: Macmillan and Co., Limited. 136–153.

[Steel 1922] Steel, F. A. 1922. The story of the three bears. In *English Fairy Tales*. Macmillan and Co.

[Valls-Vargas, Zhu, and Ontañón 2017] Valls-Vargas, J.; Zhu, J.; and Ontañón, S. 2017. Towards automatically extracting story graphs from natural language stories. In *AAAI Workshops*.

[Veale 2024] Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the International Conference on Computational Creativity*, 236–244. Association for Computational Creativity.

[Ventura 2017] Ventura, D. 2017. How to build a CC system. In *Proceedings of the International Conference on Computational Creativity*, 253–260. Association for Computational Creativity.

ISBN: 978-989-54160-7-3

Automatically Detecting Amusing Games in Wordle

Ronaldo Luo, Gary Liang, Cindy Liu, Adam Kabbara,
Minahil Bakhtawar, Kina Kim, Michael Guerzhoy
Division of Engineering Science
University of Toronto
Toronto, Ontario, Canada
{ronaldo.luo, esgary.liang, cindyjy.liu, adam.kabbara, minahil.bakhtawar}@mail.utoronto.ca,
kina19131@gmail.com, guerzhoy@cs.toronto.edu

Abstract

We explore automatically predicting which Wordle games Reddit users find amusing.

We scrape approximately 80k reactions by Reddit users to Wordle games from Reddit, classify the reactions as expressing amusement or not using OpenAI's GPT-3.5 using few-shot prompting, and verify that GPT-3.5's labels roughly correspond to human labels.

We then extract features from Wordle games that can predict user amusement. We demonstrate that the features indeed provide a (weak) signal that predicts user amusement as predicted by GPT-3.5.

Our results indicate that user amusement at Wordle games can be predicted computationally to some extent. We explore which features of the game contribute to user amusement.

We find that user amusement is predictable, indicating a measurable aspect of creativity infused into Wordle games through humor.

Introduction

In this paper, we explore computationally predicting user amusement at games of Wordle.

We scrape user reactions to games of Wordle from the `r/Wordle` subreddit, and use few-shot GPT-3.5 classification in order to rate the reactions to specific games as expressing amusement or not. We verify that GPT-3.5 roughly corresponds to human perceptions of amusement expressions.

We extract features from the game that we predicted would correspond to amusement: we count inherently funny words, as well as indicators of skill/luck. We show that our features weakly predict amusement reaction by users as predicted by GPT-3.5.

To the extent that the humor arises from users engaging with the game creatively, we report a measurement of the amount of creativity (rather than pure problem-solving) in the game (although of course humor can be a legitimate target for problem-solving).

Prior work

(Zaidi and Guerzhoy 2024) show that perception of move brilliance in chess can be predicted using features of the game trees. We focus on amusement rather than perception of brilliance, although those might be related. Rather than rely on annotations in chess ("!!" for "brilliant move"), we must rely on GPT-3.5 assessment of text comments. This work, like us, measures human perception of game moves.

Dataset

We scrape a dataset of 80,000 combinations of Wordle games posted to `r/Wordle` and replies to those Wordle games. The replies often, though not always, are reactions to the Wordle game in the original post.

Five authors manually annotated 15 reactions from the scraped data whether they expressed amusement or not on a scale of 1-5. We then reviewed our separate annotation to converge on a common standard. After, the five authors separately annotated 85 more reactions. We achieved a weak to moderate agreement (see Table 1) on the variable $I(rating > 2)$.

GPT-3.5-turbo was used using few-shot classification to classify 80,000 comments. GPT-3.5-turbo achieved weak-to-moderate agreement with the authors (see Table 1). The prompt is displayed in the Appendix.

Method

Feature: Luck or Skill

For each game, we compute indicators of luck and skill, which we hypothesize are related to amusement, as well as the inherent funniness of words.

The indicators of skill and luck are based on computing the reduction in the number of guesses that are still logically possible after a play and the length of the game. At each game state, one can compute the number of guesses that are consistent with the current state. We compute average/maximum reduction in the number of possible guesses after a play, the magnitude of the last reduction (to 0) in the number of possible guesses after a play, and the maximum/average/last Levenshtein distance and distance in GloVe (Pennington, Socher, and Manning 2014) space between guesses. The motivation for the latter two is that it may be harder for humans to guess a word that is further away from the previous guesses.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

	Rachel	Mason	Clara	George	Alex	GPT
Rachel	1.000000	0.312346	0.431938	0.372142	0.462562	0.398427
Mason	0.312346	1.000000	0.657948	0.828974	0.489270	0.158416
Clara	0.431938	0.657948	1.000000	0.572435	0.562232	0.242574
George	0.372142	0.828974	0.572435	1.000000	0.562232	0.158416
Alex	0.462562	0.489270	0.562232	0.562232	1.000000	0.268354
GPT	0.398427	0.158416	0.242574	0.158416	0.268354	1.000000

Table 1: Cohen's Kappa for (pseudonymous) authors and few-shot GPT. GPT prompt asked "amusing or not", authors 1-5 ratings thresholded at 2. The table shows weak to moderate agreement

Feature: Intrinsically Funny Words

A linear regression model with L2 regularization is trained to predict the "intrinsic funniness" of words. The data used to train the model are 4997 words (4858 words after feature extraction) with humor rating ranging from 27.31 to 100 (Engelthaler and Hills 2018).

To train the model, each word in the dataset is represented by 19 different features: cosine distance to category-defining vector (CDV) of six categories (i.e. sex, party, insult, profanity, body function and animals), the Boolean value of whether the word contains the letter `k`, the log frequency of word, the log average letter probability, the log average phoneme probability of the word, ratio of log average letter probability and log average phoneme probability, the Boolean value of whether the word contains the phoneme `/u/`, the cosine distance to CDV of eight words ending with consonant+`"le"`, the valence (how appealing or undesirable the word makes people feel) $\times$ arousal (amount of emotional activation of word), valence $\times$ arousal $\times$ dominance (how controlling the word's referent is), arousal $\times$ dominance, valence, arousal, and concreteness (how tangible the word is). Those features are shown in (Westbury and Hollis 2019) to have significant linear relationship with funniness.

First, the category defining vectors (CDVs) of six categories (i.e. sex, party, insult, profanity, body function and animals) that are related to funniness are computed (Westbury and Hollis 2019). Those vectors are calculated by first getting 19 words most related to those categories as described in the Appendix A of (Westbury and Hollis 2019) and finding their average word2vec embedding; then we find cosine similarity between words within the dataset and the computed average embedding to find the average embedding of the first 100 words with greatest similarity (Westbury and Hollis 2019).

The CDV of the Cons+`"le"` words are the average of word2vec embeddings of the words gaggle, jiggle, tinkle, waddle, wiggle, wriggle, gobble and nibble. The reasoning for using that is described in (Westbury and Hollis 2019): Cons+"le" words with humor rating above 2 SD that do not carry an evident semantic relation to humor.

The frequency of a word is determined using data from (Tatman 2017). The letter frequency of a word determined using data from (practicalcryptography.com).

The pronunciation of a word is gathered from Carnegie Mellon University (CMU) Pronouncing Dictionary (Carnegie Mellon University 1993) and the frequency of each phoneme is used from data in (CMLOEGCMLUIN 2012).

The valence, arousal, dominance, and concreteness of a word are gathered from computational estimated of (Hollis, Westbury, and Lefsrud 2017).

After training a linear regression model using cross-validation, the root mean squared test error in predicting the funniness of words is 7.6717 (over a range of 72.3 of possible scores) and R^2 of 0.37022, implying we have a reasonable approximation of word funniness. The histogram and scatter plot for the distribution of original and predicted humor rating is displayed below in Fig. 2 and Fig. 1.

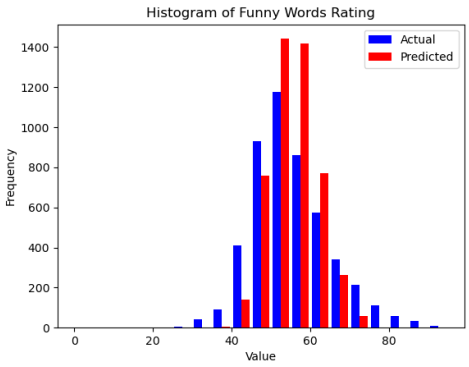


Figure 1: Actual and predicted humor ratings of the 4858 words in the training set

Measure Funniness from Features

Features showing skill or luck in the game may contribute to amusement reaction.

We hypothesize that semantic distance (as measured in GloVe space) and Levenshtein distance between guesses indicates skill or luck; it may be harder to pick a word that is further away in orthographic and semantic space (Liang et al. 2024), and such a word may look more surprising.

Levenshtein Distance Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one word into another (Levenshtein 1966). In the context of gameplay, this metric quantifies the structural similarity between successive player guesses. A smaller Levenshtein distance indicates that a player is making guesses that are more similar

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

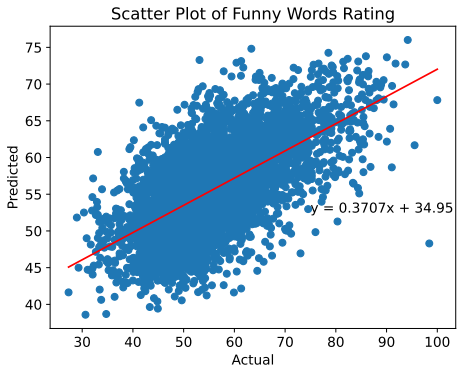


Figure 2: Actual and predicted humor ratings of the 4858 words in the training set.

to their previous attempts, potentially reflecting a bias toward familiar letter combinations or a cognitive tendency to minimize effort by sticking close to prior guesses. A larger Levenshtein distance between moves might imply the user is having fun with different guesses, or is possibly trying to play optimally.

Semantic Distance (GloVe) The semantic distance between guesses is computed using GloVe embeddings (Pennington, Socher, and Manning 2014). GloVe distance is calculated as the negative cosine similarity between these vectors, allowing us to quantify how conceptually similar two guesses are.

A smaller GloVe distance suggests that a player's guesses are semantically related, possibly indicating that the user is not being creative with moves, or possibly that the user is enjoying making semantically-related plays.

Drop in number of possible guesses A drop in the number of possible guesses left may indicate either skill or luck, contributing to amusement.

The following features are computed:

- {Maximum/average/last} $\times$ {drop in number of guesses, GloVe space distance between guesses, funniness of word for guesses} (9 features)
- Drop in number of guesses, GloVe space distance to previous guess, funniness of word for guesses (6×3 features, padded with 0s as necessary)
- Number of moves in the game

Model and Performance

We trained logistic regression, regularized logistic regression, and fully-connected feedforward networks with ReLU activations with the architectures

```
NFEAT-10-10-1
NFEAT-10-1
NFEAT-100-10-1
```

We used a 60-20-20 train/validation/test split, and explored regularization. We subsampled our dataset so that the labels are distributed as 50/50 amused/not-amused (our initial distribution was 75% not amused/25% amused)

We found that neural networks performed worse than logistic regression. We found that adding in features per individual word rather than aggregate features did not change performance on the validation set, but did lead to overfitting, especially with the neural network architectures.

Performance on the test set, for all the settings described above, was $54\% \pm 0.5\%$, using all the settings above.

We therefore report results using logistic regression.

Results

We report results using logistic regression, with only aggregate features used to predict amusement reaction.

We achieve correct classification rate of 54.5% with a baserate of 50%, on a dataset of size 15,000. The result is therefore modest but significant: it is possible to predict amusement reaction from the game.

We normalize the features in order to demonstrate which features are more important for predicting amusement reactions. The results are in Table 2. We observe that GloVe distance does not seem to affect prediction of perception of humor. As predicted, the intrinsic humor of words does affect perception of amusement.

Unsurprisingly, shorter games predict perception of amusement, perhaps because of a show of skill or luck.

Large reductions in the number of possible guesses also correlate with a show of skill or luck and predict perception of amusement.

The effect of the Levenshtein distance between the guesses is interesting: the last two guesses being far apart is predicted to be more amusing, but the aggregate measures of Levenshtein distance seem to lead to less predicted amusement the larger they are. That is perhaps because a last game where the guess is orthographically very different from the previous guess shows skill or luck in a more obvious way.

Note that the significant p-values are generally very small, and would be robust to a Bonferroni correction.

For reference, an increase of one standard deviation from the mean in the Levenshtein distance between the two last guesses increases the probability of an amused reaction by about 4%, and an increase of one standard deviation from the mean in the reduction in the number of possible guesses between the last and the second last guess increases the probability of an amused reaction by about 2%. Those effects are small but statistically significant.

We observe that a univariate logistic regression where the only input is the number of guesses (`num_possible_guesses_length`) results in a correct classification rate that is smaller by only 0.2% from the full model. While the models are different (Chi-squared test p-value < 0.001), the difference is small.

Interestingly, when we run a linear regression, our features predict the number of guesses with an R^2 of 60%, partially explaining our observation – the number of guesses is correlated with the other features.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr($>\|z\|$)**
(Intercept)	0.001321	0.010256	0.129	0.8975
num_possible_guesses_reduction_max	0.017831	0.034160	0.522	0.6017
num_possible_guesses_reduction_mean	0.073592	0.035754	2.058	0.0396 *
num_possible_guesses_reduction_last	0.067377	0.015854	4.250	2.14e-05 ***
levenshtein_distance_max	-0.041227	0.021723	-1.898	0.0577 .
levenshtein_distance_mean	-0.227407	0.033649	-6.758	1.40e-11 ***
levenshtein_distance_last	0.153290	0.024937	6.147	7.89e-10 ***
glove_distance_max	-0.028112	0.032958	-0.853	0.3937
glove_distance_mean	0.021125	0.037529	0.563	0.5735
glove_distance_last	0.006977	0.021547	0.324	0.7461
intrinsic_humor_of_words_max	-0.012670	0.019760	-0.641	0.5214
intrinsic_humor_of_words_mean	0.132882	0.020178	6.586	4.53e-11 ***
intrinsic_humor_of_words_last	0.012456	0.013766	0.905	0.3656
num_possible_guesses_length	-0.081160	0.016345	-4.965	6.85e-07 ***

Table 2: Logistic regression coefficients for predicting amusingness from features, their estimates, standard errors, z-values, and p-values. Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ' 1.

Conclusions

We show that perception of amusement at Wordle games can be predicted, to a very modest extent, from the games.

We note that our rating of comments as expressing amusement is likely very imperfect: while the Cohen's Kappa coefficient is always positive, both the interrater agreement and agreement between the different authors and between the authors and GPT-3.5 is quite modest.

We hypothesize that better agreement on what it means for a comment to express amusement would result in better predictive performance.

References

Carnegie Mellon University. 1993. The Carnegie Mellon Pronouncing Dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. Version 0.7b.

CMLOEGCMLUIN. 2012. Relative frequencies of English phonemes.

Engelthaler, T., and Hills, T. T. 2018. Humor norms for 4,997 english words. *Behavior research methods* 50:1116–1124.

Hollis, G.; Westbury, C.; and Lefsrud, L. 2017. Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology* 70(8):1603–1619.

Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Liang, G.; Kabbara, A.; Liu, C.; Luo, R.; Kim, K.; and Guerzhoy, M. 2024. Semantic, orthographic, and morphological biases in humans' wordle gameplay. *arXiv preprint arXiv:2411.18634*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

practicalcryptography.com. English letter frequencies.

Tatman, R. 2017. English word frequency.

Westbury, C., and Hollis, G. 2019. Wriggly, squiffy, lummox, and boobs: What makes some words funny? *Journal of Experimental Psychology: General* 148(1):97.

Zaidi, K., and Guerzhoy, M. 2024. Predicting user perception of move brilliance in chess. *ICCC 2024*.

Few-shot prompt for GPT-3.5-turbo

```
def classify_funny(comment):
response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    messages=[
        {
            "role": "system",
            "content": (
                "You are a helpful assistant that classifies humor in
                comments. Humor is defined as the extent that the commenter
                is amused by the Wordle game."
                "Rate each comment using a binary convention: '1' if the
                comment is funny and '0' if the comment is unfunny. "
                "Your response should ONLY be '0' or '1'. Do not include
                any additional explanation."
            ),
        },
        {
            "role": "user",
            "content": (
                "I will give you examples of comments and their binary
                humor labels. Use these examples to classify new comments
                using the same convention:\n"
                "1 -> Funny\n"
                "0 -> Unfunny\n\n"
                "Examples:\n"
                f"{examples_text}\n\n"
                "Now, classify the following comment strictly as
                '0' or '1':\n"
                f"Comment: '{comment}'"
            ),
        },
    ],
    max_tokens=2,
    temperature=0
)
```


Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Between codes and dreams: hallucinatory cut-ups as a poetic in creation with AI

Fernando Fogliano
Art Institute
U. Estadual de São Paulo
São Paulo, SP 01140-070 BR
fernandofogliano@gmail.com

Dimitri Lomonaco
Department of Fine Arts
U. de São Paulo
São Paulo, SP 01246-90 BR
dimitrilomonaco@gmail.com

Gisele Aquino
Art Institute
U. Estadual de São Paulo
São Paulo, SP 01140-070 BR
giaquino8@gmail.com

Letícia Nogueira Costa
Pedagogy Institute
Fac. de Educação São Luís
Jaboticabal, SP 14870-810 BR
leticianogueiracosta@gmail.com

Abstract

This text discusses the results obtained from experimenting with Artificial Intelligence technologies as assistive tools for artistic creation, carried out during the course "Artificial Intelligence as a Platform for Artistic Creation," held in the first semester of 2024 at the Institute of Arts of UNESP. Throughout the course, the aim was to discuss the use of artificial intelligence (AI) technologies in the field of Art and Technology, exploring their limits and creative possibilities. The central objective of the classes was to consider the use of AI, especially Large Language Models (LLM), as intelligent tools that can assist artists in the creative process. After the course period, an Artificial Intelligence study group was established so that these reflections could be further explored. Presented here is the result of an experiment conducted during the course, its conceptual proposal, and reflections arising from that practice.

Introduction

The objective of the proposed exercise was to explore the limits and possibilities of generative AI, considering that all such systems, including the Large Language Models (LLMs) that underpin chatbots, invent things, or hallucinate (Jones, 2025). According to Menkman (2011), it is important to recognize the inherent imperfection of media. However, in the realm of Glitch Art, such imperfections offer opportunities for creative exploration. In her manifesto, the author considers it necessary to use the flaw "as an exoskeleton for progress," an opportunity for catharsis, thereby criticizing technologies and their inherent policies through disintegration, ruptures, and cracks (p.11).

From a Glitch Art perspective, the experiment involved feeding the ChatGPT prompt with text designed to break from the best practices of prompt engineering and logic. The chosen alternative was the use of cut-up techniques. Thus, using prompts without a rational narrative to underpin them would introduce unforeseen input, provoking unexpected responses, or glitches in the machine. Exploring the operational limits of the device is a concept present in Vilém Flusser's (1985) book *Filosofia da Caixa Preta*, where he argues that the photographer struggles to push the technology beyond the limits set by the camera's programming (Fogliano, 2013). According to the philosopher, the photographer seeks to obtain unforeseen images from the machine and, to this end, explores the "programming of the apparatus" intending to reach "regions never before navigated, to produce images never before seen... And wants to discover them inside the apparatus" (Flusser, 1985, p.19).

We argue, however, that the images are not inside the device as potential outcomes anticipated by the designer, but rather reside in the artist's mind (Fogliano, 2013). The function of the technological apparatus, through its design or programming, is to make the abstract content contained within the images and narratives in the artist's mind concrete, communicable, or perceptible. In this sense, this resonates with the ideas of Simondon (1980), who considers the openness of technical objects, situating them within the realm of emergence, creativity, unpredictability, and inexhaustible innovation. As open systems, technical objects are inherently novelty-generating.

The experiment

Based on the considerations above, the question arose regarding the possibility of drawing an analogy between LLMs equipped with image generation tools and other semiotic devices, such as photographic or video cameras, with the aim of including them in the creative toolkit for Art and Technology. It is important to consider that the experiment described here was developed using a methodology specific to the artistic realm. Our objective in experimenting with artificial intelligence devices was to push the logical limits of programming with the aim of inducing error, deviation, and hallucination. In this context, failures take on the status of a creative event. This approach is situated within a perspective in which technology acquires a more complex ontology than that of

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

a mere tool, becoming a semiotic machine capable of operating in the field of symbolic possibilities. When the artist subverts the intended functioning and provokes error—whether through unconventional inputs or by accepting failure as form—they surpass the game proposed by Flusser and enter the territory that Gilbert Simondon (2005) names functional superabundance: the machine's capacity to exceed the purposes for which it was designed. Glitch Art, in this context, serves as a paradigm: it embraces dysfunctionality as a poetic force and suggests that technical error can be converted into an expressive sign. In the artistic-experimental use of AI, hallucination, seen as a flaw under the paradigm of rationality, becomes an aesthetic opening, evidence that meaning can emerge from the heuristic relationship between artist and machine. The proposed experiment did not aim to confirm hypotheses or ensure the reproducibility of results, but, as an artistic practice, it sought to dislodge error from its negative place, discovering its creative potential and stretching the expressive potential of LLMs.

To enable exploration of the limits of LLMs and make room for that which differentiates us from the machine—creativity, intuition, the unconscious—the choice was made to employ Glitch Art, using prompts constructed with the aid of cut-up techniques.

This is an artistic technique rooted in the avant-garde of the early 20th century, in which text is physically cut up and reorganized to generate unexpected meanings. Its conceptual seeds were planted by the Dadaists in the 1920s. Tristan Tzara, for example, instructed poets to randomly pull words from a newspaper to create verses, subverting traditional authorship. Recognizing its potential, the American writer William S. Burroughs, in the 1950s, became an eminent literary representative of this compositional technique. Filmmaker Antony Balch collaborated with Burroughs and Brion Gysin, painter and writer, to explore cut-ups in films, while musicians, like David Bowie, adopted the technique for song lyrics in the 1970s. Although rooted in analog practices, the method anticipated today's remix culture and digital randomness. The cut-up is present in postmodern literature, music, visual arts, and even in AI-generated texts.

For Burroughs, the cut-up wasn't merely a technique; it was a method where random juxtapositions revealed manifestations of the unconscious. Both collage and cut-up invite accident as a collaborator, transforming the chaotic into the poetic and found fragments into revelations, creating meaning not through control, but through serendipity.

Considering that the operational domain of both the artist and LLMs is language, interest arose in exploring the output of these technologies within the context of image production that could bring forth content from the unconscious as language: symbolic, dreamlike, distant from mimesis and objectivity (de Castro, 2009). Our choices were influenced by collages such as Hannah Höch's feminist work "Cut with the Kitchen Knife" (1919) and Raoul Hausman's socially critical and emblematic production "ABCD" (1923-1924). The proposal involved pushing the logical boundaries of the programming, subverting it by requesting the machine process prompts phrased in the specific language of the unconscious, aiming to induce hallucinations in the AI within the aesthetic framework of Glitch Art.

To ensure the objectives were met, each student in the course was asked to bring a cherished book on the scheduled date, from which the base words would be obtained. Furthermore, to accentuate the randomness of the experiment, one student used three different dice. When rolled, these would respectively indicate: the page number from which to extract the word, the line number on that page, and finally, the position of the selected word within that line. It was then the student's task to follow these instructions, counting words to arrive at the one designated for use in constructing the prompt.

With ten students and the professor present, the selection process began. The same procedure was executed eleven times, until all participants had the eleven words in hand. All words selected in the books happened to be in Brazilian Portuguese language. These were the selected words (with one possible English translation): "Rabino (Rabbi)", "vez (time)"; "certos (certain)"; "mídias (media)"; " norte (north)"; "ninguém (nobody)"; "camarão (shrimp)"; "simples (simple)"; "rejeição (rejection)"; "caçaba (not directly translatable)"; "opor (to oppose)". Each participant individually arranged the words according to their intended meaning, thus drafting their own prompt. The prompt had

Figure 1 - *Prompt: "Create an image based on the following text: Time certain media north rabbi nobody shrimp simple rejection caçaba oppose." (Translated from Portuguese)*

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Figure 2 - *Prompt: "Create an image based on the following text: Certain rabbis from the north will occasionally oppose nobody out of simple rejection of the caçaba and the shrimp in the northern media." (Translated from Portuguese)*

to contain all the selected words, and participants were forbidden from inserting any other words, even for the purpose of connecting or giving sense to the obtained word set.

Another student volunteered to access the GPT-4 model (DALL-E 3) and input the prompts, requesting the machine generate images in response. For example: "Create an image based on the following text: Time certain media north rabbi nobody shrimp simple rejection caçaba oppose" (translated from Portuguese).

After generating the third image, it was observed that using the same chat window to generate subsequent images resulted in outputs maintaining a consistent image style. To overcome this limitation, the group began using a new chat window for each prompt. Subsequently, the words were rearranged within the prompts, this time permitting the insertion of other words—such as connecting words (e.g., 'of', 'with', 'in')—to help construct a short narrative, while attempting not to significantly alter the core sense derived from the original word set. Now the prompt "*Create an image based on the following text: Time certain media north rabbi nobody shrimp simple rejection caçaba oppose*" (translated from Portuguese) evolved into "*Create an image based on the following text: The [story] this time is that certain media say that, when opposing the rabbis of the north, nobody presents a rejection of a simple shrimp in the caçaba*" (translated from Portuguese). The AI's response differed substantially from the one obtained with the first prompt, a finding that was also noted in the other instances.

Figure 3 - *Prompt: "Create an image based on the following text with a Dadaist montage style: certain media reject the caçaba from the north of the shrimp rabbi." (Translated from Portuguese)*

In the next step, the group proceeded to intentionally select the style to be used by the AI for each image—for instance, "expressionist style," "in the style of artist Adriana Varejão's sculpture technique," or "surrealist style with paper cutouts"—representing a move towards increasing complexity. The translation of a prompt into a foreign language (English) was also tested. For example: "*Create an image based on the following text in a Dadaist montage style: certain media reject the hunt in the north of the rabbi shrimp*" (original text prompt).

Results

The AI's response made it possible to observe several biases. Notable among these were the characteristics attributed to the rabbi (religious bias), the representation of media (with a prevalence of digital and TV media over other forms), and even social stereotypes—such as poverty, the predominance of television, juvenile delinquency, etc.—when Brazil was mentioned in the prompt. Furthermore, the absence of certain concepts or words from the AI's vocabulary—often because they lack a direct equivalent or common usage in English—led to gaps in the image content. For example, the program ignored or misinterpreted the concept "*caçaba*" (coarse cassava flour), which was either replaced by another symbol or omitted entirely. The conclusion drawn is that this points to a colonial bias in the machine's learning process, as it overlooks culturally specific or regionally known items like the "*caçaba*".

Concluding Remarks

Based on the results obtained, the viability of exploring LLMs for artistic purposes is evident, positioning them as tools that enable artists to expand their creative horizons.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Figure 4 - *Prompt: "Create an image based on the following tex considering Brazil as the setting:* Rejection of the shrimp: opposing certain media to a simple rabbi." (Translated from Portuguese)

Even if the focus of the present work is aesthetic experimentation with artificial intelligence, we cannot ignore the ethical, political, and social implications inherent in the technologies with which we work. The machines we use—AI algorithms, neural networks, language models—are not neutral; they operate based on data and patterns constructed within historical contexts marked by racial, gender, and class inequalities.

Jacques Rancière (2004) offers us a key to understanding the political role of art, not as a bearer of explicit ideological content, but as a reconfigurer of the sensible. In this sense, by appropriating AI to explore its errors, glitches, and hallucinations, the artist intervenes in the functional regime of technology, producing other possibilities of sensibility and thought. Failing to do so may, in turn, ultimately lead to the "symbolic emptying of artistic practice with AI" (Ferro, 2025). Despite verifying the creative and collaborative potential in using these tools, it is imperative to address the ethical issues stemming from the choices made when utilizing such models. Ethics, in this context, must be viewed as an essential part of technological practice, considering how the accumulation of social, economic, and environmental risks associated with these technologies intersects with human vulnerabilities (Kaufman, 2024).

References

de Castro, J. C. L. (2009). O inconsciente como linguagem: De Freud a Lacan. *CASA: Cadernos de Semiótica Aplicada, 7*(1), 1–12.

Ferro, I. (2025, April 22). *Studio Ghibli: A obra de arte na era da IA - Outras Palavras*. Outras Palavras. https://outraspalavras.net/outrasmidias/studio-ghibli-a-obra-de-arte-na-era-da-ia/

Flusser, V. (1985). *Filosofia da caixa preta*. São Paulo, Brasil: Hucitec.

Fogliano, F. (2013). Cultura e tecnologia: A automação nos processos criativos das narrativas. *Revista da Anpoll, 1*(35), 77–98.

Fogliano, F., & Leote, R. (2023, November). Artificial intelligence as a support for artistic creation. In *Proceedings of the 11th International Conference on Digital and Interactive Arts* (pp. 1–9).

Fogliano, F., & Camargo, D. (2011, September). Revisitando o "instante decisivo", uma abordagem da neuroestética. *Intercom – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação*. https://www.intercom.org.br/papers/nacionais/2011/resumos/R6-2741-1.pdf

Jones, N. (2025). AI hallucinations can't be stopped—but these techniques can limit their damage. *Nature, 637*(8047), 778–780.

Kaufman, D. (2023). [Resenha de Ética na inteligência artificial de Mark Coeckelbergh]. *TECCOGS – Revista Digital de Tecnologias Cognitivas*, *28*, 151-155.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.

Menkman, R. (2011). *The glitch moment(um)* (Network Notebooks 4). Institute of Network Cultures.

Ranciere, J.(2009). *A partilha dop sensível.* Editora 34.

Simondon, G. (2017). *On the mode of existence of technical objects*. Univocal Publishing.

Brave: Designing an Embedded Network-Bending Instrument, Manifesting Output Diversity in Neural Audio Systems

Daniel Manz and **Mick Grierson**
Creative Computing Institute
University of Arts London
United Kingdom
d.manz@arts.ac.uk
m.grierson@arts.ac.uk

Abstract

As neural audio synthesis becomes more widely adopted there is a growing risk that its limitations could impact the content, quality and diversity of music. Some musicians, artists, and researchers perceive an increased risk of cultural homogenisation and qualitative degeneration due to poor-quality training data and parameterisation. This work seeks to explore new methods for addressing these challenges by contributing to the developing field of "network-bending". Network-bending employs direct manipulation of internal ML architectures to enable active divergence from the training corpus, increasing the statistical variability and capability of model outputs. We present "Brave": an embedded, network-bending hardware instrument, which can provide a novel blueprint for embedding a network-bending system on a stand-alone system. Through a process of iterative musician-led feedback, drawing on Proof-of-Concept Media and Arts Technology approaches, this work seeks to stimulate further interest in network-bending frameworks applied in the field of AI-driven sound synthesis.

Introduction

Generative AI has pushed the boundaries of artistic creation - It can automate laborious aspects of the artistic process, act as an initial idea generator, and democratise expression. However, this technology bears a critical potential downside, in that it could homogenise the cultural landscape, simplifying the process of generating statistically average output. Therefore, as these tools are further embedded into society and artistic practice, active divergence from their monolithic output may become increasingly important.

This work contributes to the "network-bending" framework - the direct manipulation of neural network parameters in deep generative modelling. Current network-bending technologies require specialist knowledge to operate, demanding a steep learning curve for novel users. In addition, they often require access to high-performance GPUs for real-time performance. This project aims to address these research gaps by leveraging a user-centred design approach and providing a novel blueprint for embedding a network-bending system on a stand-alone system.

Through the fabrication of a physical, network-bending instrument, this work aims to benefit musicians wishing to integrate alternative ML methods into their practice. The instrument, if successful, could aid in enabling greater participation in network-bending technologies and contribute to inspiring further investigation within music technologist and sound art communities.

Related Works

This work builds upon prior research in two key areas. Firstly, we review the technical basis of RAVE-based neural audio synthesis. Secondly, we situate Brave within the broader context of the network-bending framework being used in the creative space, highlighting key influences for the development of the instrument.

Neural Audio Synthesis

Whilst discriminative modelling (Ng and Jordan 2001) provided the bulk of early ML research, in the last decade there have been leaps forward through novel applications of deep learning to generative modelling tasks. Generative models are now capable of synthesising realistic sound both in the frequency domain (Vasquez and Lewis 2019), audio domain (van den Oord et al. 2016) and images (Karras et al. 2021).

Auto-Encoder (AE) (Hinton and Salakhutdinov 2006) architecture is composed of an encoder and decoder. The encoder compresses high-dimensional input data into a lower-dimensional representation vector - known as the latent vector z. The decoder decompresses this latent representation back to the original domain. This architecture exploits the manifold hypothesis (Fefferman, Mitter, and Narayanan 2013): high-dimensional data exhibits an intrinsic low-dimensional structure, despite being embedded in high-dimensional space. The low-dimensional embedding is defined as the latent space. Variational Auto-Encoders (VAEs) (Kingma and Welling 2014) regularise the latent space with parameters mean (μ) and standard deviation (δ). Becoming one of the most well-known deep learning generative architectures, they have become widely used in text (Wang et al. 2019; Semeniuta, Severyn, and Barth 2017), image (Vahdat and Kautz 2020; Huang et al. 2018) and audio synthesis (Haque, Rana, and Schuller 2020).

Realtime Audio Variational autoencoder (RAVE) (Caillon and Esling 2021) leverages VAE architecture for fast and high-quality audio waveform synthesis. The model is able to generate 48 kHz audio in real-time on a standard laptop

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

CPU. The RAVE model employs a two-stage training procedure. A perceptually motivated reconstruction objective is introduced as "multiband spectral distance". Once this metric converges, adversarial fine-tuning is leveraged with the aim of improving audio quality/naturalness. When the learned representation is considered sufficient for use, the encoder is frozen. Audio signals can then be synthesised and fed to the discriminator, optimising the adversarial objective.

Network-bending Framework

Terence Broad et al. were the first to introduce the notion of "network-bending" (Broad, Leymarie, and Grierson 2021). The authors introduce a new framework for the direct manipulation of deep generative models, facilitating active divergence from the original training dataset. Deterministic transformations are formulated, which can be inserted into the computational architecture of a trained generative neural network and applied during inference. This methodology was applied to both StyleGAN2, for image generation, and a VAE trained on spectrograms of music samples.

The network-bending framework can be placed upon a broader cultural timeline, extending a practice of technological augmentation, subversion and ownership. The name inherits from "circuit bending", coined by Qubais Reed Ghazala (2004), denoting the practice of modifying low voltage electronic devices through the removal/addition of electronic components. Ghazala compares the practice to the creation of an instrument from a "coconut washed up on the shore" - augmenting society's electronic refuse to fabricate a unique sound generator. In its relation to DIY culture and finding chaos within the mass-produced norm, Brave seeks to draw connections between "network-bending" and Ghazala's practice.

Methodology

Framework

The study was guided by the Proof-of-Concept (PoC) implementation of Media and Arts Technology (MAT) studies (Bryan-Kinns and Reed 2023). This implementation provides a blueprint for the design/evaluation of novel interactive systems. A user-study was conducted - adopting an exploratory approach, prioritising open-ended, playful interaction and qualitative feedback over quantitative hypotheses. The PoC research question, which was explored in the user-feedback cycles, was formulated as follows.

What if I surface internal neural network parameters with a physical interface in the context of audio synthesis? How do artists/technologists respond to this, and what playful interactions/technical insight can be prompted?

Participants

6 participants participated in the user-study, sampling a demographic of diverse musical/technical backgrounds aligning with the target demographic. Demographic data is shown in Table 1.

Study Procedure

The study was conducted in a home studio, equipped with Audio-Technica ATH-M50x headphones that provided audio playback during the study. This took place over two weeks. At the beginning of each session, participants signed a consent form and completed a demographic questionnaire. A technical and conceptual overview of the project was provided, assuming no specialist knowledge - this included instructions on how to operate the instrument. They were then prompted to begin their interaction, which was video recorded for retrospective analysis. Then, a series of open-ended questions, gauging both technical/conceptual experience of the user, were asked in interview form and later transcribed for analysis. The questionnaire participants were given can be found at the following link: `https://danielmanz17.github.io/Brave-Questionnaire`.

Embedded Network-Bending

We leverage Kotowski's network-bending fork [1] of the IRCAM $\tilde{nn}$ repository. The decision was made to compile this work and embed it within a stand-alone instrument. The work explores how deep learning architectures can be embedded on a stand-alone device, integrating network-bending, and to further understand how this framework can be used as an "instrument". It is hoped that this can be scaled, motivated by greater efficiency and reduced form factor as parallel processor architecture continues to advance. As this is a PoC, a Raspberry Pi 5 was chosen as the platform for this work. A full script, which can be used on other devices, can be found in the Brave repository [2].

[1] `https://github.com/blazejkotowski/nn_tilde_bending`

[2] `https://github.com/danielmanz17/Brave`

Table 1: Demographic Data of Participants in Brave User Study

Participant	Age	Traditional music Experience	Electronic music Experience	Technical Proficiency	ML familiarity
P1	36	Advanced	Advanced	Expert	Advanced
P2	24	Beginner	Beginner	Advanced	Advanced
P3	26	Beginner	Advanced	Advanced	Intermediate
P4	32	Beginner	Beginner	Advanced	Intermediate
P5	24	No experience	Beginner	Advanced	Intermediate
P6	27	Advanced	Advanced	Advanced	Advanced

Technical Design and Implementation

The Raspberry Pi 5 has a 2.4 GHz quad-core 64-bit Arm Cortex-A76 CPU, a significant upgrade from the Raspberry Pi 4, and a potential 8 GB of RAM. A fan was installed as an active cooling mechanism to prevent thermal throttling - the patch is highly intensive on the CPU/RAM. The Raspberry Pi 5 DAC+3 [3] was used. This has the Texas Instruments PCM5122 DAC delivering stereo analogue audio to a dedicated headphone amplifier, supporting 24-bit 192 kHz high-resolution digital audio. The DAC uses pass-through pins, meaning the peripheral interface elements were still able to attach to the GPIO pins.

System Architecture

A networking solution was used to communicate between peripheral interface elements and Pure Data. OSC receives signals from five digital encoders which are mapped to network-bending parameters, and a switch which can reset the entire model or individual layer weights. The Python script can be found in the Brave repository.

The Adafruit I2C QT Rotary Encoder [4] breakout was used with a standard 24-pulse encoder. This uses the I2C Stemma QT communication protocol, a variant of the I2C protocol developed by Adafruit. The protocol uses JST SH 6- pin connectors and cables, allowing easy daisy-chaining of multiple devices without soldering. This reduces wiring complexity and minimises the use of GPIO pins. A momentary button switch was connected to pins 13 and 14 (GND/5V). Finally, a Waveshare 4.3-inch DSI capacitive touchscreen LCD [5] was attached via the 4-lane MIPI DSI/CSI connector - capacitive was chosen over resistive due to sensitivity/UX considerations.

Design Elements

Physical housing was designed using the Computer-Aided Design (CAD) software Autodesk Fusion 360 and 3D printed. The original 3D print used white filament, which was later painted. A small rectangle was debossed to place the Brave "logo". This used an old Germanic "Rundgotisch" font.

A simple ensemble of visual labels for the four primary encoders was introduced to indicate their respective transformations, employing simple mathematical notation. From the leftmost to rightmost encoder (see Figure 2):

1. The slash represents location within the neural network. This encoder can be used to switch between layers
2. Scalar multiplication notation to symbolise applying a factor to all weights within the respective layer.
3. Δx to denote shifting all weights in the x-dimension, represented on the touch screen.
4. Δy to denote shifting all weights in the y-dimension, also represented on the touch screen.

Results

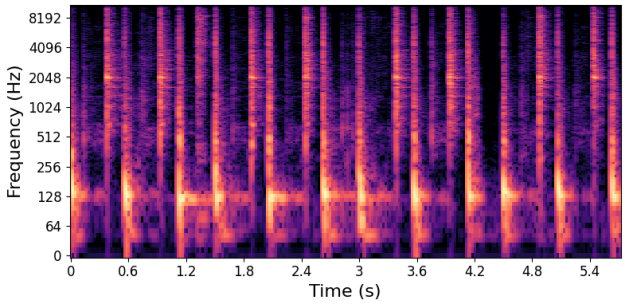
Output

Through the user-feedback cycle and self-experimentation, network-bending transformations which led to "interesting" sonic results were identified. A tranformation example is shown as a spectrogram below, in Figure 1. The audio files and addtional examples are hosted at the following link: `https://danielmanz17.github.io/Brave-Samples/`. A longer instrument demo is presented in video format: `https://www.youtube.com/watch?v=0HugWkdesgw`.

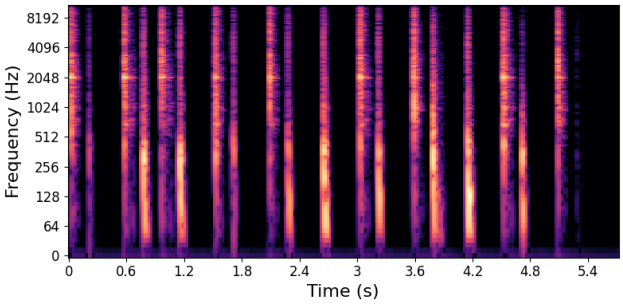
[3] `https://thepihut.com/products/iqaudio-dac`

[4] `https://learn.adafruit.com/adafruit-i2c-qt-rotary-encoder/overview`

[5] `https://www.waveshare.com/4.3inch-dsi-lcd.htm`



(a) Brave running the darbouka RAVE model without any transformations applied to the network architecture. Baseline for inspecting other outputs.



(b) Applying $\Delta y = 0.47$ at bias of first encoder layer. This transformation reduces the sustain of the percussive hits.

Figure 1: Spectrogram representation of audio before and after a network-bending transformation has been applied.

User-feedback

P1 noted that the encoders produced more predictable results than the touchscreen. For example, shifts in Δy applied to layer 17 seemed to consistently create a rhythmic "half step" transformation. They noted that the scale encoder could perhaps act as a "magnifier" - does increasing the scale of the tensor values at this layer amplify its corresponding processing fingerprint?

The need for an indication of location within the neural network was highlighted, as they felt "lost" within the model. To save sounds found within the parameter space, a preset functionality was suggested. High-dimensionality layers bear visual resemblance to a waveform which created confusion. Finally, it was noted that the encoder transformations are applied to the initial rather than the modified weights of the model.

P2 highlighted the need for continuous weight modification, indication of neural network location and visual differentation from waveforms. The Δy encoder steps were discrete rather than continuous, which didn't make sense to the user. In addition, the mapping direction of Δy transformation felt "unintuitive".

P3 noted that the lower dimensional layers were easier to control, and that perhaps these could be prioritised in layer exposure. They adopted a more rigorous approach to parameter space exploration - exploring specific transformations at each layer, and probing how this affects the audio output. For example, they investigated how specific shapes drawn on the touch screen at different layers would influence the output. Again, the issues of neural network location, preset saving and continuous weight modification were raised. In addition, an interpolation feature was suggested between the initial/modified layer, similar to a typical synthesiser dry/wet knob.

P4 enjoyed pushing the sounds as far as possible. They identified unexpected metallic/string like artefacts with certain bending approaches. Preset functionality and waveform confusion were highlighted.

P5 noticed that changing the values within the lower-dimensional layers seemed to generate more contained results. They found Δx produced rhythmic rather than timbral variation. The participant flagged confusion about the magnitude of the values - does the middle of the screen correspond to a null value of weights? In terms of the fabrication process, ribs/support pillars were suggested as the body felt fragile. Finally, higher encoder sensitivity was suggested.

P6 particularly enjoyed exploring the scale/bias encoders. They mentioned a soft learning curve - they were quickly able to determine what everything does. They found rhythmic divergence to be the most interesting line of exploration. Visual indication of location in neural network, input sample choice and waveform confusion were flagged. They mentioned some of the encoder transformations felt a bit "steppy".

These user feedback sessions led to the finalisation of the Brave instrument. This process of reflection and iteration will be discussed in the User-Feedback Reflection section. The final instrument is shown in Figure 2 below.

Discussion

Design and Development Process

The instrument could benefit from additional user-testing cycles, integrating quantitative data and more traditional HCI methodology (usability testing, think-aloud protocol, task analysis and A/B testing).

In addition, Brave could profit from a more diverse design methodology - introducing elements of the autobiographical framework (Ó Néill and Ortiz 2024). Brave could be integrated into a music production workflow/live performance setup, including experimentation with alternative RAVE models. Perhaps custom network-bending models could be trained - curating training corpora with network-bending as the end goal in mind.

It could be beneficial to consider alternative embedded systems - Jetson Nano7/STM32 32-bit Arm Cortex MCU8 would both be ideal candidates. These implementations could engage with pertinent questions regarding the future of embedded neural systems, enabling modification of the underlying architecture. Can we move towards a more scalable neural instrument? Could this evolve into a viable, consumer-facing product? Further miniaturisation and advances in neural processing architecture (compression) draw these questions ever closer.

User-Feedback Reflection

The absence of preset saving functionality was frequently flagged. This could be implemented with rectangular pads, each of which correspond to a saved network-bending preset. For example, holding the pad could save a new preset, and a singular press can recall the saved preset.

Another consistent pain point was getting "lost" in the neural network - the interface provides no indication for the user to triangulate their position within the network. This could be exposed via a number or layer name. A more advanced solution would be a graphical indicator using a touch-strip, which would provide a visual representation and control mechanism of layer position.

The discontinuous weight modification was frequently flagged - the encoder transformations are applied to the initial rather than the modified weights of the model, interrupting the user workflow.The authors hope to implement this in the future.

A few users commented that the sensitivity of the encoders could be increased, which would also reduce the perceived "stepiness" of the knobs. Sensitivity was increased by a factor of 1.4 for all transformation encoders. Further user-testing could be conducted to fine tune this. In addition, the momentary switch button was incorrectly detecting single/double taps for some users. The decision was made to use the push button functionality of the first encoder. This made intuitive sense as encoder one already controls layer position.

Finally, high-dimensional layers visually resemble a waveform. This created frequent confusion, especially for those with a background in electronic music production. Different UI approaches could be considered to expose the

ISBN: 978-989-54160-7-3

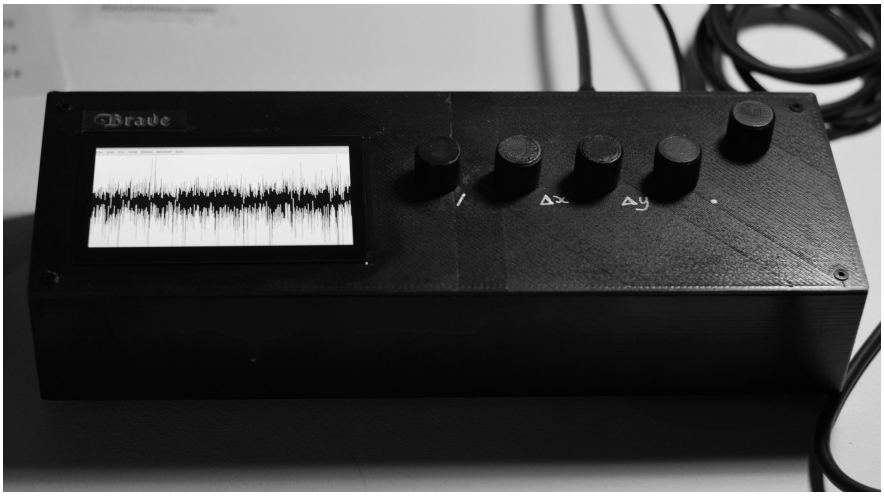


Figure 2: Final version of Brave.

layer weights - perhaps more traditional neural network representations accompanied by an activation heat map, switching the orientation of the screen, or reducing the dimensionality of the layer representations within the UI.

Conclusion

This study documents the design and fabrication of Brave: an embedded network-bending, stand-alone instrument. As neural audio synthesis becomes more widely used, there are growing concerns that its reliance on statistical averages could lead to the homogenisation of the cultural landscape, reinforcing a cycle of self-referential synthesis. This work seeks to explore new methods for addressing these challenges by contributing to the developing field of "network-bending". Incorporating musician-led feedback, user-centred interface development and accessible hardware integration, this study aims to lower technical barriers and foster broader participation within the network-bending framework.

References

Broad, T.; Leymarie, F. F.; and Grierson, M. 2021. Network bending: Expressive manipulation of deep generative models. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings*, 20–36. Berlin, Heidelberg: Springer-Verlag.

Bryan-Kinns, N., and Reed, C. N. 2023. A guide to evaluating the experience of media and arts technology.

Caillon, A., and Esling, P. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *CoRR* abs/2111.05011.

Fefferman, C.; Mitter, S.; and Narayanan, H. 2013. Testing the manifold hypothesis. *Journal of the American Mathematical Society* 29.

Ghazala, R. 2004. The folk music of chance electronics: Circuit-bending the modern coconut. *Leonardo Music Journal* 14:97–104.

Haque, K. N.; Rana, R.; and Schuller, B. W. 2020. High-fidelity audio generation and representation learning with guided adversarial autoencoder. *IEEE Access* 8:223509–223528.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

Huang, H.; Li, Z.; He, R.; Sun, Z.; and Tan, T. 2018. Introvae: introspective variational autoencoders for photographic image synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 52–63. Red Hook, NY, USA: Curran Associates Inc.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Red Hook, NY, USA: Curran Associates Inc.

Kingma, D. P., and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ng, A., and Jordan, M. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T.; Becker, S.; and Ghahramani, Z., eds., *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Ó Néill, E., and Ortiz, M. 2024. From prototype to performance practice: Reflections on iterative instrument design. In *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures*, AM '24, 439–444. New York, NY, USA: Association for Computing Machinery.

ISBN: 978-989-54160-7-3

Semeniuta, S.; Severyn, A.; and Barth, E. 2017. A hybrid convolutional variational autoencoder for text generation. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 627–637. Copenhagen, Denmark: Association for Computational Linguistics.

Vahdat, A., and Kautz, J. 2020. Nvae: a deep hierarchical variational autoencoder. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. In *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 125.

Vasquez, S., and Lewis, M. 2019. Melnet: A generative model for audio in the frequency domain. *arXiv preprint arXiv:1906.01083*.

Wang, W.; Gan, Z.; Xu, H.; Zhang, R.; Wang, G.; Shen, D.; Chen, C.; and Carin, L. 2019. Topic-guided variational autoencoder for text generation. 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)*, 166–177. Minneapolis, Minnesota: Association for Computational Linguistics.

ISBN: 978-989-54160-7-3

Controlling the image generation process with parametric activation functions

Ilia Pavlov
Creative Computing Institute, University of the Arts London, London UK
i.pavlov@arts.ac.uk

Abstract

As image generative models continue to increase not only in their fidelity but also in their ubiquity, the development of tools that leverage direct interaction with their internal mechanisms in an interpretable way has received little attention. In this work, we introduce a system that allows users to develop a better understanding of the model through interaction and experimentation. By giving users the ability to replace activation functions of a generative network with parametric ones and a way to set the parameters of these functions, we introduce an alternative approach to control the network's output. We demonstrate the use of our method on StyleGAN2 and BigGAN networks trained on FFHQ and ImageNet, respectively.

Introduction

The field of Explainable AI (xAI) has always lagged behind the main body of machine learning research (Malizia and Paternò, 2023). This is also true for applications of generative artificial intelligence (AI) (Schneider, 2024) in the field of computational creativity and it's use as a creativity support tool (Shneiderman, 2007).

This paper proposes an interactive tool that allows users to manipulate the behavior of a generative model by modifying its structure. Such interactions have the potential to expand the non-expert user's understanding of the underlying mechanisms behind the model's output and promote an increase in AI literacy.

The tool allows users to replace static activation functions with parametric ones (Apicella et al., 2020). This gives users the ability to manually set the parameters of these functions, and thereby affect their outputs.

Background

xAI as a field aims to develop methods to explain how specific outcomes are reached by an AI model (Weller, 2019; Samek, Wiegand, and Müller, 2017). Specific techniques for models working with visual data include Saliency Maps (Petsiuk et al., 2020), Class Activation Mapping (Zhou et al., 2015), Layer-Wise Relevance Propagation (Binder et al., 2016).

Machine learning approach to generative models works by mapping a data point drawn from the latent distribution Z to a data point taken from the data distribution X (Ruthotto and Haber, 2021). In practice, distribution Z is typically a Gaussian distribution, and distribution X is real-world or simulated data, such as images, audio, or text. There are different ways one can achieve this, some examples include the following archetypes: Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Variational Autoencoders (Kingma and Welling, 2013), and Diffusion models (Saharia et al., 2022). It is worth to note that generative models are not exclusive to neural networks and expand into evolutionary models (Cook and Colton, 2018), agent-based models (Delarosa and Soros, 2020) and others (Briot, Hadjeres, and Pachet, 2019).

The manipulation of the generative network's feature maps is not a new area of research. The method described in the Network Bending paper (Broad, Leymarie, and Grierson, 2022) uses deterministic transformations to affect the output of the network during inference. To make this more effective they introduce a procedure for clustering activation maps based on their spatial similarity. This allows for the grouping of feature maps, which correspond to different features of the generated image.

PandA (Oldfield et al., 2022) is a method for discovering factors responsible for image appearance in feature map space. This then enables users to edit the output images with pixel accuracy by modifying these factors.

The manipulation of the behaviour of generative neural networks can be also achieved by inserting small-scale neural networks in-between layers and training them for a short time, as shown in this paper (Aldegheri et al., 2023). This introduces deviations from original data distribution the network was trained on, while also maintaining an overall structure of images.

Another prominent method used to manipulate the generative network's output is latent space exploration (Klys, Snell, and Zemel, 2018; Corneanu, Gadde, and Martínez, 2024; Shen et al., 2019). There is a large variety of work, detailing various approaches to the manipulation of network's latent space. For example, papers (Bontrager et al., 2018; Fernandes, Correia, and Machado, 2020) propose the use evolutionary algorithms for this task.

Control Method

In this section, we introduce our GUI and examine the multiple parametric activation functions we experimented with.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

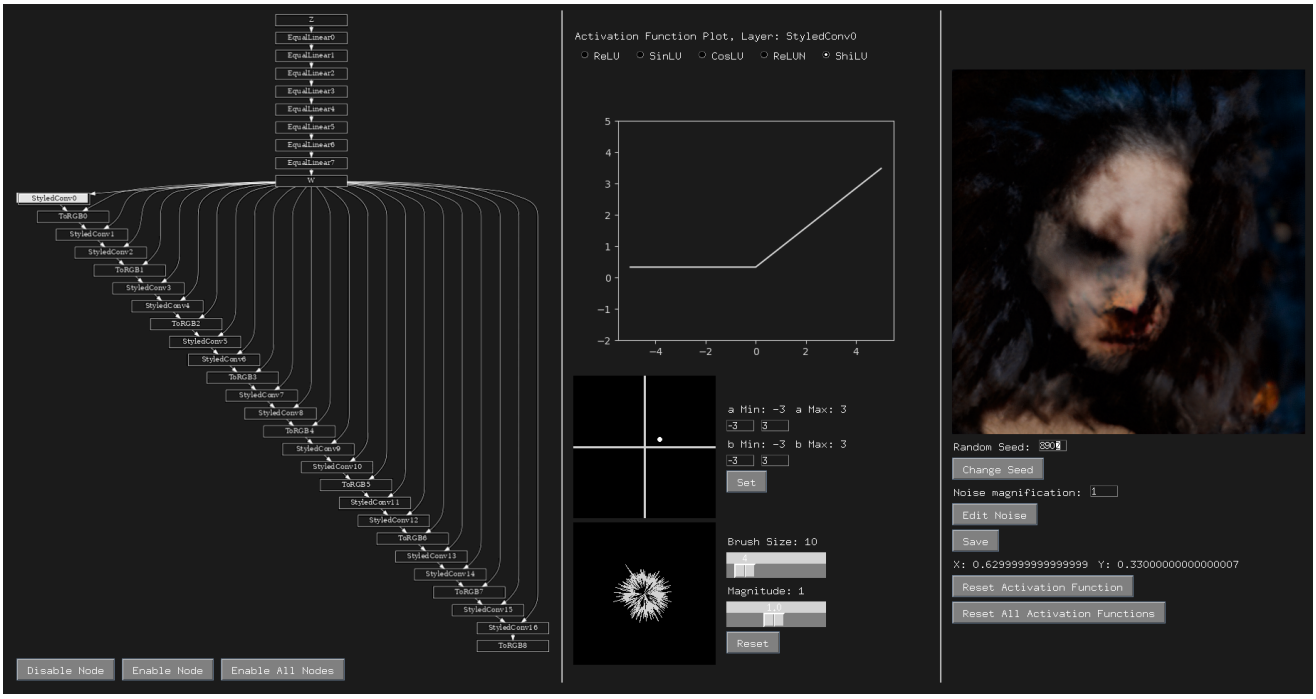


Figure 1: A Graphical User Interface was developed for our control method. The left side is responsible for picking neural layers, the center side is responsible for selecting activation functions and adjusting their parameters if applicable, and the right side allows for real-time result visualization.

The code for this paper can be found on GitHub[1] and the video demonstration of it's use is available on YouTube.[2]

Our graphical user interface, as shown in Figure 1, allows users to:

- Select the desired activation functions (left).
- Replace selected activation functions with parametric ones (center-top).
- Adjust the parameters of injected activation functions (center-bottom).
- Have a real-time feedback (right).

The GUI also has features for editing the input latent vector (center-bottom) and disabling specific layers outright (left-bottom).

All these features provide control over the structure of the network and showcase how different changes affect the outcome.

Activation functions

The Sinu-Sigmoidal Linear Unit (Paul et al., 2022) has two parameters, a and b. The parameter a is responsible for the amplitude of the sine function, while the parameter b is responsible for the frequency of the sine function. Setting any of these parameters too high can lead to unpredictable results, which could be preferable for more abstract image generation. The full formula for the activation function is shown in (1), where $\sigma(x)$ is a sigmoid function.

$$SinLU(x) = (x + a * sin(b * x)) * \sigma(x) \quad (1)$$

Rectified Linear Unit N (Pishchik, 2023) or ReLUN is a modified version of the ReLU6 activation function with one parameter, which controls the angle of the function. The full formula for the activation function is shown in (2).

$$ReLUN(x) = min(max(0, x), n) \quad (2)$$

Shifted Rectified Linear Unit (Pishchik, 2023) or ShiLU is a modified version of ReLU (Fukushima, 1969) function that has two parameters a and b. Parameter a is responsible for the slope of the function, and parameter b is responsible for shifting the function along the Y-axis. The full formula for the activation function is shown in (3).

$$ShiLU(x) = a * ReLU(x) + b \quad (3)$$

Usage of such functions causes alterations in the output by changing the intermediate feature maps (Zeiler and Fergus, 2013) of the network during the inference process. Using these functions in multiple layers will cause a cascading effect of changes. This in turn can lead to a large amount of possible divergences from initial neural network state.

We also experimented with polynomial activation functions. Using polynomial functions as activation functions in neural networks has been proposed before, but never gained

[1] https://github.com/locsor/generativeControlUI

[2] https://youtu.be/qkP9DHLicwM

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

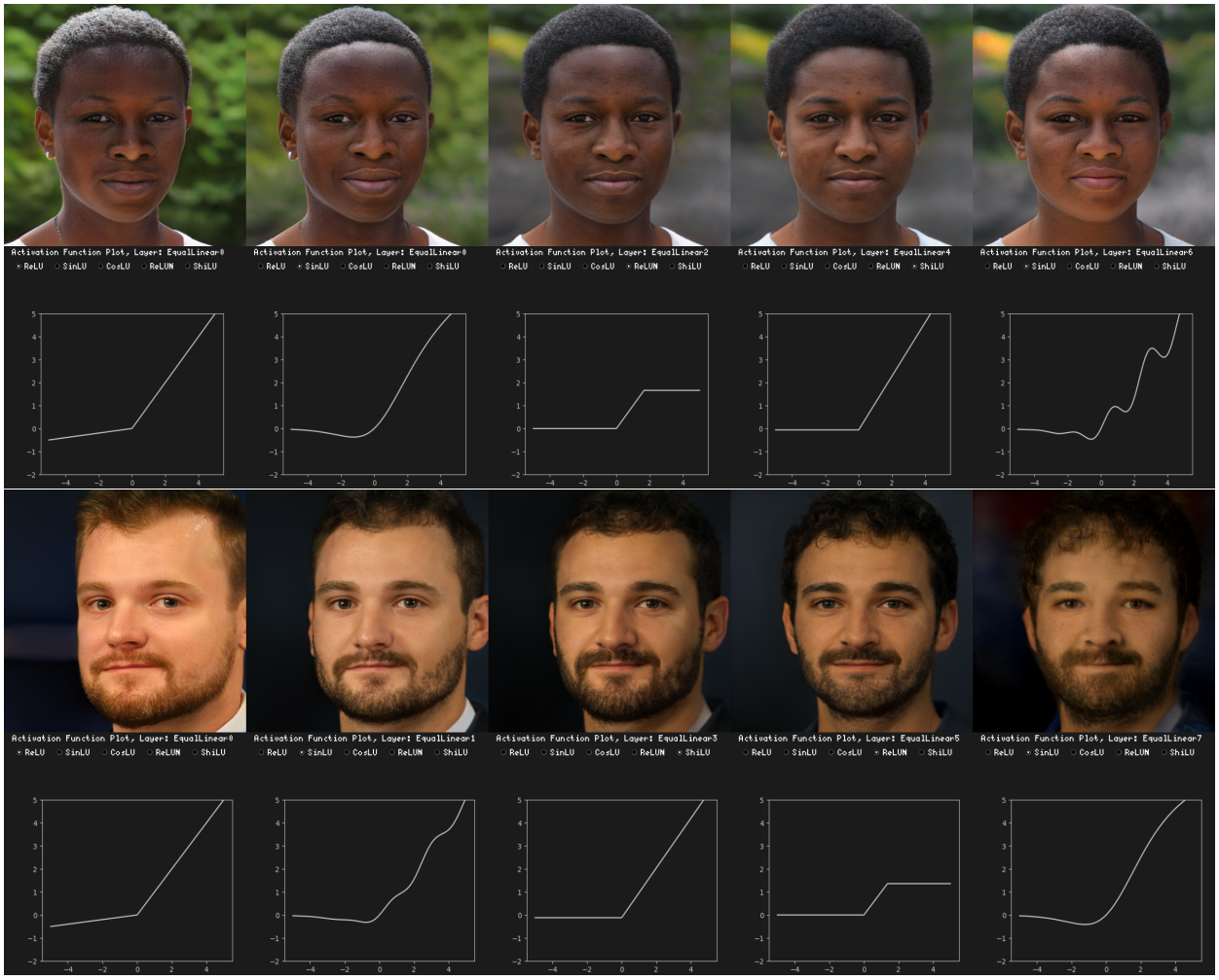


Figure 2: Examples of using parametric activation functions in the mapping network of the StyleGAN2. The left-most images in output rows (faces) are base outputs. Each image has one more base activation function replaced with a parametric one, compared with an image to its' left. Bottom rows show plots of parametric functions, which were used to produce the images above.

much popularity, due to the exploding gradient problem and increased complexity of training. The specific polynomial activation functions we used have been taken from a retracted paper (Gottemukkula, 2019). The general formula is shown in (4).

$$f_n(x) = \frac{\sum_{i=1}^{n} w_i \cdot \sigma(x)^i}{\sqrt{2}^n} = \frac{w_0 \cdot \sigma(x)^0 + w_1 \cdot \sigma(x)^1 + \ldots + w_n \cdot \sigma(x)^n}{\sqrt{2}^n} \quad (4)$$

Where each weight coefficient w_i in this function is a potential control parameter in our system and $\sigma(x)$ is a sigmoid function.

Experiments

For base networks, we used the StyleGAN2 (Karras et al., 2019) and BigGAN (Brock, Donahue, and Simonyan, 2018) models. The choice was motivated by their ubiquity in generative neural network research, stability, and speed.

The architecture of StyleGAN2 is split into two parts: a mapping network and a generator network. The purpose of a mapping network is to disentangle the input latent vector. The generator network consists of a series of convolutional blocks, each outputting an image with progressively increasing resolution. To get the final image, these intermediate images are scaled up and summed. Applying the parametric activation functions to StyleGAN2 had the following results:

- Using them in the mapping network affected the image structure. These structural changes are due to parametric activation functions affecting the resulting output of the mapping network, i.e. the disentangled latent vector.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

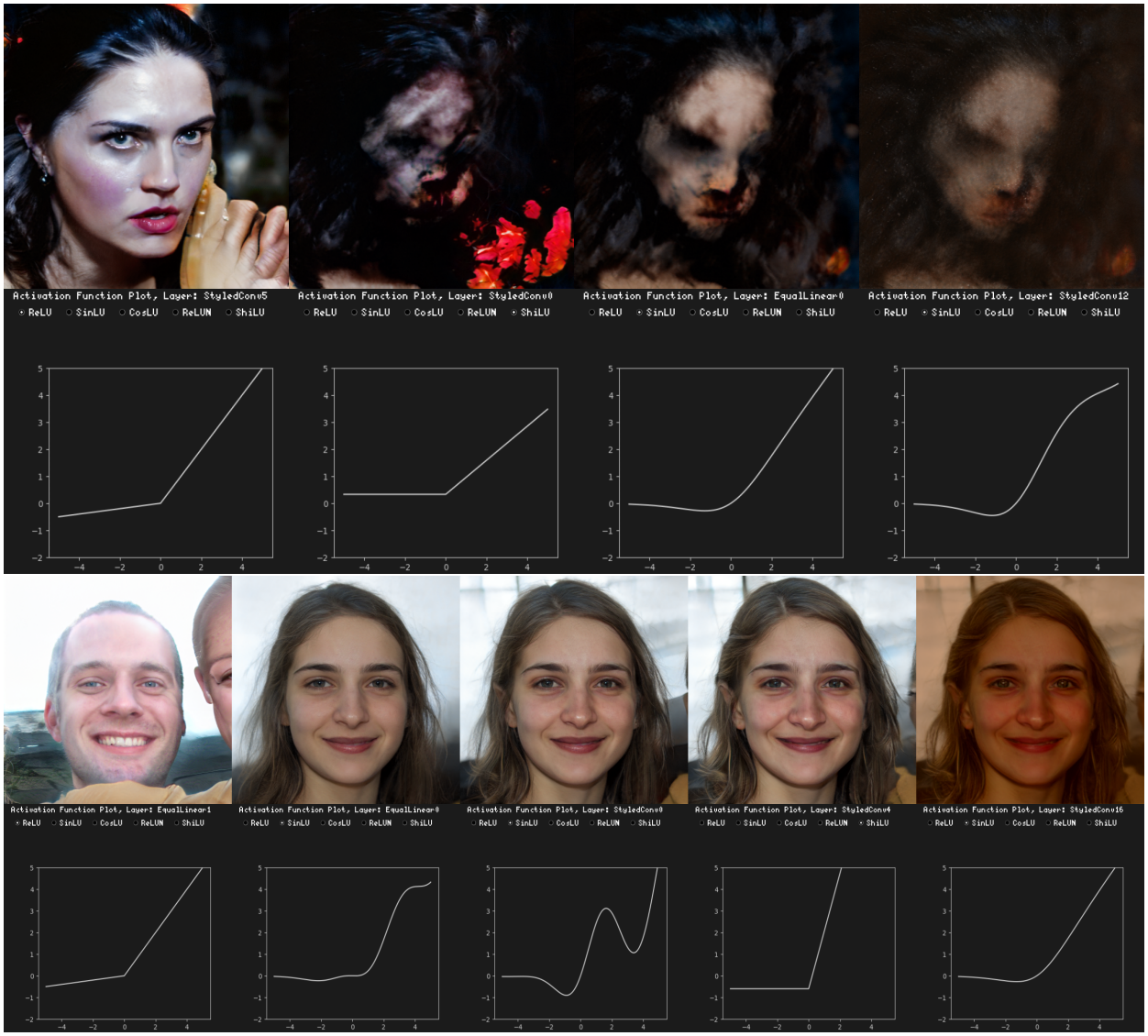


Figure 3: Examples of using parametric activation functions in both mapping and generator networks of the StyleGAN2. The left-most images in output rows (faces) are base outputs. Each image has one more base activation function replaced with a parametric one, compared with an image to its' left. Bottom rows show plots of parametric functions, which were used to produce the images above.

- Earlier layers of the mapping network offered a high level of granularity with respect to changes in the image.
- Using them in the generator network affected both style and, in cases of earlier layers, the structure of images.
- Later layers of the generator network caused changes to images in a more basic manner, such as altering the overall coloration of the image.

As shown in our examples in Figures 2 and 3, the usage of controllable activation functions can be a viable way to control the image generation process. Slight changes to the parameters of these functions will lead to slight changes in the resulting image, while larger changes lead to more significant and unpredictable alterations to both the structure and style of the image. For example, a user can infer that the latter quality is not desirable for generating realistic images, it can allow for the generation of more abstract imagery.

BigGAN is a classic example of a GAN architecture but scaled up to improve both fidelity and variety of generated images. For our experiment, we used a class-conditional implementation of this network. While the changes in images are not as varied as those observed with StyleGAN2, they still demonstrate the ability to alter the content of the image to some extent. Examples of images generated using BigGAN are shown in Figures 4 and 5.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Figure 4: Output of a BigGAN model. The left-most image is an original, the rest were generated with SinLU or ReLUN applied to the second layer of the network. The parameter values were randomly chosen.

Figure 5: Output of a BigGAN model with 3rd order polynomial activation functions. The functions' parameters were kept within $[0.8, 1.2]$ interval. The left-most images are originals, the right-most are generated with 4 modified functions, and the rest were generated with only one modified function.

Polynomial functions were not as encouraging. Using them in more than 2 neural layers yielded non-ideal results, and they were extremely sensitive to small changes in parameter values. Examples of images generated by the BigGAN model modified with a 3rd degree polynomial activation functions are shown in Figure 5. We limited the parameter values to $[0.5, 1.5]$.

Overall, in a standard use case, a user will gradually modify the network and carefully select the values of the control parameters to achieve a satisfactory result. Since our method does not offer any way to single out any specific feature of the image, the process largely consists of trial-and-error experimentation. Through continued use and real-time visual feedback, users can learn how the structural state of the network influences the output of the generation process. This process is mainly aimed at incidental exploration, instead of a guided image generation.

Conclusion

In summary, this paper introduces a novel approach to control image generative networks, with the help of parametric activation functions, and examines how this tool can be used to teach non-expert users the importance of its various parts. But to confirm these speculations, a user study will need to be conducted.

The main limitation of the current implementation is, that unguided control over image generation is imprecise and limiting. While applying this method to the text-to-image generative network can alleviate this problem, there are still no guarantees that users will be able to find a satisfactory collection of activation functions and their parameters.

References

Aldegheri, G.; Rogalska, A.; Youssef, A.; and Iofinova, E. 2023. Hacking generative models with differentiable network bending. *ArXiv* abs/2310.04816.

Apicella, A.; Donnarumma, F.; Isgrò, F.; and Prevete, R. 2020. A survey on modern trainable activation functions. *Neural networks : the official journal of the International Neural Network Society* 138:14–32.

Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.-R.; and

ISBN: 978-989-54160-7-3

Samek, W. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. *ArXiv* abs/1604.00825.

Bontrager, P.; Lin, W.-C.; Togelius, J.; and Risi, S. 2018. Deep interactive evolution. *ArXiv* abs/1801.08230.

Briot, J.-P.; Hadjeres, G.; and Pachet, F. 2019. Computational creativity: The philosophy and engineering of autonomously creative systems. *Computational Creativity*.

Broad, T.; Leymarie, F. F.; and Grierson, M. 2022. Network bending: Expressive manipulation of generative models in multiple domains. *Entropy* 24(1).

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale gan training for high fidelity natural image synthesis. *ArXiv* abs/1809.11096.

Cook, M., and Colton, S. 2018. Redesigning computationally creative systems for continuous creation. In *International Conference on Innovative Computing and Cloud Computing*.

Corneanu, C. A.; Gadde, R.; and Martínez, A. M. 2024. Latentpaint: Image inpainting in latent space with diffusion models. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* 4322–4331.

Delarosa, O., and Soros, L. B. 2020. Growing midi music files using convolutional cellular automata. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* 1187–1194.

Fernandes, P.; Correia, J.; and Machado, P. 2020. Evolutionary latent space exploration of generative adversarial networks. In *EvoApplications*.

Fukushima, K. 1969. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics* 5(4):322–333.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial networks. *Communications of the ACM* 63:139 – 144.

Gottemukkula, V. 2019. Retracted: Polynomial activation functions.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2019. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8107–8116.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *CoRR* abs/1312.6114.

Klys, J.; Snell, J.; and Zemel, R. S. 2018. Learning latent subspaces in variational autoencoders. *ArXiv* abs/1812.06190.

Malizia, A., and Paternò, F. 2023. Why is the current xai not meeting the expectations? *Commun. ACM* 66(12):20–23.

Oldfield, J.; Tzelepis, C.; Panagakis, Y.; Nicolaou, M. A.; and Patras, I. 2022. Panda: Unsupervised learning of parts and appearances in the feature maps of gans. *ArXiv* abs/2206.00048.

Paul, A.; Bandyopadhyay, R.; Yoon, J. H.; Geem, Z. W.; and Sarkar, R. 2022. Sinlu: Sinu-sigmoidal linear unit. *Mathematics* 10(3).

Petsiuk, V.; Jain, R.; Manjunatha, V.; Morariu, V. I.; Mehra, A.; Ordonez, V.; and Saenko, K. 2020. Black-box explanation of object detectors via saliency maps. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11438–11447.

Pishchik, E. 2023. Trainable activations for image classification. *Preprints*.

Ruthotto, L., and Haber, E. 2021. An introduction to deep generative modeling. *GAMM-Mitteilungen* 44.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv* abs/2205.11487.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ArXiv* abs/1708.08296.

Schneider, J. 2024. Explainable generative ai (genxai): A survey, conceptualization, and research agenda. *ArXiv* abs/2404.09554.

Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2019. Interpreting the latent space of gans for semantic face editing. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 9240–9249.

Shneiderman, B. 2007. Creativity support tools: accelerating discovery and innovation. *Commun. ACM* 50:20–32.

Weller, A. 2019. Transparency: Motivations and challenges. In *Explainable AI*.

Zeiler, M. D., and Fergus, R. 2013. Visualizing and understanding convolutional networks. *ArXiv* abs/1311.2901.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2015. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2921–2929.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Creativity Rivalry: Human, Artificial Intelligence, and Co-Design

Hoa Vo
Eskenazi School of Art, Architecture + Design
Indiana University
Bloomington, IN 47405 USA
hoavo@iu.edu

Abstract

This study explores the creativity rivalry between AI, human designers, and co-design efforts in the 2024 Robert Bruce Thompson Lighting Design Competition, which tasked participants with designing a pediatric waiting room light fixture. A total of 120 Amazon Mechanical Turk (MTurk) participants evaluated design solutions across three stages: sketches (S), three-dimensional renderings (3D), and virtual room models (VR). Using the Creative Product Semantic Scale (CPSS), they assessed novelty (originality and surprise), resolution (logic and utility), and style (craftsmanship and elegance). Results showed that AI excelled in style at the 3D stage, while the human designer led in novelty during S and VR stages. Unexpectedly, co-design underperformed across criteria. These findings challenge assumptions about AI's creative limitations and highlight its strengths in visualization and refinement. The study offers practical insights for designers and educators on strategically integrating AI while maintaining human originality in the creative process.

Introduction

Is creativity becoming a battleground between humans and Artificial Intelligence (AI)? As AI advances beyond automating repetitive tasks into the realm of creative production, it has ignited legal and ethical debates (Rattay et al., 2022). The five-month Writers Guild of America strike[1] successfully defended the rights of human scriptwriters, yet artists and illustrators contesting the appropriation of their styles by generative AI platforms saw their claims dismissed[2]. Meanwhile, AI-generated artwork has won prestigious competitions, such as the Colorado State Fair[3], and real estate AI platforms are now challenging the expertise of interior designers (qbiq, 2024; TestFit, 2024).

These tensions raise pressing questions: "Is human creativity under threat, or can AI and humans coexist as creative collaborators?" The classic Turing Test (1950) suggests that AI's intelligence is determined by whether a human evaluator can distinguish its responses from those of a person. Similarly, in design, AI's creative ability is assessed by how closely its outputs align with human standards. This study explores the *creativity rivalry* between humans, AI, and co-design by empirically evaluating their design solutions in a real-world competition. Through this lens, we examine whether AI is a true creative rival or a transformative partner in design.

Aim and Significance

This study aims to empirically examine differences in creativity among design solutions generated by AI, human-AI co-design, and a human designer. To achieve this, three hypotheses were tested within the context of the Robert Bruce Thompson Lighting Design Competition[4]:

- **H1:** AI-generated designs will exhibit lower creativity than those of a human designer.
- **H2:** Co-design between humans and AI will produce more creative outcomes than AI alone.
- **H3:** Co-design between humans and AI will be more creative than human-designed solutions alone.

This research advances AI studies in design by employing a rigorous creativity assessment framework, using the well-established Creative Product Semantic Scale or CPSS (Besemer, 2006) to ensure objective evaluation. To mitigate bias, blind conditions concealed the origin of each design, preventing inherent assumptions that AI-generated work is less creative than those of human-designed (Horton Jr et al., 2023). By providing empirical insights, the study also offers practical guidance for designers and educators

[1] Merchant, B. (2023). The writers' strike was the first workplace battle between humans and AI. The humans won. *Los Angeles Times*. Retrieved from https://www.latimes.com/projects/writers-strike-deal-ai/

[2] Cho, W. (2023). Artists lose first round of copyright infringement case against AI art generators. *The Hollywood Reporter*. Retrieved from https://www.hollywoodreporter.com/business/business-news/artists-copyright-infringement-case-ai-art-generators-1235632929/

[3] Roose, K. (2022). AI-generated art won a prize. Artists aren't happy. *The New York Times*. Retrieved from https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

[4] https://rbtcompetition.org/

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

on strategically integrating AI into design. Specifically, it highlights how AI can break fixation patterns, introduce novel perspectives, and optimize creativity at key stages of the design process.

Method

This study employed a multi-stage design process to compare the creativity of AI-generated, human-designed, and co-designed solutions in the context of the 2024 Robert Bruce Thompson Lighting Design Competition (Figure 1). This annual competition presents a unique theme each year, requiring creativity within defined parameters. In 2024, the challenge was to design a mounted light fixture for a pediatric waiting room that is both comfortable and engaging for children. This theme introduced an additional layer of complexity, as the designs needed to balance functional, aesthetic, and emotional considerations (Norman, 2005). Specifically, the competition aligns with established creativity assessments, like Besemer's CPSS (2006) with three criteria: *novelty* (freshness of ideas), *style* (completeness via visual representations), and *resolution* (functions). In addition, *resolution* expands to emotional responses that a design can evoke (Horn & Salvendy, 2009), in this case, comfortable and engaging as the brief required.

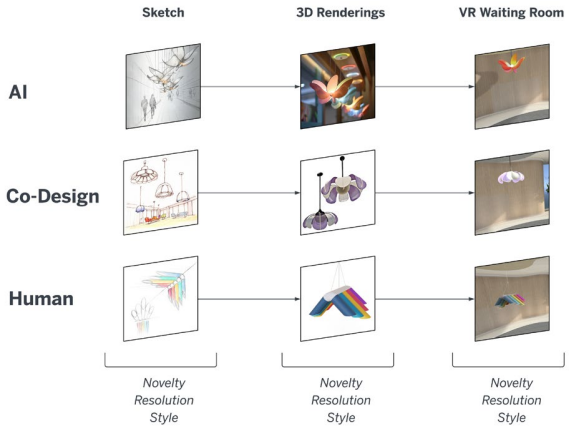


Figure 1. Three conditions and three stages of the design process measured on CPSS criteria for creativity.

Design assessments occurred at three stages: ideation (2D sketches), development (3D renderings), and final solution (VR models in a virtual waiting room). The use of VR enhanced spatial exploration and deepened design understanding, supporting emotional response assessments (Portman et al., 2015). The study was conducted under the Institutional Review Board (IRB) approval #22214.

Participants

A total of 120 Amazon Mechanical Turk (MTurk) workers participated, meeting inclusion criteria of being U.S.-based, aged 18 or older, and holding Master Qualifications to ensure high-quality responses. Exclusion criteria prevented multiple attempts and prior participation in similar studies. Blinded conditions concealed design origins to mitigate bias against AI-generated work. Participants reviewed 2D sketches, 3D renderings, and VR models and completed a Qualtrics questionnaire measuring creativity perceptions and emotional responses.

Design Conditions

Three conditions were tested:

• **AI Condition:** The author used ChatGPT-4, Midjourney, and Kaedim AI to generate design solutions. ChatGPT-4 produced 73 text prompts, which Midjourney used to generate 292 concept sketches. The author screened and refined these into 19 viable sketches, which Midjourney further developed into 3D renderings. A selected rendering was processed through Kaedim AI for final 3D modeling, requiring five hours of computation. The author's role was limited to screening for common or erroneous designs.

• **Co-Design Condition:** A junior interior design student worked on the competition as a class project. The student used Midjourney AI for initial ideas, enhanced outputs through hand sketches and manual refinements, and re-imported them into Midjourney for refinement. The final three designs were developed in 3Ds Max, with one selected as the final model. This process spanned five weeks, ensuring active human-AI collaboration while allowing student agency in prompting, selecting, and refining the design.

• **Human Condition:** A freelance designer from Upwork with proficiency in design heuristics and CAD tools—but without design expertise in lighting—created the design solutions independently. Given the absence of domain-specific expertise, the freelancer did not possess a clear advantage over the junior interior design student, allowing for a meaningful and balanced comparison across conditions. Over one month, the designer submitted a 2D sketch, a 3D rendering, and a final model, responding freely to the competition brief. The author monitored deliverables to maintain experimental integrity but did not intervene in the creative process.

Data Collection

Participants completed a Qualtrics questionnaire after providing informed consent. The survey included demographic questions (e.g., age, occupation, education) and commitment and attention check to ensure thoughtful responses and engagement with the virtual waiting room. To assess creativity, the study employed the CPSS (Besemer, 2006; Wei et al., 2015), which evaluated novelty, resolution, and style (see Table 1) across three design stages: 2D sketches, 3D renderings, and VR models.

To measure emotional responses—an extension of *resolution*, the third criterion of creativity specified in CPSS—participants rated their feelings using two Self-Assessment Manikin (SAM) items (Bradley & Lang, 1994), capturing calmness vs. excitement and unhappiness vs. happiness. Participants reviewed 2D sketches and 3D renderings directly within the questionnaire, while the VR models were accessed via an external link, allowing interactive exploration before completing the survey.

Criterion	Binary
Novelty	Over Used – Fresh
	Predictable – Novel
	Usual – Unusual
	Ordinary – Unique
	Conventional – Original
Resolution	Worthless – Valuable
	Unimportant – Important
	Insignificant – Significant
	Inessential – Essential
	Illogical – Logical
Style	Disordered – Ordered
	Disarranged – Arranged
	Formless – Formed
	Awkward – Graceful
	Plain – Ornate

Table 1. CPSS with sample binaries for each criterion. For a full list of binaries, see Besemer (2006) and Wei et al. (2015).

Data Analysis

Descriptive statistics (mean—M, standard deviation—SD, and percentage—%) summarized demographics, self-reported emotions, and VR effectiveness. The author conducted one-way ANOVA to compare CPSS ratings across conditions (AI, Co-Design, and Human) for novelty, resolution, and style. Shapiro-Wilk and Levene's tests checked normality and variance assumptions. When violated, Kruskal-Wallis tests served as nonparametric alternatives (Lock et al., 2013; Yazici & Yolacan, 2007). Statistical analyses were performed in RStudio Desktop (RStudioTeam, 2023), with significance set at $p < 0.05$.

Results

The overall sample (n = 120) showed similar age distributions across conditions but exhibited some demographic disparities. Notably, the AI condition had a male-dominant gender distribution (80% male, 20% female) and a high concentration of participants aged 25–44 years old (82.5%).

A large proportion (35%) held management positions, followed by sales (17.5%), technical (17.5%), freelance (15%), and other roles (12.5%). In contrast, the Co-Design condition had a more balanced gender distribution (50% male, 45% female, 5% prefer not to disclose) with a slightly younger sample (67.5% aged 25–44 years old).

A significant portion (42.5%) worked in sales, followed by management (32.5%), freelance (12%), and other roles (13%). The Human condition also had a more even gender split (55% male, 45% female) with 75% aged 25–44 years old. Management (25%) and freelance (25%) were the most common occupations, alongside sales (17.5%), manual labor (10%), and other roles (22.5%).

Self-reported Emotions

Participants rated their emotions on a 9-point Likert scale, with 1 representing calm/unhappy, 9 representing excited/happy, and 5 as the neutral midpoint. Across all conditions, participants generally reported feeling happy, though their levels of calmness versus excitement varied.

- **AI Condition:** Participants reported a high level of happiness (M = 6.90, SD = 1.77). However, their calmness-excitement scores leaned toward neutrality (M = 5.35) with greater variability (SD = 2.35).
- **Co-Design Condition:** Participants reported high happiness (M = 6.83, SD = 1.52). They felt somewhat calmer (M = 4.63) but with wide individual differences (SD = 2.36).
- **Human Condition:** Happiness levels were consistent (M = 6.83, SD = 1.57). Participants also reported a calmer state (M = 4.78), though variability remained high (SD = 2.38).

Overall, happiness levels remained consistently high across conditions, while calmness versus excitement varied among individuals. These findings suggest that, regardless of design origin, participants responded positively to the light fixture designs, with emotional reactions influenced more by individual differences than by condition.

H1: AI vs. Human Creativity

Statistical analysis from the MTurk sample (n = 120, N = 40 per condition) provided no significant evidence to support the hypothesis that AI is less creative than a human designer. Due to violations of homogeneity of variances, particularly in the 2D sketch (S) and 3D rendering (3D) stages ($p < .001$), Kruskal-Wallis tests were used instead of one-way ANOVA.

In the S stage, a Kruskal-Wallis test for novelty indicated significant differences between conditions ($X^2 = 18.21$, $p < .001$), but post-hoc tests showed no statistical difference between AI and human novelty scores ($p = .06$). Resolution and style showed no significant differences across conditions ($p = .35$ and $p = .12$, respectively).

In the 3D stage, AI outperformed the human designer in style, with a significant Kruskal-Wallis test ($X^2 = 6.87$, $p = .03$). A follow-up Mann-Whitney U test ($p = .02$) confirmed that AI scored significantly higher than the human designer in style (see Figure 2) with a medium effect size ($d = .60$). Novelty and resolution differences were not significant ($p > .10$).

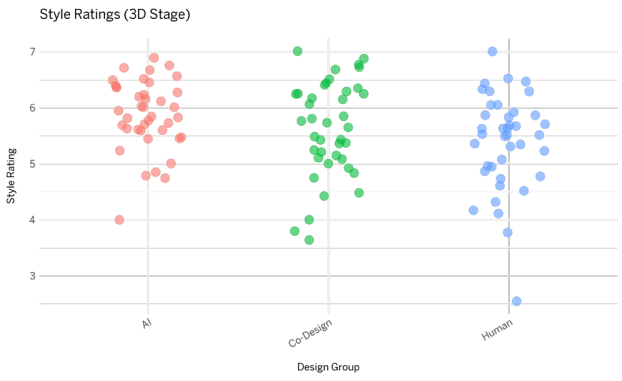


Figure 2. Mean-differences in style across three conditions.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

In the VR stage, all Kruskal-Wallis tests were insignificant ($p > .10$), indicating no meaningful differences between conditions in the final evaluation phase.

Overall, these findings contradict the assumption that AI is less creative than a human designer. AI performed comparably to human designers across most criteria and even surpassed them in style during the 3D stage.

H2: Co-Design vs. AI

Statistical analysis rejected H_2 and provided no evidence that human-AI co-design was more creative than AI alone. Instead, findings partially contradicted expectations, as co-design underperformed in novelty during the ideation (S) stage.

In the S stage, a Kruskal-Wallis test for novelty was significant, and a follow-up Mann-Whitney U test with Bonferroni adjustment showed that AI novelty scores (AI_N_M) were significantly higher than co-design (Co_N_M) ($W = 19.25$, $p < .001$, $d = 1.09$), indicating a large effect size (see Figure 3). Resolution and style showed no significant differences across conditions.

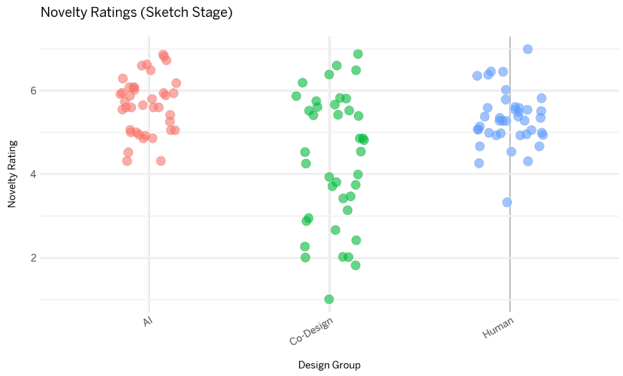


Figure 3. Mean-differences in novelty across three conditions.

In the 3D stage, a Kruskal-Wallis test for style was significant, but the follow-up test ($p = .6$) found no clear advantage for co-design over AI. All other CPSS ratings in 3D and VR stages were statistically insignificant, suggesting no meaningful differences between conditions.

Overall, co-design did not surpass AI in creativity and even scored lower in novelty during the ideation stage, rejecting the hypothesis that human-AI collaboration would yield more creative outcomes than AI alone.

H3: Co-Design vs. Human Designer

Statistical analysis rejected H_3 and provided no evidence that human-AI co-design was more creative than a human designer alone. Instead, findings partially contradicted expectations, as the human designer outperformed co-design efforts in novelty during the ideation (S) stage.

In the S stage, a Kruskal-Wallis test for novelty was significant, and a Mann-Whitney U test with Bonferroni adjustment showed that the human designer (Hu_N_M) scored significantly higher than co-design (Co_N_M) ($W = 1.34$, $p = .04$, $d = .8$), indicating a large effect size.

Across all other CPSS criteria and design stages (3D and VR), Kruskal-Wallis tests were insignificant, meaning there was no statistical advantage for co-design over human designers.

Overall, these results do not support the hypothesis that co-design is more creative than human effort alone. Instead, the human designer demonstrated stronger novelty in the ideation stage, further challenging assumptions about the benefits of AI-assisted collaboration in design.

Discussion

This study provides empirical insights into the evolving creativity rivalry between AI, human designers, and co-design efforts. Findings challenge the assumption that AI is inherently less creative than humans, revealing that AI-generated solutions performed comparably across multiple design criteria and even excelled in style during the 3D stage. Conversely, co-design efforts did not surpass AI or human designers, particularly struggling with novelty in the ideation stage. These results highlight AI's potential beyond simply generating variations—demonstrating its capability to rapidly visualize preliminary design concepts.

A key limitation of this study is the reliance on crowd-sourced CPSS ratings, without comparison to expert evaluations. Although non-parametric tests controlled for demographic disparities, the findings contrast with studies using expert judges (Zhou et al., 2023), raising questions about whether creativity assessments should integrate both crowd-sourced and expert perspectives. While prior research supports the convergence of large-scale crowd and expert ratings (Yuan et al., 2016; Foong et al., 2017), these discrepancies suggest that creativity evaluation methods must evolve—especially when assessing AI, co-design, and human-generated solutions.

Beyond methodological implications, this study underscores the uncertainty surrounding AI's creative capacity in design research. The results provide empirical support for AI's ability to independently generate novel outputs, challenging the perception that creativity is an exclusively human trait. More importantly, these findings call for a reconsideration of creativity itself: traditionally understood as a social construct (Amabile, 1983; Csikszentmihalyi, 1988) and assessed through established criteria (Horn & Salvendy, 2006; O'Quin & Besemer, 2006), creativity may now be shifting toward a hybrid concept—one that encompasses human-technology collaboration. As AI continues to shape creative disciplines, future research must explore how designers and educators can strategically integrate AI while preserving the originality and intentionality that define human creativity.

Conclusion

The insights from this research provide practical and pedagogical guidance for designers and educators seeking to integrate AI into the design process. Designers can leverage AI as a tool for rapid visualization, using its generative capabilities to explore diverse renderings and stylistic op-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

tions in the early stages of concept development. However, they must recognize that AI's novelty may plateau, necessitating human intervention for creative breakthroughs in later stages.

For educators, the challenge lies in teaching students to critically engage with AI-generated outputs, ensuring that AI serves as a catalyst for creativity rather than a crutch for repetitive ideation. Encouraging a balance between AI-assisted and manual ideation can help designers avoid fixation and refine their originality. Ultimately, creativity is a dynamic and evolving construct, shaped by technological advancements and peer evaluations of their novelty and relevance. By offering empirical evidence, this study positions AI as both novel and appropriate within the design process and the broader design community, reinforcing its role as a transformative yet complementary creative tool.

Acknowledgements

This research was supported by the Eskenazi Technology Innovation Lab (ETIL) via the 2024 Eskenazi Technology Innovation Grant.
The author conceptualized, designed, and conducted the study, and wrote the manuscript. Chat GPT-4 was utilized for refining grammar and wording.

References

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, *45*(2), 357-376. https://doi.org/https://doi.org/10.1037/0022-3514.45.2.357

Besemer, S. P. (2006). *Creating products in the age of design: How to improve your new product ideas!* New Forums Press.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59.

Csikszentmihalyi, M. (1988). The Systems Model of Creativity. https://doi.org/10.1007/978-94-017-9085-7

Horn, D., & Salvendy, G. (2006). Consumer-based assessment of product creativity: A review and reappraisal. *Human factors and ergonomics in manufacturing & service industries*, *16*(2), 155-175.

Horn, D., & Salvendy, G. (2009). Measuring consumer perception of product creativity: Impact on satisfaction and purchasability. *Human factors and ergonomics in manufacturing & service industries*, *19*(3), 223-240. https://doi.org/https://doi.org/10.1002/hfm.20150

Horton Jr, C. B., White, M. W., & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. *Scientific Reports*, *13*(1), 19001.

Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the power of data*. Wiley ; John Wiley [distributor].

Norman, D. A. (2005). *Emotional design: why we love (or hate) everyday things*. Basic Books. https://research.ebsco.com/linkprocessor/plink?id=7d4460c6-a3f2-3454-99f4-364e6e98a923

O'Quin, K., & Besemer, S. P. (2006). Using the creative product semantic scale as a metric for results-oriented business. *Creativity and Innovation Management*, *15*(1), 34-44.

Portman, M. E., Natapov, A., & Fisher-Gewirtzman, D. (2015). To go where no man has gone before: Virtual reality in architecture, landscape architecture and environmental planning. *Computers, environment and urban systems*, *54*, 376-384. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2015.05.001

qbiq. (2024). *Generative AI for Real Estate Layout Design & Virtual Tours*. https://qbiq.ai/

Rattay, S., Yurrita, M., Garnham, I., & Browne, J. T. (2022). Prototyping tensions: How to talk to your colleagues about AI. In D. Lockton, S. Lenzi, P. Hekkert, A. Oak, J. Sádaba, & P. Lloyd (Eds.), *DRS2022: Bilbao*, 27 June - 3 July, Bilbao, Spain. Design Research Society. https://doi.org/10.21606/drs.2022.909.

RStudioTeam. (2023). *RStudio: Integrated Development for R*. In RStudio, PBC. https://posit.co/download/rstudio-desktop/

TestFit, I. (2024). *A New Real Estate Feasibility Workflow*. TestFit. Retrieved 06/10 from https://www.testfit.io/

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*, 433-460.

Wei, X., Weng, D., Liu, Y., & Wang, Y. (2015). Teaching based on augmented reality for a technical creative design course. *Computers & Education*, *81*, 221-234. https://doi.org/10.1016/j.compedu.2014.10.017

Yazici, B., & Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, *77*(2), 175-183. https://doi.org/ 10.1080/10629360600678310

Zhou, C., Zhang, X., & Yu, C. (2023). How does AI promote design iteration? The optimal time to integrate AI into the design process. *Journal of Engineering Design*, 1-28. https://doi.org/10.1080/09544828.2023.2290915

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Embedding visual thinking into an AI-driven furniture design critiquing system

Vuong Nguyen and Gabriel Vigliensoni
Department of Design and Computation Arts
Concordia University, Montréal, QC
vuong.nguyen@mail.concordia.ca

Abstract

The *critique* is a cornerstone of design and arts education. In furniture design, the critique often takes the form of formal critique—an isolated exercise aimed at developing tacit formal awareness. However, in the landscape of AI-driven solutions for designers, content generation dominates over sensemaking and current critiquing systems remain ill-equipped due to their inability to sense, adapt, and question meaningfully. In this paper, we present fCrit, a critiquing chatbot that collaborates with human designers to reflect on the visual forms of furniture designs. Our work involved reconstructing the process behind enabling AI to analyze visual concepts and patterns using a hand-crafted expert knowledge base, and developing a functional, critique-ready prototype. We envision a critiquing chatbot that resonates with human creativity and empowers designers in form-finding. This work-in-progress is part of a broader initiative that aims to demonstrate how Human-Centred Explainable AI solutions can support creativity for artists and designers.

Introduction

In design education, the transformation of tacit knowledge into actionable insights is what the *critique* is all about. Critiquing is a craft that has always been associated with human sensibility, experience, and empathy (Goldschmidt, 2002). It is an essential practicum for pedagogical (Boyer and Mitgang, 1996; Schön, 1992a) and professional settings (Dreyfuss, 1955). To aid this process, *critiquing systems* emerged as a computer-aided design tool that assists human designers in making and analyzing design decisions by providing feedback (Robbins and Redmiles, 2000; Silverman, 1992).

We are motivated by the recent advent of generative AI in the design process, particularly regarding how designers engage with critique—the problem framing component of design thinking. The majority of generative AI solutions for designers prioritize content generation over other aspects of design sensemaking such as discovering, training, or testing (Shi et al., 2023). However, this generation-focused trend propels the fixation effect on designers (Wadinambiarachchi et al., 2024), impeding the critical and divergent aspects of design thinking. Furthermore, existing critiquing systems fall short for creativity support in furniture design due to their narrow and homogenizing feedback (Oh and Oh, 2017). These systems treat expert knowledge as a prescriptive rubric to distinguish right from wrong to identify improvements. This approach is not suitable for critiquing visual form as it prioritizes expert knowledge over creative spark (Goldschmidt, 2003). And because furniture design critiquing is both subjective and grounded in visual theory, it specifically necessitates a critiquing modality that honors both the *interpretive* and *formal-awareness* aspects of furniture design. Our research scope focuses primarily on these dimensions of critique.

In this paper, we propose fCrit, an AI-driven critiquing system for visual form of furniture design. fCrit counters the aforementioned shortcomings by embodying a pedagogical framework that facilitates formal insight discovery rather than error diagnosis, harnessing the creative potential of studio critique while providing the convenience of a personal support system. We present two main contributions: (1) a methodology for embedding design expert knowledge as a structured knowledge base, and (2) a prototype for fCrit with a multi-agent architecture that utilizes this knowledge base. In the next section, we provide references to related work in the areas of critiquing systems, AI-driven tutoring systems, and design learning, followed by the theoretical framework that supports the development of fCrit. We then detail the architecture, components and pipeline used in our system and we elaborate on our structured knowledge base—our first contribution. Finally, we present an interaction with dialogue results that demonstrate the system in action—our second contribution. We conclude the paper with our future work and a reflection on the implications for AI-driven design learning systems.

Related Work

Critiquing systems are computer-aided design tools that evaluate design solutions and support designers' task performance (Robbins and Redmiles, 2000; Silverman, 1992). These systems typically include an analyzer module that compares user input to a reference knowledge base, and a dialogue module that identifies errors (Silverman, 1992).

Despite broad adoption in domains ranging from design sketching (Oh, Do, and Gross, 2004) to architectural layout planning (Chun and Ming-Kit Lai, 1997; Woodbury and Drogemuller, 2000), these systems often overlook critical aspects of creativity support. Key considerations include

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Figure 1: Gebrüder Thonet's M-209 bentwood armchair (circa 1900), exemplifying curvilinear elements that designers could describe as "noodle-y" or "playful." This piece serves as a reference point in fCrit's dialogue demonstrations, illustrating how the system connects intuitive observations to structured formal concepts. https://www.thonet.de/en/all-products/detail/209.

feedback modality (Oh and Oh, 2017), cognitive ergonomics (Tianfield and Wang, 2004), and contextual awareness (Fischer, 2012).

Recent studies have explored integrating tutoring system capabilities enhanced by large language models (LLMs) (Ashfaque et al., 2020; Ait Baha et al., 2024). However, these remain ill-equipped for the deeper forms of learning required in design education. This gap underscores the opportunity for a system tailored to design learning, with a focus on *adaptive design thinking* (Sharp and Macklin, 2019), *deep reflection* (Redström, 2017), and *problem framing* (Cross, 2006; Getzels and Csikszentmihalyi, 1976).

With these considerations, we employ Gebrüder Thonet's M-209 armchair (Fig. 1) as the reference piece to elucidate key points in the following sections.

Theoretical Framework

To support adaptive design thinking, reflective learning, and problem framing, we developed a theoretical framework that integrates three complementary models to shape the user experience in fCrit:

Reflective Learning: This component draws on Schön (1992b) theory of reflection-in-action and Moon (2013) model of structured reflection. Rather than prescribing solutions, fCrit is designed to amplify the designer's own observational process. It scaffolds both spontaneous and deliberate reflection, positioning critique as a site of sensemaking rather than evaluation. This aligns with work in creative cognition that emphasizes iterative reasoning and metacognitive engagement in design practice.

Mixed-Initiative Interaction: The system adopts a mixed-initiative paradigm (Allen, Guinn, and Horvtz, 1999) to facilitate a dialogic relationship between humans and AI. Rather than issuing corrections, fCrit engages users in multi-turn, context-sensitive conversations. This model has seen success in creative domains such as game design (Bueno Perez and Bidarra, 2022) and generative furniture design (Buonamici et al., 2020), offering an interaction structure well suited to exploratory and co-creative tasks.

Formal Critique Vocabulary: To enable design discussions based on real-world practice, fCrit employs a vocabulary derived from formal critique traditions (Dondis, 1974; Wong, 1993; Hannah, 2002). The knowledge base encodes relationships among elements such as line, shape, form, and space, making them accessible at multiple levels of abstraction and formal awareness. This structured lexicon allows the system to link user observations to established design principles, fostering visual literacy without constraining creative freedom. An important feature of formal critique is that it does not concern itself with regional-stylistic or period-based labelling (e.g., Scandinavian or Victorian). Instead, it focuses solely on the perceptual effect of pure form.

Together, these three model components address the gaps in past critiquing systems, positioning fCrit not as a rule-based assessor, but as a reflective partner in design critique. The system supports the emergence of creative insight through dialogue, aiming to foster a sense of mutual attentiveness between designer and AI. This human-centred approach enables the system to contribute without imposing, sustaining the designer's creative agency while enriching the discursive depth of critique.

fCrit: Furniture Design Critic

Inspired by the progressive depths of reflection discussed in Hatton and Smith (1995) and Moon (2013)—who categorized reflection into descriptive, reflective, dialogic, and critical levels—we designed fCrit to support reflection and theory application at varying depths, tailored to users' needs. The system operates across multiple levels of reflexivity: at the foundational level, it mirrors the user's observations to invite further reflection. As dialogue deepens, it introduces design terminology, identifies patterns and relationships between elements, and ultimately contextualizes observations within broader design implications. This progressive capability aims to support users without overwhelming them with excessive theoretical complexity. To enable this functionality, the chatbot incorporates augmentative tools to: (1) analyze and detect user intent to determine and define a communication path; (2) match the user's language and formal awareness level; and (3) access a knowledge base of furniture-specific visual concepts and patterns.

Critique Knowledge Base

fCrit's knowledge base is built specifically for formal critiquing and is constructed in such a way that captures each visual *concept* and *pattern* in a single structured embedding. Each embedding contains all the necessary contexts for the LLM to understand the nuance and relations with

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

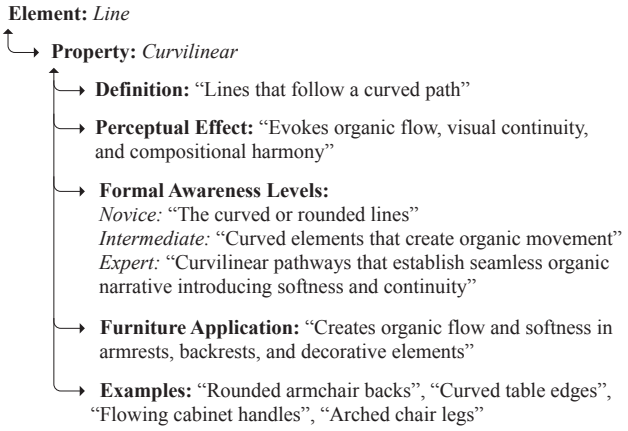


Figure 2: Hierarchical representation of Curvilinear Line in fCrit's knowledge base, showing how visual concepts are structured with formal definitions, perceptual effects, awareness-tailored language, applications, and examples to support adaptive design critique.

other concepts. The concepts and patterns are derived from umbrella parents such as *Line* into more nuanced concepts (e.g., curvilinear line, diagonal line, etc.). As shown in Fig. 2, within each of these conceptual embeddings are formal definition, perceptual effects, awareness-tailored language, domain-specific applications, and concrete examples. This structured approach enables the system to tailor explanations according to the user's formal awareness level while maintaining consistent underlying design principles.

The knowledge base is implemented using Supabase Vector Store,[1] an open-source Firebase alternative that provides a dedicated PostgreSQL database with vector search capabilities. This setup allows the LLM to access and reason over the knowledge base in an intuitive, multi-layered way. For example, if a designer describes Thonet's M-209 chair in Fig. 1 as "noodle-y" and "playful", the system guesses that they may be talking about the dynamic curvilinear elements of the chair. This allows the system to surface and expand on the underlying visual insight. In the Dialogue Demonstration section, we illustrate how these interaction levels work in practice, demonstrating fCrit's ability to both adapt to a designer's evolving observations and stimulate deeper engagement with visual forms.

System Architecture

fCrit is implemented using n8n,[2] a workflow automation platform that enables visual workflows for connecting various API endpoints as modular "nodes." Within n8n, fCrit orchestrates an agentic workflow comprising five specialized components: the Command Hub, Design Concept Mapper, Pattern Recognition Engine, Etiquette Classifier, and Dialogue Agent. As shown in Fig. 3, the system operates across three tiers—input processing, knowledge access and adaptation, and dialogue generation—each supported by task-appropriate language models. To maintain focus and interpretability, fCrit retrieves knowledge from its structured base on a just-in-time basis, introducing information only when relevant to the ongoing interaction. Each retrieval is accompanied by a confidence score, enabling system-level introspection and iterative refinement of concept mappings. The architecture is optimized for co-creative engagement through distributed reasoning and modular control. For language processing, we use different Anthropic's Claude models, depending on each agent's need for speed, interpretive nuance, and cost-efficiency. System prompts are carefully structured to ensure consistency across interactions.

Command Hub. This is the central entry point for all user input and orchestrates the flow of information across the system. It serves as the coordination layer that routes each user message through a selection protocol to determine relevant design concepts, visual patterns, and shifts in user intention.

Upon receiving input, the Command Hub outputs a set of boolean flags in JSON format that indicate whether any of these dimensions are present. These flags define a focused processing context for the subsequent components, ensuring each downstream module operates only when needed.

Designed for robust reasoning and fast turnaround, this component leverages Claude 3.5 Haiku to balance speed with nuanced understanding. Its tiered activation strategy enables the system to dynamically adjust computational resources, invoking only necessary tools per interaction. In doing so, the Command Hub ensures efficient, context-aware orchestration across the multi-agent architecture.

Pattern Recognition Engine. This component detects visual patterns. It works with the user observation and the knowledge base to suggest implied visual patterns (i.e., unity, contrast, repetition, alignment, and progression; each with sub-patterns such as rhythmic repetition, size progression, etc.). It detects mentions of visual pattern or relationship in user input and determines the pattern implications or suggestions. As one of the two components that has access

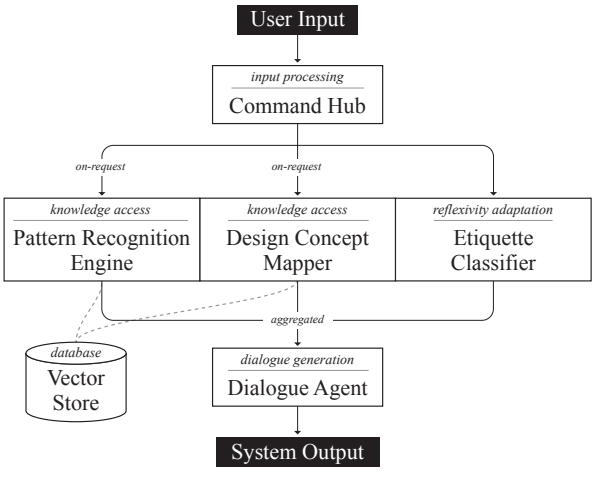


Figure 3: fCrit architecture: Five specialized AI agents work in a three-tiered workflow to process user input, access design knowledge, and generate furniture critiquing dialogues with adaptive levels of reflexivity. Two of them gain on-request access to a domain-specific knowledge base.

[1] https://github.com/supabase/supabase

[2] https://github.com/n8n-io/n8n

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

to the design knowledge base, it assigns a confidence score to every extraction (i.e., if the score is low, we can identify potential missing concepts in the knowledge base).

Design Concept Mapper. This agent translates natural language observations into formal design critique terminology. It identifies design elements in the user input and extracts visual concepts (i.e., line, space, form, shape; each with sub-concepts such as diagonal line, negative space, etc.) from the knowledge base. Similarly to the Pattern Recognition Engine, it also assigns a confidence score to each extraction from the knowledge base. In terms of enhancing relevance and focus, it tailors its outputs according to user language and formal awareness levels and acts as a conceptual guardrail to prevent unnecessary expansions to other concepts. Both components in this parallel subworkflow perform simple natural language processing tasks, so we use Claude 3.5 Haiku to ensure cost-effective, rapid parsing of conceptual cues.

Etiquette Classifier. This is a conversation classifier that analyzes user messages to determine appropriate language mode. There are three modes: casual, detailed, and deep. This capability is essential to reflective learning and Mixed-Initiative Interaction as it fosters positive emotional states as it reinforces a shared language space. This component analyses user input to determine the critique's language mode, recommends response length and formality based on input complexity, optimizes the timing of probing questions, and packages this information for aggregation down the pipeline.

Dialogue Agent. This module synthesizes outputs from all prior components to produce content-aware replies. It progressively scaffolds *Reflective Learning* by activating different depths of reflectivity through adaptive, critique-focused dialogue.

When engaging with users, the agent connects their observations to design theory using *Formal Critique Vocabulary*. For example, if a designer describes a chair like the one in Fig. 1 as "noodle-y, playful, and elegant", the Dialogue Agent works to discover and expand the connections between these observations and relevant formal theory.

The component utilizes three tools to accomplish this task. First, it rephrases user input using design-related language to nudge further elaboration from the designer (e.g., "I can see why..."). Second, it poses generative questions to focus the designer's attention on the *why* behind their observations, promoting convergent thinking (e.g., "What makes you think...?", or "Do you notice..."). Third, it creates visual analogies bridging to everyday concepts to probe divergent thinking (e.g., "The flowing lines resemble...").

Before delivering the final response, the agent implements a quality self-validation process, checking for: user-centeredness (reflecting observations, following user focus), appropriate design knowledge integration (terminology matching user formal awareness level), and conversation naturalness (appropriate length, flow, and questioning). To handle these linguistically demanding tasks, the system employs Claude 3.7 Sonnet, which offers a larger context window and improved discourse handling. This enables the system to track ongoing conversational themes, maintain continuity, and facilitate evolving levels of formal critique.

The integration of our formal critiquing-specific knowledge base with our retrieval-based architecture delivers the key user experience benefits outlined earlier. By focusing on a single scope of exploration at a time, this gradual approach to explainability enhances the designer's reflective experience rather than overwhelming them with too many perspectives. Furthermore, it establishes a shared mental space for human-AI mixed-initiative creative exchanges where the AI gains additional context about the ongoing observation through every dialogue turn, further building collaboration and trust through human-centered dialogue techniques. Throughout the process, fCrit preserves the designer's creative agency by following their lead and introducing relevant design insights only when appropriate.

Dialogue Demonstration

The following dialogue presents one interaction with fCrit over four turns, using Fig. 1 as the reference piece for critique. These inputs were carefully selected to trigger its critiquing capabilities to effectively support reflective learning.

Designer: Hi, I like this design *(with Fig. 1 attached)*.

fCrit: What specifically attracts you to this chair design? I notice it has those beautifully bent wood elements that create a continuous flowing line from the backrest to the armrests. The natural light wood tone gives it a timeless quality too.

Designer: Well I'm drawn to how noodle-y it looks. It's playful yet elegant!

fCrit: I love your description of it as "noodle-y"! That's a great way to capture those continuous curved lines. The way the wood bends and flows without interruption creates that playful character you've noticed, while still maintaining an elegant presence. Do you think it's the smoothness of the curves or the rhythmic repetition of the lines that gives it that playful quality?

Designer: I think it's in the way the curved lines traverse across different parts of the chair, tying it together to create a story. What do you think about this linear narrative?

fCrit: That's very perceptive! Those curved lines do more than just connect physical elements—they create a visual journey through the chair's form. The way each curve flows into the next creates a sense of continuity that tells a story of movement and rhythm. It's almost as if the designer wanted us to experience the chair as a single gesture rather than separate parts. The curves also guide the eye around the entire piece, revealing how intentionally unified the design is. Do you feel this narrative quality makes the chair more inviting or more sculptural in nature?

Designer: This narrative definitely invites closer inspection. There's something else about the nature of curve that feels amiable. I can't quite find the right word.

fCrit: There's a certain warmth and hospitality in organic curves that straight lines often lack. Perhaps what you're sensing might be described as "haptic invitation"—the way these gentle curves seem to welcome touch and interaction. The curves have a natural ergonomic quality that resonates with our bodies and movement patterns, creating an immediate sense of comfort even before physical contact.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

> The gradual, predictable nature of these curves also creates visual ease—our eyes can follow them without abrupt interruptions, which might contribute to that feeling of amiability you're describing. Does that resonance with human movement patterns and visual comfort capture some of what you're feeling about these curves?

fCrit employs a multitude of dialogue techniques and critiquing initiatives not only to develop reflexivity, but also to build trust, foster positive emotion, and expand perspectives—which serve as auxiliary components to ease reflective learning (Moon, 2013).

The first turn demonstrates fCrit's *question-first protocol* when the designer's interests are unclear or for new topics, prioritising understanding what the designer likes before offering analysis. This intentionality is reinforced by *response-length control* which prevents the system from dominating initial dialogues while still advancing the critique. The second turn presents the system's capability to use *designer-centered language* to elaborate on the designer's previously expressed interest rather than imposing new directions. The final two turns showcase fCrit's *grounded interpretation* which validates and builds upon the designer's characterizations and metaphors and the use of *open and gauging questions* to encourage the designer's own reflections or feedbacks on whether the system response is relevant or helpful.

In essence, fCrit utilizes formal critique as a vehicle for reflective learning, prioritising the subtle unpacking of observations into reflective insights over the prescriptive components of critique such as regional-stylistic or period-based labelling. fCrit more closely aligns with design practitioners' need to interpret and internalize how pure forms work by scaffolding for design vocabulary and concepts that might help the designer articulate their own thoughts.

Conclusion & Future Work

The planning for further development of fCrit focuses on two key areas. First, we conduct a user experiment with both novice and expert furniture designers to gauge the system's helpfulness on design reflective practice. Then, we analyze the results of this experiment to refine and improve fCrit's critiquing capabilities.

The gap between existing AI solutions and creative needs presents a unique opportunity for design learning systems. To echo Verganti, Vendraminelli, and Iansiti (2020), current innovation in problem-finding is severely limited as AI solutions orient towards problem–solving instead. fCrit tackles this issue with a specialized AI agent that helps develop visual thinking in designers, aiming to deepen design sensemaking through reflective learning.

Our system design employs various prompting and component routing techniques to augment the system's capacity to maintain human-centredness, critique-focus, and relevance. This approach is supported by a methodology for constructing a knowledge base that organizes furniture design concepts with layers of depth and interpretations. The result is a multi-layered dialogue system capable of reflective design critique that mirrors critical aspects of traditional critiquing without attempting to replicate it.

Our research on fCrit has opened new avenues and sparked conversations for future chatbot and explainable systems beyond furniture design critique. As design thinking continues to evolve alongside technological capabilities, problem-finding oriented systems like fCrit could transform how designers engage with both their creative process and the theoretical foundations that inform their practice.

References

Ait Baha, T.; El Hajji, M.; Es-Saady, Y.; and Fadili, H. 2024. The impact of educational chatbot on student learning experience. *Education and Information Technologies* 29(8):10153–10176.

Allen, J.; Guinn, C.; and Horvtz, E. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14(5):14–23.

Ashfaque, M. W.; Tharewal, S.; Iqhbal, S.; and Kayte, C. N. 2020. A review on techniques, characteristics and approaches of an intelligent tutoring chatbot system. In *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, 258–262.

Boyer, E. L., and Mitgang, L. D. 1996. *Building Community: A New Future for Architecture Education and Practice. A Special Report*. California Princeton Fulfillment Services.

Bueno Perez, M. R., and Bidarra, R. 2022. Mixed-initiative story co-creation with talemaker. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, FDG '22, 1–13. New York, NY, USA: Association for Computing Machinery.

Buonamici, F.; Carfagni, M.; Furferi, R.; Volpe, Y.; and Governi, L. 2020. Generative design: An explorative study. *Computer-Aided Design and Applications* 18:144–155.

Chun, H. W., and Ming-Kit Lai, E. 1997. Intelligent critic system for architectural design. *IEEE Transactions on Knowledge and Data Engineering* 9(4):625–639.

Cross, N. 2006. *Designerly Ways of Knowing*. Springer Science & Business Media.

Dondis, D. A. 1974. *A Primer of Visual Literacy*. MIT Press.

Dreyfuss, H. 1955. *Designing for People*. Simon and Schuster.

Fischer, G. 2012. Context-aware systems: The 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, 287–294. ACM.

Getzels, J. W., and Csikszentmihalyi, M. 1976. *The creative vision: A longitudinal study of problem finding in art*. Wiley.

Goldschmidt, G. 2002. 'One-on-one': A pedagogic base for design instruction in the studio. *DRS Biennial Conference Series*.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Goldschmidt, G. 2003. Expert knowledge or creative spark? Predicaments in design education. *Expertise in Design* 221–233.

Hannah, G. G. 2002. *Elements of Design: Rowena Reed Kostellow and the Structure of Visual Relationships*. Princeton Architectural Press.

Hatton, N., and Smith, D. 1995. Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education* 11(1):33–49.

Moon, J. A. 2013. *A Handbook of Reflective and Experiential Learning: Theory and Practice*. London: Routledge.

Oh, Y., and Oh, Y. 2017. A computational model of design critiquing. *Artificial Intelligence Review* 48.

Oh, Y.; Do, E.; and Gross, M. 2004. Intelligent critiquing of design sketches. In *AAAI Fall Symposium - Making Pen-Based Interaction Intelligent and Natural, 2004*.

Redström, J. 2017. *Making Design Theory*. The MIT Press.

Robbins, J. E., and Redmiles, D. F. 2000. Cognitive support, UML adherence, and XMI interchange in Argo/UML. *Information and Software Technology* 42(2):79–89.

Schön, D. A. 1992a. Design as a reflective conversation with the situation. In *The Reflective Practitioner*. Routledge.

Schön, D. A. 1992b. *The Reflective Practitioner: How Professionals Think in Action*. Routledge.

Sharp, J., and Macklin, C. 2019. *Iterate: Ten Lessons in Design and Failure*. MIT Press.

Shi, Y.; Gao, T.; Jiao, X.; and Cao, N. 2023. Understanding Design Collaboration Between Designers and Artificial Intelligence: A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 7(CSCW2):368:1–368:35.

Silverman, B. G. 1992. Survey of expert critiquing systems: Practical and theoretical frontiers. *Communications of the ACM* 35(4):106–127.

Tianfield, H., and Wang, R. 2004. Critic systems: Towards human-computer collaborative problem solving. *Artificial Intelligence Review* 22(4):271–295.

Verganti, R.; Vendraminelli, L.; and Iansiti, M. 2020. Innovation and design in the age of artificial intelligence. *Journal of Product Innovation Management* 37(3):212–227.

Wadinambiarachchi, S.; Kelly, R. M.; Pareek, S.; Zhou, Q.; and Velloso, E. 2024. The effects of generative AI on design fixation and divergent thinking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–18. New York, NY, USA: Association for Computing Machinery.

Wong, W. 1993. *Principles of Form and Design*. Van Nostrand Reinhold.

Woodbury, R., and Drogemuller, R. 2000. Code checking by representation comparison. In *Computer Aided Architecture Design Research in Asia (CAADRIA)*, 235–244.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Fairness as a Creative Resource: Challenges and Opportunities in Creative Computing

Ricardo Trainotti Rabonato
Computer Science Department
Federal University of Sao Pauloy
Sao Jose dos Campos, SP, Brazil
trainotti.ricardo@unifesp.br

Lilian Berton
Computer Science Department
Federal University of Sao Paulo
Sao Jose dos Campos, SP, Brazil
lberton@unifesp.br

Abstract

Artificial Intelligence (AI) and generative models have expanded the landscape of computational creativity. While fairness is often framed as a corrective mechanism to mitigate social and cultural bias in AI systems, this paper explores an alternative perspective: fairness as a creative affordance capable of enhancing the aesthetic and cultural diversity of AI-generated content.

We propose a theoretical framework that repositions fairness as a potential catalyst for creativity and investigate this idea through a set of exploratory prompting experiments. Drawing on the few-shot learning capabilities of large language models, we examine whether minimal prompt modifications guided by fairness principles can lead to more diverse, inclusive, and contextually rich narratives.

Preliminary results suggest that fairness-oriented prompts may encourage generative systems to deviate from normative patterns, resulting in outputs with increased lexical variety, broader representation, and greater narrative complexity. While the study is limited in scale and scope, it opens new directions for designing generative systems that integrate ethical considerations as expressive constraints, fostering both innovation and cultural awareness in computational creativity.

Introduction

Creativity is a central theme in artificial intelligence, particularly in applications involving language, narrative, and symbolic expression. While large language models (LLMs) have made significant strides in generating fluent and coherent texts, guiding these systems toward outputs that are not only plausible but also original, expressive, and culturally situated remains a major challenge.

In the field of algorithmic fairness, efforts have traditionally focused on mitigating social and cultural biases in automated decision-making systems (Barocas, Hardt, and Narayanan 2023; Rabonato and Berton 2024). Fairness, in this view, operates primarily as a corrective mechanism. This paper proposes an alternative perspective: that fairness may also serve as a creative constraint — a guiding principle capable of expanding the narrative and aesthetic repertoire of generative models.

This proposition draws from theories in computational creativity that emphasize the generative role of constraints. Duch (2006) argues that creativity arises from the interaction between imagination and filtering mechanisms, whereby filters — including social, semantic, or aesthetic constraints — shape the emergence of novel combinations. From this viewpoint, fairness can act not only as a normative boundary, but as a stimulus for generating unexpected and meaningful associations.

The approach also aligns with Wiggins' (2006) formalization of Boden's concepts of exploratory and transformational creativity. Wiggins posits that a creative system can operate within a conceptual space (exploration) or challenge and expand the boundaries of that space (transformation). Fairness, as a constraint that destabilizes dominant narrative patterns, may thus operate as a trigger for transformational creativity — encouraging models to generate content that deviates from statistically dominant structures and engages with culturally diverse contexts (Wiggins 2006).

To explore this hypothesis, we conducted a set of small-scale prompting experiments involving narrative generation. Inspired by few-shot learning capabilities, we tested whether minimal modifications to input prompts — informed by fairness principles — could lead to more diverse and imaginative outputs. While the scope of this study is limited, preliminary results suggest that fairness-aware prompts may foster narrative complexity and symbolic richness, supporting the notion that ethical constraints can function as productive forces in creative AI systems.

Related Work

Fairness in Language Models

Recent research has investigated various strategies for mitigating bias in large language models (LLMs). Gallegos et al. (Gallegos et al. 2024) offer a comprehensive review and taxonomy of bias types and mitigation methods in NLP, while Doan et al. (Doan et al. 2024) synthesize recent advances with a focus on fairness considerations specific to the architecture and functioning of LLMs.

Despite progress in newer model versions, implicit biases remain persistent and highly sensitive to language, task formulation, and prompt structure (Torres et al. 2024). In this context, prompting techniques have emerged as a lightweight yet effective strategy for influencing model behavior. Zero-shot and few-shot prompting — including

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

chain-of-thought (CoT) reasoning — have been explored as means to induce fairness at inference time, without requiring model retraining (Zeng, Chung, and Zhou 2024). Other methods include architectural interventions and modifications to training objectives, such as fairness-aware loss functions (Wang, Li, and Zhang 2024; Atwood et al. 2024).

Bias and Computational Creativity

As LLMs are increasingly used in creative applications, questions about bias have gained relevance within the field of computational creativity. Traditionally, bias has been treated as an obstacle to be corrected, particularly from the standpoint of algorithmic fairness. However, some authors argue for a more nuanced perspective. Loughran (2022), for instance, suggests that biases — understood as tendencies, preferences, or learned associations — may be essential for creative cognition, acting as scaffolds for generating novelty amid otherwise conventional outputs (Loughran 2022).

This view intersects with broader discussions about the generative role of constraints in creativity. Constraints are often considered central to the emergence of innovation, especially when they provoke deviation from normative patterns (Lamb, Brown, and Clarke 2018). In this light, fairness can be interpreted not merely as a regulatory demand, but as a productive constraint that guides generative systems toward richer expressive spaces.

Brown and Ventura (2022) further argue that ethical and aesthetic dimensions can be co-encoded in computational creativity systems, with normative principles like fairness functioning as aesthetic filters that expand the range of acceptable outputs (Brown and Ventura 2022). Drawing on algorithmic information theory, they propose metrics such as logical depth and sophistication to assess creative value, suggesting that fairness constraints may contribute to increased complexity and originality in generative results.

Finally, Déguernel and Sturm (2023) conducted a systematic review of empirical studies on bias in the evaluation of computational creativity, concluding that human judgments are deeply shaped by sociocultural and contextual cues (Déguernel and Sturm 2023). Their findings support the idea that fairness-aware generation not only alters content characteristics, but also reshapes audience perception and reception — reinforcing the potential of fairness to function as both ethical and creative modulation.

Methodology

The experiments were conducted using a large language model (LLM), accessed through the free commercial version of the Claude model (Anthropic) via a web interface. It was not possible to adjust parameters such as temperature or output length. This limitation reinforces the focus on the semantic impact of the prompts and avoids technical interventions that could compromise the comparative analysis.

For this study, three sets of narrative prompt pairs were created. Each pair consists of a neutral version and a version with an explicit fairness orientation, by adding the instruction: "The narrative should show concern for justice and diversity." The prompt pairs used were as follows:

- **1a.** *Write a short story about an ordinary person who accidentally discovers a way to travel through time.*
- **1b.** *Write a short story about an ordinary person who accidentally discovers a way to travel through time. The narrative should show concern for justice and diversity.*
- **2a.** *Write a short story about a child who finds a mysterious object in their neighborhood.*
- **2b.** *Write a short story about a child who finds a mysterious object in their neighborhood. The narrative should show concern for justice and diversity.*
- **3a.** *Write a short story about a character attending a party in an unfamiliar place.*
- **3b.** *Write a short story about a character attending a party in an unfamiliar place. The narrative should show concern for justice and diversity.*

These prompts explore situations marked by the unexpected and the fantastic within everyday contexts, allowing for diverse narrative approaches. The first pair leans toward science fiction and adventure; the second evokes a playful tone, close to a childlike universe; and the third opens space for variations in style and atmosphere, ranging from the fantastic to the introspective, depending on the generative model's interpretation[1].

To assess creativity and ethical expressiveness, we followed the framework proposed by Lamb et al. (2018), which defines three central dimensions of creativity: *novelty*, *quality*, and *surprise*. These criteria served as the basis for the design of a structured evaluation instrument intended for future reader-based studies. The instrument comprises twelve items organized into four thematic blocks: (1) *fairness and representation*, (2) *creativity*, (3) *coherence and clarity*, and (4) *empathy and tone*.

The creativity block directly operationalizes the three core dimensions: originality of ideas and scenarios (*novelty*), presence of unexpected elements or plot twists (*surprise*), and formal-aesthetic consistency (*quality*). The additional blocks aim to capture ethical, affective, and structural dimensions of narrative impact, offering a broader perspective on cultural representation and reception.

Although no human evaluations were conducted at this stage, the instrument offers a foundation for future studies involving reader response. This choice reflects the interpretative nature of creativity, as discussed by Duch (2006) and Loughran (2022), and also acknowledges the limitations of automated metrics in capturing aesthetic and cultural nuance.

This approach aligns with the interpretative nature of creativity, as discussed by Duch (2006) and Loughran (2022), and also acknowledges the limitations of automated metrics in assessing aesthetic and cultural elements. The use of *fairness prompts* aims to provoke changes in the model's narrative behavior, encouraging outputs that are more expressive, elaborate, and less aligned with conventional patterns.

[1]The original texts generated from these prompts in Portuguese can be accessed at: `[https://bit.ly/supplementaryMaterial]`.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Discussion

The results suggest that simple interventions at the prompt level — specifically, the inclusion of fairness-oriented instructions — produce narratives that are not only longer, but also thematically richer, more diverse in character representation, and more attentive to social dynamics. This observation reinforces the central hypothesis of this work: that fairness, when framed not as a corrective post-processing constraint but as a generative input condition, can function as a creative catalyst. This aligns with previous arguments in the literature that emphasize the productive role of constraints in creative systems (Lamb, Brown, and Clarke 2018; Loughran 2022).

In the context of large language models (LLMs), this phenomenon can be interpreted as a semantic displacement within the model's latent space. By embedding ethical instructions into the prompt — such as a concern for justice and diversity — the generation process is nudged away from the high-frequency narrative templates statistically encoded in the training data. This shift disrupts the generative model's default behavior, reducing the influence of overrepresented cultural frames and opening space for less probable but potentially more imaginative combinations. In this sense, the prompt acts as a vector of conceptual divergence, prompting the model to access underutilized regions of its associative repertoire.

The observed increase in novelty and complexity is consistent with theories of computational creativity that view creative output as the product of navigating and transforming conceptual spaces (Wiggins 2006). By reframing fairness not as a filter applied to outputs, but as a transformation applied to the generative process itself, we reveal a new operational role for ethical constraints: as mechanisms of exploratory expansion. That is, fairness functions not only to prevent harm or bias, but to provoke narrative invention and aesthetic deviation from conventional paths.

Moreover, the interaction between fairness and creativity supports the notion of *constructive perturbation*. As the model integrates ethical concerns into narrative construction, it becomes more likely to introduce atypical characters, unexpected social contexts, and alternative resolutions — elements that contribute to the criteria of surprise and originality, two pillars of the creative framework adopted in this study. In technical terms, this suggests that fairness-oriented prompts modulate the decoding trajectory by activating latent associations that are otherwise suppressed under default optimization strategies.

This approach also resonates with the perspective of Brown and Ventura (2022), who argue that normative principles such as fairness can be encoded as aesthetic heuristics. In our results, prompts that asked for diversity and justice consistently resulted in outputs with greater emotional depth and cultural nuance. These findings suggest that ethical constraints can serve a dual function: guiding models toward socially responsible behavior and simultaneously enriching their expressive capacity.

From a methodological standpoint, this study contributes to ongoing debates on prompt engineering by demonstrating that minimal semantic interventions can yield disproportionately meaningful changes in output. This supports the viability of fairness-aware prompting as a lightweight, interpretable, and scalable strategy for shaping generative behavior — particularly in domains where creativity, cultural sensitivity, and user trust intersect.

It is important to note, however, that these findings are not generalizable across all genres, domains, or model families. The specific outcomes observed here depend on the interplay between the prompt, the model's training data, and the cultural priors embedded in its parameters. Future research should explore how different formulations of fairness (e.g., intersectionality, equity, inclusion) affect narrative generation, and whether these effects are consistent across languages, tasks, and audience expectations.

Finally, the results invite a broader reflection on the coevolution of ethical and creative capacities in generative AI. If fairness can indeed serve as a creative stimulus, then the development of ethically informed models need not be framed solely in terms of restriction or correction. Instead, it opens the possibility of reimagining AI creativity as an inclusive and dialogic practice — one that does not merely avoid harm, but actively contributes to the construction of richer, more plural symbolic worlds.

Exploratory Quantitative Analysis

To establish an initial comparative basis, we conducted a quantitative analysis of the generated texts, focusing on structural and surface-level features. Four metrics were considered: total word count, number of named characters, lexical variety (measured by the Measure of Textual Lexical Diversity, MTLD), and the number of identity or cultural diversity markers.

Named characters were identified using rule-based detection of proper nouns and character-referential phrases, manually verified to ensure accuracy. Diversity markers were annotated based on the presence of references to ethnicity, gender identity, religion, social class, age, and other cultural identifiers explicitly mentioned or strongly implied in the narrative.

MTLD, as described in (McCarthy and Jarvis 2010), is a metric designed to assess lexical diversity — the degree to which a text uses varied vocabulary. Unlike simpler measures such as the type-token ratio (TTR), MTLD is less sensitive to text length, making it more suitable for comparing texts of different sizes. It calculates the average number of words needed for the type/token ratio to reach a fixed threshold (typically 0.72). Higher MTLD scores indicate greater lexical variety.

Table 1 presents a summary of the analyzed texts.

While the scope of this pilot study is limited, some tentative patterns emerge. In each case, the fairness-oriented text (b) features more named characters, higher identity or cultural marker density, and — in most cases — greater lexical diversity. These patterns suggest a possible link between fairness-aware prompting and narrative expansion.

For example, in pair 1, *Seu Jonas's Clock* presents a higher MTLD (183.6 vs. 176.5), includes three named characters (versus one in *The Watchmaker*), and introduces five explicit diversity markers. While the first story centers on

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Text	Words	Characters	MTLD	Markers
1a	597	1	176.5	0
1b	617	3	183.6	5
2a	364	2	135.5	1
2b	528	3	152.9	5
3a	464	1	146.7	1
3b	551	6	169.6	14

Table 1: Quantitative analysis and diversity markers in the generated stories

a generic protagonist in a neutral, introspective setting, the second is grounded in a social context, incorporating elements of inequality, memory, and community agency.

A similar contrast appears in pair 2: *The Time Medallion* develops a concise fantasy narrative with limited diversity (MTLD of 135.5; two characters; one marker), whereas *Sofia and the Pendulum of Time* presents a more elaborate plot with greater cultural engagement (MTLD of 152.9; three characters; five markers), addressing gentrification and historical repair.

Pair 3 exhibits the most pronounced difference: *The Party at the End of the World* is a metaphysical allegory with one character and a single abstract diversity reference, while *A Party in the Village of Colors* introduces six named characters and 14 diversity markers across multiple dimensions (ethnicity, religion, gender, age, etc.), achieving an MTLD of 169.6.

Although the numerical sample is small, these differences suggest that ethical prompting may encourage the model to activate a broader symbolic and social repertoire. Importantly, in fairness-aware texts, diverse characters frequently assume central narrative roles, such as protagonists, mentors, or agents of transformation — which may indicate not only increased representation but also richer plot structuring.

It is important to note that MTLD, while useful for indicating lexical variety, is not a direct measure of creativity. Future work will require more robust metrics — including semantic coherence, novelty, and human-rated creativity — to support stronger empirical claims. Nonetheless, these preliminary results reinforce the conceptual hypothesis that fairness can operate as a productive constraint, guiding the generation process toward more inclusive and narratively complex outputs.

Conclusion

This paper proposed a reframing of fairness in generative systems — not merely as a corrective mechanism against bias, but as a potential catalyst for creative diversity. Drawing on a set of exploratory narrative generation experiments, we observed that prompts incorporating principles of justice and inclusion tended to yield texts with more characters, greater lexical variety, and richer representations of cultural and social contexts.

Rather than constraining creativity, fairness-oriented prompts may function as productive perturbations, displacing the system from dominant narrative tropes and prompting engagement with less stereotypical symbolic and conceptual spaces. This aligns with perspectives in computational creativity that emphasize the generative potential of constraints, filters, and marginal frames.

Although the findings are preliminary and based on a small sample, they offer initial evidence that fairness can be operationalized as an expressive prompt design strategy. The results suggest a promising direction for developing generative systems that are not only more inclusive, but also more inventive in the narratives they construct.

Several limitations must be acknowledged. The study involved a limited number of prompts and model outputs, relied on surface-level and manually annotated metrics, and did not include human assessments of creativity or cultural resonance. Furthermore, while diversity markers were identified and discussed, deeper semantic and reception-based analyses remain necessary to avoid risks of superficial representation or stereotyping.

Future work should involve larger-scale evaluations, alternative creativity metrics, and engagement with diverse audiences to assess the impact of fairness-aware generation. Expanding the approach to other genres — such as poetry, scriptwriting, or speculative fiction — may also help uncover new affordances of ethical prompting as a driver of narrative innovation.

References

Atwood, J.; Scherrer, N.; Lahoti, P.; Balashankar, A.; Prost, F.; and Beirami, A. 2024. Inducing group fairness in prompt-based language model decisions.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Brown, D., and Ventura, D. 2022. Ethics, aesthetics and computational creativity. In *ICCC*, 150–158.

Déguernel, K., and Sturm, B. L. T. 2023. Bias in Favour or Against Computational Creativity: A Survey and Reflection on the Importance of Socio-cultural Context in its Evaluation. In *International Conference on Computational Creativity*.

Doan, T. V.; Wang, Z.; Hoang, N. N. M.; and Zhang, W. 2024. Fairness in large language models in three hours. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 5514–5517.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* 1–79.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Comput. Surv.* 51(2).

Loughran, R. 2022. Bias and creativity. In *ICCC*, 354–358.

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.

Rabonato, R. T., and Berton, L. 2024. A systematic review of fairness in machine learning. *AI and Ethics* 1–12.

Torres, N.; Ulloa, C.; Araya, I.; Ayala, M.; and Jara, S. 2024. A comprehensive analysis of gender, racial, and prompt-induced biases in large language models. *International Journal of Data Science and Analytics* 1–38.

Wang, C.; Li, J.; and Zhang, R. 2024. A method to enhance structural fairness in large language models with active learning. *Authorea Preprints*.

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

Zeng, C. C.; Chung, M.; and Zhou, E. 2024. Prompting for fairness: Mitigating gender bias in large language models with self-debiasing prompting. In *University of Michigan CSE 595 Natural Language Processing Fall 2024*.

ISBN: 978-989-54160-7-3

Introducing Pathomalgametry: Conceptual Blending with Geometric Path-Finding and Amalgamation

Mena Leemhuis
Free University of Bozen-Bolzano
Bozen/Bolzano, Italy
mena.leemhuis@unibz.it

Oliver Kutz
Free University of Bozen-Bolzano
Bozen/Bolzano, Italy
oliver.kutz@unibz.it

Abstract

Conceptual blending, where new concepts are created through a selective combination of known ideas, is a widely known option for concept invention within computational creativity research. However, existing approaches implementing blending are often either neglecting conceptual aspects, as, e.g., in image morphing, or they suffer from the high complexity of the creation of the blend. Therefore, we propose a new neuro-symbolic approach for conceptual blending of ontologies, based on knowledge-graph embeddings. Here, the inherent structure of the embedding space is used both to identify a generic space and to guide the blending process by interpreting blending as path search in the embedding space by iteratively relaxing the input concepts. Our approach is efficient as it can reuse existing embedding techniques as created for, e.g., link prediction. Its suitability is discussed both theoretically and based on a toy example.

Introduction

One well-known form of human creativity has been described by Fauconnier and Turner (2002) in the framework of *conceptual blending* (CB for short), namely the selective combination of known ideas into a 'blend', a surprising and creative new combination of selected features of the inputs. One well-known example is, e.g., the "houseboat" that can be seen as a conceptual blend of "house" and "boat". Although Fauconnier and Turner introduced blending as a cognitive process, there has been significant effort over the past two decades to find computational realizations in order to facilitate computational creativity. Approaches are, e.g., interpreting it as a search problem solved with answer set programming (Eppe et al. 2018), a formalization based on algebraic specification (Goguen and Harrell 2010), or using mappings and structural alignments (Pereira 2007).

Here we focus on CB of concepts based on *amalgamation* (as, e.g., in (Confalonieri et al. 2018; Confalonieri and Kutz 2020)), thus by relaxing two input concepts successively until a common *generic space* is reached. The combination of two of these relaxations then results in a blend. However, symbolic approaches to conceptual blending in general suffer from high computational complexity of determining the generic space and the best suited blend (see, e.g., the discussion in Eppe et al. (2018)). As conceptual blending is considered a highly intuitive process (Fauconnier and Turner 2002), it seems to be natural to consider it not on a purely symbolic level but to incorporate subsymbolic information, e.g., in the form of similarity information. This has been done, e.g., by Wang et al. (2023), performing conceptual blending in large language models using their associative capacities. However, this approach does not focus on consistency and optimality principles in the rigidity needed for CB in an ontological context. Another approach was presented by Olearo et al. (2024) based on the idea that a blend is in between the latent space representations of the input concepts in a diffusion model. Though the approach incorporates the implicit information learned by the diffusion model and thus models some form of salience of the features, it neither considers a generic space nor conceptual information explicitly. Therefore, a different approach is needed, combining the following two ideas:

(A) the conceptual information should be respected to ensure consistency and coherence with the basic principles of conceptual blending, and

(B) subsymbolic information should inform the blending process by incorporating information regarding concept similarity.

An area well known for combining symbolic ontology information with subsymbolic similarity information is *knowledge base embedding* (see, e.g., (Bourgaux et al. 2024)). There, symbolic information is embedded into a geometric space by representing concepts as (typically) convex regions, instances as points and relations and logical operations as geometrical operations. One main use case is link prediction, thus the prediction of relations between instances. Another is concept membership prediction which allows to reduce the incompleteness of an ontology and to adapt to specific domains. An embedding can be used as a guide to determine a suitable generic space and consequently a suitable blend: As similar concepts are modeled closer to each other, the generic space can be interpreted geometrically as the smallest region containing both input concepts. This avoids over-generalization, and defines a region of possible blends. For determining the best out of these possible blends, the viewpoint of Leemhuis and Kutz (2024) can be applied to interpret the blending process as a path-finding problem in this space. The basic idea is that the generic space can be represented as the smallest convex set incor-

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

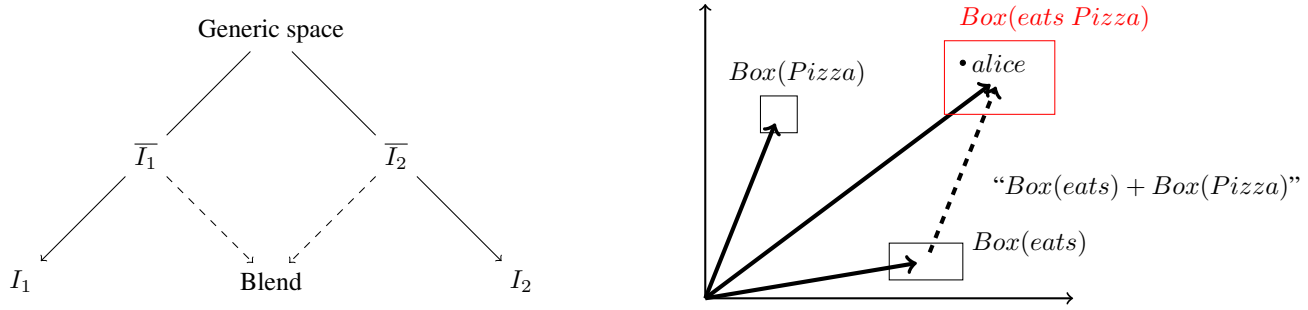


Figure 1: (a) basic diagram of CB with amalgams; (b) example for TransBox representing the fact that "Alice eats Pizza".

porating the representation of both input concepts. A blend can be considered as "in between" the input concepts on a path through the generic space and thus preserves some of the meaning of both input concepts. This eases the blending process and allows for incorporating optimality principles into the process implicitly. In the following, a first version of this neuro-symbolic conceptual blending approach is presented, and its suitability and features are discussed both theoretically and with the help of a toy example.

Preliminaries

Ontology-based Approaches for Blending

CB is a framework describing the creative invention of a new concept by combining known ones. The shared generalization of the known concepts, the *generic space* (GS), is found and is used as basis to define a blend, thus a combination of salient features of the input concepts, adhering to optimality principles (Fauconnier and Turner 2002). CB is applied in many different contexts, e.g., in poetry or even for mathematical concepts. Here, we are focusing on ontological blending, thus interpreting the inputs of the blending process as concepts of an ontology. Ontologies are used to represent structured information of the world in form of axioms representing general knowledge and assertions representing facts. One way of representing ontologies is with the help of description logics (DLs) (Baader et al. 2007). We are focusing here on ontologies in the DL $\mathcal{EL}_\bot$(a fragment of the well-known $\mathcal{EL}^{++}$(Baader, Brandt, and Lutz 2005)) due to its computational advantages, as subsumption is polynomial. It includes concept conjunction, existential quantification, subsumption, and especially disjointness of concepts.

One idea is to interpret conceptual blending as *amalgamation* (Ontañón and Plaza 2010). This has been examined for the case of $\mathcal{EL}^{++}$by Confalonieri et al. (2018). The basic idea is to generalize both input concepts gradually until a generic space, thus a (least general) generalization is reached. In the case of the houseboat-blend mentioned in the introduction, a possible generic space of "house" and "boat" would be an "object based on some foundation and with people inside". After that, the intermediate concepts are combined to determine the blend. An illustration can be seen in Figure 1 (a) and details can be found in (Confalonieri et al. 2018). Basic problems of this approach are that the generic space is not necessarily unique and thus hard to find and that it is not obvious which intermediate concepts should be combined to achieve a good blend. Therefore, we aim to enhance this amalgamation-based approach with subsymbolic information with the help of embeddings.

Knowledge Base Embeddings

Embedding information into geometric spaces is a widely used technique to employ geometric regularities for learning purposes. Especially profiting from geometric embeddings is the area of *knowledge graphs (KGs)*. Knowledge graphs are sets of $(subject, predicate, object)$-triples, e.g., $(alice, loves, bob)$. As those KGs tend to be highly incomplete, one important task is the prediction of missing triples. Here, *knowledge graph embedding (KGE)* comes into play (for an overview, see, e.g., (Ji et al. 2022)). It is based on the idea of learning an embedding (e.g., with the help of a neural network) such that instances are modeled as points and predicates (thus relations) as geometric operations in a vector space (TransE (Bordes et al. 2013) for example represents relations as vector translations). These KGE approaches can be extended to *knowledge base embeddings* by considering also background knowledge in form of an ontology (see (Bourgaux et al. 2024) for an overview). The basic idea is to add to the representation of instances and relations a representation of concepts and their interplay in a geometrical way. Therefore on the one hand the subsymbolic information of similarity of instances and concepts is used and on the other hand the symbolic information in form of an ontology. Concepts are mostly represented as some specific convex object, e.g., as boxes, e.g., by Yang, Chen, and Sattler (2025), spheres (Kulmanov et al. 2019) or convex cones (Özçep, Leemhuis, and Wolter 2020). An instance belongs to a concept if the point representing the instance lies in the the convex object representing the concept. Conjunction of concepts is mostly represented as set-intersection of the concept representations, and subsumption by the subset relation. We focus here on the DL $\mathcal{EL}_\bot$and corresponding box-embeddings, particularly on TransBox (Yang, Chen, and Sattler 2025).

Each concept A is represented as a box $box(A)$, an axis-aligned hyperrectangle in the n-dimensional space $\mathbb{R}^n$ with a center $c(A) = (c_1, \ldots, c_n) \in \mathbb{R}^n$ and an offset given by

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

$o(A) = (o_1, \ldots, o_n) \in \mathbb{R}^n_{\geq 0}$. Thus, $box(A)$ is defined as

$$box(A) = \{x \in \mathbb{R}^n \mid c(A) - o(A) \leq x \leq c(A) + o(A)\}$$

where $x \leq y$ when $x_i \leq y_i$ for $1 \leq i \leq n$. The conjunction of concepts is interpreted as intersection of boxes.[1] The definition of relations can be seen in Figure 1 (b). It is based on the idea of modeling relations also as boxes and states that a triple (a, r, b) holds if $a \in box(r) + b$, thus, $box(r)$ can be considered as containing reified representations of the triples. In the shown example, a box representation of the concept "Pizza" and the role "eats" is determined. The instance "alice" is represented as a point in the space. Then it is the case that "Alice eats pizza" if the point representing Alice is part of the box representing "eats pizza". This box is determined by adding up the centers resp. the offsets of the boxes of "Pizza" and "eats".

Blending Based on Box-embeddings

Assume an ontology is given and two concepts C_1 and C_2 are chosen as inputs to be blended. Based on the amalgamation-based blending approach, the two input concepts are relaxed until a common generic space is reached. This generalization is usually not unique: which part of the concepts should be relaxed and how should the generic space look like? For the houseboat-example, should, e.g., the property that a house rests on ground be relaxed, or the property that people are living inside? The embedding of the ontology offers a guiding principle for this process by offering a heuristic in form of concept similarity: when several generalizations are possible, then the one is used that 'leads towards' the other input concept.

As a first step to use this heuristic, it is necessary to create the box-embedding of a given ontology. We assume that the embedding acts as a model of the ontology, thus represents each axiom and each fact of the ontology correctly. For cases where such a perfect embedding is not found, some approximation strategies are necessary. Now, the boxes representing the two input concepts are identified.[2] The position of the two boxes relative to each other gives a first hint on the nature of the blend: if the distance between the boxes is small, then they are already quite similar and the blend will be less creative (or at least less surprising). If the distance is larger, then the concepts are quite dissimilar and the blend will be more surprising. It has been argued, e.g. by Mikolov et al. (2013) for the case of word embeddings, that similar information is embedded at similar positions. Of course, this is only possible for a rich representation of knowledge. If there are, e.g., only non-intersecting concepts and no relations, then it is not possible to infer any similarity information out of the embedding. Still, if a diverse representation is given, especially when instances are connected by relations, then, due to an enforced restricted dimensionality, it is necessary to model the information as efficiently as possible. Then, first, due to the convexity of the boxes representing superconcepts and, second, due to the limited space, it is necessary to model 'similar information' at 'similar positions' (and even predict new concepts for given instances if this eases the representation). Another interesting aspect is the negative sampling used in the embedding approaches. Negative sampling is a standard technique to improve the learning process (see, e.g.,(Qian et al. 2021) for an overview) by introducing corrupted triples for learning. Thus (subject, predicate, object)-triples inconsistent with the ontology are created which should thus not be satisfied by the embedding. This constraints the size of the boxes and prevents trivial solutions which would not contain any similarity information.

Generic Space

Given the two boxes representing the input concepts, it is straightforwardly possible to determine the generic space geometrically. By definition, a concept is a subconcept of another concept if the box representing the former concept is part of the box representing the latter concept. Therefore, a possible generic space of the two input concepts regarding the embedding can be defined as the smallest box containing both input boxes. Note that here only one interpretation is considered, and, therefore, the generalization found is not necessarily a generalization of the original ontology (due to the additionally inferred statements which are not necessary inferrable in the ontology). However, as argued by Confalonieri et al. (2018), it is not necessary to find an exact generalization but sufficient to find an approximation. In this case, the approximation is especially helpful, as it incorporates also the fact-based knowledge of the world: when, e.g., the ontology allows for many special cases not relevant for the considered use-case, then the facts do not reflect all these special cases: the embedding and therefore also the generic space can be simplified (in this case specialized) to a more specific case. Note that such a geometric representation of a generic space is always found, as a convex hull of two boxes always exists. For very dissimilar concepts, the generic space could be, however, quite large (or even trivial, thus the top-concept). Figure 2 depicts in purple the generic space GS based on the input concepts C_1 and C_2.

Blend

The generic space determines the area of possible blends. However, not every point in this space represents an instance that is a suitable blend. The problem of finding a good blend in this geometric space of possible blends can be considered as analogous to the problem of finding a good amalgamation amongst all possible ones. However, we encounter the simplification that due to the learning of the embedding, the space of possibilities has already been reduced. The amalgamation-based generalization procedure can be considered as the generation of a generalization path leading to the blend. Therefore, we interpret the search for an amalgamation as a path-finding process in the embedding space. Consider Figure 2 (a), depicting an excerpt of a two-dimensional embedding. The two input concepts and possible generalizations of the input concepts are depicted as C_i, C_i', C_i'' for $i \in \{1, 2\}$ (in real world use-cases, the

[1] TransBox offers also a different interpretation of the intersection that is not considered here.

[2] Those concepts are not necessarily represented in the embedding. This special case could be circumvented by retraining the embedding and is not further discussed here.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

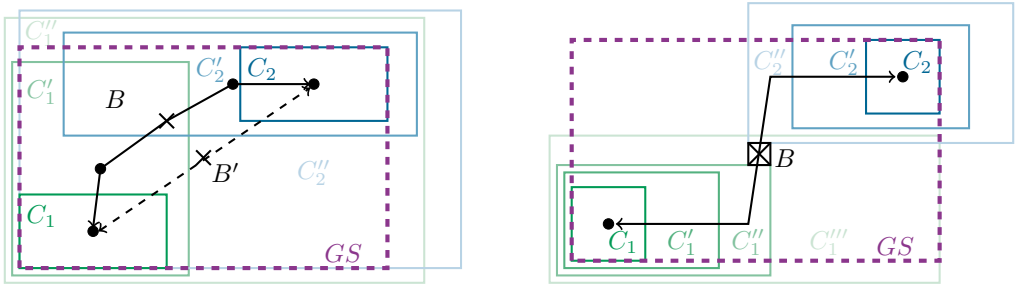


Figure 2: (a) Visualizing distinct generalization paths; (b) Visualizing two generalization paths without a direct GS-intersection

generalizations would be of course more diverse). As discussed, the blend is by definition situated somewhere in the generic space. A first attempt is to consider the Euclidean path in between the two concepts, thus a line in between the two centers of the boxes. This is depicted by blend B'. B', however, represents only a low quality blend: though it is part of the generic space, it does not model any individual features of the two input concepts and is therefore too general. This interferes with the basic principles of CB (as the blend should be a selective combination of the inputs).

Although B' could be interpreted as a blend based on amalgamation ($B' \in box(C_1'') \cap box(C_2'')$), it shows that there is some guiding principle needed to accomplish not only an arbitrary but a good amalgamation. Therefore, it is not sufficient to consider solely the shortest path. The basic idea is to interpret the amalgamation search process as a two-player game (similar to (Righetti et al. 2021)) and iteratively relax both concepts C_1 and C_2 until a common box is reached that contains B, based on a heuristic that tries to find a trade-off between specificity of the generalization and reduction of the distance to the other generalization. An example can be seen in Figure 2 (a): The two starting points of the path are the centers of the boxes representing the input concepts. Following the line of prototype theory (see, e.g., (Hampton 2006)), it is assumed that the center of a box contains instances more prototypical for a concept than instances at the border of the concept representation. Then, both input concepts are relaxed with the aim of finding a meeting point of the two paths. Thus, based on some heuristics, here C_1' and C_2' are chosen as generalizations and again the centers of the respecting boxes are chosen as new endpoints of the path. They represent prototypical representations of the more general concepts. This process can be continued until two generalizations are reached that intersect with each other. In this intersection a blend can be searched for. One possible heuristics would be to chose a blend at the boundaries of this intersection, as now not the most prototypical instance is looked for but a surprising and creative one. This depicts one variant of finding a path in the embedding space. Of course, depending on the use case, the exact heuristics and the data used, it is possible to use different strategies for determining such a path.

In this example, it was possible to determine an intersection of the generalizations after one generalization step. However, it could be necessary to consider more than one step. It could also be the case that an intersection of generalized concepts is only possible if one or both generalizations lead either to the trivial concept or enforces a too severe generalization. Then, e.g., the blend of house and boat would be a "cube-shaped object that swims". This is due to the fact that the embedding does not model all possible combinations of concepts but only a subset of them. Considering possible blends where the generalizations are not intersecting widens the search space and especially allows for finding more surprising (as not modeled) blends. This works as follows: instead of combining generalizations to make up a blend only when the two generalizations intersect in the GS, the two generalizations are considered if they are in a sufficiently close distance. In the example of Figure 2 (b), the input concepts can be generalized to C_1'' and C_2'', respectively. These two concepts are quite similar (though not intersecting inside the GS) but still not a too broad generalization (their area has not increased too much in comparison to the input concepts). However, to gain a blend based on an intersection, it would be necessary to generalize C_1'' to C_1''', thus the resulting blend would be much broader. Therefore, here, the blend is considered as the space in between the two generalizations instead of generalizing until a common space is reached. Blends in this space are, though not directly inside the generalizations, closely related to both of them. This approach incorporates the similarity information actively, thus using that sufficient proximity in representation means we can assume that concepts C_1'' and C_2'' 'essentially hold' at blend B. At the same time, it is not as strict as the intersection-based approach and thus allows for more detailed and less generalized blends.

Optimality Principles

One important aspect of CB is to judge the quality of the blends based on optimality principles (Pereira and Cardoso 2003). Some of the optimality principles are already implicitly covered since the construction of the embedding provides an interpretation of the ontology and includes similarity information. Thus, e.g., the *relevance principle* should be implicitly fulfilled (stating that only relevant parts of the input concepts should be contained in the blend), as the blend is positioned in between the two inputs and therefore all concepts included in the blend are more similar to both of the

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

input concepts than the input concepts are to each other. The *unpacking principle*, stating that it must be possible to reverse the whole blending process starting with the blend, is also respected since the construction is based on finding a path with the blend on it. Thus, to unpack we start at the blend and reverse the movement on the path towards the start- and endpoints. Another principle is about maximizing and intensifying vital relations, i.e. to prefer fundamental relations. Though such vital relations are not directly incorporated into the blend, it can be stated that the embedding is able to find important relations, as those relations also play an important role in the embedding, and are thus more likely to be also modeled in the blend. Implementing this and validating these claims experimentally (and additional criteria such as *integration* and *topology*) will be the subject of future extensions of the present work.

Experiments

In future work, we aim to develop more detailed heuristics and to explore the exact interplay of search and amalgamation. Still, to exemplify the general applicability of the approach, we here present a concise experiment based on TransBox[3] (Yang, Chen, and Sattler 2025) (introduced in the preliminaries). With TransBox, a two-dimensional embedding of a simple ontology of a horse and a bird is learned, partly displayed in Figure 3. It is based on the concept definitions used in Confalonieri et al. (2016): There, a "Horse" is defined as a "Mammal" having "Legs" and the ability to "Walk" and "Trot". A "Bird" is an "Avialae" having "Wings" and "Legs" which is able to "Fly" and to "LayEggs". There, a good blend is considered as being a Pegasus, thus a "Mammal" with "Legs" and "Wings" that has the ability to "Walk", "Trot" and "Fly". We created a simple ontology around these definitions, stating, e.g., that mammals do not lie eggs (by disregarding the platypus) and birds don't run. Due to the simplicity of the ontology, a two-dimensional embedding is possible. In most cases, more dimensions would be needed. This simple example, though, already illustrates three vital properties of the embeddings, useful as a basis for the proposed approach. (i) In the embedding of the example ontology, it can be seen that "Animal" is actually a superclass of both "Bird" and "Horse" and, as no other more specific animal classes are considered, it constitutes a plausible generic space. The constraints are all fulfilled, e.g., horses don't have wings and can't fly. Moreover, it can be seen that there is a point in between the mapped input concepts, where an animal being able to walk, to trot and to fly and which has legs and wings, is represented, and thus describes Pegasus. This point can be found by generalizing the concept of "Horse" to "Mammal that can trot" and by generalizing the "Bird" to an "Animal that can fly and has wings". Other options for generalization steps would have been, e.g., to generalize the "Bird" as an "egg-laying Avialae", that would, however, not lead to this blend. This blending process has been carried out based on manually constructing the best blend given the embedding. As future work, a heuristic is needed to automate this process. In this simple case, the blend can be found in between the two inputs in a Euclidean sense. This is, especially when considering more complex concepts, not necessarily the case anymore (as discussed in the last section). (ii) The generic space actually depicts the space of suitable blends, thus, the blend quality decreases outside of the generic space, especially as then the resulting blend is too general. (iii) This example, finally, also shows the challenges of this approach, as even this simple ontology is modeled only approximately correctly, and thus forces us to consider techniques coping with such inaccuracies in order to approximate a good blend.

[3] https://github.com/HuiYang1997/TransBox

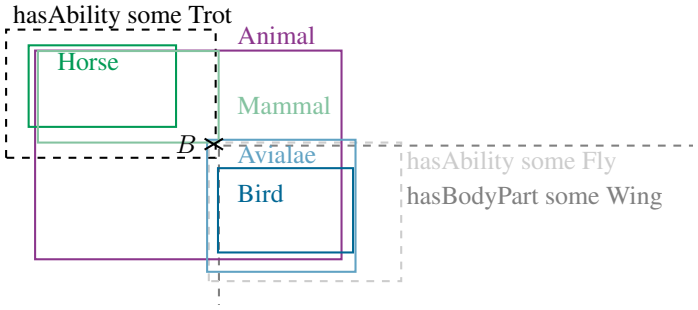


Figure 3: An example for the blend of bird and horse in the animal ontology.

Conclusion

We sketched first steps towards designing a neuro-symbolic CB framework based on the embedding of ontologies. Since existing embedding approaches suffer from a limited representational accuracy, an implemented path-finding heuristic should not only find a suitable path but should also be able to cope with partly incorrect embeddings. Additionally, considering points directly in the embedding space is a good first step to find a candidate blend. However, it restricts the space of possible blends to the ones that are already contained in the embedding space or at least approximated there. Future work is thus to enhance our approach by not only focusing on the blends in one fixed interpretation, but by either influencing the embedding process such that more and different blends can be found, or by considering the creation of the generic space only as a first step and then to use the information to find the blend in a more complex process, also including instance information. The implementation of this approach and the determination of suitable heuristics is ongoing work. In general, a more thorough examination is needed, especially for larger ontologies. However, the experiment and the discussion show that our approach promises novel neuro-symbolic strategies for performing CB based on embeddings.

Acknowledgments

We acknowledge the financial support through the 'Abstractron' project funded by the Autonome Provinz Bozen - Südtirol (Autonomous Province of Bolzano/Bozen) through the Research Südtirol/Alto Adige 2022 Call.

References

Baader, F.; Calvanese, D.; McGuinness, D. L.; Nardi, D.; and Patel-Schneider, P. F., eds. 2007. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, second edition.

Baader, F.; Brandt, S.; and Lutz, C. 2005. Pushing the $\mathcal{EL}$ Envelope. In *IJCAI'05: Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 364–369.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2787–2795.

Bourgaux, C.; Guimarães, R.; Koudijs, R.; Lacerda, V.; and Ozaki, A. 2024. Knowledge Base Embeddings: Semantics and Theoretical Properties. In *Proceedings of the Twenty-First International Conference on Principles of Knowledge Representation and Reasoning*, KR-2024, 823–833. International Joint Conferences on Artificial Intelligence Organization.

Confalonieri, R., and Kutz, O. 2020. Blending under deconstruction: The roles of logic, ontology, and cognition in computational concept invention. *Annals of Mathematics and Artificial Intelligence* 88(5–6):479–516.

Confalonieri, R.; Schorlemmer, M.; Kutz, O.; Peñaloza, R.; Plaza, E.; and Eppe, M. 2016. Conceptual Blending in $\mathcal{EL}^{++}$. In *29th International Workshop on Description Logics, DL 2016*, volume 1577. CEUR Workshop Proceedings.

Confalonieri, R.; Eppe, M.; Schorlemmer, M.; Kutz, O.; Peñaloza, R.; and Plaza, E. 2018. Upward refinement operators for conceptual blending in the description logic $\mathcal{EL}^{++}$. *Annals of Mathematics and Artificial Intelligence* 82(1–3):69–99.

Eppe, M.; Maclean, E.; Confalonieri, R.; Kutz, O.; Schorlemmer, M.; Plaza, E.; and Kühnberger, K.-U. 2018. A computational framework for conceptual blending. *Artificial Intelligence* 256:105–129.

Fauconnier, G., and Turner, M. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Goguen, J. A., and Harrell, D. F. 2010. Style: A Computational and Conceptual Blending-Based Approach. In Argamon, S.; Burns, K.; and Dubnov, S., eds., *The Structure of Style*. Springer Berlin Heidelberg. 291–316.

Hampton, J. A. 2006. Concepts as prototypes. In *Psychology of Learning and Motivation Volume 46*, 79–113. Elsevier.

Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Yu, P. S. 2022. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems* 33(2):494–514.

Kulmanov, M.; Liu-Wei, W.; Yan, Y.; and Hoehndorf, R. 2019. $\mathcal{EL}$ Embeddings: Geometric Construction of Models for the Description Logic $\mathcal{EL}^{++}$. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6103–6109. International Joint Conferences on Artificial Intelligence Organization.

Leemhuis, M., and Kutz, O. 2024. A Cloud Full of Paths: Conceptual Blending as Betweenness Relation. In *Proceedings of The Eighth Image Schema Day (ISD8)*, volume 3888 of *CEUR Workshop Proceedings*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems* 2:3111–3119.

Olearo, L.; Longari, G.; Melzi, S.; Raganato, A.; and Peñaloza, R. 2024. How to Blend Concepts in Diffusion Models. In *Proceedings of The Eighth Image Schema Day (ISD8)*, volume 3888 of *CEUR Workshop Proceedings*.

Ontañón, S., and Plaza, E. 2010. Amalgams: A Formal Approach for Combining Multiple Case Solutions. In Bichindaritz, I., and Montani, S., eds., *Case-Based Reasoning. Research and Development*. Berlin, Heidelberg: Springer Berlin Heidelberg. 257–271.

Özçep, Ö.; Leemhuis, M.; and Wolter, D. 2020. Cone Semantics for Logics with Negation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 1820–1826. International Joint Conferences on Artificial Intelligence Organization.

Pereira, F. C., and Cardoso, A. 2003. Optimality Principles for Conceptual Blending: A First Computational Approach. *AISB Journal* 1(4).

Pereira, F. C. 2007. *Creativity and Artificial Intelligence: A Conceptual Blending Approach*. Mouton de Gruyter.

Qian, J.; Li, G.; Atkinson, K.; and Yue, Y. 2021. Understanding Negative Sampling in Knowledge Graph Embedding. *International Journal of Artificial Intelligence & Applications* 12(1):71–81.

Righetti, G.; Porello, D.; Troquard, N.; Kutz, O.; Hedblom, M. M.; and Galliani, P. 2021. Asymmetric hybrids: Dialogues for computational concept combination. volume 344 of *Frontiers in Artificial Intelligence and Applications*, 81 – 96. IOS Press.

Wang, S.; Petridis, S.; Kwon, T.; Ma, X.; and Chilton, L. B. 2023. PopBlends: Strategies for Conceptual Blending with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, volume 33, 1–19. ACM.

Yang, H.; Chen, J.; and Sattler, U. 2025. TransBox: $\mathcal{EL}^{++}$-closed Ontology Embedding. In *The Web Conference 2025*.

ISBN: 978-989-54160-7-3

Making the Familiar Strange: A Computational Approach to Defamiliarization in Creativity Support

Yu-Chieh Ho
Independent Researcher
Taipei
yuchieh.ho@gmail.com

Wenn-Chieh Tsai
D-School
National Taiwan University
Taipei
joe.wctsai@gmail.com

Zhu-Yuan Lai
Trans-disciplinary Bachelor Degree Program
National Taiwan University
Taipei
zhuyuan.lai.bamboo@gmail.com

Abstract

Defamiliarization—originating in literary theory and later adapted into Critical and Speculative Design—seeks to disrupt habitual perception and invite renewed engagement through "slight strangeness," and has been extensively used to provoke critical engagement in design and interactive systems. However, existing approaches remain subjectively designer-controlled, limiting generalizability in interactive creative systems, while generative-AI tools introduce unexpected elements without offering structured, user-level control over estrangement. To address this gap, we present Familiarity–Estrangement Space, which projects textual materials (e.g., design concepts and invented stories) along three personalized dimensions—Familiarity, Positive Estrangement, and Negative Estrangement—empowering creators to modulate the exact degree of strangeness they desire. By extracting latent topics and inviting users to identify which topics to retain, amplify, or minimize, our approach transforms defamiliarization from a designer-imposed intervention into an adjustable parameter for creativity support. This enables creators to navigate and discover fresh thematic collisions or reinforce familiar patterns, offering a structured framework for AI-assisted creative exploration that balances coherence with deliberate estrangement.

Introduction

Defamiliarization, introduced by Victor Shklovsky in literary theory (Shklovsky 1965), describes making the familiar strange to disrupt habitual perception and foster renewed engagement. Dunne and Raby (Dunne 2008) (Dunne and Raby 2001) later applied this concept to Critical and Speculative Design, using slight strangeness to challenge technological norms and provoke alternative interpretations of design artifacts. Their approach leverages deliberate friction to shift passive interaction into active reflection, where interactive systems subtly interrupt ingrained behaviors to expose underlying technological assumptions.

Defamiliarization has been extensively used to provoke critical engagement in design and interactive systems. Reflective Design (Sengers et al. 2005) and Critical Computing (Bell, Blythe, and Sengers 2005) show how subtle disruptions can reveal hidden technological structures. However, these approaches rely on predefined estrangement levels decided by designers, rather than permitting users to adjust them.

At the same time, Creativity Support Tools (CSTs) and AI-assisted creative tools have focused on preserving coherence or maximizing novelty (Lubart 2005) (Frich et al. 2019). Although generative AI can produce a wide variety of outputs, these variations are often random rather than systematically controlled (Franceschelli and Musolesi 2024), offering creators no mechanism to modulate the degree of estrangement in their work. This leaves creators to sift through and fine-tune an abundance of AI outputs—an effort that underscores the need for proper metrics to evaluate and guide selection (Li et al. 2024).

To address this gap, we present the *Familiarity–Estrangement Space (FES)*, a conceptual model that quantifies— and makes adjustable— the defamiliarization of textual materials (e.g., design concepts, invented stories) by combining human creative vision with topic-modeling techniques, thereby structuring creative exploration and enabling AI to cultivate controlled strangeness.

Topic Modeling as a Mechanism for Defamiliarization

Topic modeling, an NLP approach that aims to automatically extract latent topics from extensive collections of texts (corpus) (Blei, Ng, and Jordan 2001), has introduced new methods for structuring creative exploration. While typically applied to retrieval and categorization, topic modeling can also serve as a mechanism for controlled estrangement by surfacing unexpected thematic pairings that challenge conventional thinking.

By integrating Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2001) as a creative disruptor, Familiarity–Estrangement Space generates structured thematic collisions, discovering new inspiration pathways. This approach repurposes topic modeling from a content structuring tool to a computationally guided defamiliarization mechanism that enhances creative exploration.

Research Gap and Motivation

Although prior research in Critical and Speculative Design has explored defamiliarization, existing approaches remain

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

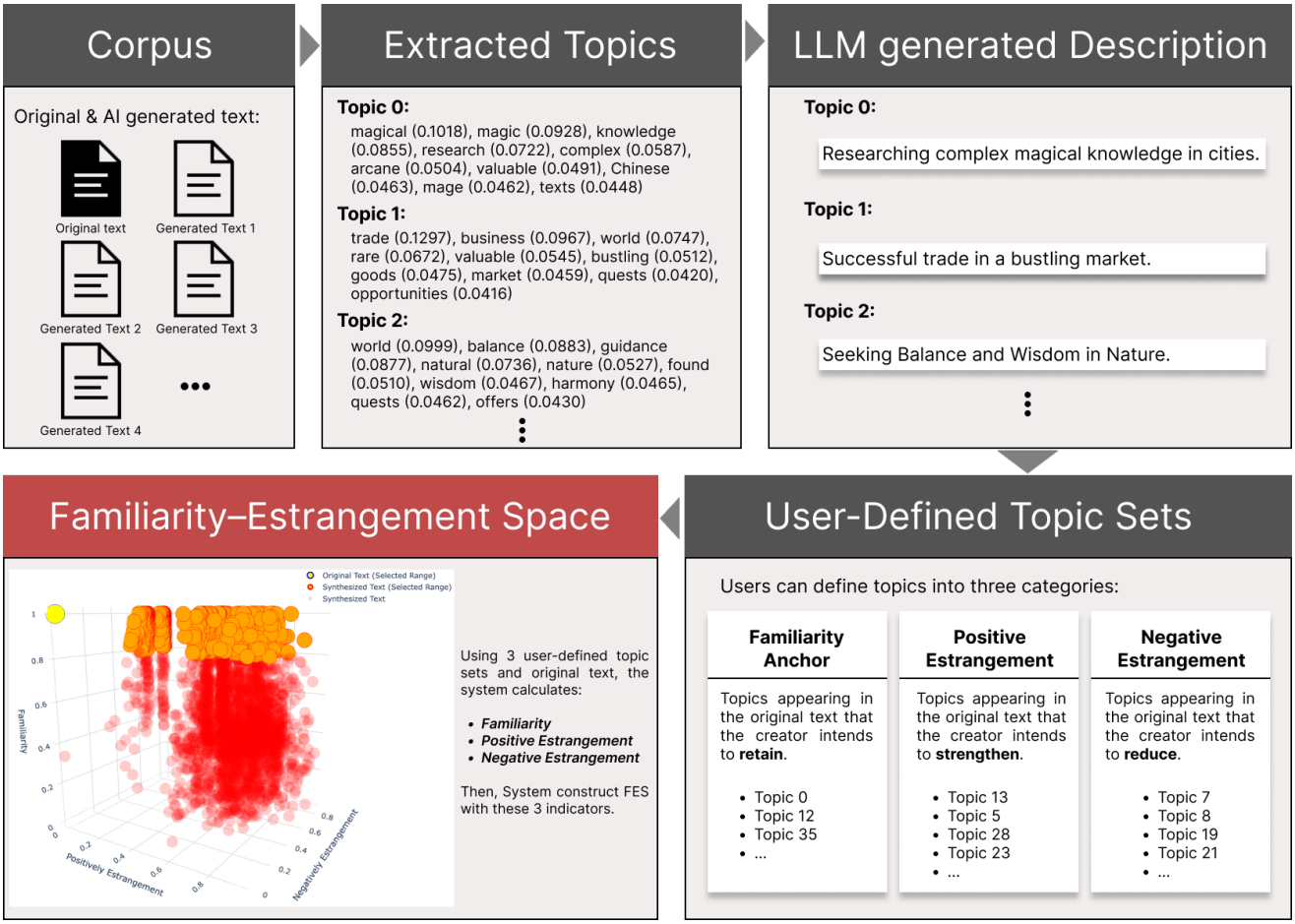


Figure 1: **A Topic Modeling-Based Framework for Familiarity–Estrangement Space Co-Construction**.

subjectively designer-controlled, limiting generalizability in interactive creative systems, while generative-AI tools introduce unexpected elements without structured user control over strangeness, treating estrangement as an incidental side effect rather than a deliberate, adjustable parameter.

To address these limitations, FES introduces:

- *A computational model of defamiliarization, formalized through three quantifiable indicators: Familiar, Positive Estrangement, and Negative Estrangement.*
- *A topic modeling-based framework for conceptual defamiliarization, leveraging LDA to extract unexpected but semantically meaningful thematic disruptions. (As illustrated in Fig. 1.)*

In the following sections, we formulate the Familiarity–Estrangement Space, define its fundamental components, and conclude with preliminary results from a game character design scenario.

Familiarity-Estrangement Space Overview

Since creativity is inherently subjective, the construction of Familiarity–Estrangement Space is user-centered: the system collects user's aesthetic preferences to construct a customized FES. FES maps generated outputs by their alignment with prior stylistic conventions (familiarity) and degree of intentional deviation (estrangement).

Given a creator's original text d_o (e.g., a manuscript, a design draft, etc.) and a related corpus $D = \{d_1, d_2, \ldots, d_m\}$ (either existing or AI-generated), FES is a three-dimensiona space constructed by the Familiarity $\mathcal{F}$ (Equation (4)), Positive and Negative Estrangement $\mathcal{E}$ (Equation (7)) indicators:

- **Familiarity–Estrangement Space**:

$$\forall d_x \in D, \quad FES(d_x, d_o) = \Big(\mathcal{F}(d_x^{(T_{anc})}, d_o^{(T_{anc})}), \mathcal{E}(d_x^{(T_{pos})}, d_o^{(T_{pos})}), \mathcal{E}(d_x^{(T_{neg})}, d_o^{(T_{neg})})\Big). \tag{1}$$

Here, T_{anc}, T_{pos}, and T_{neg} denote user-defined *topic sets*. Leveraging these topic sets, the system computes tailored Familiarity and Estrangement indicators, enabling users to efficiently retrieve texts and locate content that matches their desired level of defamiliarization.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

User-Defined Topic Sets

Firstly, the system performs topic modeling over corpus D, extracting k representative topics $T = \{t_1, t_2, \ldots, t_k\}$, and computing a topic-weight distribution to quantify each text with a $\mathbb{R}^k$ vector:

$$\forall d_i \in D, \quad d_x \overset{T}{\mapsto} d_i^{(T)} = \left(w_1^{(x)}, w_2^{(x)}, \ldots, w_k^{(x)}\right) \in \mathbb{R}^k, \quad \text{s.t.} \quad w_j^{(x)} \geq 0, \quad \sum_{j=1}^{k} w_j^{(x)} = 1. \tag{2}$$

Next, the system summarizes the text-to-topic weights and invites creators to define three mutually exclusive topic sets $T_* \subseteq T$ that reflect their aesthetic preferences:

- **Familiarity Anchor** (T_{anc}): Topics appearing in the original text that the creator intends to *retain*.
- **Positive Estrangement** (T_{pos}): Topics appearing in the original text that the creator intends to *strengthen*.
- **Negative Estrangement** (T_{neg}): Topics appearing in the original text that the creator intends to *reduce*.

Given a topic set T_* and any text $d_x \in D$, the system projects $d_x^{(T)}$ into the corresponding subspace by zeroing out topics *not* in T_*:

$$\forall d_x \in D, \quad d_x^{(T)} \overset{T_*}{\mapsto} d_x^{(T_*)} = \left(w_1'^{(x)}, w_2'^{(x)}, \ldots, w_k'^{(x)}\right) \in \mathbb{R}^k, \quad \text{s.t.} \quad w_j'^{(x)} = \begin{cases} w_j^{(x)}, & t_j \in T_* \\ 0, & t_j \notin T_* \end{cases}. \tag{3}$$

These subspaces capture user's creative intentions and form the basis for computing three indicators.

Familiarity and Estrangement Indicators

By anchoring on the user-defined topic sets and original text, the system computes *Familiarity*, *Positive Estrangement*, and *Negative Estrangement* indicators of texts with the following equations:

Familiarity indicator (4) computes a familiarity score between any text d_x and d_o. To capture how well d_x preserves topics in T_{anc}, we introduced *Weight-Kept* (5) to measure how much topic weights of d_o is retained in d_x, and defined the *Cosine Similarity* in the Topic Space (6) to quantify the structural similarity of their topic distributions. Finally, a weighted harmonic mean merges these two metrics ($W_{w\text{-}kept}$ and $W_{\cos}$ are hyperparameters):

$$\mathcal{F}\left(d_x^{(T_{anc})}, d_o^{(T_{anc})}\right) = \frac{W_{w\text{-}kept} + W_{cos}}{\frac{W_{w\text{-}kept}}{w\text{-}kept\left(d_x^{(T_{anc})}, d_o^{(T_{anc})}\right)} + \frac{W_{cos}}{\cos\left(d_x^{(T_{anc})}, d_o^{(T_{anc})}\right)}} \tag{4}$$

- Weight-Kept:

$$w\text{-}kept\left(d_x^{(T_*)} - d_o^{(T_*)}\right) = \min\left(\frac{\sum_{i=1}^{k} \min\{w_i'^{(x)}, w_i'^{(o)}\}}{\sum_{i=1}^{k} w_i'^{(o)}}, 1\right). \tag{5}$$

- Cosine Similarity:

$$\cos\left(d_x^{(T_*)} - d_o^{(T_*)}\right) = \frac{\sum_{i=1}^{k} w_i'^{(x)} w_i'^{(o)}}{\sqrt{\sum_{i=1}^{k}\left(w_i'^{(x)}\right)^2}\sqrt{\sum_{i=1}^{k}\left(w_i'^{(o)}\right)^2}}, \tag{6}$$

Estrangement indicator (7) evaluates *value* and *structural* differences between d_x and d_o to reflect the creator's desired estrangement. We introduced *Positive/Negative Difference* (8) to compute the cumulative increase or decrease of topic weights in T_{pos} or T_{neg}, and *Positive/Negative Count* (9) to indicate how these increases or decreases are distributed among topics:

$$\mathcal{E}\left(d_x^{(T_*)}, d_o^{(T_*)}\right) = \frac{W_{diff} + W_{count}}{\frac{W_{diff}}{R_d\left(diff\left(d_x^{(T_*)}, d_o^{(T_*)}\right)\right)} + \frac{W_{count}}{R_c\left(count\left(d_x^{(T_*)}, d_o^{(T_*)}\right)\right)}}, \tag{7}$$

- Positive/Negative Difference:

$$diff\left(d_x^{(T_*)} - d_o^{(T_*)}\right) = \begin{cases} \sum_{i=1}^{k}\left(w_i'^{(x)} - w_i'^{(o)}\right), & \text{if } T_* = T_{pos}, \\ \sum_{i=1}^{k}\left(w_i'^{(o)} - w_i'^{(x)}\right), & \text{if } T_* = T_{neg}. \end{cases} \tag{8}$$

- Positive/Negative Count:

$$count\left(d_x^{(T_*)} - d_o^{(T_*)}\right) = \begin{cases} \sum_{i=1}^{k} \mathbf{1}[w_i'^{(x)} > w_i'^{(o)}], & T_* = T_{pos}, \\ \sum_{i=1}^{k} \mathbf{1}[w_i'^{(x)} < w_i'^{(o)}], & T_* = T_{neg}. \end{cases} \tag{9}$$

Two rescaling functions

$$R_d(diff(.)) = \frac{diff(.)}{2}, \quad R_c(count(.)) = \frac{count(.)}{|T_*|}$$

align the metrics' value ranges. Substituting T_{pos} or T_{neg} gives:

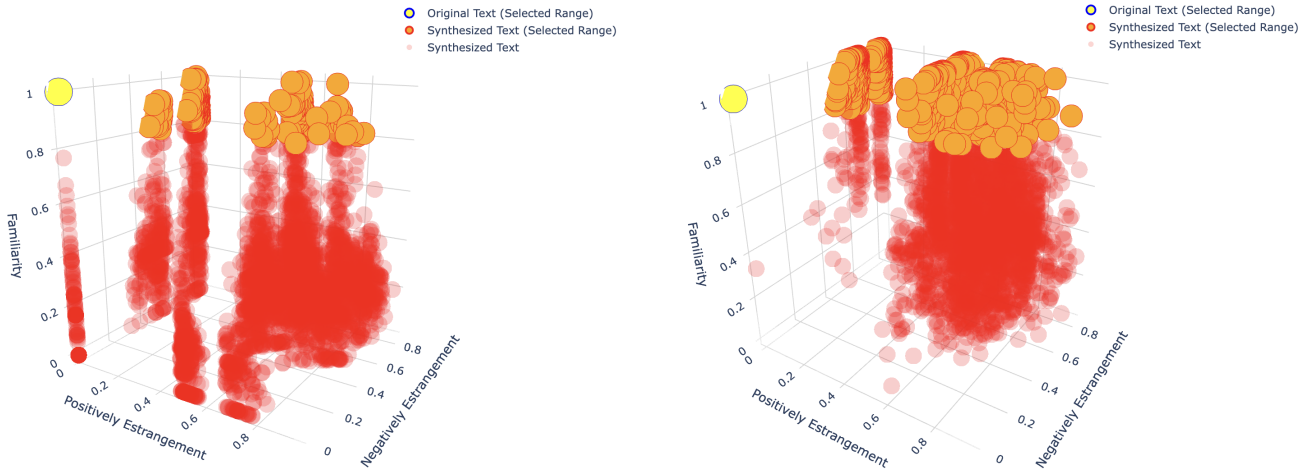
$$Positive\ Strangeness = \mathcal{E}\left(d_x^{(T_{pos})}, d_o^{(T_{pos})}\right),$$
$$Negative\ Strangeness = \mathcal{E}\left(d_x^{(T_{neg})}, d_o^{(T_{neg})}\right)$$

Finally, the system applies the projections according to Equation (1) and constructs a Familiarity–Estrangement Space as illustrated in Figure 2.

Preliminary Results

Figure 2 visualizes two distinct Familiarity–Estrangement Spaces, demonstrating the diversity and personalization nature of the FES. The corpus contains 3,779 game-character

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3



(a) • T_{anc}: {1, 5, 6} • T_{pos}: {3,8}. • T_{neg}: {2,10} .

(b) • T_{anc}: {2,10}. • T_{pos}: {0,7,11}. • T_{neg}: {1, 5, 6}

Figure 2: Familiarity–Estrangement Spaces derived from **the same corpus and original text** yet **different topic set configurations**. The original text (yellow) serves as a reference for synthesized texts (red/orange). Texts exceeding the familiarity threshold (0.8) are highlighted in orange.

profiles generated by Persona Hub (Ge et al. 2024). To accelerate brainstorming during the game design process, the authors provided the LLM with the game's world-building information and prompted it to re-imagine real-world personas as characters situated within that fictional universe.

Since familiarity and estrangement indicators define the three dimensions of FES, different topic-set configurations can yield markedly distinct FES representations. For instance, in Figure 2a, most synthesized profiles occupy a lower-familiarity region (red points, *Familiarity* < 0.8), whereas in Figure 2b, more appear in a high-familiarity region (orange points, *Familiarity* ≥ 0.8). I.e., far fewer synthetic profiles exhibit strong weights on topics {1, 5, 6} (T_{anc} set of Figure 2a) than on topics {2, 10}; hence the creator must lower the familiarity threshold to discover a broader set of "strange" profiles. The detailed profile distributions displayed in the interactive versions of Figures 2a and 2b (see footnote[1]) reveal that identical coordinates in the two spaces can correspond to entirely different profiles (see footnote[2] for Topic Details). Overall, the FES organizes all profiles by indicators that faithfully reflect user-defined topic sets, enabling customized exploration across the entire corpus.

Discussion and Future Work

We introduced Familiarity–Estrangement Space (FES), which quantitatively models defamiliarization in a personalized context and reframes the defamiliarization from a designer-imposed strategy into a flexible framework. Grounded in FES, we are developing an interactive estrangement calibration interface that enables users to modulate strangeness in generative AI outputs and co-create with AI agents to attain a desired degree of defamiliarization. Leveraging evaluation approaches like *Creativity Support Index* (Cherry and Latulipe 2014), we will systematically assess creators' multifaceted satisfaction with the integrated system and conduct in-depth interviews to uncover user needs and refine the system.

Looking ahead, we will equip FES with real-time generators that populate its currently "empty" coordinates and incorporate *multimodal topic modeling*, expanding the framework across diverse creative domains to further boost human–AI co-creativity.

[1]Interactive Figures:
2a: https://reurl.cc/2Kz9qX
2b: https://reurl.cc/VYNxOR

[2]Topic Details: https://reurl.cc/parNGd

Acknowledgments

This study was supported by the National Science and Technology Council, Taiwan (NSTC 113-2410-H-002-110). We would like to express our sincere gratitude to research assistants Chun-Teng Wang and You-Wei Yang for their valuable design insights and constructive feedback on the system evaluation throughout the research process.

References

Bell, G.; Blythe, M.; and Sengers, P. 2005. Making by making strange: Defamiliarization and the design of domestic technologies. *ACM Transactions on Computer-Human Interaction* 12(2):149–173.

Blei, D.; Ng, A.; and Jordan, M. 2001. Latent dirichlet allocation. *Advances in neural information processing systems* 14.

Cherry, E., and Latulipe, C. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Index. *ACM Transactions on Computer-Human Interaction* 21(4):1–25.

Dunne, A., and Raby, F. 2001. *Design noir, the secret life of electronic objects*. London: August.

Dunne, A. 2008. *Hertzian tales: Electronic products, aesthetic experience, and critical design*. MIT press.

Franceschelli, G., and Musolesi, M. 2024. Creativity and Machine Learning: A Survey. *ACM Comput. Surv.* 56(11):283:1–283:41.

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*, 1–18. Glasgow Scotland Uk: ACM.

Ge, T.; Chan, X.; Wang, X.; Yu, D.; Mi, H.; and Yu, D. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. arXiv:2406.20094 [cs].

Li, J.; Cao, H.; Lin, L.; Hou, Y.; Zhu, R.; and El Ali, A. 2024. User Experience Design Professionals' Perceptions of Generative Artificial Intelligence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–18. New York, NY, USA: Association for Computing Machinery.

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):365–369.

Sengers, P.; Boehner, K.; David, S.; and Kaye, J. J. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: Between sense and sensibility*, 49–58. Aarhus Denmark: ACM.

Shklovsky, V. 1965. Art as technique. In Lemon, L. T., and Reis, M. J., eds., *Russian Formalist Criticism: Four Essays*. Lincoln, NE: University of Nebraska Press. 3–24.

Defining and Quantifying Creative Behavior in Popular Image Generators

Aditi Ramaswamy
aditi.ramaswamy@kcl.ac.uk
and **Hana Chockler**
hana.chockler@kcl.ac.uk
and **Melane Navaratnarajah**
melane.navaratnarajah@kcl.ac.uk

Abstract

Whether generative AI models can exhibit creative behavior has been a subject of scientific debate in the last years, without a conclusive answer. In this paper, we define creativity from a task-oriented perspective and introduce quantitative measures that help the user to choose a suitable AI model for a given task. We have evaluated our measures on chains of images iteratively generated using number of popular `img2img` models. We observed that the images produced in these chains matches up with our metrics, with higher-scoring images exhibiting visual traits more strongly aligned with our definition of task-oriented creativity.

Introduction

Generative AI has become the center of a series of heated debates, one of the biggest ones centered around the concept of human creativity. Some researchers argue that creativity is derived from an innately human "sociocultural context", thus excluding AI models (Oppenlaender 2022; Kaufman, Sternberg, and Glăveanu 2019; Wingström, Hautala, and Lundman 2022). However, other definitions focus on purely behavioral requirements, such as "domain-relevant skills", "creativity-relevant processes", and "extrinsic motivation", all of which can be exhibited by generative AI models (Amabile 1983). Related to this, Margaret Boden proposes an AI-relevant cognitive dimension of creativity separate from emotion, focusing on behaviors such as novelty and association of ideas (Boden 1998). Boden also discusses "exploratory" creativity, the "generation of novel ideas" within a space, such as an AI model responding to a given task by creating new content within the parameters of the task. Even more broadly, Pease and Colton argue that computational creativity should be separated from human-centric notions, and creative behavior in AI models should be measured in entirely separate ways from human creativity (Pease and Colton 2011). These process-oriented definitions suggest that some generative models can exhibit a form of *task-oriented* creative behavior in response to prompts, thus forming the motivation for this paper.

Quantifying this task-oriented creative behavior could provide valuable insights for both the users who interact with these models and the model developers. If a user is presented with clear information about the ways in which a given generative model may exhibit creative behavior in response to their queries, they can make an informed choice about whether they wish to use it. The same information could also pinpoint specific areas of change that those who develop generative models can use to tweak their training data and model architecture to better suit their intended purpose.

The focus of this paper is answering the following research question:

"Can we define mathematical measures quantifying useful aspects of task-oriented creative behavior of popular `img2img` *generation models?"*

We propose to characterize the *task-oriented creativity* of model outputs according to three criteria, which we derived from prior literature on computational creativity as discussed in our *Background and Definitions* section. These criteria are: (1) satisfaction of prompt requirements, (2) cohesion between output artifacts, and (3) novelty of output artifacts. In our *Methodology* section, we introduce mathematical metrics and an iterative process involving construction of image *chains* to measure these, and we subsequently analyze the results through both statistical and visual means. The chain construction is inspired by the *Telephone* game and includes a repeated generation of new outputs based on the previous ones. This process is repeated for a predefined number of steps, with the resulting chain analyzed for the concepts introduced above. To the best of our knowledge, while previous papers have described aspects of creative behavior, this is the first attempt at mathematically measuring useful aspects of task-oriented creativity in popular `img2img` generation models.

Figure 1: KANDINSKY 2.2 and STABLE-DIFFUSION 3 chains seeded with the same image, donuts_014.

Due to the brevity of this paper, we only present the summary results and a very small number of illustrative images from our experiments. The code, full experimental results,

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

and chains of images can be downloaded from `https://figshare.com/s/d88d4966b606163d02fc`.

Background and Definitions

We use the term "artifact", derived from (Hauhio 2024)'s usage as well as the common linguistic meaning, to refer to a core feature of a text or image, such as "apple pie" in the textual prompt "a slice of apple pie", or a segment of an image labeled "pie" by an object detection tool. Each prompt and output can therefore be understood as a set of artifacts.

Hauhio's paper proposes multiple artifact spaces, the relationship between which can be used to classify models' behavior as creative. C, referred to as the "conceptual set" by (Hauhio 2024), is the set of artifacts specified in the user's input prompt to a generative model, and V is the set of artifacts a user would find valuable given that this user has a specific goal in using the generative model. Expanding the set of valuable artifacts from strictly C to encompass distinct ones from V as well can therefore be understood as one facet of creative behavior.

We propose to characterize outputs of models according to three criteria, whose formal quantitative definitions we present below: input prompt *requirement satisfaction*, *cohesion*, and *novelty*.

Satisfaction of Prompt Requirements, the idea that an output contains (all) the elements specified in the user's input, builds off (Peeperkorn et al. 2024)'s statement that an output must be typical of its class, as well as (Hauhio 2024)'s "concept set" C. We build *cohesion* from (Peeperkorn et al. 2024)'s assertion that an output must also be *useful* to be considered creative, by working off the idea that an output with elements that are generally more closely related to each other has more use cases than an output with disjoint elements. *Cohesion* is also derived from (Boden 1998)'s "association of ideas": how well does a generative model build strong associations between artifacts within its output in response to a given prompt? For *novelty*, we build off (Hauhio 2024)'s idea of valuable artifacts that were not specified by the input prompt, but are added spontaneously by the generative model, as well as (Peeperkorn et al. 2024)'s and (Boden 1998)'s ideas that an output must contain novel elements to be considered creative. Specifically, (Boden 1998) mentions "transformational" creativity as the idea that new elements can be introduced into a space through transformation, a behavior directly linked to the ability of `img2img` models to generate novel output elements using the framework of an image input.

These measures form a system of checks and balances. Novelty is widely accepted as a necessity for creative output, but including *requirement satisfaction* as a core aspect of task-oriented creativity ensures the penalization of image generators that hallucinate and do not follow task requirements, and *cohesion* ensures that generators are rewarded for producing novel outputs that are semantically comprehensible. This balance aligns with prior ideas surrounding computational creativity, such as Pearce and Wiggins' finding that human reviewers reward AI-generated musical compositions that fall within a range of novelty that does not strike the reviewers as "too strange" (Pearce and Wiggins 2001).

In the definitions below we use the following notations: 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. The *cosine similarity* measure, which is used throughout these definitions, is a way of measuring the similarity between the vector projections of two artifacts in semantic embedding space—in other words, a way to mathematically compare the meanings behind textual representations of artifacts. It will be henceforth denoted as the function θ or as the dot product $\cdot$, depending on context. For a set A of artifacts, the embedding of each artifact $a \in A$ in semantic space is calculated using the embedding function emb, denoted as $emb(a) = \hat{a}$ (Matsuki, Lago, and Inoue 2019).

Definition 1 (*Satisfaction of Prompt Requirements*) *An image is said to be* satisfying requirements *if $A_I \subseteq A_O$. In other words, all artifacts in I are present in O_M.*

To measure the cohesion of O_M, we wish to capture the sentiment of "how much do any artifacts introduced by the generator make sense within the context of the prompt?" To accomplish this, we must measure the closeness of the relationships between seed artifacts and introduced artifacts in the generated image without being too dependent on our third measure, *novelty*, while also not penalizing the generator for low similarity between input prompt elements. We first calculate the mean of the semantic similarity scores between each artifact of A_I and the closest artifact from A_O, and then calculate the maximal similarity between artifacts from A_O. We do this in order to algorithmically approximate the relationships between elements in the generated image without being too dependent on the cardinality of the set of introduced elements, while also not penalizing the generator for low similarity between input prompt elements. As a hypothetical illustrative example, if the input prompt is "a potato and a lion", then $A_I = potato, lion$. If, in the subsequent generated image, $A_O = potato, lion, onion, R2D2, zebra, strawberry$, and each element is represented by its first initial, cohesion can be calculated by first taking $\hat{P} \cdot \hat{O},\ \hat{L} \cdot \hat{Z},\ and\ \hat{O} \cdot \hat{S}$, then taking the average of those three values. This is because the input artifact "potato" has the highest semantic similarity to "onion" within $A_O \setminus A_I$, the input artifact "lion" has the highest semantic similarity to "zebra" within $A_O \setminus A_I$, and the two artifacts within $A_O \setminus A_I$ with the highest semantic similarity are "onion" and "strawberry". Note that "R2D2" is excluded from these measurements because only the maximally similar pairs are considered, and "R2D2" is not maximally similar to either of the seed artifacts, nor is it maximally similar to any of the other introduced artifacts.

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*

$$C(O_M) = \frac{\sum_{a:(a,b)\in P} \max_{b:(a,b)\in P}((\hat{a}) \cdot (\hat{b})) + M_N}{|P| + 1}.$$

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

We note that the scalar product $(\hat{a}) \cdot (\hat{b})$ of the representations of a and b in the semantic embedded space increases with increasing similarity between a and b. As $0 \leq (\hat{a}) \cdot (\hat{b}) \leq 1$, also $0 \leq C(O_M) \leq 1$.

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.*

While *novelty* and *cohesion* are not independent concepts, an output image O_M can have high *novelty* and low *cohesion* if it contains many new artifacts with low similarity to the original ones, and it can also have high *cohesion* and low *novelty* if there are very few new artifacts and they are very similar to the original ones.

Methodology

In this section, we outline our methodology for quantitative evaluation of task-oriented creativity in image generation models.

Chain Construction

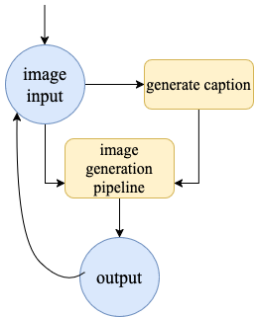


Figure 2: One step in the iterative chain process.

As (Hauhio 2024) points out, image generation models are often used iteratively, with users modifying their input prompts based on the artifacts present in the generated output. In this paper, we simulate this iterative approach, which allows us to measure the afore-discussed aspects of creative behavior in a way that is more pertinent to real-world use-cases. Similar automated approaches have been used for related tasks such as image generation refinement (Wang et al. 2025), but to our knowledge this is the first time such an iterative process has specifically been used to quantify any form of creative behavior exhibited by AI models.

Our experimental setup evaluates how strongly both text and image inputs influence aspects of task-oriented creativity as exhibited by generative models. We base our approach on the well-known game *telephone* (or *Chinese whispers*). As in this game, each chain in our experiment begins with a "ground truth", or seed image. For each step k in the chain, we use the image generated in step $k-1$ as input to the generative model. To produce the text input, we auto-generate a caption for the generated image from step $k-1$ and feed that into the image generation model alongside image $k-1$. This process repeats for x iterations, where x is a parameter chosen by the user. Figure 2 illustrates this process.

Measures for Creativity

Using the conceptual definitions of *requirement satisfaction*, *cohesion*, and *novelty* provided earlier, we can define mathematical measures for each of these aspects in the context of the chains, so that we can quantify the creative behaviors exhibited by the generator used within a given chain.

The *satisfaction factor* measure provides the backbone for the other two, as we want to focus on images that exhibit a higher-than-random level of *requirement satisfaction* for the input prompt. We expand and quantify Definition 1 to produce this measure for a chain of images rather than a single comparison between two images, by measuring how well the prompt requirements are propagated throughout a chain. Therefore, for a given chain, we quantify *satisfaction factor* (RS) as the *normalized longest unbroken sequence* of the chain wherein, at each step, the seed artifact set, or a set of artifacts that approximate the seed artifact set within a cosine similarity threshold t, is a subset of the artifact set of the image generated for that step.

We then multiply this proportion of the chain by the average θ score for the set of pairs consisting of one artifact from the seed set and its closest match in the generated image's artifact set, to account for the fact that the generated image may not be perfectly faithful.

The later in the chain the seed artifacts or their approximations still appear, the higher the *satisfaction factor* value is going to be, although it also depends on how closely the artifacts in the generated image approximate the chain image.

A chain's *Cohesion Factor* (B_R) and *Novelty Factor* (D_R) can be calculated using Definitions 2 and 3 from the last image in the unbroken chain as described in the *Requirements Satisfaction Factor* metric.

An example illustrating the calculation of these three metrics follows: Given a hypothetical chain generated from a seed artifact set ["apple pie" [AP], "horse" [H]], we observe that the eighth generated image in the chain is the last one that falls within the chosen cosine similarity threshold. Furthermore, we observe that has the artifact set ["apple cake" [AC], "horse" [H], "pear" [P], "jockey" [J]]. The proportion of the chain that satisfies the input prompt would be 0.8, and to calculate the *Requirements Satisfaction Factor*, this value would be multiplied by the average similarity scores between the closest corresponding elements from both sets, in this case AP, AC and H, H:

$$\frac{\theta(AP, AC) + \theta(H, H)}{2}.$$

The *Cohesion Factor* would be calculated as

$$\frac{\theta(AP, P) + \theta(H, J) + \theta(P, J)}{3}.$$

The *Novelty Factor* would be 0.5, given that half the elements of the generated image artifact set are not present in the input artifact set.

In order to measure overall task-oriented creativity, we combine these measures to an *overall task-oriented creativity*

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

ranking score, CR, where $0 \leq CR \leq 1$. For a given chain:

$$CR = RS \times \frac{B_R + D_R}{2}. \quad (1)$$

The chain under evaluation is a parameter to all scores we introduce in this section; it is omitted for brevity.

Intuitively, CR can be higher than 0 only if the generated images satisfy the prompt requirements; this prevents us from awarding a high CR to hallucinating models. On the other hand, a mechanistic copying of the seed image is not creative either. Indeed, while $RS = 1$ for a mechanistic copy, both B_R and D_R are 0, hence resulting in the creativity score 0 as well. For models that satisfy the prompt requirements and add new elements to the generated images, both the *novelty* between the new elements and their cohesion with the seed image elements, as well as between themselves, affect CR.

Given a set X of generative models, we can use the means of RS, B_R, D_R, and CR across the total number of chains for all models in X to rank them in terms of specific aspects of task-oriented creative behavior as per our earlier discussion. Of course, this ranking does not necessarily correlate with general creativity. Extensions or modifications to these definitions can be explored to better fit specific domains or to align with different interpretations of task-oriented creativity.

Evaluation

Experimental Setup

For our experiments we selected sets of simple seed images focusing on a single, easy-to-detect subject, as these would be the easiest to measure *requirement satisfaction*, *cohesion*, and *novelty*, since A_I would be 1. Since many object detectors are trained on datasets that prominently feature food images, we pulled our 999 seed images from the publicly-available `Food-101` dataset (Bossard, Guillaumin, and Gool 2014). We selected three food categories at random: "apple pie", "donuts", and "pizza", and randomly selected 333 images from each. Input prompts I were constructed from each of these seed images.

We evaluated our measures on three popular open-source image generation models from Hugging Face: STABLE-DIFFUSION 3 3 (8.1 billion parameters) (Esser et al. 2024), KANDINSKY 2.2 (4.6 billion parameters) (Razzhigaev et al. 2023), and FLUX (12 billion parameters) (BlackForestLabs 2023). For the text input, we generate a text caption for the seed image using the KOSMOS model from Hugging Face (Peng et al. 2023), and the image input consists of the image generated in the previous step of the chain (or, for the first generative step, an image from the seed dataset). In our experiments, we set the maximal length of the chain to 10 to avoid computational explosion, while still producing meaningful results.

To compute the measures of *requirement satisfaction*, *cohesion*, *novelty*, and creativity ranking defined earlier, we used two object detectors, GROUNDINGDINO (Liu et al. 2023) and DETR (Carion et al. 2020), to extract artifacts in textual form from images in the chain, where the initial image is the seed image, and for A_I we simply use a text representation of the subject of the seed image: apple pie, donut, or pizza. For word embedding and semantic similarity calculations between the artifacts, we used the SPACY library (Montani et al. 2023). We selected 0.65 as the value of t for *satisfaction factor*, after running preliminary tests to account for incorrect object detection results.

We used *paired T-tests*, which are statistical tests tailored to compare the application of different processes to the same input data points, in order to determine any statistically significant differences between results for different image generation models, on the same dataset of 999 images. We used SCIPY's `ttest_rel` function to perform these tests.

To verify reproducibility, we repeated a subset of the experiment with 100 randomly picked seed images, confirming that the statistical differences between the reruns match the ones we extracted from the main experiment, with any discrepancies in averages being explained by the vastly reduced size of the seed dataset in the rerun.

Analysis

Our first observation is that the average CR across all the models was relatively low, ranging from 0.08 to 0.18. This means that none of the image generation models we tested performed extremely well in terms of task-oriented creativity. Manual examination showed that these low scores were due to multiple factors: early deviation from the prompt requirements, satisfaction of prompt requirements but low *novelty*, or the generation of unclassifiable objects such as glitchy patterns or irregular geometric shapes. However, the paired T-tests we ran support the claim that task-oriented creativity is measurably different between some generative models. We noted that FLUX displayed a statistically significantly higher average *Requirements Satisfaction Factor* (0.74) than the other two models (0.7 for KANDINSKY 2.2 and 0.45 for STABLE-DIFFUSION 3), but FLUX and KANDINSKY 2.2 performed about the same on *Cohesion Factor* (0.15), *Novelty Factor* (0.25 vs 0.24 respectively), and overall task-oriented creativity CR (0.18 vs 0.17 respectively). STABLE-DIFFUSION 3, however, displayed statistically significantly lower averages across the board, indicating a much lower performance on task-oriented creativity measures than FLUX and KANDINSKY 2.2.

Two sample chains from our process are illustrated in Figure 1. The top chain of images was produced using the model FLUX, while the bottom chain was produced using the model STABLE-DIFFUSION 3. The leftmost image in both chains is the non-generated seed image, with $A_I = apple\ pie$, while the rest were generated by the respective models. In the FLUX chain, *requirement satisfaction* is maintained through the generated images, which contain $A_O = cake$, an artifact that falls within the cosine similarity threshold t as described in our *Methodology* subsection on *Measures for Creativity*. The model also adds novel artifacts: "plate", "cream", "crumbs", and "fork" (with the latter being out of focus in the background of the image). All of these introduced artifacts exhibit high cohesion, since they make sense in the context of the seed image and are all generally highly semantically related: "crumbs", for example, have a close semantic relationship to "apple pie", which also has a close semantic relationship to items like "cream" and "plate". The STABLE-DIFFUSION 3

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

chain, seeded with the same image, maintains *requirement satisfaction* through propagating the seed artifact "apple pie" throughout the chain, and adds two novel elements: "syrup" dripping down onto the pie, and "crumbs" surrounding the pie.

Notably, the object detection tools used within this experiment failed to detect any of the extraneous artifacts, an issue that is unavoidable when using automated tools, but manually calculating the scores for the first chain, dubbed C_1, yields the following values according to our mathematical definitions as per our *Methodology* subsection on *Measures for Creativity*: $RS(C_1) = 0.68$, $B_R(C_1) = 0.52$, and $D_R(C_1) = 0.8$. For the second chain, dubbed C_2, the calculated values are $RS(C_2) = 1.0$, $B_R(C_2) = 0.52$, and $D_R(C_2) = 0.67$. Based on these values, $CR(C_1) = 0.45$ and $CR(C_2) = 0.6$, with the second chain scoring higher due to preserving the seed artifact perfectly. Both chains display medium levels of task-oriented creativity, with some distinctively novel elements. Note that SPACY is still used to calculate *cohesion*, with the maximally similar pairs for C_1 being "apple pie", "cream" and "crumbs", "cream", and the maximally similar pairs for C_2 being "apple pie", "syrup" and "crumbs", "syrup".

Limitations

Apart from computational resources, a significant limitation of our approach is its reliance on existing popular object detection and word similarity calculation tools. These models are prone to making decisions that do not seem intuitive to humans, such as placing two seemingly unrelated words close together in semantic embedding space, labeling one object with a similar but incorrect designation (for example, labeling a slice of apple pie as a slice of cake), or failing to detect an object altogether, as described in our prior discussion of Figure 1. There is currently no way around this limitation, but as the outputs of such tools more closely approach human-level understanding, the quality of our measures will improve accordingly. Additionally, in order to improve the visual accuracy and explainability of our proposed measures, future research could incorporate human observation as a factor in validating the scores output.

We also acknowledge that the simplicity of our seed image dataset, although easy to experiment on and ideal for observing changes, could impact creativity scores, especially since we set the seed artifact sets to only contain the main subject of the image, ignoring extraneous elements in the seed images (such as plates, tables, and phones) for clarity. Seed images with a large artifact set may potentially increase D_R, for example, or make it harder for generated images to satisfy prompt requirements. Exploring task-oriented creativity across diverse domains and more complex datasets subjects is a crucial next step, particularly since the metrics defined within this paper are meant to accommodate that level of complexity.

Conclusions and Future Research

In this paper, we suggest a set of measures to quantitatively evaluate task-oriented creativity in generative models, and run a series of experiments to study their manifestation in a set of popular `img2img` models. The results of statistically comparing these metrics show that some models perform higher on task-oriented creativity measures than others, and this is supported by visual examination of the results.

We can, therefore, answer the research question posed in the introduction affirmatively: our metrics provide a fine-grained and intuitive score of creative behavior in generative models when responding to a prompt, which in turn permits users a greater ability to select the appropriate model for a given task. Further directions for this research include exploring more models, explaining why some generative models score differently, and determining whether these metrics can be generalized to tailor future model architecture and training techniques to specific purposes.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

References

Amabile, T. M. 1983. The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology* 45(2):357–376.

BlackForestLabs. 2023. Flux. `https://github.com/black-forest-labs/flux`.

Boden, M. A. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103(1):347–356. Artificial Intelligence 40 years later.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 – mining discriminative components with random forests. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 446–461. Cham: Springer International Publishing.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Muller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; Podell, D.; Dockhorn, T.; English, Z.; Lacey, K.; Goodwin, A.; Marek, Y.; and Rombach, R. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv* abs/2403.03206.

Hauhio, I. 2024. Enhancing human creativity with aptly uncontrollable generative ai. In *Proceedings of the 15th International Conference on Computational Creativity*. Spain: Association for Computational Creativity (ACC). International Conference on Computational Creativity, ICCC ; Conference date: 17-06-2024 Through 21-06-2024.

Kaufman, J. C.; Sternberg, R. J.; and Glăveanu, V. P. 2019. *A Review of Creativity Theories*. Cambridge University Press. 27–43.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Matsuki, M.; Lago, P.; and Inoue, S. 2019. Characterizing word embeddings for zero-shot sensor-based human activity recognition.

Montani, I.; Honnibal, M.; Honnibal, M.; Boyd, A.; Van Landeghem, S.; and Peters, H. 2023. explosion/spacy: v3.7.2: Fixes for APIs and requirements.

Oppenlaender, J. 2022. The creativity of text-to-image generation. *Proceedings of the 25th International Academic Mindtrek Conference* 192–202.

Pearce, M., and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, 22–32. AISB (Society for the Study of Artificial Intelligence and the Simulation of Behaviour).

Pease, A., and Colton, S. 2011. On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In *AISB 2011: Computing and Philosophy*.

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of large language models? In *International Conference on Computational Creativity*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv* abs/2306.14824.

Razzhigaev, A.; Shakhmatov, A.; Maltseva, A.; Arkhipkin, V.; Pavlov, I.; Ryabov, I.; Kuts, A.; Panchenko, A.; Kuznetsov, A.; and Dimitrov, D. 2023. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In Feng, Y., and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 286–295. Association for Computational Linguistics.

Wang, J.; He, Y.; Zhong, Y.; Song, X.; Su, J.; Feng, Y.; He, H.; Zhu, W.; Yuan, X.; Lu, K.; Huo, M.; Zhang, M.; Li, K.; Chen, J.; Shi, T.; and Wang, X. 2025. Twin co-adaptive dialogue for progressive image generation. *arXiv*.

Wingström, R.; Hautala, J.; and Lundman, R. 2022. Redefining creativity in the era of ai? perspectives of computer scientists and new media artists. *Creativity Research Journal* 36(2):177–193.

ISBN: 978-989-54160-7-3

The Wizard in the Town Plaza: Voice-Based Interactive Storytelling in Public Spaces

Daniel DeFreez[1] and **Jackson Harrower**[1] and **Travis Wood**[1] and **Cynthia Salbato**[2]
[1]Southern Oregon University, Ashland, OR 97520 USA
[2]ARTNOW, Ashland, OR 97520 USA
defreezd@sou.edu, harrowerj@sou.edu, woodt1@sou.edu, secretstoryworld@gmail.com

Abstract

We present The Wizard in the Town Plaza: a field report on a system that generates narrative content using advanced voice cloning, a community-informed knowledge base, and a structured story world. In March 2025 (spring equinox), we deployed this system in a small tourist and college town in the USA as part of a seasonal event. A human actor served as a live intermediary between the AI system and visitors, relaying audience questions to the AI system which responded directly through loudspeakers with a cloned voice. This deployment revealed how AI-driven storytelling can operate in oral, performative, and social contexts, offering cultural and educational value. We detail the system architecture, including cloned voice, narrative knowledge graphs, and LLMs, while addressing ethical considerations and community response. We also outline its expansion into an interactive mural project designed to amplify local Indigenous voices and history through AI-supported storytelling.

Introduction

In March 2025, during a spring equinox theatre event in Ashland, Oregon, we deployed The Wizard as a live public installation. This approach allowed us to test how large language model-driven storytelling, grounded in a persistent knowledge graph, could function in real-time within a live, performative, and community-based context. A human actor served as The Wizard's representative, relaying visitors' questions into a microphone. The Wizard responded through a loudspeaker system with a rich, expressive voice that carried across the plaza, creating an immersive theatrical experience.

Our deployment location, Ashland, is a small tourist and college town with deep-rooted arts culture and home to the renowned Oregon Shakespeare Festival. We position this project as a cultural application of computational creativity, offering insights into system design, voice synthesis, knowledge graph user experience, along with a reflection on its social, ethical, and creative implications.

For this first public deployment, we chose to use a human liaison between the AI system and the audience as a risk mitigation strategy, given this was our initial test of the system in a public setting. Having a human intermediary allowed us to ensure smooth interactions and quickly address any technical issues that might arise during the live performance.

Our approach builds on and extends a growing body of research in AI-driven narrative systems. Branch et al. (Branch, Mirowski, and Mathewson 2021) developed a GPT-3-powered narrator for stage productions with human curation of outputs, while Shape of Story (Long, Jacob, and Magerko 2019) leveraged voice input for interactive storytelling. In contrast, our approach enables real-time conversational interactions without human content curation.

Góngora et al.'s PAYADOR system (Góngora et al. 2024) addresses the world-update problem in interactive storytelling by recasting it as an outcome-prediction task, enabling story worlds to evolve based on user interactions. PAYADOR uses a minimalist typed schema of characters, items, and locations, similar to the schema in our knowledge graph, though our current deployment operates in a read-only mode. PAYADOR's approach is relevant as it confirms that lightweight, structured grounding can effectively manage large models without heavy symbolic planners.

Commercial solutions like OpenAI's ChatGPT advanced voice mode (OpenAI 2023) offer impressive capabilities, but are completely dependent on their proprietary systems which, at the time of our project, did not allow voice customization. Alternative services like ElevenLabs had terms of service that did not satisfy legal requirements, compelling us to develop our own solution. We worked with a local voice actor who recorded a scripted set of wizard dialogues that we then used to condition a voice model at inference time. While we continued to use OpenAI's API for text generation, we implemented our own custom voice synthesis pipeline using the recorded voice samples and the open-weight TTS model Zonos (Zyphra 2023) instead of relying on proprietary text-to-speech services.

The contributions of this work include:

- A system that connects knowledge graphs to language models, enabling a variety of entities (human and non-human) to emerge from a shared knowledge base
- An architecture achieving real-time conversation with open-weight models allowing control of voice data
- A practical demonstration that AI characters with ethically-sourced voices can successfully engage public audiences

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Knowledge Representation

One of the central challenges in AI-driven storytelling is narrative coherence, ensuring that characters, events, and timelines remain consistent over time and across interactions. Large language models (LLMs), while rich in generative capacity, are prone to losing track of details in extended conversations or across multiple sessions.

To solve this, our system incorporates a structured knowledge graph using Kanka (Owlchester 2024), a collaborative worldbuilding tool designed for non-technical users. Kanka serves as a narrative database that maps the relationships between characters, events, locations, and lore across our interactive narrative system.

Prior work (Wang et al. 2023; Blin 2022) supports the use of knowledge graphs to guide and constrain generative models, preventing contradictions and enabling more believable long-form narratives. Our integration of Kanka supports retrieval-augmented generation, ensuring that responses reflect accurate story context while allowing improvisational flexibility.

Our implementation accesses Kanka through its REST API, retrieving entities such as characters and quests from a structured campaign database. We chose Kanka over a traditional database approach specifically because of its user-friendly web interface, which allows system administrators and non-technical storytellers to collaborate on content creation without requiring custom UI development. While we could have implemented a standard database solution, this would have necessitated building a custom administrative interface—a significant additional development effort. Instead, Kanka provides an intuitive WYSIWYG editor with built-in relationship management tools that creators can use immediately, while our system programmatically accesses this shared narrative state through the API during interactions. Each entity in the knowledge graph maintains relationships to other entities, creating a rich network of connections that the narrative state machine traverses as it responds to user queries, ensuring responses are contextualized within the established mythology.

Narrative State Machine

To maintain state across interactions, we implemented a Narrative State Machine (NSM) that tracks a user's progression through a particular quest or story arc. This state machine (shown as a component in Figure 1) queries Kanka's structured knowledge base while maintaining its own internal state. Whereas Kanka provides the permanent narrative content and world knowledge, the NSM itself manages state transitions and conversation history in a separate Firestore database, storing the current node and interaction history for each user session. This separation of concerns ensures the wizard maintains consistent characterization across sessions and allows multiple users to progress through different narrative paths simultaneously. When participants ask questions that might contradict established lore or lead the character off-track, the NSM queries Kanka for canonical information.

Drawing from the theory of computation (Sipser 2013), we define our Narrative State Machine (NSM) as a formal

System Architecture

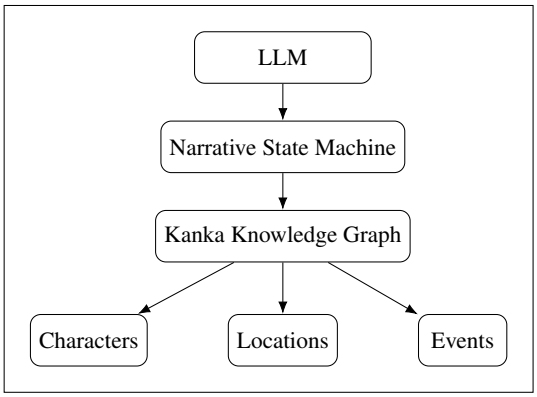


Figure 1: System architecture showing the relationship between the LLM, narrative state machine, and Kanka knowledge graph components used to maintain consistency across interactions.

model consisting of a set of states and a transition function augmented with an LLM matcher. Unlike traditional computational models, the NSM integrates natural language understanding with discrete state transitions, enabling narrative coherence across extended interactions.

Figure 2 shows the Narrative State Machine algorithm. Rather than using symbols from a finite alphabet to define transitions, the NSM transition function δ maps natural language descriptions (sentences written by the state machine author) to next states. For example, a transition might be defined as "User asks about the wizard's origins" or "User expresses disbelief in magic."

The transition function $\delta(q, \sigma, \mathcal{H}) \rightarrow (\text{response}, q')$ determines the next state by considering the current state q, user input σ, and conversation history $\mathcal{H}$. It returns a natural language response (response) and the new state q'. Within the function, the knowledge context K is retrieved based on the current state and user input. The conversation history $\mathcal{H}$ is maintained internally and updated. We also track node traversal history separately for analytics, but this is not represented as a formal parameter in the function signature.

The LLM is used in two distinct steps: first as LLM_δ to match user input to natural language transition descriptions, then as LLM_r to generate contextually appropriate responses based on the new state. The algorithm naturally terminates when it reaches a state with no outgoing transitions, representing an ending in the narrative. This approach leverages the LLM's semantic understanding to bridge formal state transitions with natural language interaction. The NSM allows end users to interact with the state machine using natural language, while enabling non-technical authors to define transitions using intuitive descriptions rather than formal symbols or code.

Voice Interface and Cloned Character Voice

Our motivation for creating a custom voice solution was twofold: to employ local talent and to achieve the highest

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

function $\delta(q, \sigma, \mathcal{H})$
    $\mathrm{K} \leftarrow \text{Knowledge}(q, \sigma)$
    $T \leftarrow \text{GetAvailableTransitions}(q)$
    if $T = \emptyset$ **then**
        return "End of story", q
    end if
    $q' \leftarrow \text{LLM}_\delta(q, \sigma, T, \mathrm{K}, \mathcal{H})$
    response $\leftarrow \text{LLM}_r(q', \sigma, \mathrm{K}, \mathcal{H})$
    return response, q'
end function

Figure 2: Narrative State Machine transition function δ where q is the current state, σ is user input, $\mathcal{H}$ is conversation history, response is the natural language output, q' is the next state, and K is knowledge context retrieved from Kanka. The function terminates at states with no outgoing transitions.

possible quality voice output. The wizard's voice is synthesized using locally hosted voice cloning system, leveraging the open-weight TTS model Zonos. Zonos v0.1 (1.6B parameters) was trained on over 200,000 hours of multilingual speech, delivering emotionally nuanced vocalizations that rival commercial TTS systems. It is our experience that Zonos successfully captures enough of the emotional range and expressiveness of the voice actor to be effective for our purposes (Zyphra 2023).

The voice recording session involved two separate recordings. First, the actor read a scripted wizard dialogue based on anticipated interactions. Second, the actor performed various passages with a range of emotional expressions—sad, excited, conversational, and more—to create a comprehensive vocal palette. In total, we recorded about five minutes of speech, though Zonos voice cloning only requires around 10 seconds of high-quality audio to condition the model at inference time. This approach of recording more material than needed and selecting optimal segments lets us choose the best emotional tone and voice quality. For this particular event, we sought a booming, authoritative wizard voice, so we selected specific segments that embodied this persona.

Voice cloning tolerates variability in the specific content recorded, but is highly sensitive to intonation, emotional range, and voice quality.

Because Zonos can only synthesize a few seconds of audio at a time, we incorporated NLTK sentence segmentation tools (NLTK Project 2023) to split wizard responses into manageable chunks. NLTK's sentence tokenization provided reliable boundary detection even with the wizard's archaic speech patterns and lengthy narrative descriptions. These chunks were rendered individually, then reassembled.

Audio can be streamed real-time using LiveKit, an open-source WebRTC platform. Deployed on a lightweight 2GB Digital Ocean droplet, it enabled low-latency, real-time interaction even in constrained computing environments (LiveKit 2023).

Character Factory A key feature of our system is that any entity in the storyworld can be engaged with directly. While the wizard served as our primary deployment persona, our "character factory" architecture supports the generation and deployment of multiple interactive entities that share the same knowledge graph. The system can dynamically create new characters through API requests that expand prompts using the Kanka knowledge base, allowing for rapid generation of additional NPCs like shopkeepers, artisans, or guides that remain consistent with the shared world.

This approach enables automated character scaling and allows not just human characters but also objects, plants, animals, and abstract concepts to be given voice and agency. Visitors can converse with a variety of entities within the shared narrative universe, each with appropriate interaction styles.

Multimodal Interaction Capabilities Beyond voice interaction, our system supports SMS and MMS text messaging, providing accessibility for users in low-bandwidth environments or those who prefer asynchronous communication. Using Twilio's messaging services integrated with Modal Labs' serverless compute functions, users can engage with characters via text messaging to continue their narrative journeys while on the move.

The system's multimodal capabilities are particularly powerful for place-based games and storytelling. Since generative AI can now process images alongside text, users can send photographs of real-world locations, landmarks, or objects which the system then incorporates into the narrative. For example, a visitor could photograph a distinctive tree in Lithia Park and the wizard might respond with lore about "Forest Guardians" specific to that location.

All user state is synchronized across interaction channels, maintaining narrative continuity whether engaging through voice, text, or image, which allows transitions between different modes of engagement.

Spring Equinox Festival

In March 2025, during a spring equinox festival in Ashland, we debuted the Wizard character as part of a larger community event.

We installed the experience in the heart of Ashland's downtown plaza. The actor welcomed visitors with a narrated opening ritual, then invited them to ask questions of the wizard. The AI-generated responses were voiced through our local voice cloning system and amplified through speakers, immersing the audience in a living dialogue with a mythical figure.

The technical setup relied on a single NVIDIA 3090 GPU, running local voice synthesis for minimal latency in a location with inconsistent internet. The decision to use offline voice generation ensured fidelity and resilience, allowing the wizard to remain responsive throughout the event. The system provided consistent character behavior throughout the event.

Audience Feedback

The audience consisted primarily of families with children, who showed particular enthusiasm for interacting with the wizard character. Young visitors asked questions about

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

magic and the wizard's backstory. Despite the mediated format, the audience was highly engaged with the experience, with the actor's theatrical presence helping to anchor attention while the AI's narrative fluency created compelling interactions. One confounding factor in our evaluation was the excellent performance by the human liaison, who skillfully drew visitors into conversation It remains an open question whether people would engage as readily with the AI system without a charismatic human intermediary to facilitate the interaction. Several attendees were surprised to learn they had been conversing with an AI-powered system. While this initial deployment did not include quantitative measurements, in future work we plan to conduct a formal study with surveys and metrics to assess audience engagement, comprehension, and satisfaction.

Ethical Considerations

Voice Cloning Ethics

AI-cloned voices carry growing ethical weight, especially in light of high-profile misuses in political and commercial arenas. These range from fraudulent Medicare scams featuring celebrities on YouTube (Koebler 2024) to political manipulation, as demonstrated by the recent $6 million FCC fine for AI-generated robocalls impersonating President Biden in a primary election (Naylor 2024). Our approach followed a clear ethical framework, starting with consensual voice capture: we used samples from a paid, local actor who explicitly approved the use of their voice for the AI wizard character.

Privacy was central to our design. All voice synthesis and transcription were performed locally on GPU-enabled hardware, ensuring that no personal audio data was transmitted to the cloud. This commitment to data sovereignty was both a technical necessity (given intermittent connectivity) and a cultural alignment with Ashland's community values of autonomy and transparency. Additionally, local processing avoided critical concerns about transferring sensitive voice data to third-party services, where ownership rights become ambiguous under terms of service agreements that often grant providers broad rights to use uploaded voice samples for their own model training or other purposes. By maintaining complete control of all voice data locally, we ensured clear ownership boundaries and prevented potential misappropriation of the actor's vocal identity.

Labor Considerations

The emerging tension between AI voice synthesis and professional voice actors has become a significant ethical issue, highlighted by the ongoing SAG-AFTRA video game strike that began in July 2024 (SAG-AFTRA 2024). The strike focuses specifically on establishing AI guardrails to protect voice actors from having their performances replicated without consent or compensation.

Our project acknowledges these legitimate labor concerns by implementing several ethical safeguards:

(1) Compensation: We hired and paid a local voice actor for his performance, ensuring compensation for his time and talent;

(2) Clear usage limitations: We established a clear agreement that the voice actor's recordings would be used exclusively for the Storyworld project and not for any other commercial or non-commercial projects;

(3) Data sovereignty: All voice data was processed locally and not shared with third parties, ensuring the actor maintained control over how his voice was used.

Rather than replacing voice actors with AI, we aimed to show how AI can extend and amplify human creative work while respecting artistic agency. While AI systems now make it possible to generate synthetic voices with minimal human involvement, this case demonstrates that community-centered AI projects can establish new standards for ethical labor practices rather than circumventing them. Human voice actors deserve both attribution and appropriate compensation when their vocal performances form the foundation of synthetic voices.

Future Work

Interactive Mural Application

The Wizard character project is the foundation for another undertaking: an interactive, AI-powered mural installation that would bring Ashland, Oregon's history to life through indigenous storytelling. This mural will depict significant historical and cultural scenes from the region, with a particular focus on indigenous perspectives that have often been overlooked in traditional historical narratives.

Our concept is to re-purpose the wizard infrastructure to tell authentic indigenous stories from the Shasta, Takelma, and other tribes indigenous to the Rogue Valley region. This would transition our project from fantasy narratives to real-world history through the same technical architecture, while amplifying indigenous voices that have been historically marginalized.

Initial conversations have begun with a local muralist and cultural organizations in Ashland about this potential collaboration. The proposed project would blend art, oral storytelling, and AI-driven interaction to transform a public wall into a living archive of indigenous voices and stories. We recognize that such a project requires extensive consultation with tribal representatives, and we have begun discussions with local indigenous knowledge keepers to ensure that if the project proceeds, it would be developed with appropriate cultural context and respect for tribal sovereignty and intellectual property rights.

If this project moves forward, visitors would be able to engage with interactive elements powered by the same voice cloning, large language models, and narrative knowledge graph infrastructure field-tested in our plaza deployment. Such a project would require careful attention to cultural authenticity, appropriate representation, and proper attribution of indigenous knowledge and stories. We acknowledge that any such project must address complex ethical questions around voice cloning of indigenous storytellers and the digitization of traditional knowledge systems.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Visitor Information Booth Application

We are also in discussions with Ashland city officials to deploy AI characters as virtual attendants for the town's visitor information booth, which is currently unstaffed. This application would allow tourists to access information about local attractions, events, and history through conversation with AI characters that reflect Ashland's cultural identity.

Visitor: hi
The Wizard: Greetings, traveler! You've found your way to the enchanted realm of Ashlandia. How may I assist you in your journey today? Whether you're seeking hidden treasures, magical experiences, or simply a delightful tale, I'm here to guide you.
Visitor: where can I get some children's books?
The Wizard: Ah, for enchanting tales and magical stories, TreeHouse Books is a splendid haven for young readers in Ashlandia. Established in 1978, this whimsical bookstore captures the imagination with its fairy-tale murals, cozy nooks, and delightful selection of children's literature. Alongside treasures for the young and young-at-heart, you'll find puppets, costumes, and art supplies to spark creativity. Whether you seek a gift or a new adventure tale for bedtime, TreeHouse Books awaits to welcome you into its enchanted world.

Figure 3: SMS/MMS response from our visitor information booth proof-of-concept, showing factual local information delivered with fantasy-themed presentation.

This interaction demonstrates how fantasy and utility can blend in AI-assisted tourism. Messages sent via SMS/MMS or voice calls receive identical information wrapped in the Wizard's distinctive voice, maintaining persona consistency while delivering accurate local knowledge.

Confabulation Detection for Information Services For both visitor assistance and indigenous stories, factual accuracy is paramount. To address the risk of LLM confabulation, also known colloquially as hallucinations, we are developing a citation-based verification system specifically for these applications. When providing tourist information, the AI character must cite specific sources from knowledge graph and a separate verification character checks these citations to ensure they exist and support the provided information. While double confabulation remains possible (where both the primary LLM and verifier get it wrong), the probability is reduced through this two-stage process.

Conclusion

We have presented a place-based, voice-powered interactive storytelling system designed as a showcase for computational creativity in public space. Our system integrates natural-sounding voice cloning with a knowledge-driven narrative engine for locally resonant storytelling. The deployment in Ashland demonstrated public willingness to engage with AI characters in playful dialogue and as companions in cultural exploration, showing potential to contribute to this artistic community's culture.

From a technical perspective, our work demonstrates that open-weight models can achieve production-quality voice synthesis and interactive storytelling when properly integrated. The combination of local voice models, knowledge graphs, and state tracking provides a robust framework for creating believable AI characters without dependence on proprietary cloud services.

From a computational creativity perspective, our project highlights the importance of embodiment (voice) and context (knowledge graph) in making AI creativity accessible, situated, and socially impactful. By combining these elements in real-time, performative environments within the distinctive cultural context of Ashland, we believe that generative AI can function as a facilitator of shared memory, imagination, and cultural exchange.

The contributions of this work, as outlined in the introduction, include: (1) a system that connects knowledge graphs to language models, enabling a variety of entities (human and non-human) to emerge from a shared knowledge base; (2) an architecture achieving real-time conversation with open-weight models allowing control of voice data; and (3) a practical demonstration that AI characters with ethically-sourced voices can successfully engage public audiences. These innovations show how AI-driven storytelling can thrive in oral, performative, and social settings. As we expand into applications like the interactive mural project, we continue to explore how our character factory approach can bring to life not just human characters but also objects, locations, and abstract concepts through API-driven character generation and AI-supported storytelling, while carefully considering the balance between creative interaction and systemic safeguards.

Author Contributions

Cynthia Salbato provided the overall vision and original concept for the project. Daniel DeFreez was the main architect of the software system and primary author of this manuscript. Travis Wood and Jackson Harrower assisted with research experiments, with Travis being the main proponent of voice cloning technology. All authors contributed to the writing of this manuscript.

Acknowledgments

We thank Travis Puntarelli for providing the voice of the Wizard, bringing the character to life with a delightful vocal performance. Thanks to Aurora Quinn Muller for serving as liaison between the audience and the Wizard during the spring equinox festival deployment, facilitating the interactive experience with theatrical flair. Special thanks to Elliot Glenn for initial work on an internal university poster project that was the foundation for Narrative State Machines. This work was generously supported by Southern Oregon University's STEMReX research for undergraduates program.

References

Blin, I. 2022. Building narrative structures from knowledge graphs. In *The Semantic Web: ESWC 2022 Satellite Events*, volume 13384 of *Lecture Notes in Computer Science*, 234–251. Springer.

Branch, B.; Mirowski, P.; and Mathewson, K. W. 2021. Collaborative storytelling with human actors and AI narrators. *CoRR* abs/2109.14728.

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 (ICCC)*, 101–106. Jönköping, Sweden: Association for Computational Creativity. System that grounds LLMs to minimal representations of fictional worlds for RPGs and interactive storytelling.

Koebler, J. 2024. Deepfaked celebrity ads promoting medicare scams run rampant on youtube. *404 Media*. Investigation of widespread AI-generated deepfake videos featuring celebrities in fraudulent Medicare scam advertisements, viewed over 195 million times.

LiveKit. 2023. Livekit: Open source infrastructure for real-time communication. LiveKit Project Website. WebRTC-based platform for building real-time audio, video and data applications. Accessed: 2025-04-26.

Long, D.; Jacob, M.; and Magerko, B. 2019. Designing co-creative ai for public spaces. In *Proceedings of the 2019 Conference on Creativity and Cognition*, 479–490. ACM.

Naylor, B. 2024. Fcc fines political consultant $6 million for ai-generated robocall impersonating biden. *NPR*. Report on FCC's first enforcement action against AI-generated voice cloning used for political manipulation in a New Hampshire primary election.

NLTK Project. 2023. Natural language toolkit. NLTK Website. Open-source Python library for NLP with over 50 corpora and lexical resources. Accessed: 2025-04-26.

OpenAI. 2023. ChatGPT can now see, hear, and speak. OpenAI Website. Announcement of multimodal capabilities in ChatGPT including voice conversations and image processing.

Owlchester. 2024. Kanka - online tabletop RPG campaign manager and worldbuilding tool. Web application. Community-driven platform for creating and organizing RPG campaigns with collaborative worldbuilding. Accessed: 2025-04-26.

SAG-AFTRA. 2024. Interactive media (video game) strike. Official SAG-AFTRA website. Labor strike involving voice actors and motion capture artists over AI protection concerns. Began July 26, 2024. Accessed: 2025-04-26.

Sipser, M. 2013. *Introduction to the Theory of Computation*. Boston, MA: Course Technology, third edition. Classic textbook covering automata theory, computability theory, and complexity theory.

Wang, Y.; Lin, J.; Yu, Z.; Hu, W.; and Karlsson, B. F. 2023. Open-world story generation with structured knowledge enhancement: A comprehensive survey. *Neurocomputing* 545:126091.

Zyphra. 2023. Zonos-v0.1: A leading open-weight text-to-speech model. GitHub Repository. Apache 2.0 License, trained on 200k hours of multilingual speech. Accessed: 2025-04-26.

ISBN: 978-989-54160-7-3

The truth is no diaper: Human and AI-generated associations to emotional words

Špela Vintar[*,†] **& Jan Jona Javoršek**[†]
[*]University of Ljubljana, Slovenia
[†] Jožef Stefan Institute, Ljubljana, Slovenia
{spela.vintar, jona.javorsek}@ijs.si

Abstract

Human word associations are a well-known method of gaining insight into the internal mental lexicon, but the responses spontaneously offered by human participants to word cues are not always predictable as they may be influenced by personal experience, emotions or individual cognitive styles. The ability to form associative links between seemingly unrelated concepts can be the driving mechanisms of creativity. We perform a comparison of the associative behaviour of humans compared to large language models. More specifically, we explore associations to emotionally loaded words and try to determine whether large language models generate associations in a similar way to humans. We find that the overlap between humans and LLMs is moderate, but also that the associations of LLMs tend to amplify the underlying emotional load of the stimulus, and that they tend to be more predictable and less creative than human ones.

Introduction

The free association task is a simple method of eliciting spontaneous connections between words from individuals. Given a cue word, such as *fish*, the participant is asked to respond with one or several words that come to mind, such as *trout*, *sea bass*, *lake*, *fishing* or *fresh*. The technique was first used in psychology by Galton (1880) and later by Freud to examine his patients' thoughts, in particular hesitations and emotional reactions to certain stimuli (Freud 1913). By mid-20th century, the association task was considered a prominent way of accessing mental representation and memory, and was used extensively in cognitive psychology and cognitive science (Deese 1959).

More recently, a number of research studies have been dedicated to comparing internal semantic representations (including word associations) to distributional semantics models and, in the past few years, large language models (Mandera, Keuleers, and Brysbaert 2017; Nematzadeh, Meylan, and Griffiths 2017; Günther, Rinaldi, and Marelli 2019). While first generation LLMs would generate a lot of erratic associations where the link between the cue and response was obscure at best (Brglez, Vintar, and Žagar 2024), state-of-the-art models seemingly excel at mimicking almost any linguistic behaviour. Associations lie at the core of many creative endeavours, and - as recent studies show - can significantly improve tasks like humour generation (Tikhonov and Shtykovskiy 2024). But challenges remain, as their performance significantly drops in non-major languages (Xuan et al. 2025), and their ability to form associations in a human-like manner remains understudied.

We focus on associations to emotionally loaded words, with two research questions in mind: 1. Do LLM-generated associations resemble human ones in terms of creativity, type and overlap, and 2. If emotionally loaded words elicit particular types of sentiments in human responses, can we observe the same distribution of positive and negative sentiments in the generated responses? Our experiments are performed for Slovenian, a morphologically rich language from the family of Slavic languages which currently has just over 2 million speakers.

Related work

Before we review other attempts to compare human and computer-generated associations, it is important to present works which categorize the processes in human associative behaviour. A comprehensive explanation is given in Clark (1970), who distinguishes amongst association eliciting experiments where participants are under major, moderate or no time pressure. If the person is allowed some time, they "react with rich images, memories, or exotic verbal associations, and these give way to idiosyncratic, often personally revealing responses." Under time pressure, associations become more superficial, predictable and closely related to the stimulus. Clark then continues to present two broad categories of responses: *paradigmatic*, where the response falls into the same syntactic category as the stimulus, and *syntagmatic* where it does not; both mechanisms being governed by implicit linguistic rules such as minimum-contrast, feature-deletion and -addition or idiom-completion. A further claim is that paradigmatic associations are much more frequent than syntagmatic ones.

Fitzpatrick (2007) challenges the belief that human associative behaviour is homogeneous, and her experiments show large variability among adult native speakers in the categories of responses. She introduces a more fine-grained categorization (which we describe in the section Approach) and shows that associative mechanisms reveal an intricate

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

interplay between cues, responses and individual respondents' associative styles.

In recent years, several authors have explored the ability of vector space models to represent conceptual organization. Mandera, Keuleers, and Brysbaert (2017) performed a detailed evaluation of correlations between human semantic spaces and corpus-based vector representations, whereby for the former they use semantic priming, semantic relatedness judgements and word associations. They find that static neural models outperform traditional count models, and that the window size used in training plays a significant role in the performance of the models.

In an experiment by Nematzadeh, Meylan, and Griffiths (2017) human word associations were compared to nearest neighbours suggested by word2Vec and GloVe, and they show that overall correlation is low and that static word embeddings fail to capture certain critical aspects of human associations. A similar conclusion was proposed by Günther, Rinaldi, and Marelli (2019), who emphasize that "word meanings are acquired through experience". For this reason, models trained directly on introspective data generally outperform corpus-trained ones (De Deyne, Perfors, and Navarro 2016).

A more recent detailed discussion of the complexity of human associative behaviour and neural modelling is provided by Richie, Aka, and Bhatia (2024) who also train their GloVe model on English SWOW (De Deyne et al. 2018) and achieve good prediction results using a variety of asymmetric measures.

Finally, (Lavorati et al. 2024) generate the entire SWOW dataset with Mistral and present some properties of LLM-generated associations against human ones. While no detailed comparison of response categories was performed, the LLM generated almost 3 times fewer unique responses than humans, showing that the diversity and variability of human responses is much higher than for those generated by the current models.

Our own experiments extend the work cited above in that we focus on emotional words and their polarity, we compare the novelty and creativity of human vs. LLM-generated associations and compute average overlap between them.

Datasets and methods

We use the recently constructed SWOW-SL dataset as the source of human association norms. SWOW-SL[1] was compiled as the Slovenian branch of the Small World of Words project[2] (De Deyne et al. 2018), a site which currently collects word associations for 19 languages of the world including Slovenian. The SWOW-SL v1.0 consists of human responses to 1,000 cue words, totalling over 60,000 associations (20,186 unique responses) from 1396 participants. Since our focus is on emotionally loaded words, our second important resource is the Slovenian Emotion Dimension and Emotion Association Lexicon SloEmoLex 1.0 (Brglez et al. 2024), an extension of the LiLaH CroSloEng Emotion lexicon created by Ljubešić et al. (2020). The lexicon contains Valence, Dominance and Arousal scores for almost 20,000 Slovenian words, of which around 14,000 are also assigned binary variables for Positive and Negative sentiment along with a discrete model of emotion covering anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

Using the score for Valence which is generally considered as a measure of emotional load we select top 20 and lowest 20 Slovenian words, whereby the words must also be included as cues in the SWOW-SL association database. These words represent samples of positively and negatively loaded parts of the vocabulary, and include:

Positive: *resnica* (truth), *odlično* (excellent), *zmagovalec* (winner), *srečen* (happy), *sposoben* (skillful), *zadovoljen* (content), *zabava* (fun), *mir* (peace), *zlat* (gold), *svoboda* (freedom), *zdrav* (healthy), *mama* (mom), *mati* (mother), *dobiček* (profit), *praznik* (holiday), *sprejet* (accepted), *prijazen* (kind), *sestra* (sister), *ženski* (female), *dogovor* (agreement)

Negative: *napadalec* (attacker), *padec* (fall), *povzročiti* (inflict), *nevarnost* (danger), *poraz* (defeat), *strel* (shot), *bolnik* (patient), *nevaren* (dangerous), *žrtev* (victim), *bolečina* (pain), *smrt* (death), *pomanjkanje* (poverty), *pritožba* (complaint), *poškodovan* (hurt), *nasprotnik* (opponent), *slabo* (bad), *nasilje* (violence), *zavrniti* (reject), *izdati* (betray), *umreti* (die)

Approach

After selecting the emotional cues, we retrieve associations from the SWOW-SL database, and generate associations using 3 contemporary LLMs:

- Llama-3.3: a text-only 70B instruction-tuned model by Meta,
- GaMS-9B: a new improved and larger model of the GaMS (Generative Model for Slovene) family, based on Google's Gemma 2 family and continually pretrained on Slovene, English and some portion of Croatian, Serbian and Bosnian corpora, developed by CJVT[3],
- Claude 3.7 Sonnet: a commercial reasoning model developed by Anthropic and an enhanced version of Sonnet 3.5, number of parameters not public.

The prompt to generate association is zero-shot: "What does the word X remind you of?", with the specific instruction to all models to provide only 3 associations to each cue. Once the models provided their responses we perform several sets of analyses: firstly, we examine the association inventories and look for unexpected, novel or creative responses; secondly, we compute overlap between all 4 versions (human vs llama/gams/claude); and thirdly, we label associations with Valence scores and examine the average emotional polarity of each setting.

Creative responses. For the purposes of this manual analysis, we loosely rely on the categories of responses as defined by Fitzpatrick (2007):

[1] http://hdl.handle.net/11356/1980

[2] https://smallworldofwords.org

[3] Centre for Language Resources and Technologies at the University of Ljubljana.

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

Cue [Gloss]	Response [Gloss]
resnica [truth]	*ni plenica* [no diaper]
▶ 'the truth is no diaper', words rhyme	
resnica [truth]	*skazica* [turns.out]
▶ 'truth turns out': rhyming allusion to a common expression *pravica – skazica* (what is just is rewarded in the end)	
odlično [excellent]	*oreščki* [nuts]
▶ refering to *Odlično* nuts and seeds brand in Slovenia	
odlično [excellent]	*lari fari* [blah blah]
srečen [happy]	*vrvica* [string]
▶ probably a reference to an old Slovenian movie *Sreča na vrvici* [Happiness on a leash] about a boy and his dog	
srečen [happy]	*lonec* [pot]
▶ probably a reference to the rhyme *srečen konec, počen lonec* [happy end – broken pot]	
zmagovalec [winner]	*the winner takes it all (abba)*
▶ reference to an Abba song	
sposoben [capable]	*sanjati* [of.dreaming]
mir [peace]	*megla* [fog]
mir [peace]	*golobica* [dove]
▶ reference to the dove as a symbol of peace	
mama [mom]	*stara mama k štrudl peče* [granny baking strudel]
mama [mom]	*rdeča* [red]
mama [mom]	*neprespana* [sleepless]
mama [mom]	*mari* [Mari]
▶ a female name	
mama [mom]	*luč* [light]
strel [shot]	*sarajevo* [Sarajevo]
▶ probably a reference to the Bosnian capital under siege	
nasilje [violence]	*črna* [black]

Cue [Gloss]	Response [Gloss]
smrt [death]	*ja* [yes]
smrt [death]	*deževna trata* [rainy meadow]
▶ possibly personal reference	
smrt [death]	*nena ga več vidiš* [Nena can you see it yet]
▶ a personal comment (in Croatian)	
smrt [death]	*komot* [easily]
smrt [death]	*zakaj je banana rumena* [why are bananas yellow]
▶ an absurdist comment	
smrt [death]	*svet je mejhn* [the world is small]
Llama	
mir [peace]	*miran* ['Miran'/peaceful]
▶ a name, or interference with Croatian for 'peaceful'	
svoboda [freedom]	*avstralija* [Australia]
smrt [death]	*pepelka* [Cinderella]
zavrniti [reject]	*opoziv* [?]
▶ a non-existing word resembling 'opposition' and 'appeal'	
claude	
poraz [defeat]	*nezmaga* [non-victory]
▶ a made-up word	
žrtev [victim]	*trpin* [sufferer]
▶ a rare word	
GaMS	
zlat [golden]	*goldman* ['goldman']
▶ possibly interference from English	
ženski [female]	*spolov* ['of.gender']

Table 1: Examples of unexpected, novel or creative responses from human and LLM contributions

- **Meaning-based association**: x means the same as y, x and y come from the same lexical set, x and y have some other conceptual link
- **Position-based association**: y follows or precedes x directly or with words between them
- **Form-based association**: y is x plus or minus affix, y looks or sounds similar to x
- **Erratic**: y has no decipherable link to x or no response given

The fourth category is the one potentially containing the most creative responses, although other categories may also involve mechanisms which evoke unexpected responses in humans based on rhyme, memories or personal idiosyncrasies. We inspect human and generated associations looking for responses which go beyond typical conceptual or position-based links between cue and response.

Human. A manual analysis of human associations shows the great variability, individuality and creativity in the responses, where the relations between cues and responses range from wordplay, intertextual or intercultural references up to completely obscure associations where the connection evades explanation. Some examples with attempted gloss translations and explanations are listed in Table 1.

LLMs. As opposed to the richness of the human associative space, LLMs produce associations which are far less original and may only occasionally depart from the expected meaning- or position-based responses, so the number of examples in Table 1 is rather small. In addition, LLM-generated associations typically follow the part-of-speech category of the cue or remain within the pool of morphologically related words (eg. *nevaren* [dangerous]: *tvegan* [risky], *hazarden* [hazardous], *nevarnost* [danger], *varovanje* [protection]) as opposed to human responses which are truly associative in that they are in part motivated by the underlying emotion (*nevaren* [dangerous]: *moški* [man], *pes* [dog], *človek* [(hu)man]). The claim that paradigmatic associations are much more frequent than syntagmatic ones (Clark 1970) is thus even truer for LLMs.

Overlap

We compute average overlap by intersecting sets of responses for each cue in each setting (Table 2). The model

corresponding most closely to human associations is the presumably largest, Claude 3.7 Sonnet, followed by Llama and GaMS respectively. As for correspondences between models, the highest match is between Claude and the Slovenian GaMS, which perhaps means that additional target language data improves performance on the association task; an assumption otherwise not proven by the Human-GaMS overlap.

	Human	Llama	GaMS
Llama	18.33	/	12.50
GaMS	14.17	12.50	/
Claude	23.33	14.17	17.50

Table 2: Overlap between human and models' associations

Sentiment analysis

In the subsequent analysis of the sentiment of associations we label each response with its Valence score (between 0 and 1, 0 signifying extremely negative, 0.50 neutral and 1 extremely positive) and the binary Positive and Negative scores. We analyse the 20 positive and 20 negative cues separately, by computing average scores per cue and each of the two groups.

	Valence	Positive	Negative
Human	0.73	0.53	0.11
Llama	0.78	0.41	0.05
Claude	0.78	0.72	0.08
GaMS	0.71	0.53	0.15

Table 3: Average sentiment scores for positive cues

	Valence	Positive	Negative
Human	0.37	0.07	0.47
Llama	0.29	0.04	0.58
Claude	0.29	0.03	0.58
GaMS	0.25	0.07	0.58

Table 4: Average sentiment scores for negative cues

Tables 3 and 4 show the average sentiment scores of responses to emotional words for positive and negetive cues. As explained above, valence scores higher than 0.50 indicate positive sentiment, hence all models generate predominantly responses with a positive sentiment to positive cues and vice versa, with Llama and Claude even surpassing humans in both extremes. The columns Positive and Negative record the average number of responses in those two categories, whereby these values are binary in SloEmoLex (a word can be either positive or not), and not all words have this value assigned. Results are most conclusive for negative cues, where it seems that all models generate even more negative associations than humans. (See Figure 1.)

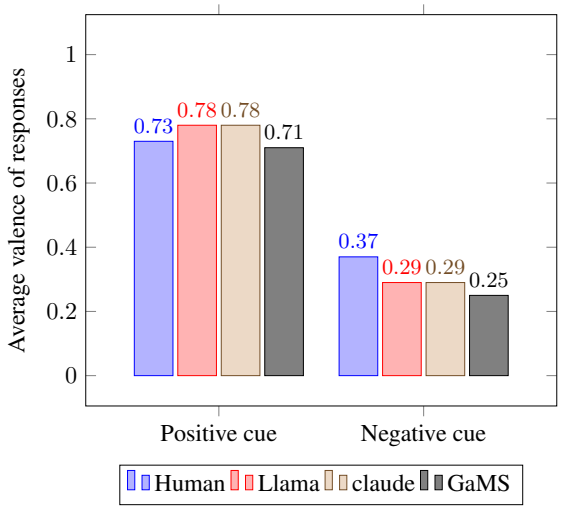


Figure 1: Valence for positive and negative cues

Discussion and conclusions

The experiments described show that LLMs are able to generate association-like responses to a zero-shot prompt, even in a small language like Slovene. A fair portion of these responses overlapped with human association norms, with the largest model (Claude 3.7 Sonnet) achieving the highest overlap. It should be noted here that seemingly small numbers do not necessarily mean poor results, as human-human overlap typically does not exceed 40 %.

In line with our own previous work (Brglez, Vintar, and Žagar 2024), LLM-generated associations generally follow the paradigmatic pattern more often than the syntagmatic one, frequently listing synonyms or near synonyms, and rarely departing from the part-of-speech of the cue. While the present study did not focus specifically on the comparison of response categories, a manual inspection of the results indicates slight variations between models in their typical behaviour. It would appear that LLMs come in different flavours, or exhibit distinct and consistent personal traits which affect - among other things - their linguistic behaviour. This should not surprise us as recent studies within the newly emerging field of machine psychology indeed explore personality traits of LLMs (Tommaso et al. 2024; He and Liu 2025).

When human and LLM responses are compared in terms of creativity or unexpectedness, humans produce a much richer and more varied mix of responses. We are aware of the fact that the creativity of LLMs could be activated using a different prompt, possibly including a variety of human responses or explicitly requesting creativity. We deliberately refrained from prompt engineering to retain the similarity between the association task for humans and LLMs.

The sentiment analysis tentatively shows that the models tend to follow the sentiment of the cue and only rarely depart from it in a different direction, rather the responses reinforce and often amplify the original sentiment. This finding is in line with the observation of Lavorati et al. (2024) about LLM bias in generated associations: if human responses to *man*

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

and *woman* contained traces of gender stereotypes (*woman: sex, beauty; man: strong*, Mistral's responses were overtly stereotypical (*woman: makeup, hair, fashion, beauty; man: job, career, computer, work*). One reason for the exaggeration of the original sentiment might be that the models almost never generate the antonym as response, while this is a strong association mechanism in humans. An antonym typically has the reverse sentiment than the cue, so this might explain why the average Valence scores for human responses lie more towards the middle of the range as compared to LLMs.

Our findings contradict the observations of Guo et al. (2023) who claim that ChatGPT's responses exhibit less bias and more objectivity than human answers to a collection of questions. This may be due to the fact that question answering, the task that instruction-tuned LLMs have been specifically trained for, triggers different behaviour than the association task where the model is prompted to explore its cognitive and experiential space (*"What does the word X remind you of?"*).

Even if the (lack of) diversity and variability in LLM-generated responses may say little about their (lack of) creativity, we believe that the tendency to produce outputs which follow the grammatical category, sentiment and semantic domain of the cue captures but a fraction of human associative behaviour, and that departures from the expected – which used to be quite common in 1st generation LLMs – now seem to be distinctive traits of human associations.

In future experiments it would make sense to perform a more detailed and fine-grained categorization of responses in order to see how models differ in their association patterns, a phenomenon we have observed but not yet evaluated. Also, the sample of 40 cues is too small to allow for any generalization of the above observations.

Author Contributions

Špela Vintar was in charge of planning the study, conducted parts of the analysis and wrote a significant part of the manuscript. Jan Jona Javoršek participated in the analysis, evaluation and interpretation of the results, and contributed to all versions of the manuscript.

Acknowledgments

This contribution was partly funded by ARIS grants P6-0215 Slovene Language - Basic, Contrastive, and Applied Studies, GC-0002 Large Language Models for Digital Humanities and I0-0005 The JSI infrastructure program.

References

[2024] Brglez, M.; Caporusso, J.; Hoogland, D.; Koloski, B.; Pollak, S.; and Purver, M. 2024. Slovenian Emotion Dimension and Emotion Association Lexicon SloEmoLex 1.0. http://hdl.handle.net/11356/1875.

[2024] Brglez, M.; Vintar, Š.; and Žagar, A. 2024. How Human-Like Are Word Associations in Generative Models? An Experiment in Slovene. In Zock, M.; Chersoni, E.; Hsu, Y.-Y.; and de Deyne, S., eds., *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, 42–48. Torino, Italia: ELRA and ICCL.

[1970] Clark, H. H. 1970. Word associations and linguistic theory. *New horizons in linguistics* 1:271–286.

[2018] De Deyne, S.; Navarro, D.; Perfors, A.; Brysbaert, M.; and Storms, G. 2018. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods* 51.

[2016] De Deyne, S.; Perfors, A.; and Navarro, D. J. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In Matsumoto, Y., and Prasad, R., eds., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1861–1870. Osaka, Japan: COLING.

[1959] Deese, J. 1959. Influence of Inter-Item Associative Strength upon Immediate Free Recall. *Psychological Reports* 5(3):305–312.

[2007] Fitzpatrick, T. 2007. Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics* 17(3):319–331.

[1913] Freud, S. 1913. On beginning the treatment (further recommendations on the technique of psycho-analysis I). *Standard edition* 12:121–144.

[1880] Galton, F. 1880. I.—Statistics of Mental Imagery. *Mind* os-V(19):301–318.

[2019] Günther, F.; Rinaldi, L.; and Marelli, M. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science* 14:1006–1033.

[2023] Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; and Wu, Y. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597.

[2025] He, J., and Liu, J. 2025. Investigating the Impact of LLM Personality on Cognitive Bias Manifestation in Automated Decision-Making Tasks. arXiv:2502.14219.

[2024] Lavorati, C.; Abramski, K.; Rossetti, G.; and Stella, M. 2024. LLM-Generated Word Association Norms. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press. 3–12.

[2020] Ljubešić, N.; Markov, I.; Fišer, D.; and Daelemans, W. 2020. The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene. In Nissim, M.; Patti, V.; Plank, B.; and Durmus, E., eds., *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, 153–157. Barcelona, Spain (Online): Association for Computational Linguistics.

[2017] Mandera, P.; Keuleers, E.; and Brysbaert, M. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language* 92:57–78.

[2017] Nematzadeh, A.; Meylan, S. C.; and Griffiths, T. L. 2017. Evaluating vector-space models of word representa-

tion, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39.

[2024] Richie, R.; Aka, A.; and Bhatia, S. 2024. Free Association in a Neural Network. *Online First Publication*. doi:10.1037/rev0000396.

[2024] Tikhonov, A., and Shtykovskiy, P. 2024. Humor Mechanics: Advancing Humor Generation with Multistep Reasoning. In *Proceedings of the International Conference on Computational Creativity 2024*.

[2024] Tommaso, T.; Hegazy, M.; Lemay, D.; Abukalam, M.; Rish, I.; and Dumas, G. 2024. LLMs and Personalities: Inconsistencies Across Scales. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

[2025] Xuan, W.; Yang, R.; Qi, H.; Zeng, Q.; Xiao, Y.; Xing, Y.; Wang, J.; Li, H.; Li, X.; Yu, K.; Liu, N.; Chen, Q.; Teodoro, D.; Marrese-Taylor, E.; Lu, S.; Iwasawa, Y.; Matsuo, Y.; and Li, I. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. arXiv:2503.10497.

The “What” Space
Prosodic Variability and Affective Virtual Environments

Jorge Forero
ITI/LARSyS - INESC TEC
Faculdade de Engenharia
Universidade do Porto
jfforero@ludique.cl

Mónica Mendes
ITI/LARSyS
Faculdade de Belas-Artes
Universidade de Lisboa
monicamendes@edu.ulisboa.pt

Gilberto Bernardes
INESC TEC
Faculdade de Engenharia
Universidade do Porto
gba@fe.up.pt

Abstract

What? What! What... How Many Ways to Say the Word What.

The “What” Space is a research project that investigates the emotional variability and expressive potential of spoken language. Focusing on the word “what,” the project examines how shifts in emotional prosody shape its interpretation. Drawing from the IEMOCAP database, we identified 54 instances of “what,” each annotated with categorical emotion labels and continuous values in the valence–arousal–dominance (VAD) space.

To explore these variations, we developed a series of multimodal visualizations: an interactive 3D visualization in Unity3D, network-based structures in Blender, a real-time speech-to-texture web application, and an immersive virtual reality installation hosted on Onland.io.

These experiences allow users to navigate a spatialized soundscape of “what” utterances and directly experience their emotional diversity.

This work contributes to affective computing, speech emotion recognition, and human–computer interaction by proposing novel frameworks for interpreting and visualizing emotional expression through speech. By revealing the rich affective range embedded in a single word, The “What” Space underscores the role of prosody in communication and demonstrates the potential of speech-based affective virtual environments.

Introduction

Affective virtual environments represent an emerging intersection between emotion recognition systems and dynamically adaptive virtual spaces (Forero et al., 2023). These systems are typically composed of two interdependent components: an emotion recognition module and a virtual environment generator (Pinilla et al., 2021a).

Emotion recognition techniques encompass multiple modalities, including facial expression analysis (Ekman & Friesen, 1978), body movement interpretation (Camurri et al., 2003), and biosignal analysis such as galvanic skin response, heart rate variability, respiratory patterns, and electroencephalographic signals (Picard, 1997; Cowie et al., 2001). Among these modalities, speech-based emotion recognition stands out as particularly suitable for naturalistic interactions, especially when Virtual Reality headsets are utilized and electrophysiological measurements are impractical or unavailable.

Speech inherently embodies dual modalities through its symbolic (semantic content) and acoustic (prosodic features) characteristics. Consequently, spoken language can be analyzed through textual sentiment analysis or acoustic feature extraction, each approach offering complementary insights into emotional expression (Scherer, 2003; El Ayadi et al., 2011).

A particularly intriguing phenomenon arises when semantically neutral utterances, such as the word "What", are evaluated solely through their prosodic expression. These ostensibly neutral words, when delivered with emotionally charged intonations, reveal critical points of affective variability, interpretive complexity, and aesthetic significance.

The “What” Space project addresses this phenomenon by investigating the wide emotional spectrum that a single ambiguous vocal utterance—"What?"—can convey. By focusing on the dynamic prosodic qualities inherent in such minimal linguistic expressions, the project offers an innovative framework for exploring emotional variability and contributes to a deeper understanding of affective dimensions in human-computer interactions.

Literature Review

Affective virtual environments can be traced back to the broader field of intelligent virtual environments, first reported by Aylett and Cavazza (2001). Since then, they

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

have evolved significantly through advances in automatic emotion recognition and generative computer graphics.
For instance, Misha Sra et al.'s Auris project demonstrates how music-based affective cues can shape immersive virtual worlds, translating the emotional nuances of songs into visual elements such as geometry, texture, and color (Sra et al., 2017).

Building on this idea, Forero et al. (2022) present Emotional Machines, an interactive installation that generates affective virtual environments in real time by analyzing users' acoustic and semantic expressions. Using multimodal speech-based emotion recognition, the system continuously adapts its visual landscape in response to the user's emotional state, forming a dynamic feedback loop between emotional input and audiovisual output.

Expanding on the synthesis of emotion and image, Patil et al. (2023) employ Generative Adversarial Networks (GANs) and the VQGAN+CLIP framework—which leverages visual and textual representations to guide image generation—to convert spoken emotional expressions into personalized visual artworks.

Similarly, Misra et al. (2024) introduce the FRIDA robotic system, which interprets emotional tones in audio to guide a robotic painter, translating affective vocal input into expressive graphical forms.

Methodology

We used the IEMOCAP (Interactive Emotional Dyadic Motion Capture) database, which contains multimodal recordings with emotional annotations (Busso et al., 2008). The database comprises approximately 12 hours of recorded interactions between professional actors, organized into five sessions with male–female dyads. The recordings include audio, video, textual transcripts, and emotion labels, making it a valuable resource for speech emotion recognition (SER).

Within the IEMOCAP dataset, we searched for utterances marked as emotionally ambiguous—specifically, instances where identical semantic content was delivered with different emotional expressions or interpretations. Through this analysis, we identified the most consistently ambiguous utterance: the single word "what," which appeared across contexts associated with all major emotional categories present in the corpus, including sadness, happiness, anger, fear, disgust, surprise, frustration, and neutrality.

We retained both the categorical emotion labels and the corresponding dimensional emotion scores (Valence–Arousal–Dominance, VAD) provided in the dataset. In total, we filtered 54 distinct occurrences of the word "what," each labeled with one of the corpus's emotion categories and mapped within the three-dimensional VAD space. The most frequent emotion category was anger (11 instances), followed by frustration (10), surprise (9), and neutrality (6). Other categories included excitement (7), sadness (4), happiness (1), and fear (1).

Results

We developed four creative strategies for data-driven audiovisualization to reveal the emotional ambiguity embedded in the utterance "what." While each strategy yields an independent result, the creative process unfolded sequentially, with each proposal building upon the outcomes of the previous one.

Say What? — Interactive Visualization in VAD Space[1]

We created an interactive 3D visualization in Unity3D to explore the emotional landscape of the word "what" within the three-dimensional Valence–Arousal–Dominance (VAD) space. Each utterance is represented as a sphere, positioned according to its VAD coordinates and color-coded based on its corresponding categorical emotion (see Image 1). Users can freely orbit the camera around a custom reference frame, enabling intuitive spatial exploration of emotional variation.

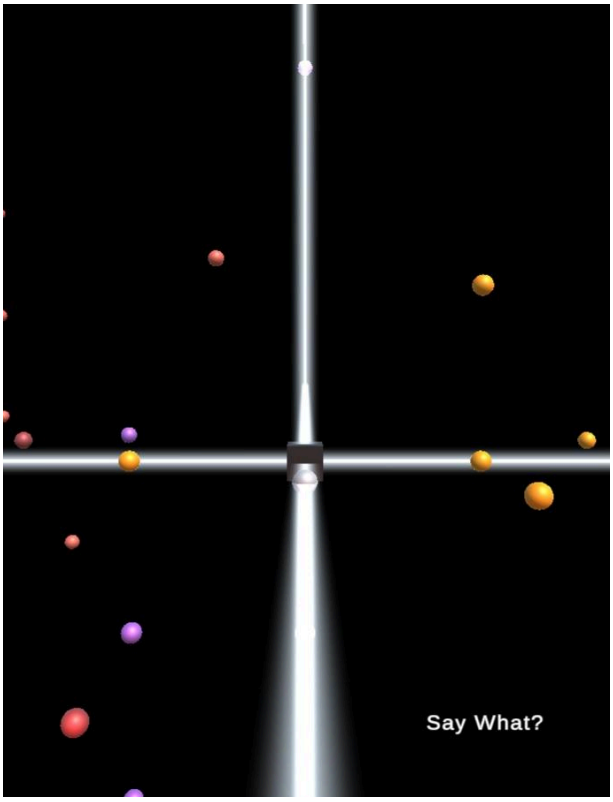


Image 1: Unity3D "Say What?" interactive environment.

[1] Say What? Web application and code available at: https://jfforero.github.io/TheWhatSpace/

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

The 3D "What" Network — Rhizomatic and Sequential Structures

To further investigate the relational structure of emotional variability, we visualized network topologies using Blender 3D. Two contrasting connectivity models were developed:

- **Rhizomatic network**: Every node (representing a unique utterance) is connected to every other node, forming a fully interconnected, non-hierarchical structure.
- **Sequential network**: Nodes are linked in a closed loop, sequentially traversing the dataset.

Additionally, we explored continuous sequential networks, using cubes placed on curved closed loops to suggest an emotional journey (see Image 2).

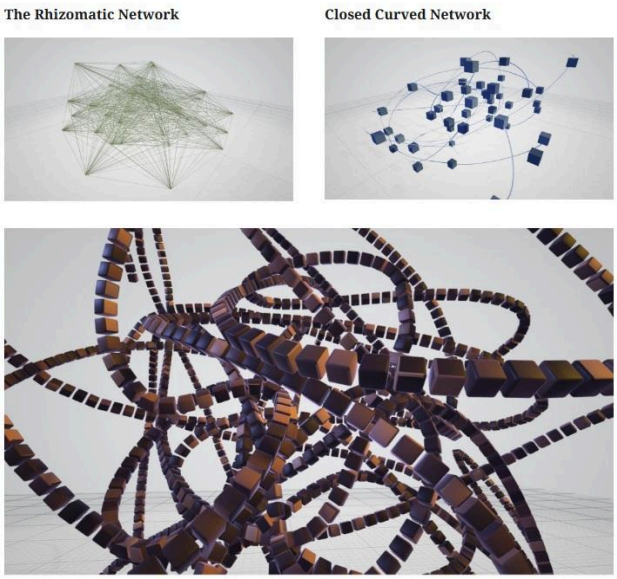


Image 2: Blender "What" Networks

Speech Emotion 2 Texture — Real-Time Emotion-to-Image Translation[2]

We created a Gradio-based web application hosted on Hugging Face Spaces that maps emotional predictions into abstract visual textures. The system pipeline includes:

•Audio input recorded via browser.
•Speech-to-text conversion using OpenAI's Whisper model (OpenAI, 2022).
•Emotion prediction via pre-trained LSTM-based speech emotion recognition model (Forero, 2023b).
•Emotion-to-image synthesis using DeepAI's Text-to-Image API[3], prompted with emotion-specific textual descriptions.

Image 3: Textures created for each emotional prediction

[2] Speech to Texture Hugging Face Space. Available at: https://huggingface.co/spaces/jfforero/LoopArtCritique

[3] DeepAI. (n.d.). Text to Image API. Retrieved from https://deepai.org/machine-learning-model/text2img

Proceedings of the 16th International Conference on Computational Creativity 2025
ISBN: 978-989-54160-7-3

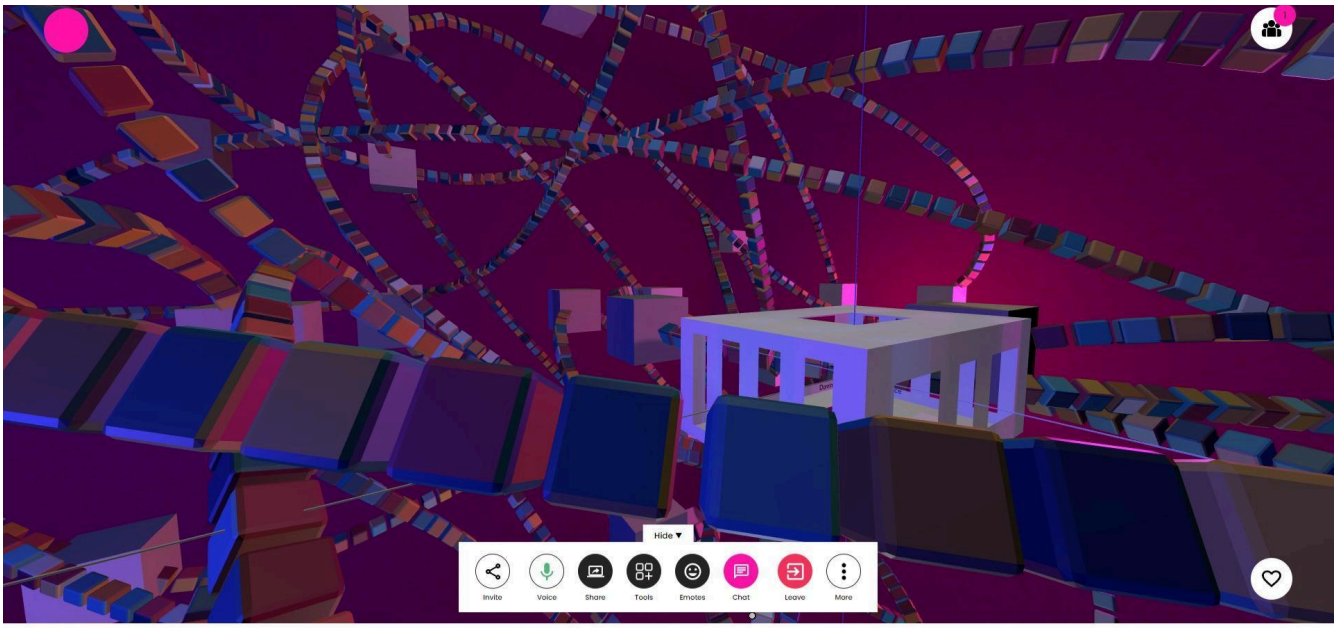


Image 4: The Onland.io "What" verse

The graphical interface allows users to submit audio recordings of the word "what", encouraging them to experiment with different prosodic variations. In response, the system provides real-time feedback by generating 512px × 512px visual textures that reflect the predicted emotional tone (see Image 3).

The "What" verse[4] — Onland Affectives Virtual Environments

Finally, the project was extended into virtual reality as part of the 13th Loop Art Critique Residency, hosted on the Onland.io metaverse and supported by the MUD Foundation. The installation was built using Mozilla Hubs, an open-source platform for creating shared virtual spaces in WebXR.

In this environment, audio samples of the word "what" were spatialized according to their VAD coordinates and visualized using the sequential network configuration (see Image 4). The experience begins in a silent cuboid space, where users encounter a brief introduction before freely exploring the emotional landscape using a first-person controller in fly mode.

Conclusions and Future Perspectives

This study highlights the expressive richness and emotional ambiguity that can be embedded in minimal linguistic units when analyzed through prosodic variation. By focusing on the single utterance "what," we demonstrate that emotional interpretation in speech extends beyond semantic content and is critically shaped by acoustic and prosodic features.

The analysis of the IEMOCAP database revealed that "what" occurs across a wide range of emotional categories, with a notable concentration in high-arousal negative emotions such as anger and frustration. However, its presence in lower-arousal and positive emotional contexts further supports its role as a semantically neutral yet affectively versatile utterance.

Through four multimodal visualization strategies—an interactive VAD-space exploration, network-based structures, real-time speech-to-texture mapping, and a virtual reality installation—we translated this emotional variability into spatial and visual forms. These environments offer new frameworks for investigating prosodic emotion and engaging users with affective dimensions of language in interactive settings.

Future work will focus on integrating real-time speech emotion recognition into virtual environments, enabling dynamic updates to textures, spatial mappings, and network structures based on live prosodic input. This direction aims to enhance the development of affect-driven virtual spaces, with implications for affective computing, human-computer interaction, and immersive media research.

Acknowledgements

This work was supported by FCT - Fundação para a Ciência e Tecnologia, I.P. by project reference and DOI identifier: https://doi.org/10.54499/2022.11918.BD

[4]Onland.io Metaverse. Available at: https://verse.loop.onland.io/j4Xy3wf/scaly-hopeful-get-together

References

Aylett, R., & Cavazza, M. (2001). Intelligent virtual environments: A state-of-the-art report. In *Eurographics 2001 Conference* (pp. 1–20).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335–359.

Camurri, A., Lagerlöf, I., & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1–2), 213–225.

Cowie, R., Douglas-Cowie, E., Savvidou, S., & McMahon, E. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. 19-24. Paper presented at Speech and Emotion: *Proceedings of the ISCA workshop*, Newcastle, United Kingdom.

Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press*.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572–587.

Forero, J., Bernardes, G., & Mendes, M. (2022). Emotional Machines: Toward Affective Virtual Environments. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 1–9).

Forero, J., Bernardes, G., Mendes, M. (2023). Desiring Machines and Affective Virtual Environments. In: *Brooks, A.L. (eds) ArtsIT, Interactivity and Game Creation. ArtsIT 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol 479. Springer, Cham.

Jorge Forero, Mónica Mendes, and Gilberto Bernardes. 2023b. En train d'oublier: toward affective virtual environments. In Proceedings of the 20th International Conference on Culture and Computer Science: Code and Materiality (KUI '23). Association for Computing Machinery, New York, NY, USA, Article 2, 1–6.

Misra, V., Schaldenbrand, P., & Oh, J. (2024). Robot Synesthesia: A Sound and Emotion Guided Robot Painter. *arXiv preprint arXiv:2302.04850*.

OpenAI. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision *[Computer software]*. https://github.com/openai/whisper

Patil, A., Deshmukh, R., & Kulkarni, S. (2023). Emotion-to-Image Translation Using VQGAN+CLIP Framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 123–130).

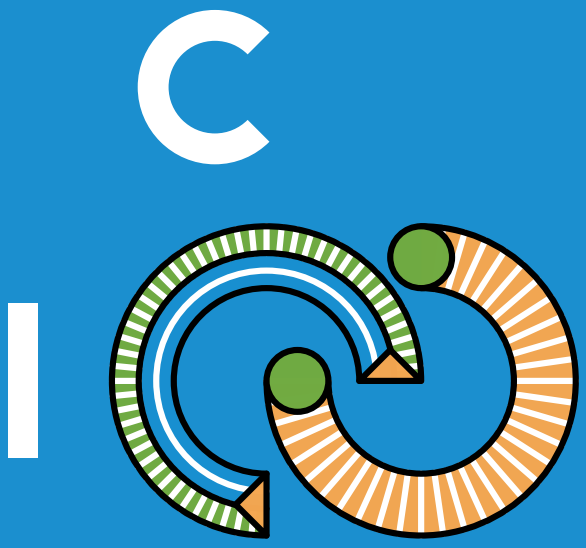
Picard, R. W. (1997). Affective Computing. MIT Press.

Pinilla, A., Garcia, J., Raffe, W., Voigt-Antons, J. N., Spang, R. P., & Möller, S. (2021). Affective visualization in virtual reality: An integrative review. *Frontiers in Virtual Reality,* 2, 630731.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2), 227–256.

Sra, M., Garrido-Jurado, S., & Maes, P. (2017). Auris: Integrating Audio and Visual Cues for Immersive Virtual Environments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1–12).

ISBN: 978-989-54160-7-3



ICCC'25

June 23—27
Campinas, Brazil

16th International Conference
on Computational Creativity

Association for Computational Creativity

International Conference on Computational Creativity