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.

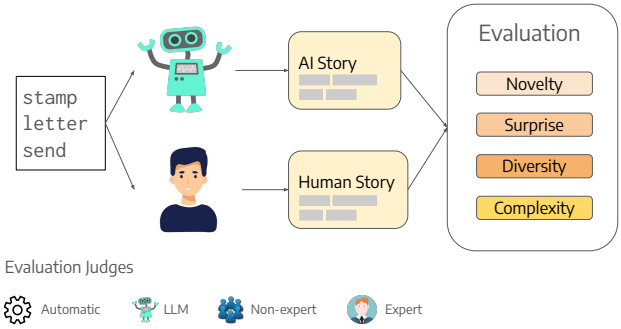


Figure 1: Our study setup illustrated with an example. Both humans and LLMs are asked to write a creative short story using three cue words and evaluated on several creativity metrics by human (experts vs non-experts) and LLM judges.

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-

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.

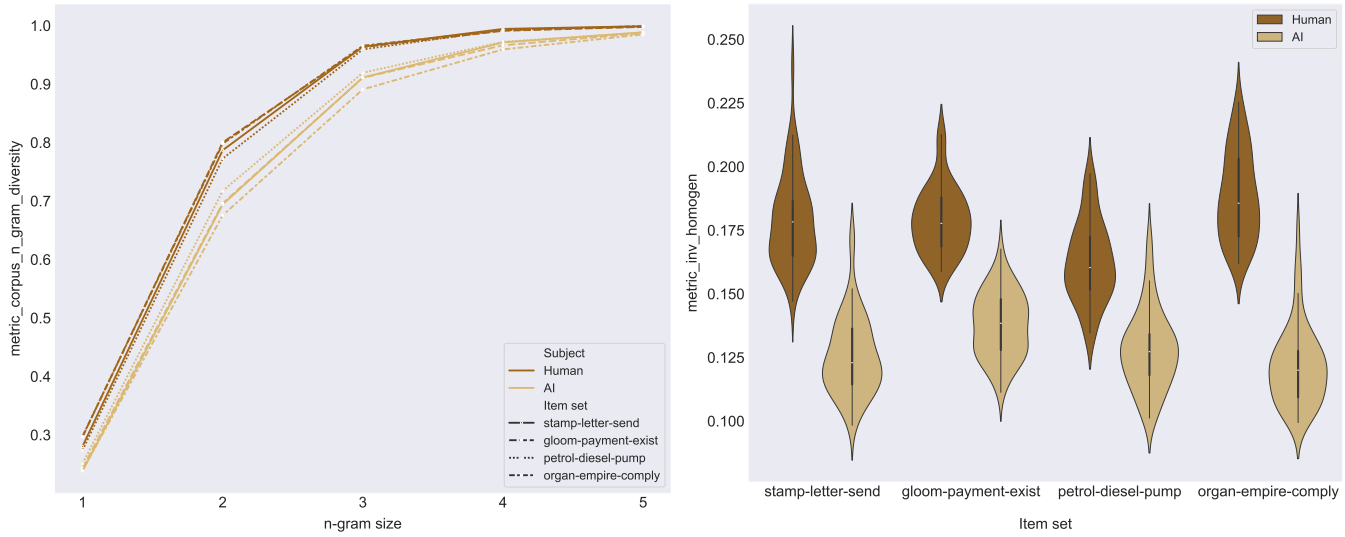
Related Work

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.



(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/

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:

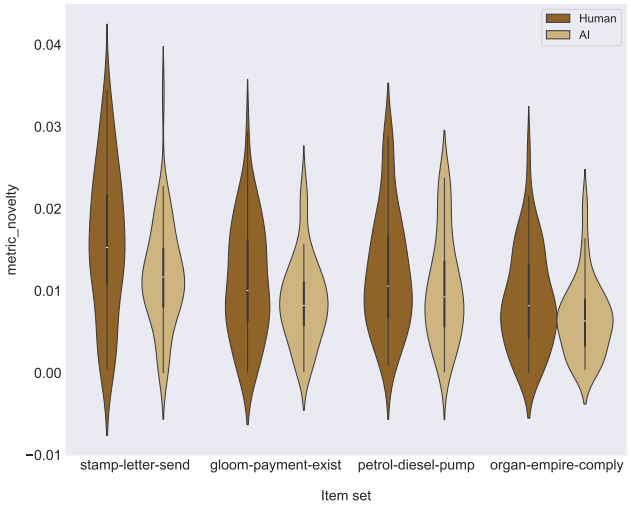


Figure 3: Novelty scores across all item sets.

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.

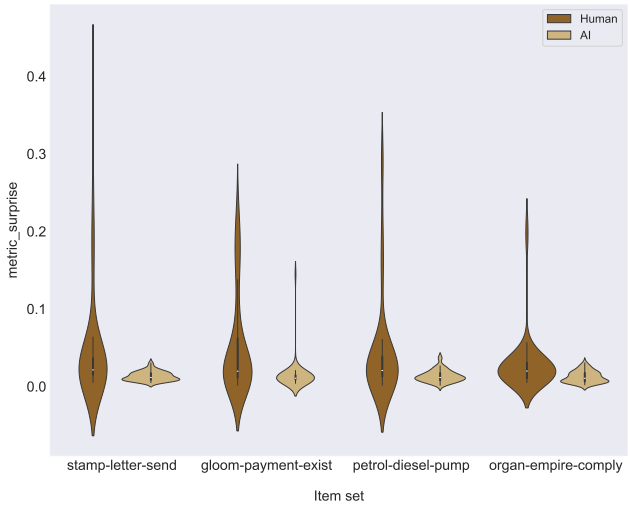


Figure 4: Surprise scores across all item sets.

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.

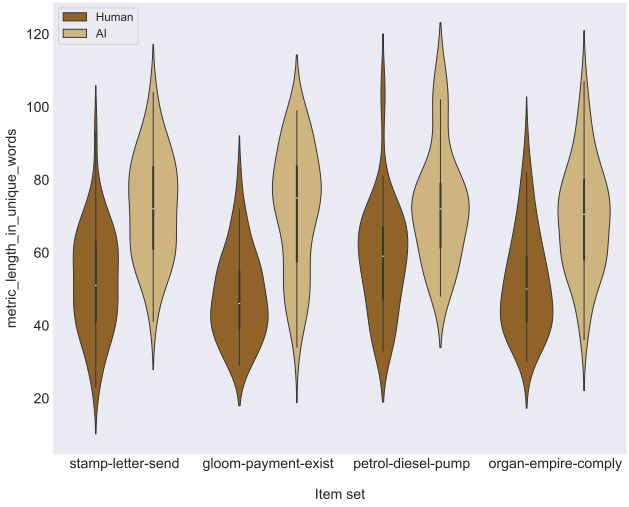


Figure 5: Lexical complexity scores across all item sets measured by story length in number of unique words.

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

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.

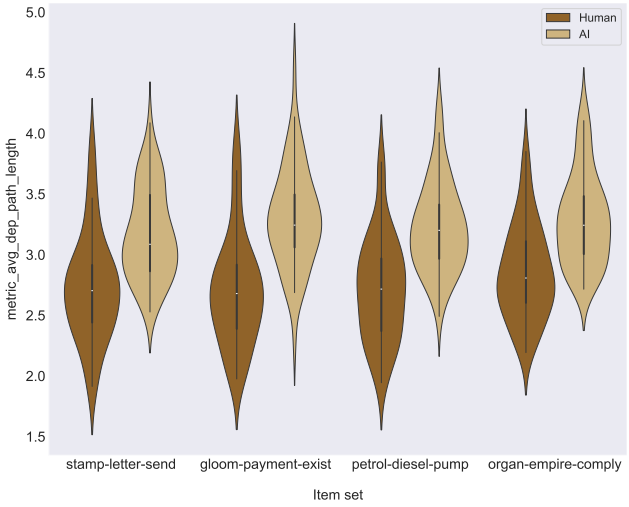


Figure 6: Syntactic complexity scores across all item sets measured by average dependency path length.

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 ($p < 0.01$).

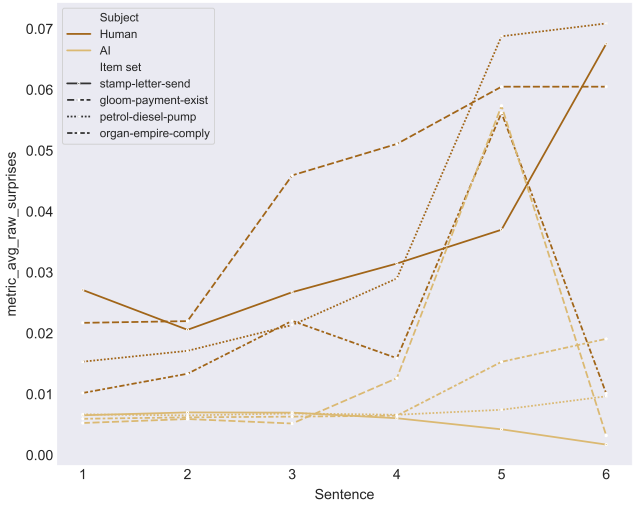


Figure 7: Surprise profile scores across sentence positions averaged over all stories.

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-

	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).

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

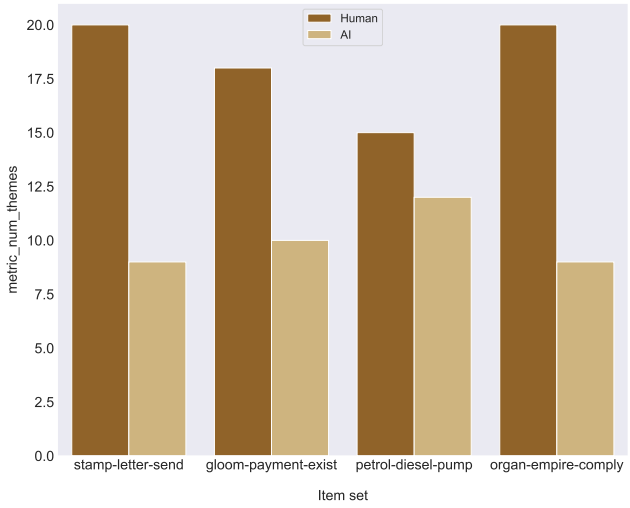


Figure 8: Number of themes for human and AI stories.

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).

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-

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.