



# Echoes in AI: Quantifying lack of plot diversity in LLM outputs

Weijia Xu<sup>a,1</sup> , Nebojsa Jotic<sup>a</sup> , Sudha Rao<sup>a</sup>, Chris Brockett<sup>a</sup>, and Bill Dolan<sup>a</sup>

Edited by Jeffrey Ullman, Stanford University, Stanford, CA; received March 17, 2025; accepted July 9, 2025

With rapid advances in large language models (LLMs), there has been an increasing application of LLMs in creative content ideation and generation. A critical question emerges: can current LLMs provide ideas that are diverse enough to truly bolster collective creativity? We examine two state-of-the-art LLMs, GPT-4 and LLaMA-3, on story generation and discover that LLM-generated stories often consist of plot elements that are echoed across a number of generations. To quantify this phenomenon, we introduce the *Sui Generis* score, an automatic metric that measures the uniqueness of a plot element among alternative storylines generated using the same prompt under an LLM. Evaluating on 100 short stories, we find that LLM-generated stories often contain combinations of idiosyncratic plot elements echoed frequently across generations and across different LLMs, while plots from the original human-written stories are rarely recreated or even echoed in pieces. Moreover, our human evaluation shows that the ranking of *Sui Generis* scores among story segments correlates moderately with human judgment of surprise level, even though score computation is completely automatic without relying on human judgment.

LLM | text generation | deep learning | creativity | AI

Rapid advances in large language models (LLMs) have spurred an ongoing debate on the usefulness of these models for tasks that require human-level creativity. On the one hand, there are works that highlight the strengths of LLMs in creative writing (1, 2), poetry generation (3), idea generation (4, 5) and even creative thinking (6). On the other hand, there has been research arguing that LLM creativity is much weaker than human creativity (7) and that LLM-generated stories are identifiably bad (8, 9).

One recent study (10) finds that while the use of an AI assistant in writing appears to enhance the creativity of individual writers, it also reduces the collective diversity of novel content produced by multiple writers. Similarly, a user study on argumentative essay writing (11) finds that writing with LLMs reduces the diversity of content produced by a group of users. These findings reflect a phenomenon that is by now familiar\* to teachers whose students use LLMs for help with writing assignments: While any individual LLM output might seem compelling and novel, reading through multiple texts produced by the same prompt can be a deflating experience. A narrative element that seems strikingly innovative and creative when encountered in the first output may begin to seem, by the time it is encountered in 10 more outputs, more like the product of a deterministic process than human-like creativity.

An example of this phenomenon is shown in Table 1, which shows a small subset of 100 short story continuations generated by GPT-4† given the first part of *Give It Up*, a short story by Franz Kafka‡ (see the full list of LLM-generated continuations in *SI Appendix*). None of GPT-4's continuations resemble Kafka's own ending for the story, in which a policeman disconcertingly tells the direction-seeking narrator to "Give it up!" and abruptly turns away. Instead, in 50 out of 100 generations, the policeman gives instructions to take the second left; in 18/100 to take the second right; and in others walks with the protagonist to show them the way. In 16/100 of the generations, a bakery is mentioned as a landmark. These echoes, as we refer to them here, repeat surprisingly frequently across generations at all semantic levels and not necessarily in the same order. The previously observed reduction in lexical diversity of LLM outputs (12) corresponds to what we view as semantically lower-level echoes, such as the frequent use of word "bakery." However, lexical diversity cannot be used to explain or quantify echoes at the higher semantic level. Narrative developments such as the policeman's decision

## Significance

Reading through a set of texts generated by large language models (LLMs) under the same prompt can be a disappointing experience. The first might seem convincingly like a novel, creative work by a human writer. Read more, though, and the lack of diversity in these LLM-generated outputs reveals itself. We show that short stories generated in this way often contain repetitive combinations of plot elements, while human-written stories maintain a higher level of uniqueness. This research quantifies this observation and introduces an automatic metric aimed at measuring the usefulness of LLMs for creative content generation at the narrative level. We believe that this metric will help drive progress toward LLMs that generate more diverse and creative content.

Author affiliations: <sup>a</sup>Microsoft Research, Redmond, WA 98052

Author contributions: W.X., N.J., S.R., C.B., and B.D. designed research; W.X. and S.R. performed research; W.X., N.J., S.R., C.B., and B.D. analyzed data; and W.X., N.J., S.R., C.B., and B.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2025 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

<sup>1</sup>To whom correspondence may be addressed. Email: weijiaxu@microsoft.com.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2504966122/-DCSupplemental>.

Published August 28, 2025.

\*<https://edtechrce.org/educators-develop-strategies-to-detect-ai-generated-student-work/>.

†We use temperature  $\tau = 1$  as LLMs are trained with the goal of fully capturing the data distribution with  $\tau = 1$ .

‡<https://www.flashfictiononline.com/article/give-it-up/>.

to show the way or point the protagonist toward a specific direction can be paraphrased in different ways, as the example continuations show. These more abstract echoes are not merely the result of repeated tokens.

In this paper, we aim to quantify such repetitions at the *narrative* level by introducing a *Sui Generis* score.<sup>§</sup> We use story generation as a testbed. To compute the *Sui Generis* score for a story, we first ask the LLM to generate many alternative continuations for the same story given varying lengths of the story prefix as context. We then count the number of times an original story segment is echoed (at the narrative level) in the alternative continuations. We adopt the intuition that a segment is more likely to appear in other story samples and is thus less unique if it is echoed in a larger number of alternative continuations given less context from the previous plot.

We test the *Sui Generis* score on 100 stories consisting of around 3,700 segments from two datasets: WritingPrompts, which contains short stories posted on Reddit; and a set of television episode plot summaries from Wikipedia. We experiment with two state-of-the-art LLMs: GPT-4 and LLaMA-3. Our results highlight the lack of diversity in LLM outputs: LLM-generated stories are often composed of the same combinations of idiosyncratic plot elements that are echoed frequently across its own generations and even across different LLMs, while only rarely exhibiting plot elements found in human narratives sparked by the same prompts. As the *Sui Generis* score is automatically calculated without human judgment, this result provides a quantitative explanation for the previous qualitative findings that the use of an AI assistant appears to reduce output collective diversity (10). It explains the contradictory findings in previous user studies on whether LLM generations have reached human-level creativity, as our study suggests that an LLM generation may seem novel to a human who has not seen many of its generations in that domain, but may seem banal to someone more familiar with its outputs. Furthermore, our human evaluation of the surprise level of story segments shows that the *Sui Generis* score correlates moderately with the average human judgment of surprisal. We observe that the story segments with high scores often correspond to key plot elements or interesting turning points, whereas those with low scores correspond to developments that are bland or highly predictable from the context.

## 1. Related Work

LLMs are increasingly used in creative writing (4, 13). However, there is a debate on whether LLMs boost creativity. Some studies have suggested that LLM-generated content is considered more creative or preferred by users. For example, Kefford (14) finds that ideas generated by ChatGPT were more likely to be purchased than those generated by Wharton MBA students. Lee and Chung (4) find that when participants were asked to generate creative ideas for everyday purposes, the use of ChatGPT increased their creativity. On the other hand, Begus (15) finds that AI-generated narratives tend to be less imaginative and only occasionally include plot twists. Chakrabarty et al. (7) invite expert writers to rate the stories generated by LLMs vs. professional writers based on Torrance Test of Creative Writing and discover that LLM-generated stories are less creative than those of professionals.

While these works all focus on evaluating the creativity or novelty of each individual LLM output, Padmakumar and He (11) and Doshi and Hauser (10) discover that, although writers

**Table 1. Continuations generated by GPT-4 when prompted with the first part of a short story by Kafka**

**Prompt:** It was very early in the morning, the streets clean and deserted, I was walking to the station. [...] % I did not yet know my way very well in this town; luckily, a policeman was nearby, I ran up to him and breathlessly asked him the way. He smiled and said: "From me you want to know the way?" "Yes," I said, "since I cannot find it myself."

**Original Continuation:** "Give it up! Give it up," he said, and turned away with a sudden jerk, like people who want to be alone with their laughter.

**Continuation 1:** "Follow this street for two blocks, then take a left at the bakery, and you'll see the station just ahead," he instructed with a gentle nod.

**Continuation 2:** "Keep going straight for two blocks, then take a left and you'll see the station right ahead," he explained, giving a friendly nod.

**Continuation 3:** He nodded in understanding, his smile broadening as he pointed down the street, "Take the second left and cross the square; you'll see the station right ahead."

**Continuation 4:** "Go straight ahead and take your second left, pass the old library, and the station will be right before your eyes," he responded, tipping his cap.

**Continuation 5:** "Alright," he said, his smile broadening, "I'll guide you. Follow me." Grateful, I trailed behind him, my earlier panic subsiding with each step.

report an AI assistant being helpful in their creative writing, it reduces the collective diversity of content produced by multiple writers. Similarly, Anderson et al. (16) shows that ideas generated with the assistance of LLMs are as diverse as the ideas generated without LLMs at the individual level but are significantly less diverse at the group level. These findings suggest that we should examine the *distribution* of LLM creations for a *given* prompt instead of each creation individually.

A recent work on detecting novelty in LLM outputs suggests that "the text must not have been copied from the training data" (17). This definition appears too narrow and surface-level: nothing resembling the original Kafka text is found in the story continuations by GPT-4 (Table 1). Yet by this standard, the generated continuations in story must be treated as novel, even though to human readers it is the original Kafka's ending that is unusual while the generated continuations are quite conventional and lack diversity. Other existing metrics that measure diversity in LLM outputs focus either on lexical features, such as n-gram overlap between output texts (18) or key-point summarizations of the texts (11), or topic-level features, such as the semantic distance between document embeddings (10). However, lexical features are insufficient for measuring diversity, as Ghosal et al. (19) observe, "identifying novel text is not straightforward because the text may have less lexical overlap yet convey the same information," while topic-level features are too high-level to assess the novelty of creative works in which novelty may reside in detailed plots. Thus, in our work, we look beyond lexical or topic-level diversity and introduce the *Sui Generis* score to measure the uniqueness of text spans at the narrative level.

## 2. Methods

We evaluate the uniqueness of a story segment based on the "alternative continuations" generated by LLMs themselves. Formally, in a story (string)  $S$  segmented into  $n$  segments  $S = (s_1, s_2, \dots, s_n)$  (as shown in Fig. 1C with

<sup>§</sup>Sui Generis is a Latin phrase meaning "of its own kind," used in English biology and law literature to indicate something unlikely to be repeated or recreated.

Alternative Continuation  $C_8^1$ 

[Prompt: In the year 1989, a police investigator arrives at the site of an impossible accident: a 2016 Toyota Corolla is wrapped around a lamp post. Title: Incident: Time-ly Collision In the quaint and ordinarily serene town of Clearwater, Nevada, headlines of an inexplicable incident stole the buzz in the] [year 1989. Snake-hipped police investigator, Dale McQuarrie, renowned for his observational skills and analytical mind, had been summoned to decipher an extraordinary accident. A car was embraced in an implausible collision—wrapped around a lamp post, as if both were caught in a dance of destruction. But it was not just]

[the peculiar accident that captured everyone's interest. It was the car itself. The metallic, midnight-blue vehicle was labeled as a 2016 Toyota Corolla, a model and make twenty-seven years ahead of its time.]

[...][...][...]

[...] For the following weeks, McQuarrie delved into the investigation, relentlessly probing every lead, examining each thread that the accident]

[offered. He contacted Toyota, where after numerous heated exchanges between baffled and skeptical executives, he was assured that the car indeed was a future model yet to be designed. Intriguingly, the man identified as Claude Matthews, a theoretical physicist from Caltech, remained in a comatose state. He was reported missing]

End of Story Prefix

from a conference on quantum mechanics and time-space anomalies just two days before the accident. McQuarrie, with a growing suspicion that this was no ordinary case, decided to visit Matthews' lab at Caltech. He was met with resistance, but his badge and the gravity of the situation eventually granted him access.

...

A

Alternative Continuation  $C_2^1$ 

[Prompt: In the year 1989, a police investigator arrives at the site of an impossible accident: a 2016 Toyota Corolla is wrapped around a lamp post. Title: Incident: Time-ly Collision In the quaint and ordinarily serene town of Clearwater, Nevada, headlines of an inexplicable incident stole the buzz in the] [year 1989. Snake-hipped police investigator, Dale McQuarrie, renowned for his observational skills and analytical mind, had been summoned to decipher an extraordinary accident. A car was embraced in an implausible collision—wrapped around a lamp post, as if both were caught in a dance of destruction. But it was not just]

End of Story Prefix

any ordinary car. It was a 2016 Toyota Corolla, a model that wouldn't exist for another 27 years. The vehicle's sleek design and unfamiliar technology baffled the onlookers who gathered around the scene, whispering theories and snapping polaroids. Dale approached the car with a mix of caution and curiosity. The license plate was smeared with dirt, but he could make out the registration—it was from the future, too. He circled the car, taking in every detail: the airbags deployed, the dashboard lit with strange icons, and the scent of new car mixed with the acrid tang of burnt rubber and metal.

The driver's door was ajar. Inside, slumped over the wheel, was a man in his mid-thirties, unconscious but alive. He

wore clothes of a style not yet seen in the late '80s, his

smartphone clutched tightly in one hand—a device that would

seem like magic to the townsfolk of Clearwater.

...

**The ambulance arrived, and the medical team carefully extracted the man from the car...**

He pocketed the device, feeling its weight and the warmth it emitted. He needed answers, and he knew just the person who might help.

...

B

## Original Story

[Prompt: In the year 1989, a police investigator arrives at the site of an impossible accident: a 2016 Toyota Corolla is wrapped around a lamp post. Title: Incident: Time-ly Collision In the quaint and ordinarily serene town of Clearwater, Nevada, headlines of an inexplicable incident stole the buzz in the] [year 1989. Snake-hipped police investigator, Dale McQuarrie, renowned for his observational skills and analytical mind, had been summoned to decipher an extraordinary accident. A car was embraced in an implausible collision—wrapped around a lamp post, as if both were caught in a dance of destruction. But it was not just]

[...][...][...]

[...] Inside the Toyota, sat an unconscious man, his hands still gripping the steering wheel, his sandy]

[hair matted with blood. McQuarrie signaled for the paramedics, who quickly cut the man free and rushed him to the hospital, leaving McQuarrie to unravel the mystery behind the untimely vehicle. For the following weeks, McQuarrie delved into the investigation, relentlessly probing every lead, examining each thread that the accident]

[offered. He contacted Toyota, where after numerous heated exchanges between baffled and skeptical executives, he was assured that the car indeed was a future model yet to be designed. Intriguingly, the man identified as Claude Matthews, a theoretical physicist from Caltech, remained in a comatose state. He was reported missing]

[a few days before the accident, last seen working late in his lab on an undisclosed project. More questions were piled onto McQuarrie's investigation, with answers seemingly as evasive as the car's origins. Driven by an odd mixture of skepticism and desperation, McQuarrie decided to confront the impossibility, visiting Caltech.]

...

C

**Fig. 1.** Two continuations of the same story with (A) longer prefix, with  $j = 8$  segments, and (B) shorter prefix, with  $j = 2$  segments from the same original story, partially shown in (C). Segments are delineated with  $[]$ . We highlight two segments in (C),  $i = 7$  in blue and  $i = 9$  in red, that are echoed in these alternative continuations. The red one (visiting a scientist's lab) is echoed only in continuations conditioned on a long story prefix, while the blue one (discovering a driver inside the car) is echoed frequently even given a short prefix. The *Sui Generis* score more severely penalizes the echoes discovered given shorter prefix (indicating that similar plot is more likely to be repeated by the LLM).

square brackets), at any point  $j \in [1, n - 1]$ , we truncate the story to its prefix  $S_{-j} = (s_1, s_2, \dots, s_j)$  and consider replacing suffix  $S_j = (s_{j+1}, s_{j+2}, \dots)$  by sampling possible alternative continuations  $C_j$  from the model  $K$  times:

$$C_j^k \sim p_{LLM}(\cdot | S_{-j}), \quad k \in \{1, \dots, K\}. \quad [1]$$

Each string  $(S_{-j}, C_j^k)$  is thus a full story with its first part taken from the given story  $S$  (either human-written or LLM-generated) and the second part being one possible way an LLM would finish it. Fig. 1 shows examples of such continuations with different prefixes: with a)  $j = 8$  and b)  $j = 2$  segments from the analyzed story in c).

Next, to evaluate how often segment  $s_i$  is repeated across generations, we compute its echo scores by comparing it to the alternative continuations  $\{C_1^k\}, \{C_2^k\}, \dots, \{C_{i-1}^k\}$  generated from varying lengths of story prefix  $S_{-1}, S_{-2}, \dots, S_{-(i-1)}$ . Specifically, we compute the echo score  $p_{ij}$ , which is an estimated likelihood that similar plot in segment  $s_i$  appears in an alternative continuation:

$$p_{ij} \approx \frac{1}{K} \sum_k a(s_i, C_j^k), \quad [2]$$

where  $a(s, C)$  is a binary function that indicates if segment  $s$  or its analog is present in continuation  $C$ . For instance, in Fig. 1, both  $a(s_7, C_8^1)$  and  $a(s_9, C_2^1)$  should be 1. We automate this function by prompting GPT-4 using the prompt template shown in Fig. 2, as our human evaluation shows that GPT-4's judgment correlates well with human judgment on this task (Section 3.3).<sup>4</sup>

<sup>4</sup>Note that the granularity of the function  $a(s, C)$  can be easily changed to capture lower-level repetitions (e.g., the policeman often suggests to take the second left in various continuations in Table 1) by adapting the prompt template.

The echo scores  $p_{ij}$  thus signal how likely and how early in the story (indicated by position  $j$ ) the plot in segment  $s_i$  is suggested to be generated. We note that the earlier  $j$  is, the less likely it is that such plot element will be repeated by the LLM in an alternative continuation, therefore echoes induced earlier should be weighted more heavily. Thus, we compute the *Sui Generis* score of segment  $s_i$  by taking the weighted average of the negative log of echoes  $\{-\log p_{i,1}, -\log p_{i,2}, \dots, -\log p_{i,i-1}\}$ , where we give higher weights to  $-\log p_{ij}$  with smaller  $j$ 's:

$$SG_i = \frac{-\sum_{j=1}^{i-1} \lambda^j \cdot \log p_{ij}}{\sum_{j=1}^{i-1} \lambda^j} \quad [3]$$

**Here is the beginning of a story:**

[Story prefix  $S_{-j}$ ]

**Here is a continuation of the story:**

[Alternative continuation  $C_j$ ]

**Does the above continuation contain any plot segment similar to the following (disregarding details like character names, specific objects in the scene, or exact dates)?**

[Story segment  $s_i$ ]

**Let's think step by step.**

[LLM's step-by-step reasoning]

**Finally, give a YES or NO answer. So the answer is**

[LLM's final answer (YES/NO)]

**Fig. 2.** Prompt template used for estimating function  $a(s_i, C_j)$ , i.e., if the plot in segment  $s_i$  is present in continuation  $C_j$ . The texts in black are part of the prompt while the texts in pink should be generated by the LLM.

**Table 2. An example of a human-written story with both high-scored and low-scored segments**

| Score | Story segment   |
|-------|---|
| -     | <b>Prompt:</b> In the middle of the night, a piece of paper is slipped under your door. It says, "DUCK!"  |
| 13.82 | Confused, you grab the paper and open the door . Looking down the apartment complex corridor you see more and more doors open with perplexed heads poking out one by one . No trace is left of whoever put these papers there, they just seem to have appeared all at once . "Duck?" Asks that one neighbour that you have rarely ever spoken to . "Yeah I got it too," you reply, "What do you think it means?" "Dunno" he |
| 3.00  | replies as you both turn towards the now gathering crowd . You head towards the crowd, which is incredibly loud with chatter now . "What's the big deal?" "Why should we even care?" Everyone's thinking the same thing yet everyone seems inexplicably interested . Amidst the crowd you notice that   |
| 13.82 | the last door is still closed . You point in the direction and immediately the crowd goes silent . Everyone begins to walk over to the closed door when the knob turns . It's dark inside . All you can see is a dark figure inside . As it creeps forward you  |
| 13.82 | notice an eerie grin . A woman creeps out of the dark room . She looks up, straight into your eyes . Your heart beats faster and faster as she holds up her paper . It reads: "Goose."  |

Scores were computed using GPT-4. The story contains a twist ending which is scored high. The high-scored segments correspond to the key plot elements and turning points, while the low-scored one is just an extension of the previous plot.

where  $\lambda < 1$  is a constant that controls the exponential weight decay. Intuitively, the echoes that are created given shorter prefixes (e.g., the echo highlighted in Fig. 1B) influence the final score more than the echoes produced given longer prefixes (e.g., the echo highlighted in Fig. 1A).

The described procedure for computing *Sui Generis* scores involves neither human judgment nor a large database for matching the generated text. All that is needed are the LLM's own generations and its own judgment of whether a story segment or its equivalent is contained in an alternative story. Table 2 shows a story with its segment-level *Sui Generis* scores. We show in the experiments that, compared to human-written stories, segments of LLM-generated stories have lower scores; in other words, these plot segments tend to be generated repetitively by the LLM, though possibly in a different order or in different parts of the story.

A qualitative user study involving expert writers (7) indicated that LLM-generated stories may differ from human-written ones in narrative pacing: in human-written stories, a plot twist is usually carefully foreshadowed and developed in order to maintain tension, which is in line with the tendency for uniform information density in language production (20). By contrast, we observe that LLM-generated stories often rapidly accelerate over time without fully resolving the plot.

Armed with the *Sui Generis* score, we can verify this observation quantitatively. Specifically, given the story segments  $s_{1\dots n}$  through time 1 to  $n$ , we take the *Sui Generis* scores  $SG_{1\dots n}$  and compute the "drop ratio" between the scores of consecutive segments clipped by a threshold  $\theta$ :

$$\text{drop}_i = \max\left(\frac{SG_i - SG_{i+1}}{SG_i} - \theta, 0\right). \quad [4]$$

Intuitively, a high drop ratio indicates that the *Sui Generis* score decreases immediately after the peak at position  $i$ , without progressively unravelling the surprising plot.

### 3. Experimental Setup

**3.1. Datasets.** We test our *Sui Generis* score on 100 stories from two story datasets: 1) the WritingPrompts dataset (21), which contains story prompts and the corresponding human-written stories posted on an online forum<sup>#</sup> by the year of 2018, and 2) plot summaries of TV episodes in the year of 2023 crawled from Wikipedia.

For WritingPrompts, we randomly sample 50 stories and directly use the story prompts by humans for LLM story generation.<sup>||</sup> To account for the fact that

<sup>#</sup><https://www.reddit.com/r/WritingPrompts/>.

<sup>||</sup>The prompt templates used for LLM generation are listed in *SI Appendix*. We also tried using GPT-4 to generate 20 different instruction prompts and evaluated the stories generated using these prompts. Results show very small variance (with a SD of 1.1) in *Sui Generis* scores among stories generated with different instruction prompts.

the paragraph segmentation of stories generated by different human authors and LLMs can vary significantly, thus affecting the segment-level scores, we break both human- and LLM-generated stories into segments of 50 words.<sup>\*\*</sup> Examples of such segments are shown in Fig. 1 within square brackets [...]. This results in a total of 1,887 segments in the original and LLM-generated stories for WritingPrompts.

On 50 randomly sampled Wiki plot summaries, we use the first two sentences from each plot summary as the prompt for LLM story generation. We segment both human and LLM stories by a fixed length of 30 words, which is the average length of a sentence in Wiki plot summaries.<sup>††</sup> We test on a total of 1,862 segments from the original and LLM-generated stories.

**3.2. *Sui Generis* Scoring Setup.** We sample  $K = 20$  alternative continuations at each position in the story. We set  $\lambda = 0.9$  when computing the weighted average of the negative log of echoes for the *Sui Generis* score in Eq. 3. To compute the drop ratio, we set the threshold  $\theta = 0.5$ .

**3.3. LLM Setup.** We use GPT-4(22) and LLaMA-3-70B-Instruct(23)(more details in *SI Appendix*) with  $\text{top\_p} = 1.0$ , sampling temperature  $\tau = 1.0$  for story generation.<sup>‡‡</sup> When computing the *Sui Generis* score, we use the same story generation model (with  $\text{top\_p} = 0.5$  and  $\tau = 1.0$ ) to generate alternative continuations  $C_j$ , and GPT-4 with  $\text{top\_p} = 0.5$  and  $\tau = 0$  to estimate the function  $a(s, C)$  (plot entailment), i.e., if similar plot in the original story segment appears in the alternative continuations.

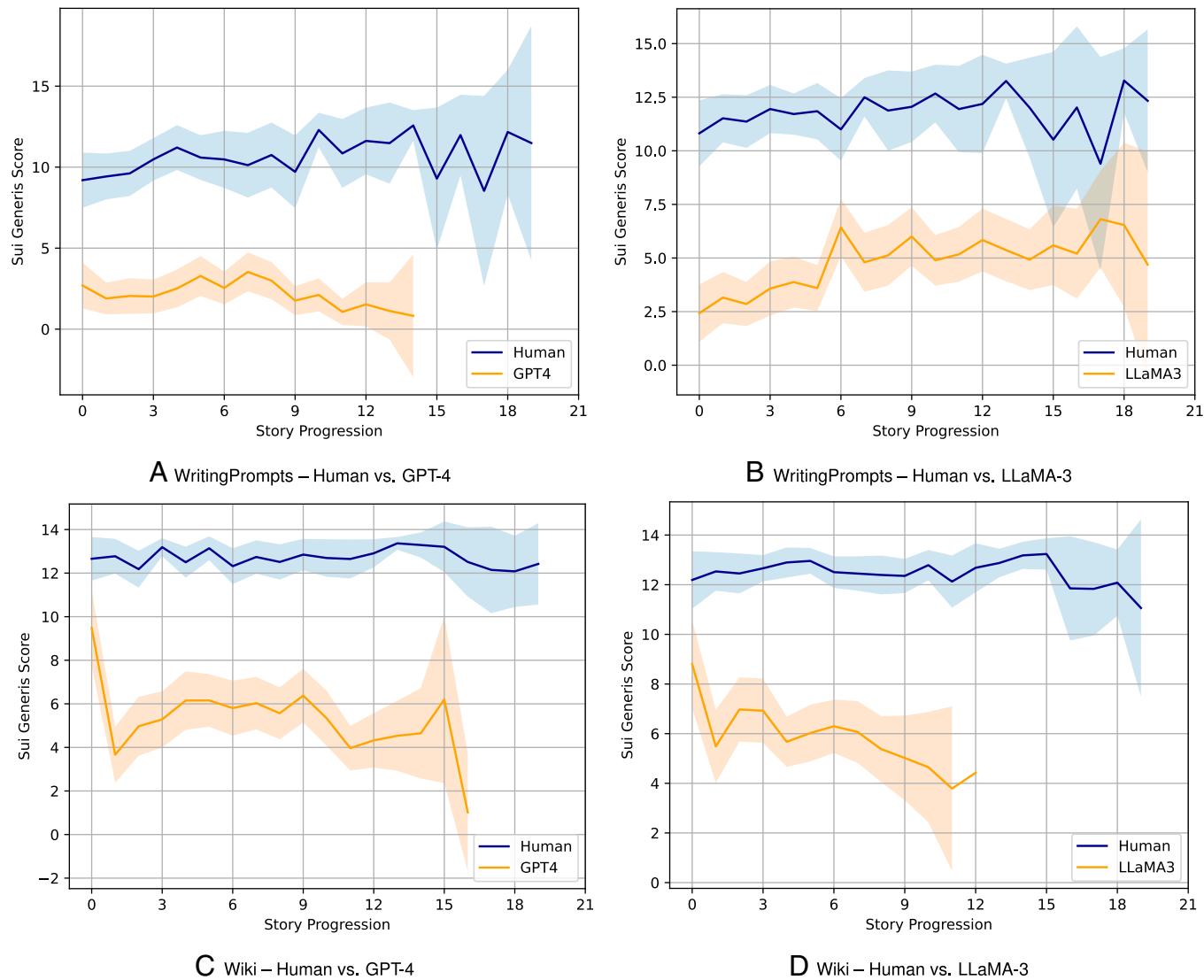
## 4. Results

**4.1. *Sui Generis* Score: Human vs. LLMs.** We first compare the average *Sui Generis* scores of story segments produced by humans vs. LLMs (including GPT-4 and LLaMA-3). As shown in Fig. 3, for both datasets, human-written stories yield significantly higher *Sui Generis* scores than LLM-generated ones. We also show the average Echo score matrix  $p_{i,j}$  on Wiki data in Fig. 4 (see the other heatmap figures in *SI Appendix*). For LLM-generated stories, the Echo rate increases as more of the previous plot is revealed (i.e., as the prefix length increases), while for human-written stories, the LLM can only occasionally predict the beginning section of the story. These results indicate that LLM-generated stories are

<sup>\*\*</sup>We chose word-level segmentation so that information is distributed evenly across segments. And the average number of words in each paragraph of WritingPrompts stories is around 50.

<sup>††</sup>Because the Wiki plot summaries are more condensed than WritingPrompts stories, and usually each sentence can cover very rich plot.

<sup>‡‡</sup>As Peeperkorn et al. (24) find that the influence of temperature on the creativity of LLMs is weak.



**Fig. 3.** Average *Sui Generis* scores of story segments generated by humans vs. GPT-4 [in (A)] or LLaMA-3 [in (B)] on WritingPrompts, and by humans vs. GPT-4 [in (C)] or LLaMA-3 [in (D)] on Wiki data at varying segment positions (x-axis). The shaded area represents the CI with confidence level  $\gamma = 0.95$ .

largely composed of idiosyncratic plot elements that are echoed across multiple LLM generations, while human-written stories often contain plot elements that land far in the tail—or not at all—of LLMs’ output distribution.<sup>§§</sup>

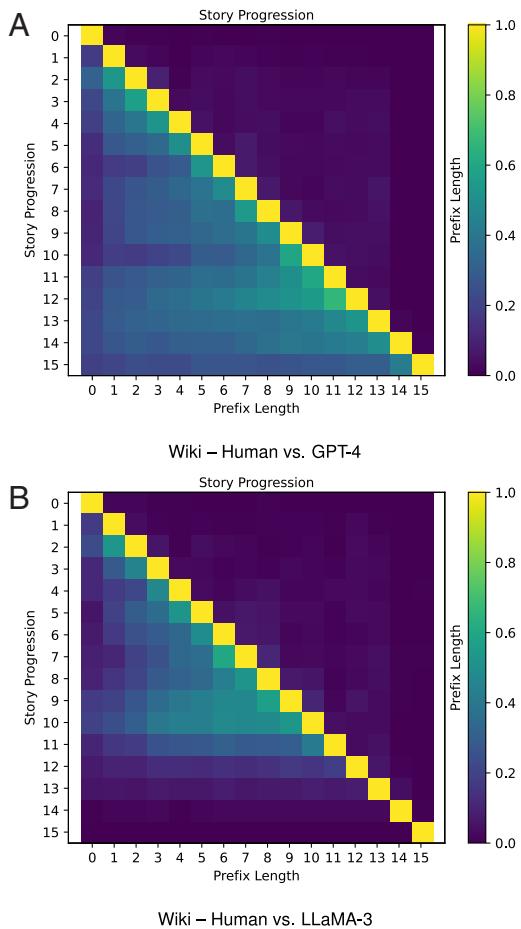
**4.2. *Sui Generis* Score: Cross-Model Scoring.** Next, we test if there are common echoes across different LLMs. To this end, we compute the cross-model *Sui Generis* scores—scores of LLaMA-3 generated stories given the alternative continuations generated by GPT-4. Results on 20 Wiki and 20 WritingPrompts stories show that the cross-model *Sui Generis* scores are slightly higher than the original scores, but the differences are small (0.6 on Wiki stories and 1.6 on WritingPrompts) compared to the gap between human and LLaMA-3 story scores ( $>6.0$ ). As shown in Fig. 5, the cross-model Echo scores distribute similarly to the original Echo scores. These results indicate that the plot elements generated by LLaMA-3 are echoed not only in its own generations but also in the alternative continuations generated by GPT-4. In other words, the plots generated by LLMs are echoed frequently across

different models, while human-generated plots are rarely echoed by any of these LLMs.

**4.3. Drop Ratio: Human vs. LLMs.** As shown in Table 3, LLM-generated stories lead to 7 to 9 higher drop ratio than human-written ones. This indicates that, in LLM-generated stories, there are more cases where the *Sui Generis* score drops immediately after a peak, while in human-written stories, the score often peaks and then stays high. We further show that this is related to the overly fast pacing in LLM-generated stories in Section 5.

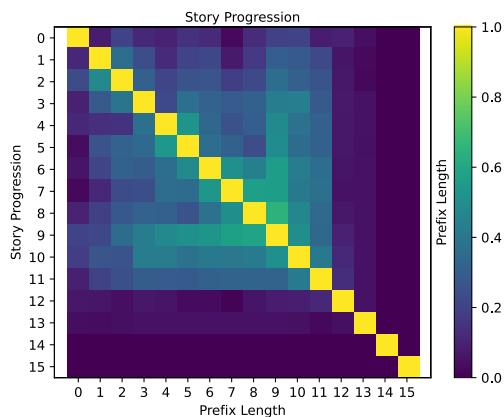
**4.4. *Sui Generis* Scores Correlate with Human Judgment of Surprise Level.** We invite four human judges, fluent in English, to evaluate the surprise level of story segments in nine randomly selected stories (159 segments) generated by LLaMA-3 based on the story prompts from WritingPrompts. The study was approved by Microsoft Research’s Institutional Review Board (case number 11035). Each human annotator in the study had signed a consent form electronically, informing them of the task and potential risks. Specifically, we ask each human annotator to read through each story segment by segment and annotate how

<sup>§§</sup>We further discuss the memorization effect in Section 6.



**Fig. 4.** Heatmap of the average Echo scores  $p_{ij}$  for the  $i$ -th story segment given story prefix  $S_j$  on 50 Wiki stories. The lower left triangle shows  $p_{ij}$  on LLM [GPT-4 in (A) and LLaMA-3 in (B)] generated stories, while the upper right triangle shows the transpose matrix of  $p_{ij}$  on human stories.

surprising each segment is from level 1 to 3 (1—most-anticipated, 2—neutral, or 3—most-surprising) based on the story so far. To enhance agreement among annotators, we give them five examples (excluded from the formal annotation task) and ask



**Fig. 5.** Heatmap of the average Echo scores  $p_{ij}$  for the  $i$ -th story segment given story prefix  $S_j$  on 20 Wiki stories by LLaMA-3. The lower left triangle shows the original Echo scores  $p_{ij}$ , while the upper right triangle shows the transpose matrix of cross-model Echo scores  $p_{ij}$  (using GPT-4 to generate alternative continuations).

**Table 3. Comparing the drop ratio (in percentage) of stories generated by humans vs. GPT-4 or LLaMA-3 on WritingPrompts (WP) and Wiki data**

|                         | Human | LLM  |
|-------------------------|-------|------|
| WP: Human vs. GPT-4     | 3.7   | 11.3 |
| WP: Human vs. LLaMA-3   | 1.7   | 9.0  |
| Wiki: Human vs. GPT-4   | 0.4   | 9.1  |
| Wiki: Human vs. LLaMA-3 | 0.5   | 7.8  |

LLM-generated stories obtain consistently higher drop ratio than human-written ones.

them to discuss their annotations to reach an agreement prior to the formal annotation task. For the formal annotation task, the interannotator agreement based on Krippendorff's alpha is moderate (0.68). Results show that segment-level *Sui Generis* scores correlate moderately with the average human judgment of surprise level: the Spearman's rank correlation between *Sui Generis* scores and the average human rating is significant<sup>11</sup> and the magnitude is 0.55. The correlation between *Sui Generis* scores and the average human rating is even stronger than that between each individual human rating and the average among the other three annotators (with Spearman's correlation between 0.38 and 0.49). The relationship between human ratings and *Sui Generis* scores is visualized in Fig. 6, where average human ratings (y axis) are shown for segments whose *Sui Generis* scores fall within a bin within a 1.4 interval around the x axis value. While humans rate segments with a wide range of mid-level *Sui Generis* scores as mildly surprising, they agree with *Sui Generis* on segments with highest and lowest scores.

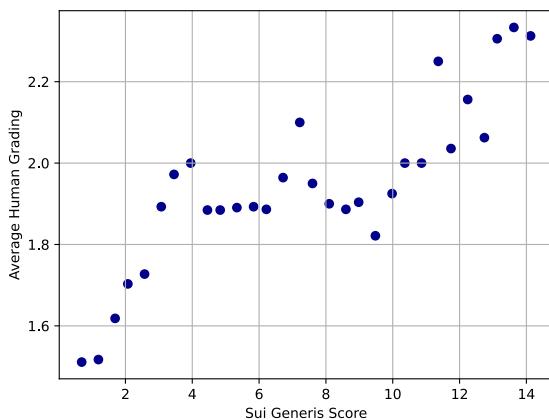
#### 4.5. The Story Prompt Itself Affects *Sui Generis* Scores for Both Human and LLM Stories.

We hypothesize that the story prompt will have similar impact on the *Sui Generis* scores of stories generated by humans and LLMs, as certain story ideas in the prompt (e.g., in Fig. 1C) simply lead themselves to more interesting and diverse continuations. To test the hypothesis, we compute the Spearman's correlation between the *Sui Generis* scores of pairs of human-written and LLM-generated stories under the same prompt. The correlation between the scores of human- and GPT-4-generated stories is 0.48 with  $P$ -value  $< 0.01$ , which indicates a moderate correlation. The correlation between human- and LLaMA-3-generated stories is 0.34 with  $P$ -value  $< 0.01$ , which indicates a weak but significant correlation. This suggests that, although human-written stories are scored significantly higher than LLM-generated ones, story prompts that lead to relatively low-scored stories by LLMs also lead to low-scored stories by humans. Thus, one can boost the *Sui Generis* score of human and LLM-generated stories by improving the story prompts; prompting is itself a creative endeavor.

#### 4.6. *Sui Generis* Captures Plot-Level Surprisal Beyond Token Likelihoods.

We further investigate how robust the *Sui Generis* score is to paraphrasing that would affect token-level surprisal measurements such as perplexity, even when there is no significant difference at the more abstract narrative level. To this end, we randomly sample 10 human-written stories and prompt LLaMA-3 to paraphrase each story without changing the semantics (which we manually verified). Next, we compute the *Sui Generis* and perplexity scores for each of the original and paraphrased stories. Results show that the *Sui Generis* score

<sup>11</sup>We use the  $P$ -value threshold  $P = 0.05$  for all the significance test in this paper unless noted otherwise.



**Fig. 6.** Average human grading of surprisal vs. the *Sui Generis* scores on story segments generated by LLaMA-3 based on the prompts from WritingPrompts. The data points are binned by an interval of around 1.4 based on the *Sui Generis* scores.

remains largely unchanged between the original and paraphrased stories (with a 4% drop on average after being paraphrased), while the perplexity score drops by 15% after being paraphrased. Additionally, the ranking of the perplexity scores among the 10 stories changes greatly after the stories are paraphrased: the Spearman’s rank correlation between the perplexity scores on the original and paraphrased stories is 0.55, and the correlation is nonsignificant. By contrast, the Spearman’s rank correlation between the *Sui Generis* scores on the original vs. paraphrased stories is both significant and much stronger (0.75). These results show that the *Sui Generis* score is far less sensitive to local wording than token-level perplexity, instead capturing plot-level surprise.

## 5. Discussion

**5.1. Comparison with Other Similarity Metrics.** We compare our prompting-based plot entailment assessment with other similarity/diversity metrics by measuring how these metrics correlate with human judgments. We conduct a human evaluation in which we invite four human judges (who are fluent in English) to judge 20 randomly sampled pairs of plot segments. The study was approved by Microsoft Research’s Institutional Review Board (case number 11035). Each human annotator in the study had signed a consent form electronically, informing them of the task and potential risks. We give each human judge the same prompt as GPT-4 (see *SI Appendix*, Figs. A1 and A2 and ask them to provide their answers among yes, no, or partially.<sup>##</sup> We then convert their answers to prediction scores between 0 and 1 (yes→1, no→0, partially→0.5) and take the average score among the four judges on each example. Next, we compute the Spearman’s correlation between the average human score and our prompting-based assessment, comparing it against five other text similarity/diversity metrics as recommended in ref. 18: compression ratio, self-BLEU (25), n-gram diversity (11), homogenization score (based on ROUGE-L) (11), and embedding-based similarity (11, 26) (using embeddings from the all-MiniLM-L6-v2 model). Results show that our prompting-based entailment assessment score correlates very strongly with the average human score with a Spearman’s correlation of 0.85. The compression ratio, self-BLEU, and n-gram diversity correlate

<sup>##</sup> The interannotator agreement based on Fleiss’s Kappa is fair (0.33).

weakly with human score with Spearman’s correlation of 0.07, 0.33, and 0.23, respectively. These metrics rely mostly on surface-form features like n-gram overlaps but fail to capture sentence-level semantics. The homogenization score and embedding-based similarity correlate moderately with the average human score (Spearman’s correlation: 0.46 and 0.50) but not as strongly as our prompting method. These results demonstrate that our prompting method is better at measuring plot entailment than existing similarity metrics. In addition, the prompting method gives us more control over the level of similarity that we want to capture, while it is difficult to control or understand the level of similarity captured by the embedding-based score.

### 5.2. What Are the Main Characteristics of Low-Scored Stories?

We examine the lowest-scored stories generated by humans and LLMs. These stories typically fall into two categories. The first includes stories that are too bland given the story prompts (*SI Appendix*, Table D1), which is more common in LLM-generated stories than human-written ones. The second class of low-scoring stories do exhibit rich plots, but the plot elements themselves are banal and are used repetitively across many LLM generations (see the examples in *SI Appendix*, Table D2).

### 5.3. What Are the Main Characteristics of High-Scored Stories?

We further examine the highest-scored stories from both humans and LLMs. The common features across human- and LLM-generated stories are that they usually contain nontraditional plots with rich details (see the example in *SI Appendix*, Table D4). Additionally, human-written stories offer more interesting twists and nonlinear storytelling (e.g., clues, distractions, interlude, and flashbacks) than LLM-generated ones, as shown in *SI Appendix*, Table D3.

### 5.4. Segment-Level *Sui Generis* Scores Can be Used to Identify Key Plot Elements.

We further examine segment-level *Sui Generis* scores and find that the high scores usually correspond to the key narrative developments and turning points. For example, in Table 2, the high-scored segments correspond to the key plot turns (e.g., when characters find that everyone has received the same note under their doors, and the twist ending where the last person who opened her door revealed the note she received), while the low-scored segment is a predictable consequence of the previous developments.

**5.5. High Drop Ratio in LLM-Generated Stories.** A user study involving expert writers (7) suggests that LLM-generated stories differ from human-written ones in narrative pacing. We therefore examine whether the high drop ratio in LLM-generated stories compared to human-written ones is related to this phenomenon. Indeed, we find that the high drop ratio in LLM stories is typically associated with overly hasty narrative pacing in LLM-generated stories, often leading to abrupt resolutions that leave key suspenseful plot elements unresolved. The result is a drop in the *Sui Generis* score (see the example in *SI Appendix*, Table D5). By contrast, human-written stories tend to have slower pacing, introduce suspenseful elements more gradually, and provide a final resolution that ties up hanging plot threads (see the human-written story under the same prompt in *SI Appendix*, Table D3).

**5.6. Computational Cost.** The number of LLM calls needed to compute the *Sui Generis* score for a story  $S$  is  $[n + K \cdot n(n-1)/2]$ , where  $K$  is the number of alternative continuations  $C_j$  sampled given each prefix  $S_{-j}$ , and  $n$  is the total number of segments

in the story  $S$ . In our experiment, running *Sui Generis* for an average-length story (10 segments, 500 words) takes 910 LLM calls, which amounts to around  $700k$  input tokens and  $700k$  output tokens through an LLM. This will cost around \$7 USD using GPT-4.1 (one of the best LLMs) based on current pricing (<https://openai.com/api/pricing/>). We anticipate that this cost will decrease rapidly in the near term given current trends (<https://epoch.ai/data-insights/llm-inference-price-trends>).

**5.7. Possible Uses of *Sui Generis* in Generation.** Beyond evaluation, the *Sui Generis* score can also be used to improve diversity through training or inference-time optimization. For instance, we can generate a story fragment by fragment, where at each step, we sample  $M$  alternative segments and keep the one with the highest *Sui Generis* score. We test this approach by sampling a story ( $S-100$ ) using GPT-4 with  $M = 100$  given a prompt from WritingPrompts.  $S-100$  obtains +5.3 higher *Sui Generis* score and is more interesting than the story generated given the same prompt in one pass (see the stories in [SI Appendix](#)), at the cost of 2,000 times more LLM calls.  $S-100$  even achieves human-level *Sui Generis* score when scoring with  $SG-20$  (where we set the number of alternative continuations  $K$  to 20). However, by scaling up  $K$  from 20 to 100 (which increases the resolution at which we can capture unique segments), we can still detect the gap between human and LLM-generated stories— $S-100$  obtains an  $SG-100$  score of 3.8, which is 6.4 points lower than the  $SG-100$  score of the human-written story given the same prompt. Note that computing the  $SG-100$  score is 20 times less expensive than generating  $S-100$ . In general, the number of LLM calls required for computing  $SG-K$  is linear in  $K$ , while the number of LLM calls for generating a story  $S-M$  that maximizes  $SG-K$  is proportional to  $K \cdot M$ .

## 6. Limitations

The WritingPrompts corpus and part of the Wiki data used in the experiment may have been in the training corpora of the LLMs evaluated. Although our results show no effect of a memorization effect for the stories we tested, it is possible that this kind of training effect might surface for other datasets that were included in the LLMs' training corpora. As the ability of LLMs to memorize long texts improves, the *Sui Generis* score will be applicable only to corpora that were not included in the training data. Additionally, the scoring method relies on the capability of an LLM to identify semantic similarities at the desired level. In our experiment, we always use GPT-4 for the entailment judgment, as our human evaluation shows that it is capable of mimicking human judgment of entailment accurately.

If one were to use our metric with a different entailment judgment model, they would first need to ensure that this model is at least as capable as GPT-4. Substituting a less capable model and using it for both generation and entailment judgment could lead to a bias, which could produce unreliable scores.

## 7. Conclusion

We introduced the *Sui Generis* score to quantify the uniqueness of LLM-generated stories at a narrative level when compared to their alternative generations. Our experiments on 100 LLM-vs. human-written stories demonstrated the lack of plot-level diversity in LLM-generated stories. Regardless of the LLM used to generate them, these stories are often composed of echoic plot elements, in contrast to the more varied and unique elements found in human-written stories. Moreover, our human evaluation showed a moderate correlation between *Sui Generis* scores and human judgments of surprise, suggesting that the *Sui Generis* score can serve as a useful tool for assessing the uniqueness and surprise level of LLM-generated content. Furthermore, the *Sui Generis* score has the potential to be extended to assess the uniqueness of sequential content across different languages and for other modalities beyond text. For example, it could be adapted to assess the uniqueness of compositions in music, visual elements in images, and temporal segments in video content by substituting the entailment function with similarity measurement in the new modality.

This work also sheds light on the societal impact of AI-generated content, as it suggests that disseminating AI-generated content at scale while being unaware of LLMs' tendency to homogenize might negatively affect cultural diversity, reducing the collective creativity of online content and the richness of educational experiences. This is reminiscent of how the lack of diversity in recommender systems has led to social polarization and has amplified social bias (27). On the bright side, the *Sui Generis* score holds promise for future work aimed at enhancing diversity in AI- or human-AI-generated content. For instance, it can be integrated with inference optimization algorithms to produce more unique outputs. Additionally, the score can help identify low-scoring segments in AI-generated content, where human input can be introduced to enrich creativity and originality; as Anderson et al. (16) suggest that by informing users that a particular piece of model output is homogeneous, it may help users resist model-induced homogenization at the group level.

**Data, Materials, and Software Availability.** The code for computing the *Sui Generis* score is deposited in Github (28).

1. A. Bellemare-Pépin *et al.*, Divergent creativity in humans and large language models. arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2405.13012> (Accessed 13 May 2024).
2. W. Orwig, E. R. Edenbaum, J. D. Greene, D. L. Schacter, The language of creativity: Evidence from humans and large language models. *J. Creative Behav.* **58**, 128–136 (2024).
3. B. Porter, E. Machery, Ai-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Sci. Rep.* **14**, 26133 (2024).
4. B. C. Lee, J. Chung, An empirical investigation of the impact of ChatGPT on creativity. *Nat. Hum. Behav.* **8**, 1906–1914 (2024).
5. C. Si, D. Yang, T. Hashimoto, Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. arXiv [Preprint] (2024). <https://arxiv.org/abs/2409.04109> (Accessed 6 September 2024).
6. E. E. Guzik, C. Byrne, C. Gilde, The originality of machines: AI takes the Torrance Test. *J. Creativity* **33**, 100065 (2023).
7. T. Chakrabarty, P. Laban, D. Agarwal, S. Muresan, C. S. Wu, "Art or artifice? Large language models and the false promise of creativity" in *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24* (Association for Computing Machinery, New York, NY, 2024).
8. M. Sato, AI-generated fiction is flooding literary magazines—but not fooling anyone. 25 February 2023, *The Verge*, 2023.
9. M. Levenson, Science fiction magazines battle a flood of ChatBot-generated stories. 23 February 2023, *The New York Times*, 2023.
10. A. R. Doshi, O. P. Hauser, Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).
11. V. Padmakumar, H. He, Does writing with language models reduce content diversity? arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2309.05196> (Accessed 1 July 2024).
12. B. Mohammadi, Creativity has left the chat: The price of debiasing language models. arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2406.05587> (Accessed 8 June 2024).
13. D. Kobak, R. González-Márquez, E. Agnés Horvát, J. Lause, Delving into ChatGPT usage in academic writing through excess vocabulary. arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2406.07016> (Accessed 11 June 2024).
14. M. Kefford, Wharton study pits ChatGPT against MBA students in creativity test. 13 September 2023, *Business Because*, 2023.
15. N. Begus, Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling. arXiv [Preprint] (2023). <https://doi.org/10.48550/arXiv.2310.12902> (Accessed 19 October 2023).
16. B. R. Anderson, J. H. Shah, M. Kreminski, "Homogenization effects of large language models on human creative ideation" in *Proceedings of the 16th Conference on Creativity and Cognition*, D. Long, J. Chan, Eds. (Association for Computing Machinery, New York, NY, 2024), pp. 413–425.

17. R. T. McCoy, P. Smolensky, T. Linzen, J. Gao, A. Celikyilmaz, How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *Trans. Assoc. Comput. Linguist.* **11**, 652–670 (2023).
18. C. Shaib *et al.*, Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2403.00553> (Accessed 1 March 2024).
19. T. Ghosal, T. Saikh, T. Biswas, A. Ekbal, P. Bhattacharyya, Novelty detection: A perspective from natural language processing. *Comput. Linguist.* **48**, 77–117 (2022).
20. T. F. Jaeger, Redundancy and reduction: Speakers manage syntactic information density. *Cogn. Psychol.* **61**, 23–62 (2010).
21. A. Fan, M. Lewis, Y. Dauphin, "Hierarchical neural story generation" in *ACL:2018:1*, I. Gurevych, Y. Miyao, Eds. (Association for Computational Linguistics, Melbourne, VIC, Australia, 2018), pp. 889–898.
22. OpenAI *et al.*, GPT-4 Technical Report. arXiv [Preprint] (2024). <https://arxiv.org/abs/2303.08774> (Accessed 4 March 2024).
23. A. Dubey *et al.*, The Llama 3 herd of models. arXiv [Preprint] (2024). <https://arxiv.org/abs/2407.21783> (Accessed 23 November 2024).
24. M. Peeperkorn, T. Kouwenhoven, D. Brown, A. Jordanous, Is temperature the creativity parameter of large language models? arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2405.00492> (Accessed 1 May 2024).
25. Y. Zhu *et al.*, "Texygen: A benchmarking platform for text generation models" in *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (2018).
26. P. J. Wang, M. Kreminski, Guiding and diversifying LLM-based story generation via answer set programming. arXiv [Preprint] (2024). <https://doi.org/10.48550/arXiv.2406.00554> (Accessed 19 July 2024).
27. L. Bojic, AI alignment: Assessing the global impact of recommender systems. *Futures* **160**, 103383 (2024).
28. W. Xu, N. Jovic, S. Rao, C. Brockett, B. Dolan, SuiGeneris. Github. <https://echoes.msr-emergence.com/>. Deposited 14 August 2025.