HHS Public Access

Author manuscript

J Creat Behav. Author manuscript; available in PMC 2024 May 02.

Published in final edited form as:

J Creat Behav. 2024 March ; 58(1): 128–136. doi:10.1002/jocb.636.

The Language of Creativity: Evidence from Humans and Large Language Models

WILLIAM ORWIG,
EMMA R. EDENBAUM,
JOSHUA D. GREENE,
DANIEL L. SCHACTER
Harvard University

Abstract

Recent developments in computerized scoring via semantic distance have provided automated assessments of verbal creativity. Here, we extend past work, applying computational linguistic approaches to characterize salient features of creative text. We hypothesize that, in addition to semantic diversity, the degree to which a story includes perceptual details, thus transporting the reader to another time and place, would be predictive of creativity. Additionally, we explore the use of generative language models to supplement human data collection and examine the extent to which machine-generated stories can mimic human creativity. We collect 600 short stories from human participants and GPT-3, subsequently randomized and assessed on their creative quality. Results indicate that the presence of perceptual details, in conjunction with semantic diversity, is highly predictive of creativity. These results were replicated in an independent sample of stories (n = 120) generated by GPT-4. We do not observe a significant difference between human and AI-generated stories in terms of creativity ratings, and we also observe positive correlations between human and AI assessments of creativity. Implications and future directions are discussed.

Keywords

artificial intelligence; creativity; large language models; semantic distance

The question of how to reliably assess the creative quality of ideas is a longstanding topic in creativity research (Amabile, 1982; Reiter-Palmon, Forthmann, & Barbot, 2019). Divergent thinking (DT) refers to the ability to generate creative ideas by combining diverse types of information. The alternative uses task (AUT) is the most common measure of DT (Guilford, 1967), in which participants are presented with an object and asked to generate unusual and creative uses for it. Studies using the AUT have traditionally relied on metrics such as *fluency*, counting the number of responses produced, or *flexibility*, the number of different categories of responses, to assess AUT performance; however, these

*Correspondence concerning this article should be addressed to William Orwig, Department of Psychology, Harvard University, 880 William James Hall, 33 Kirkland Street, Cambridge, MA 02138. williamorwig@g.harvard.edu.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

measures are highly correlated with each other and fall short in capturing the full scope of creative ideation (Acar, Ogurlu, & Zorychta, 2022; Acar & Runco, 2015). While subjective scoring methods for *originality*, the creative quality of ideas produced on the AUT, have shown evidence of convergent validity (Forthmann, Oyebade, Ojo, Günther, & Holling, 2019; Jauk, Benedek, & Neubauer, 2014), inter-rater agreement is not always high, raising issues of reliability (Barbot, 2018). Researchers in creativity assessment have sought to standardize DT performance with computational measures of semantic distance to address the subjectivity of traditional scoring methods (Beaty & Johnson, 2021; Dumas, Doherty, & Organisciak, 2020). The application of semantic distance to DT assessment is based on the associative theory of creativity (Kenett & Faust, 2019; Mednick, 1962), which characterizes creative thought as a novel and useful recombination of semantic knowledge; thus, the combination of more remotely associated concepts is considered more creative. In light of recent evidence showing that semantic distance scores of AUT responses are correlated with human ratings of creativity (Orwig, Diez, Vannini, Beaty, & Sepulcre, 2021), the present study implements a novel semantic diversity measure (Johnson et al., 2023) to overcome the limits of traditional DT assessments and characterize the degree to which creative stories connect semantically divergent concepts. Extending past work, we define semantic diversity in narrative texts and examine its relation to human judgments of creativity.

Complementary research has highlighted the contributions of episodic memory to DT. After receiving a brief Episodic Specificity Induction (ESI) – a procedure that enhances the contribution of episodic retrieval processes to a subsequent task (for review, see Schacter & Madore, 2016) –participants reliably generate more novel uses for objects on the AUT compared with a control induction (Madore, Addis, & Schacter, 2015; Madore, Jing, & Schacter, 2016). Furthermore, neuroimaging results show increased activity in memory-related brain regions, such as the hippocampus, when participants perform the AUT following the ESI, compared to following a control induction (Madore, Thakral, Beaty, Addis, & Schacter, 2019). Related studies indicate that participants generated fewer episodic details when imagining a future event and fewer ideas on a DT task, after receiving inhibitory transcranial magnetic stimulation to the angular gyrus, which led to reduced activity in connected regions including hippocampus (Thakral, Madore, Kalinowski, & Schacter, 2020). Together these findings suggest that DT may be facilitated by similar hippocampus-mediated episodic processes as autobiographical memory and imagination of future events. Perceptual details, such as sights, sounds and smells, are a central feature of episodic memory (e.g., Brunel, Labeye, Lesourd, & Versace, 2009; Conway, 2001; Saive, Royet, & Plainly, 2014; Wheeler, Petersen, & Buckner, 2000). Recent studies have shown impairment of perceptual memory retrieval following hippocampal damage (St-Laurent, Moscovitch, Jadd, & McAndrews, 2014; St-Laurent, Moscovitch, & McAndrews, 2016). Despite the growing literature on episodic contributions to creativity, the role of perceptual details in creative cognition has not yet been systematically examined. We contend that perceptual details are particularly relevant to creative writing, and may serve to illustrate the role of imagination in the construction of creative narratives.

Current theories suggest that both semantic and episodic memory processes contribute to the emergence of creative ideas (Benedek, Beaty, Schacter, & Kenett, 2023). If future imagination and creative thinking are mutually supported by episodic processes, a logical

extension is that creative stories will contain a high degree of episodic details. Indeed, one recent study showed that participants who received an ESI prior to a creative writing task produced more episodic details in their stories compared to participants in a control induction, although originality ratings of the stories were comparable following ESI and control inductions (van Genugten, Beaty, Madore, & Schacter, 2022). Episodic detail is typically quantified in memory research by tallying the "who, what, where" of a memory, as well as information that conveys the perceptual quality, or range of sensory experience, recalled. Sheldon, Gurguryan, Madore, and Schacter (2019) found that emphasizing spatial information with a modified ESI procedure prior to autobiographical memory recollection and future simulation promoted the use of episodic processes to specifically access perceptual details of that event. It may be that stories written after ESI were not rated as more original than those from a control condition (van Genugten, Beaty, Madore, & Schacter, 2022) because the induction was geared toward the more general category of episodic, as opposed to perceptual, details. We suggest that the presence of perceptual details is a salient feature of creative narratives, serving to construct a rich scene and transport the reader to an alternative time and place. Provoking sensory details with descriptions of smells, tastes, sights, and sounds may allow readers to more deeply engage with a narrative. Thus, in the present study we examine contributions of episodic memory processes to creativity, looking at the specific role of perceptual details in creative short stories.

With groundbreaking progress in natural language processing, generative language models have experienced a global surge in attention over the past year and a growing number of studies have sought to test their creative abilities. Computer science and psychology researchers alike have set out to test the limits of these sophisticated language models, by identifying tasks on which the models succeed and those on which they fail (Brown et al., 2020; Suzgun et al., 2022; Taecharungroj, 2023). A study by Stevenson, Smal, Baas, Grasman, and van der Maas (2022) tested the ability of GPT-3, an autoregressive language model from OpenAI, to generate responses on the AUT, comparing its performance to human responses in terms of fluency, flexibility, and semantic distance. Their findings indicate that humans currently outperform GPT-3 on traditional measures of creativity, though authors suggest that AI may soon achieve human-level performance. Other studies have demonstrated that in the context of poetry, GPT-3 generated text is indistinguishable from human poems (Köbis & Mossink, 2021; Liao, Wang, Liu, & Jiang, 2019). These findings have led some to explore the possibility of AI-human collaboration leading to more creative art (Hitsuwari, Ueda, Yun, & Nomura, 2023). In a related line of research, Organisciak, Acar, Dumas, and Berthiaume (2023) have demonstrated how large language models can be fine-tuned for the automated scoring of divergent thinking tasks, enabling a more nuanced evaluation of creativity. This advancement surpasses earlier computational approaches, including those based solely on semantic distance, highlighting the multifaceted utility of large language models in both generating and evaluating creative content. Given that the primary aim of this project is to identify characteristic features of creative text, we leverage large language models, in addition to human stories, as a new source of data. We explore the use of generative large language models, such as GPT-3 and GPT-4, to supplement human data collection and examine the extent to which machine-generated

stories can mimic human creativity. We extend prior work comparing output from GPT-3 and human participants in the context of short stories.

Here, we further define the features of creativity using computational linguistic tools. To overcome the limitations of conventional DT assessments such as the AUT, we assess the creative quality of short stories with an established story-writing task (Johnson et al., 2022; Prabhakaran, Green, & Gray, 2014). To assess the degree to which stories connect semantically divergent ideas, we compute *divergent semantic integration* (DSI) scores, leveraging creativity theory and distributional semantics (Johnson et al., 2022). We hypothesize that, in addition to combining semantically distant concepts, the degree to which a story incorporates perceptual details will be predictive of creativity. Additionally, developments in artificial intelligence have raised questions about the capacity of machines to be creative. Given the findings from Stevenson et al. (2022) one might expect human-written stories to be rated as more creative than those from GPT-3. Alternatively, based on findings from studies of poetry, human and GPT-3 stories might be indistinguishable in their creativity. To determine whether or not there are any substantive differences in the creative quality of text written by human participants and that of recently developed large language models, we compare the subjective creativity ratings between human and computer-generated stories. Lastly, given the appearance of GPT-4 after we initiated our study, we replicate these findings in an independent sample of stories generated by GPT-4 and examine whether current language models achieve human-level performance on this creative writing task. The findings from this study have implications for the assessment of verbal creativity and shed light on the potential role of large language models in augmenting human data collection for creative tasks.

METHODS

To test our key hypotheses, we collected creative short stories from human participants online and GPT-3. Our sample of human participants ($n = 50$) was recruited online via Prolific. This sample size is consistent with past research using a similar approach. Participant age ranged from 18 to 35 years ($M = 27.71 \pm .09$; 29 females), all of whom were native English speakers. Informed consent was obtained from all participants prior to participation, with protocol approval from the Institutional Review Board of Harvard University.

To assess creative writing, we used a computerized version of the five-sentence creative story task (Johnson et al., 2022; Prabhakaran, Green, & Gray, 2014). Participants were given a three-word prompt and asked to include all three words when writing (typing) a short story approximately five sentences in length. Participants were given a total of six prompts, presented in random order, with 5 min allotted to write each story. Prompts varied in the semantic distance between cue words, with half of the prompts containing conceptually related, semantically similar words (*stamp-letter-send, week-year-embark*, and *belief-faith-sing*) and the other half of the prompts consisting of more conceptually dissimilar, semantically distant words (*gloom-payment-exist, organ-empire-comply*, and *statement-stealth-detect*). Thus, we collected six responses from each of the 50 participants, providing a total of 300 human-generated stories.

Additionally, we used OpenAI's API to request stories from GPT-3 for each of the prompts administered to humans. Stevenson et al. (2022) conducted Monte Carlo sampling to determine which prompt and parameter settings led to the most valid responses and provided the highest snapshot creativity scores. Based on their results, we administered the short story task prompts to GPT-3's davinci-003 engine as follows. The instruction was: "In this task, you will be asked to write a short story. Try to use your imagination and be creative when writing your story. Write a short story (4–6 sentences) that includes the following words: [*stamp-letter-send*]." The only parameter settings that differed from the default were the temperature, which controls for randomness of output text (sampled from range .65–.80), the frequency penalty (set to 1, decreasing the model's likelihood to repeat the same line verbatim), and the presence penalty (also set to 1, increasing the model's likelihood to talk about new topics). We collected 50 responses from GPT-3 for each of the six prompts, providing a total of 300 GPT-3 generated stories. As a replication sample, we additionally collected 20 responses from GPT-4 for each of the prompts, providing a sample of 120 stories from GPT-4. Stories from human participants, GPT-3, and GPT-4 were rated using both qualitative and quantitative scoring techniques.

CREATIVITY ASSESSMENT

After collecting these stories, participants in an independent online sample (n = 240) were tasked with rating their creative quality. To capture a general assessment of creativity and ensure that no individual rater had too great an influence, we collect 12 ratings for each story. Our sample of raters, recruited via Prolific, consisted of individuals aged between 18 and 55 years (M = 34.58 ±.61; 154 females), all of whom were native English speakers. Raters were prompted to read each story and assign it a creativity score on a scale from 1 (very uncreative) to 5 (very creative). When rating the stories, participants were instructed not to focus too much on the length of the story or the quality of the English, but rather to consider the overall creative quality of the story. Each rater was presented with a sequence of 30 stories, randomly selected from the combined sample. Raters were not informed that stories were generated from humans and AI. A high degree of reliability was found between human ratings of creativity. The average measure ICC was 0.85 with a 95% confidence interval from 0.76 to 0.92 (F (28, 3731) = 7.38, p < .001). Attention checks were included to ensure participant engagement. We computed the mean across all ratings as a composite human rating of creativity for each story.

Additionally, we collected creativity ratings from GPT-4, leveraging its advanced natural language processing capabilities to approximate human-like assessment of the creative quality of stories. Using the identical task instructions provided to human raters, we input each short story into GPT-4, which then provided a single creativity score under the same 1–5 scale used by the sample of human raters. This scoring method was applied to each of the stories independently, ensuring that the model evaluated them based solely on their content, without the influence of previous assessments or additional contextual information. We then examine the correlation between human and GPT-4 assessment of creative quality across the full sample of short stories.

SEMANTIC DIVERSITY

We assessed the extent to which a story connects divergent ideas using DSI: a well-established and validated measure of semantic diversity (Johnson et al., 2022). To extract word embeddings from BERT, the stories are split into sentences and word embeddings are uniquely generated for each sentence. BERT generates embeddings for every word, in every sentence, in every story, reflecting a unique set of weights that indexes how much priority should be placed on each word relative to its context. DSI is then computed by taking the pairwise cosine semantic distance values between all word embeddings. Nothing (except some special characters) was removed before extracting word embeddings. The theoretical range of DSI values is from 0 to 1, though most scores tend to fall between .70 and .90. Higher DSI scores indicate that the story connects more divergent ideas.

PERCEPTUAL DETAILS

Stories were entered into the Linguistic Inquiry and Word Count program (LIWC), which performed an automated count for words falling into a variety of different psychological categories defined by an integrated dictionary (Boyd, Ashokkumar, Seraj, & Pennebaker, 2022; Pennebaker et al., 2001). Here we report results from the Perceptual Processes category, which includes words referring to the process of perceiving (e.g., “observe,” “heard,” “feeling,” “touch”). Perceptual details are computed as the number of perceptual words, relative to the total number of words, present in each story that are linked to a process statement (e.g., “I saw the new car” would count as a perceptual detail, whereas a reference to a “new car” alone would not). Higher perceptual detail scores indicate the prevalence of more such perceptual process statements in each story.

STATISTICAL ANALYSES

To examine the relationship between creativity, semantic diversity, and perceptual detail, we performed multiple linear mixed-effects models. In the first model, we estimate regression parameters for DSI and perceptual details as predictors of creativity, controlling for prompt and word count. To ensure that neither source of text was solely driving the effect, we repeated this analysis within sub-samples of human and GPT-3 stories. Then, we fit a second model to test for a possible interaction between DSI and perceptual detail. To determine whether human and GPT-3 stories differ in terms of their creative quality, we fit a third model with condition (human vs. GPT-3) as a categorical predictor of creativity. We then replicate these analyses in a new set of stories generated by GPT-4. We additionally collected creative evaluations from GPT-4 to test whether or not there was a correspondence between human and GPT evaluations of creativity. As a final step, we repeated the regression analyses with GPT ratings of creativity. Prior to analysis, predictor variables were standardized using a z-score transformation to ensure that they were on the same scale and allow for meaningful comparison. Two outliers exceeding four standard deviations from the mean were removed. R^2 values are reported as a measure of model fit. To summarize the output for our variables of interest, we report the regression parameters (β) with corresponding 95% confidence intervals (CIs). We also report Pearson correlations

(r) along with regression parameters (t- and p-statistics), with a significance threshold of $\alpha = .05$.

RESULTS

First, we performed a posterior predictive check to ensure that the model usefully mimics the observed data. Next, we performed exploratory data analysis to visualize the relationships between our variables of interest. We observed a strong positive correlation between DSI and creativity within the sample of human ($r = .56$) and GPT-3 stories ($r = .56$) (Figure 1a). Additionally, we observed a modest correlation between perceptual details and creativity within stories generated by humans ($r = .16$) and GPT-3 ($r = .25$) (Figure 1b).

As hypothesized, our analysis indicates that both DSI and perceptual details are predictive of creativity. Within the full sample, we observed that DSI was highly predictive of creativity ($\beta = .23$, 95% CI [0.18,0.27], $t = 9.55$, $p < .001$). Our model predicts that two stories that differ by one standard deviation in DSI, will differ by 0.23 points on their creativity ratings, controlling for word count and prompt. Additionally, we observed a positive association between the presence of perceptual details and creativity ($\beta = .12$, 95% CI [0.08, 0.17], $t = 5.40$, $p < .001$). Our model predicts that two stories that differ by one standard deviation in perceptual detail, will differ by of 0.12 creativity points, controlling for word count and prompt. This model explains approximately half of the observed variance in creativity ratings ($R^2 = .49$). Within the sample of human stories, we observed that DSI was highly predictive of creativity ($\beta = .32$, 95% CI [0.24, 0.39], $t = 8.63$, $p < .001$). Additionally, we observed a positive association between the presence of perceptual details and creativity ($\beta = .11$, 95% CI [0.04, 0.17], $t = 3.19$, $p = .002$). This model explains approximately half of the observed variance in creativity ratings ($R^2 = .47$). Within the sample of GPT-3 stories, we observed that DSI was highly predictive of creativity ($\beta = .18$, 95% CI [0.11, 0.25], $t = 4.90$, $p < .001$).Additionally, we observed a positive association between the presence of perceptual details and creativity (b = .09, 95% CI [0.03, 0.15], $t = 2.92$, $p = .004$). This model explains approximately 54% of the variance in creativity ratings ($R^2 = .54$).

In the second model, we observed a significant interaction between DSI and perceptual detail ($\beta = .04$, 95% CI [0.01, 0.08], $t = 2.16$, $p = .03$). This finding suggests that DSI and perceptual detail together have an effect on creativity that is greater than the effect of either variable alone. In the third model, we use the condition (human vs. GPT-3) as our main predictor of creativity. We did not observe any significant difference between human- and computer-generated stories in terms of their creative quality ($\beta = .07$, 95% CI [−0.02, 0.16], $t = 1.59$, $p = .11$). This model explains approximately 39% of the observed variance in creativity ratings ($R^2 = .39$). Our model predicts that a story written by a human will have a creativity score 0.07 points higher than a story from GPT-3 of the same length and prompt.

Within the replication sample of stories generated by GPT-4, we observed a similar set of results: creativity ratings are strongly correlated with both DSI ($r = .46$) and perceptual detail ($r = .44$). Controlling for word count and prompt, we found that both DSI ($\beta = .16$, 95% CI [0.10, 0.22], $t = 5.04$, $p < .001$) and perceptual details ($\beta = .11$, 95% CI [0.04, 0.18], $t = 2.97$, $p = .004$) are highly predictive of creativity. We did not observe a significant interaction

between DSI and perceptual details within the sample of GPT-4 stories (β = −.03, 95% CI −0.09, 0.02], t = 1.14, p = .26). When pooling human, GPT-3 and GPT-4 stories, we found that compared to human stories, both GPT-3 (t = −1.66, p = .10) and GPT-4 (t = 1.98,p = .05) tend to score lower in creativity, though this difference was not significant.

We additionally collected creative evaluations from GPT-4 to determine whether or not there was a correspondence between human and LLM evaluations of creativity. Across the full sample of stories, we observe a robust positive correlation (r = .65) between human and GPT-4 ratings of creativity. This association was relatively consistent within stories written by humans (r = .71), GPT-3 (r = .69) and GPT-4 (r = .72). The observed consistency of GPT-4's creative ratings with human benchmarks was a secondary aim of our analysis, contributing to our understanding of the capability of AI to perform subjective tasks traditionally reserved for human evaluation.

We performed the same regression analyses reported above on the GPT ratings of creativity. Within the full sample of stories, we observed that DSI was highly predictive of GPT creativity ratings (β = .25, 95% CI [0.18, 0.31], t = 7.82, p < .001). Additionally, we observed a significant positive association between the presence of perceptual details and GPT ratings of creativity (β = .11, 95% CI [0.05, 0.17], t = 3.73, p < .001). This model explains about 28% of the observed variance in GPT creativity ratings (R^2 = .28). Within the sample of human stories, we observed that both DSI (β = .30, 95% CI [0.21, 0.39], t = 6.58, p < .001) and perceptual details (β = .12, 95% CI [0.04, 0.20], t = 2.92, p = .004) were significant predictors of creativity, as rated by GPT. Within the sample of GPT-3 stories, we find that DSI was predictive of GPT ratings of creativity (β = .15, 95% CI [0.07, 0.24], t = 3.52, p < .001); however, there was not a statistically significant association perceptual details and GPT creativity ratings (β = .06, 95% CI [−0.03, 0.15], t = 1.38, p = .17). In the interaction model, we failed to observe a significant interaction between DSI and perceptual detail (β = .05, 95% CI [−0.01, 0.10], t = 1.62, p = .11). In the third model, we found that human stories were rated as significantly more creative (assessed by GPT) as compared to computer-generated stories (β = .14, 95% CI [0.02, 0.25], t = 2.36, p = .02).

Within the replication sample of stories generated by GPT-4, we observed a similar set of results: GPT ratings of creativity are positively correlated with both DSI (r = .60) and perceptual detail (r = .30). We observed that DSI was significantly predictive of GPT ratings of creativity (β = .44, 95% CI [0.34, 0.54], t = 8.67, p < .001); however, there was not a significant association with perceptual details (β = .09, 95% CI [−0.02, 0.21], t = 1.61, p = .11). Additionally, we did not observe an interaction between DSI and perceptual details within the sample of GPT-4 stories (β = −.02, 95% CI [0.11, 0.08], t = .37, p = .71). When pooling human, GPT-3 and GPT-4 stories, we found that compared to human stories,GPT-3 stories scored significantly lower (β = −.14, 95% CI [−0.25, −0.02], t = 2.36, p = .02) and GPT-4 stories scored significantly higher (β = .76, 95% CI [0.60, 0.91], t = 9.74, p < .001) on GPT ratings of creativity. These findings suggest that GPT ratings of creativity favor GPT-4 stories over human-generated stories.

DISCUSSION

The present study highlights constituent features of creative writing, specifically the role of semantic diversity and perceptual details in short stories written by humans and large language models. As expected, results showed that semantic diversity is an important predictor of creativity, explaining over half of the variance in creativity ratings. Additionally, we observed a positive association between perceptual detail and creativity, indicating that the inclusion of specific perceptual details is an important feature of creative narratives. These results suggest that both semantic and episodic memory may jointly contribute to the process of creative writing.

Converging evidence from neuroimaging and psychology studies have highlighted the distinct contributions of semantic and episodic memory to creative ideation (Beaty et al., 2020; Benedek, Beaty, Schacter, & Kenett, 2023). Semantic memory refers to the knowledge of concepts and general facts about the world, whereas episodic memory involves the recollection of specific events and experiences, including the time, place, and context in which they occurred. Semantic memory can be represented as a network structure, where related concepts are stored nearby in a semantic space. Applications of network science have led to a deeper understanding how flexibility of semantic memory structure contribute to creative thinking (Abraham, 2014; Kenett, 2018). Relatively less is known about the contributions of episodic memory to creativity. In the context of creative writing, episodic memory may enable individuals to retrieve and incorporate specific sensory details, leading to more descriptive narratives. Perceptual detail refers to the richness and specificity of sensory information in a text, and can be used to evoke vivid images and sensory experiences in the reader's mind that produce a more immersive and transportive reading experience. While the findings of van Genugten, Beaty, Madore, and Schacter (2022) did not show a reliable association between episodic detail and originality of stories, their measure of episodic detail is distinct from our assessment of perceptual detail. A proposed framework for understanding the role of memory in creative ideation (Benedek, Beaty, Schacter, & Kenett, 2023) suggests that both semantic and episodic memory play a crucial role in the creative process. Benedek, Beaty, Schacter, and Kenett (2023) contend that semantic memory could play a relatively larger role when creative thinking relies primarily on conceptual information, whereas episodic memory processes could be more relevant to tasks and strategies requiring more complex forms of imagination involving mental simulations and construction of scenes. Our findings are consistent with this framework and suggest a possible role for episodic memory processes in generating perceptual details in creative narratives.

Recent advancements in generative language models have brought this form of AI to broad attention. These models have elicited interest not only for their extraordinary capacity to produce human-like language, but also for their philosophical implications regarding the nature of language acquisition (Piantadosi, 2023). Their success challenges previous critiques of generative language models and highlights their creative potential. According to classic views of creativity, these models appear to be capable of generating responses to prompts that are both novel and useful; however, important questions remain about how the assessment of creative output may be influenced by its source. Empirical evidence suggests

that that people tend to be negatively biased against AI-generated artwork (Bellaiche et al., 2023). It may be that human engagement in the artistic process is integral to its appreciation and thus, artificial intelligence systems may never replace human creativity. Nevertheless, AI is increasingly used to help writers to generate drafts and refine language in a specific genre, suggesting that AI systems can be valuable assistants in creative pursuits (Haase & Hanel, 2023). Important questions remain about the best practice for integrating AI in the creative process.

The integration of LLMs in the assessment of creative writing holds tremendous potential for the field of creativity evaluation. In the present study, we found a strong correspondence between GPT-4 and human ratings of creativity, within the context of narrative short stories. We observed that both semantic diversity and perceptual details are key predictors of GPT-4 creativity ratings. Additionally, we found that GPT-4 evaluated its own stories as being significantly more creative than human stories, which is an unexpected finding that merits exploration in future research. Our findings not only validate the efficacy of LLMs for creative evaluation, but also highlight the impacts that AI is poised to make on the measurement of creativity in the digital age. These results complement existing research using LLMs to automatically evaluate the originality and quality of creative ideas (Luchini et al., 2023; Organisciak, Acar, Dumas, & Berthiaume, 2023), underscoring their utility in systematically evaluating elements such as originality and quality. These studies finetuned a variety of LLMs to predict human creativity ratings in divergent thinking tasks and problem-solving tasks. Though it is not the primary focus of this study, we suggest that fine-tuning of LLMs for the domain of creative writing would be a fruitful area of investigation. The present research offers new insights into the creative writing process and the role of semantic diversity and perceptual details in the writing of creative stories; however, it is essential to acknowledge the limitations imposed by the specific prompts used in our study. We recognize that the choice of prompt can greatly influence the quality and content of the generated stories, especially when using higher temperatures in generative language models. Thus, our findings should be interpreted in the context of the specific prompts employed here.

In conclusion, the present study contributes to the growing body of research on the cognitive processes that underlie creative thinking. Our results indicate that creative writing involves integrating semantically divergent concepts with perceptually descriptive information. Furthermore, this study provides evidence that GPT-3 and GPT-4 can generate stories that are comparable in creativity to those produced by humans. While we do not wish to claim that these large language models have an experience of episodic remembering as humans do, we suggest that creative writing in both humans and GPT-3/GPT-4 make use of perceptual details that are similar to sensory details typically ascribed to episodic memory in humans, as well as novel semantic associations. Future research should aim to disentangle the distinct episodic and semantic features that contribute to creative writing and explore the potential for human–AI collaboration.

Acknowledgments

This research was supported by the National Institute on Aging (AG008441 to DLS).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Abraham A (2014). Creative thinking as orchestrated by semantic processing vs. cognitive control brain networks. Frontiers in Human Neuroscience, 8, 95; doi: 10.3389/fnhum.2014.00095. [PubMed: 24605098]

Acar S, Ogurlu U, & Zorychta A (2023). Exploration of discriminant validity in divergent thinking tasks: A meta-analysis. Psychology of Aesthetics, Creativity, and the Arts, 17(6), 705–724; doi: 10.1037/aca0000469.

Acar S, & Runco MA (2015). Thinking in multiple directions: Hyperspace categories in divergent thinking. Psychology of Aesthetics, Creativity, and the Arts, 9, 41–53; doi: 10.1037/a0038501.

Amabile TM (1982). Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology, 43(5), 997–1013; doi: 10.1037/0022-3514.43.5.997.

Barbot B (2018). The dynamics of creative ideation: Introducing a new assessment paradigm. Frontiers in Psychology, 9, 2529; doi:10.3389/fpsyg.2018.02529. [PubMed: 30618952]

Beaty RE, Chen Q, Christensen AP, Kenett YN, Silvia PJ, Benedek M, & Schacter DL (2020). Default network contributions to episodic and semantic processing during divergent creative thinking: A representational similarity analysis. Neuro-Image, 209, 116499; doi: 10.1016/j.neuroimage.2019.116499. [PubMed: 31887423]

Beaty RE, & Johnson DR (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. Behavior Research Methods, 53(2), 757–780; doi: 10.3758/s13428-020-01453-w. [PubMed: 32869137]

Bellaiche L, Shahi R, Turpin MH, Ragnhildstveit A, Sprockett S, Barr N, & Seli P (2023). Humans vs. AI: Whether and why we prefer human-created compared to AI-created artwork. Cognitive Research: Principles and Implications, 8, 42; doi: 10.31234/osf.io/f9upm. [PubMed: 37401999]

Benedek M, Beaty RE, Schacter DL, & Kenett YN (2023). The role of memory in creative ideation. Nature Reviews Psychology,2(4), 246–257.

Boyd RL, Ashokkumar A, Seraj S, & Pennebaker JW (2022). The development and psychometric properties of LIWC-22. Austin,TX: University of Texas at Austin; doi: 10.1037/gdn0000195.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, … Askell A (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Brunel L, Labeye E, Lesourd M, & Versace R (2009). The sensory nature of episodic memory: Sensory priming effects due to memory trace activation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(4), 1081–1088; doi: 10.1037/a0015537. [PubMed: 19586271]

Conway MA (2001). Sensory-perceptual episodic memory and its context: autobiographical memory. Philosophical Transactions of the Royal Society B, 356(1413), 1375–1384; doi: 10.1098/rstb.2001.0940.

Dumas D, Doherty M, & Organisciak P (2020). The psychology of professional and student actors: Creativity, personality, and motivation. PLoS One, 15(10), e0240728; doi: 10.1371/journal.pone.0240728. [PubMed: 33091923]

Forthmann B, Oyebade O, Ojo A, Günther F, & Holling H (2019). Application of latent semantic analysis to divergent thinking is biased by elaboration. The Journal of Creative Behavior, 53(4), 559–575; doi: 10.1002/jocb.240.

Guilford JP (1967). The nature of human intelligence. New York: McGraw-Hill.

Haase J, & Hanel PH (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. arXiv preprint arXiv:2303.12003.

Hitsuwari J, Ueda Y, Yun W, & Nomura M (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. Computers in Human Behavior, 139, 107502; doi: 10.1016/j.chb.2022.107502.

Jauk E, Benedek M, & Neubauer AC (2014). The road to creative achievement: A latent variable model of ability and personality predictors. European Journal of Personality, 28(1), 95–105; doi: 10.1002/per.1941. [PubMed: 24532953]

Johnson DR, Kaufman JC, Baker BS, Patterson JD, Barbot B, Green AE, … Beaty RE (2023). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. Behavior Research Methods, 55(7), 3726–3759; doi: 10.3758/s13428-022-01986-2. [PubMed: 36253596]

Kenett YN (2018). Investigating creativity from a semantic network perspective. In Kapoula Z, Volle E, Renoult J, & Andreatta M (Eds.), Exploring transdisciplinarity in art and sciences (pp. 49–75). Springer Verlag.

Kenett YN, & Faust M (2019). A semantic network cartography of the creative mind. Trends in Cognitive Sciences, 23(4), 271–274; doi: 10.1016/j.tics.2019.01.007. [PubMed: 30803872]

Köbis N, & Mossink LD (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior, 114, 106553; doi: 10.1016/j.chb.2020.106553.

Liao Y, Wang Y, Liu Q, & Jiang X (2019). GPT-based generation for classical Chinese poetry. arXiv preprint arXiv:1907.00151. doi: 10.48550/arXiv.1907.00151.

Luchini S, Maliakkal NT, DiStefano PV, Patterson JD, Beaty R, & Reiter-Palmon R (2023). Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings. doi: 10.31234/osf.io/g5qvf.

Madore KP, Addis DR, & Schacter DL (2015). Creativity and memory: Effects of an episodic-specificity induction on divergent thinking. Psychological Science, 26(9), 1461–1468; doi: 10.1177/0956797615591863. [PubMed: 26205963]

Madore KP, Jing HG, & Schacter DL (2016). Divergent creative thinking in young and older adults: Extending the effects of an episodic specificity induction. Memory & Cognition, 44(6), 974–988; doi: 10.3758/s13421-016-0605-z. [PubMed: 27001170]

Madore KP, Thakral PP, Beaty RE, Addis DR, & Schacter DL (2019). Neural mechanisms of episodic retrieval support divergent creative thinking. Cerebral Cortex (New York, N.Y.: 1991), 29(1), 150–166; doi: 10.1093/cercor/bhx312. [PubMed: 29161358]

Mednick S (1962). The associative basis of the creative process. Psychological Review, 69(3), 220–232; doi: 10.1037/h0048850. [PubMed: 14472013]

Organisciak P, Acar S, Dumas D, & Berthiaume K (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. Thinking Skills and Creativity, 49, 101356; doi: 10.1016/j.tsc.2023.101356.

Orwig W, Diez I, Vannini P, Beaty R, & Sepulcre J (2021). Creative connections: Computational semantic distance captures individual creativity and resting-state functional connectivity. Journal of Cognitive Neuroscience, 33(3), 499–509; doi: 10.1162/jocn_a_01658. [PubMed: 33284079]

Pennebaker JW, Francis ME, & Booth RJ (2001). Linguistic inquiry and word count (LIWC): LIWC2001. Mahwah: Lawrence Erlbaum.

Piantadosi ST (2023). Modern language models refute Chomsky's approach to language. https://lingbuzz.net/lingbuzz/007180.

Prabhakaran R, Green AE, & Gray JR (2014). Thin slices of creativity: using single-word utterances to assess creative cognition. Behavior Research Methods, 46(3), 641–659; doi: 10.3758/s13428-013-0401-7. [PubMed: 24163211]

Reiter-Palmon R, Forthmann B, & Barbot B (2019). Scoring divergent thinking tests: A review and systematic framework. Psychology of Aesthetics, Creativity, and the Arts, 13(2), 144–152; doi: 10.1037/aca0000227.

Saive A-L, Royet J-P, & Plainly J (2014). A review on the neural bases of episodic odor memory: from laboratory-based to autobiographical approaches. Frontiers in Behavioral Neuroscience, 8, 240; doi: 10.3389/fnbeh.2014.00240. [PubMed: 25071494]

Schacter DL, & Madore KP (2016). Remembering the past and imagining the future: Identifying and enhancing the contribution of episodic memory. Memory Studies, 9(3), 245–255; doi: 10.1177/1750698016645230. [PubMed: 28163775]

Sheldon S, Gurguryan L, Madore KP, & Schacter DL (2019). Constructing autobiographical events within a spatial or temporal context: a comparison of two targeted episodic induction techniques. Memory (Hove, England), 27(7), 881–893; doi: 10.1080/09658211.2019.1586952. [PubMed: 30849029]

Stevenson C, Smal I, Baas M, Grasman R, & van der Maas H (2022). Putting GPT-3's creativity to the (alternative uses) test. arXiv preprint arXiv:2206.08932. doi: 10.48550/arXiv.2206.08932.

St-Laurent M, Moscovitch M, Jadd R, & McAndrews MP (2014). The perceptual richness of complex memory episodes is compromised by medial temporal lobe damage. Hippocampus, 24(5), 560–576; doi: 10.1002/hipo.22249. [PubMed: 24449286]

St-Laurent M, Moscovitch M, & McAndrews MP (2016). The retrieval of perceptual memory details depends on right hippocampal integrity and activation. Cortex, 84, 15–33; doi: 10.1016/j.cortex.2016.08.010. [PubMed: 27665526]

Suzgun M, Scales N, Scharli N, Gehrmann S, Tay Y, Chung HW, … Wei J (2022). Challenging BIG-bench tasks and whether chain-of-thought can solve them. ArXiv, abs/2210.09261. doi: 10.48550/arXiv.2210.09261.

Taecharungroj V (2023). "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on Twitter. Big Data and Cognitive Computing, 7(1), 35; doi: 10.3390/bdcc7010035.

Thakral PP, Madore KP, Kalinowski SE, & Schacter DL (2020). Modulation of hippocampal brain networks produces changes in episodic simulation and divergent thinking. Proceedings of the National Academy of Sciences, 117(23), 12729–12740; doi: 10.1073/pnas.2003535117.

van Genugten RD, Beaty RE, Madore KP, & Schacter DL (2022). Does episodic retrieval contribute to creative writing? an exploratory study. Creativity Research Journal, 34(2), 145–158; doi: 10.1080/10400419.2021.1976451. [PubMed: 35814526]

Wheeler ME, Petersen SE, & Buckner RL (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. Proceedings of the National Academy of Sciences, 97(20), 11125–11129; doi: 10.1073/pnas.97.20.1112.

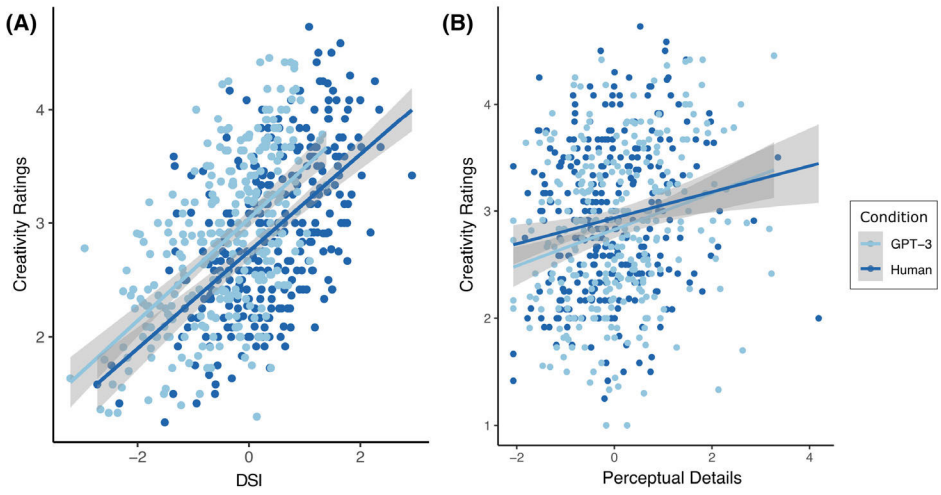

FIGURE 1.
Semantic diversity (A) and perceptual detail (B) correlate with creativity ratings.