Journal of Creativity 35 (2025) 100113

Contents lists available at ScienceDirect

Journal of Creativity

journal homepage: www.elsevier.com/locate/yjoc

Has the creativity of large-language models peaked? An analysis of inter- and intra-LLM variability

Jennifer Haase [a,*], Paul H.P. Hanel [b], Sebastian Pokutta [c]

[a] Department of Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6 10099 Berlin, Germany
[b] Department of Psychology, University of Essex, 4.704, Colchester Campus, United Kingdom
[c] Department of Mathematics TU Berlin and Zuse Institute Berlin

ARTICLE INFO

Keywords:
Generative Ai
Benchmark testing
Creativity
Large language models
LLMs

ABSTRACT

Numerous studies reported that large language models (LLMs) can match or even surpass human performance in creative tasks. However, it remains unclear if LLMs have become more creative over time and how consistent their creative output is. In this study, we evaluated 14 widely used LLMs—including GPT-4, Claude, Llama, Grok, Mistral, and DeepSeek—across two validated creativity assessments: the Divergent Association Task (DAT) and the Alternative Uses Task (AUT). We found no evidence of increased creative performance over the past 18–24 months, with GPT-4 performing worse than in previous studies. For the more widely used AUT, all models performed on average better than the average human, with GPT-4o and o3-mini performing best. However, only 0.28 % of LLM-generated responses reached the top 10 % of human creativity benchmarks. Beyond inter-model differences, we document substantial intra-model variability: the same LLM, given the same prompt, can produce outputs ranging from below-average to original. This variability has important implications for both creativity research and practical applications. Ignoring such variability risks misjudging the creative potential of LLMs, either inflating or underestimating their capabilities. The choice of prompts affected LLMs differently. Our findings underscore the need for more nuanced evaluation frameworks and highlight the importance of model selection, prompt design, and repeated assessment when using Generative AI (GenAI) tools in creative contexts.

1. Introduction

Large Language Models (LLMs) have moved out of research labs and into our everyday lives. LLMs are often marketed as *creative* due to their ability to generate text and ideas (e.g., OpenAI's GPT 4.5, OpenAI, 2025; or Grok beta, xAI, 2025). They allow users to brainstorm, draft content, and generate novel ideas with ease (Memmert et al., 2024b). Consumers have responded eagerly, with recent surveys indicating that a majority of LLM users believe that these models enhance their creativity (Pandya, 2024). However, while LLMs can facilitate idea generation (Vaccaro et al., 2024; Wan et al., 2024), their widespread adoption raises important questions about their actual creative potential and the nature of their outputs (Runco, 2023). Prior research suggested that LLM-generated ideas, although appearing individually creative, tend to lead to homogeneous outcomes across various domains, including creative writing, survey responses, and research idea generation (Anderson et al., 2024; Doshi & Hauser, 2024; Moon et al., 2024). For instance, stories written with ChatGPT assistance were more uniform than those generated independently by humans (Doshi & Hauser, 2024). Similarly, LLM-authored college essays contained fewer novel ideas than those written without LLM assistance (Moon et al., 2024). While these studies raise concerns about the creativity of LLMs, they typically focus on a single LLM, leaving unanswered the question of whether this homogeneity is unique to specific models or a broader phenomenon across different LLMs. Furthermore, it is unclear whether LLMs have become more creative since 2023, when they gained wider recognition.

The present study aimed to shed light on these questions, particularly by addressing two key questions in this context: (1) Are current LLMs more creative than earlier versions and average human baselines, and which model performs best? (2) Do LLMs generate a diverse range of ideas within a session, or do their outputs converge toward homogeneous patterns? Put differently, are the answers of LLMs stable (i.e., homogeneous)? Previous research has neglected the output variability within the same LLM. This is important because greater variability (i.e., lower stability) within the responses of the same LLM can lead to either drastically over- or underestimating their creative capabilities. Recent

* Corresponding author.
E-mail addresses: jennifer.haase@hu-berlin.de (J. Haase), p.hanel@essex.ac.uk (P.H.P. Hanel), pokutta@zib.de (S. Pokutta).

https://doi.org/10.1016/j.yjoc.2025.100113
Received 8 July 2025; Received in revised form 31 October 2025; Accepted 5 November 2025
Available online 7 November 2025

research on feature space alignment in LLMs suggested that these models exhibit structural similarities, potentially leading to homogeneous outputs across different architectures (Kleinberg & Raghavan, 2021; Wenger & Kenett, 2025). This would imply that, regardless of the specific LLM used, users may experience a collective narrowing of creative expression due to shared underlying model biases (Huh et al., 2024; Lan et al., 2025). Because expectations and framing influence human creative performance, we also tested whether prompting LLMs with the context of a "creativity test" affects their output quality, extending existing work on priming and task framing effects in LLMs (Gosling et al., 2024; Renze, 2024; Salinas & Morstatter, 2024).

To explore these issues, we systematically evaluated a diverse set of LLMs using the Divergent Association Task (DAT, Olson et al., 2021) and the Alternative Uses Task (AUT, Christensen et al., 1960), assessing both inter-model differences (creativity across models) and intra-model variance (creativity within repeated interactions with the same model). By using two tests as proxies for originality and semantic diversity, we examined which models scored highest in typical "creativity assessment" tasks and whether they encourage or constrain the generation of diverse ideas.

As LLMs become increasingly integrated into the human-AI co-creative process (Haase & Pokutta, 2024), understanding the actual breadth and depth of their creative capacities is crucial. Our findings had direct implications for model selection in practice, the design of collaborative tools, and the broader question of whether generative artificial intelligence (GenAI) systems meaningfully expand, or inadvertently narrow, the human creative landscape (Doshi & Hauser, 2024; Kleinberg & Raghavan, 2021). By revisiting and expanding previous research, we provide a comprehensive evaluation of LLM creativity, offering insights into their evolving role in human ideation and problem-solving.

1.1. Creativity of large language models

Creativity has traditionally been regarded as a uniquely human trait: one that distinguishes us from machines and automation (Miller, 2019). However, recent advances in GenAI have reignited debates about whether GenAI can exhibit creativity, particularly in fields such as literature, music, art, and problem-solving. While AI has already surpassed human capabilities in structured domains like Chess and Go (Gaessler & Piezunka, 2023; Krakowski et al., 2023) and has been used in mathematics to solve open problems (Davies et al., 2021; Mundinger et al., 2024, 2025; Swirszcz et al., 2025), it remains uncertain whether it can achieve high levels of creativity or if it simply recombines existing knowledge in novel ways (Holford, 2019; Kirkpatrick, 2023; White, 2023). Some argue that creativity remains one of the last strongholds of human superiority over AI (Holford, 2019), as it involves not only idea generation but also problem formulation, selection, and implementation (Botella & Lubart, 2016; Williams et al., 2016). Adding to the complexity of evaluating GenAI creativity is the well-documented sensitivity of LLMs to prompts. Small changes in phrasing or instruction framing can lead to substantial differences in output (Chang et al., 2024; Mizrahi et al., 2024), which complicates comparisons across studies and even within-task benchmarks.

Creativity—for humans—is defined as the ability to generate *new* and *useful* ideas (Runco & Jaeger, 2012; Plucker, 2004). Further, it is conceptualized as an interaction between cognitive abilities, environmental factors, and social validation (Amabile, 2017). High-level creativity, particularly in science and the arts, requires not only originality but also refinement, testing, and evaluation, as well as recognition, to validate an idea as a creative product (Benedek et al., 2020; Kaufman et al., 2016; Simonton, 2013). While human creativity involves free-associative thinking and problem formulation (Botella et al., 2018; Steele et al., 2018), GenAI mimics these processes through probabilistic text generation and pattern recognition (Marcus et al., 2022). Although comparable in output, as AI may produce original and useful outputs (cf. Section 2.2), some argue that it does not "create" in the human sense: the difference lies not in what is produced but in how and why it is produced (Runco, 2023).

Importantly, the discussion of AI's creative potential also addressed the ontological status of AI-generated creativity. Critics argue that GenAI, by virtue of relying on pre-existing data, is confined to "incremental creativity" and lacks the emotional depth and subjectivity that characterize human creative acts (Boden, 1998, 2009; Cropley & Cropley, 2023; Runco, 2023; White, 2023). According to this view, GenAI may simulate creativity convincingly but cannot embody the underlying processes of creative intention or self-expression.

We do not fully subscribe to such reductionist views. While it is true that LLMs are trained on existing knowledge, their ability to recombine, adapt, and contextualize information in novel ways demonstrates inherent creativity. LLMs are designed to balance factual precision with creative expression, leveraging probabilistic language modeling, flexibility, and randomness to generate content that is perceived as original and inventive (Rafner et al., 2023; Sinha et al., 2023). Empirical studies showed that GenAI-generated outputs can sometimes matched or even exceeded human performance in tasks requiring originality and elaboration (Gilhooly, 2023; Haase & Hanel, 2023). Notably, in several domains, GenAI outputs were indistinguishable from human creations, successfully fooling experts in tasks ranging from scientific abstract writing (Else, 2023) to visual art production (Haase et al., 2023).

Although the philosophical debate about whether AI merely appears creative or truly *is* creative remains unresolved (Runco, 2023; Boden, 2009) and may not be practically relevant, our focus lies on *empirical* outcomes and the *measurable* creative potential of LLMs. Humans have created with the support of technology since the development of tools (Haase & Pokutta, 2024), and we aimed to increase our understanding of how LLMs can serve as a creative support system, providing new and useful ideas to users. Thus, we briefly discuss how creativity is traditionally measured, what those measures reveal of LLMs' creative potential, and how this can be useful for the human co-creative process with LLMs in the next sections.

1.2. Measuring divergent thinking

Divergent thinking (DT) refers to the ability to generate multiple, varied, and novel ideas in response to open-ended problems. This ability is often perceived as a core cognitive process underlying creativity. Indeed, da Costa et al. (2015)'s meta-analysis showed that DT demonstrates the strongest correlation with creativity-related constructs ($\bar{r}r = 0.27$) compared to other individual difference measures. However, this moderate effect size also illustrates a critical point: DT is far from synonymous with creativity. Rather, it should be seen as an indicator of creative potential, as one aspect of a broader and more complex cognitive and motivational landscape (Runco et al., 2011). One reason for the over-identification of DT with creativity lies in the dominant use of DT measures in creativity research. The most widely used measures—such as the Alternate Uses Test (AUT; Christensen et al., 1960) and the Torrance Tests of Creative Thinking (TTCT; Torrance, 1972)—either directly assess divergent idea generation or embed association-based tasks as central components. As a result, much of what is empirically known about creativity relies on divergent idea generation.

DT tasks are typically scored along several performance dimensions: *fluency* (number of responses), *flexibility* (variety of categories), *originality* (statistical infrequency or uniqueness), and sometimes *elaboration* (amount of detail). Among these, fluency is most often reported, although originality is arguably more aligned with creative value (Silvia et al., 2008). However, fluency scores tend to correlate strongly with originality, suggesting that originality may, at least in part, go along with response quantity as well as quality.

One standard way to measure DT is via the Alternate Uses Task (AUT; Christensen et al., 1960), which asks participants to generate alternative uses for common objects. By contrast, the Divergent Association Task (DAT; Olson et al., 2021) is a more recent measure that quantifies the

semantic distance between ideas. Participants (or LLMs) are instructed to generate ten words that are as semantically different from each other as possible. Responses are then scored based on pairwise semantic distances computed from a pre-trained embedding model, yielding a scalable, language-based index of creative divergence. However, despite their widespread use, DT measures do not fully capture the complexity of creative thinking (Reiter-Palmon et al., 2019; Runco & Acar, 2012). Creativity often requires generating ideas and selecting, refining, and contextualizing them (Cromwell et al., 2022). For example, *convergent thinking* involves narrowing down multiple possibilities to identify the most effective solution and plays a key role in evaluating and implementing creative ideas (Cropley, 2006). *Emergent thinking*, on the other hand, involves exploring potential problem spaces for existing solutions, such as experimenting with new technologies to discover their applications (Cromwell et al., 2023). Recent work has also emphasized the importance of metacognitive strategies such as asking more complex questions, which can foster deeper and more creative exploration (Raz et al., 2023).

In sum, while DT remains a valuable proxy for *creative potential*, particularly in controlled experimental contexts, it represents only one dimension of creative cognition. Its explanatory power is enhanced when situated within a broader framework that includes convergent, emergent, and metacognitive thinking. Accordingly, we used DT tasks in this study not as exhaustive indicators of creativity, but as focused tools to probe one central aspect of the overarching concept of creativity.

1.3. Empirical analyses of LLMs' creativity

Recent research suggested that GenAI, and LLMs in particular, were capable of producing outputs that met established criteria for creativity—namely, *originality* and *usefulness* (e.g., Guzik et al., 2023; Haase & Hanel, 2023). In fields such as literature, GenAI-generated texts frequently matched or even surpassed human writing in fluency and coherence (Gómez-Rodríguez & Williams, 2023). Comparable evidence was emerging in other creative domains, including music (Civit et al., 2022) and visual art (DiPaola & McCaig, 2016). To empirically evaluate the creative potential of LLMs, researchers have increasingly turned to standardized psychological creativity assessments. These included DT tests such as the AUT, the DAT, and the TTCT (for an overview of studies measuring the creativity of LLMs, cf. Table 4 in the Supplementary Material). However, it is important to note that LLM performance on such tasks is highly sensitive to prompt phrasing and format. Even subtle variations in instructions can lead to substantial differences in output quality and style, as LLMs rely on learned probabilistic patterns rather than intrinsic task comprehension (Chang et al., 2024). In these tasks, GenAI systems such as GPT-3.5 and GPT-4 often performed at or above average human levels, particularly in dimensions like *fluency* and *elaboration* (Guzik et al., 2023; Haase & Hanel, 2023). Some studies even reported that GPT-4 scores in the top 1 % in *originality* and *fluency* on the TTCT (Soroush et al., 2025). However, while LLMs frequently exceed average performance, they typically fell short of the originality levels exhibited by highly creative individuals or expert human creators (Haase & Hanel, 2023; Koivisto & Grassini, 2023). Further, although such findings are striking, they should be interpreted with caution, as model performance may have reflected specific tuning to familiar prompt formats rather than generalized creative competence.

A parallel trend in this research area was the development of automated scoring systems that leverage AI and embedding-based metrics to evaluate creativity (e.g., Hadas, 2025; Organisciak et al. 2025). For example, the DAT relies on measuring semantic distance between generated words, and recent extensions include LLM-based scoring architectures (Haase et al., 2025). Similarly, systems such as the Open-Ended Creativity Scoring AI (OpenScoring; Organisciak et al., 2023) apply transformer-based models to estimate creativity in free-form outputs, offering scalable alternatives to traditional human coding (Soroush et al., 2025). These approaches increase consistency and efficiency in evaluating creative output.

2. Experimental setup

To evaluate the creative potential of contemporary LLMs, we systematically compared multiple models on two standardized DT tasks—the DAT and the AUT—assessing both inter-model performance and intra-model output variability. In addition, we tested two distinct prompt formulations for each task—one disclosing the DAT test context (aware) and one not (unaware)—to standardize input conditions across models and examine whether prompt phrasing and creative priming influence performance differences (see in the Supplementary Material "Prompt Template" for details). While this is not a prompting-focused study as many more prompt variants could (and arguably should) be explored to fully capture systematic effects, we acknowledge that prompt design is a non-trivial factor influencing LLM output, particularly in open-ended generative tasks (Runco et al., 2025). As such, we include basic prompt variations to assess the robustness of model responses. The following section describes the models evaluated, the software environment, and the methodology used to administer and score the tasks.

All experimental data were stored in JSON format. Each record contained metadata (e.g., timestamps, model identifiers, prompt variants), full results per model-object combination, and trial-level scores. The data and R code to reproduce our analyses are publicly accessible on the Open Science Framework (OSF) .[1] All models were accessed via their API using their default parameter settings through LiteLLM between 25 and 28 February 2025 for the DAT tests and 1 and 3 March 2025 for the AUT tests. Each trial was run in an isolated prompt session to avoid carryover effects or memory accumulation. To ensure reproducibility, all API calls were logged, and each model's version and associated metadata were recorded.

We used a broad range of widely used models in our experiments, with 10 models: Claude-3.5-sonnet, Claude-3.7-sonnet, DeepSeek-R1–70B, DeepSeek-R1-Distill-Qwen-7B, Gemini-pro, GPT-4.5-preview, GPT-4o, Grok-2-latest, Grok-beta, Llama-3.3–70B-Instruct, Mistral-Nemo-Instruct-2407, o3-mini, Phi-4, Qwen-2.5–7B-Instruct-1 M (summarized in Table 1 in the Supplementary Material), as research has shown that different models exhibit distinct tendencies in "behavior" even when prompted identically (Zhang et al., 2024).

2.1. Divergent association task (DAT)

LLMs were evaluated using the DAT to assess their ability to generate semantically diverse content. For scoring and evaluation, we used the official DAT website with default settings and recorded all output metrics, including semantic distance scores and percentiles.

2.2. Scoring and analysis

The evaluation process involved submitting each model's generated words to the official DAT website, recording the assigned scores, percentiles, and semantic distance matrices. The experimental process followed a structured workflow. First, the experimental setup was established by selecting the models, determining the number of trials, and defining the evaluation method. During the generation phase, each model produced 10 words based on the standardized prompt template. These words were then submitted to the DAT website for scoring, and the resulting scores, percentiles, and word matrices were recorded. Each model ran the test 100 times, unaware of previously run tests.

DAT scoring is based on semantic distance, resulting in a score that is the average semantic distance between all pairs of words (higher is better). Further, we baselined all scores with human-generated scores

[1] Data, Code and Supplementals available at: https://osf.io/e62fx/overview

for better comparison. We used a large sample of humans (n = 8907) from Olson et al. (2021). We focused, unless otherwise stated, on the percentile rank, as it allows us to directly compare LLM responses with human responses (e.g., if the percentile is > 50, the LLM would be better than the average human).

Following data collection, results were analyzed to assess model performance across the score metrics. The analysis included calculating and visualizing the average scores per model, examining score distributions, and constructing word matrices to illustrate semantic distances between generated words. Normalizing based on human performance allowed comparison of how LLM outputs align with human DT performance.

2.3. Alternate use task (AUT)

LLMs were evaluated using the AUT to assess their ability to produce original and useful ideas. For scoring and evaluation, we used the OpenScoring API[2] with default parameters. All documentation was referenced from the official API guide.[3]

2.4. Test objects

A set of sixteen common objects (brick, shoe, paper clip, button, cardboard box, pencil, bottle, newspaper, umbrella, pants, ball, tire, fork, toothbrush) was used to standardize inputs across trials. These items were selected for their general familiarity and variety in potential use contexts as used in prior literature (Christensen et al., 1960; Organisciak et al., 2023).

2.5. Scoring and analysis

The experimental setup defined the model pool, the number of trials per model-object pair, and the number of responses to be generated per prompt. During each trial, the model was prompted to generate alternative uses for a specific object using one of the predefined prompt templates. Generated responses were then submitted to the OpenScoring API, which returned evaluations for originality for each response and an overall creativity score for the full response set. As a next step, a percentile ranking against human responses was calculated as a benchmark, based on n = 151 from Hubert et al. (2024).

2.6. Performed experiments

In the following section, we provided an overview of the experiments we conducted.

1. **Divergent Association Task (DAT)**: For each of the 14 LLMs, we conducted 100 independent trials using each of the two prompt variants described in Section 3.2.1: the 'DAT aware' prompt and the 'DAT unaware' prompt. This resulted in a total of 2800 DAT evaluations across all models and prompt conditions.
2. **Alternate Use Task (AUT)**: For each of the 14 LLMs, we conducted 4 separate trials where each model was instructed to generate 100 alternative uses per trial. Each trial used once the 'Practical and Feasible' prompt and once the 'Creative and Unconventional' prompt. This resulted in 56 prompt-model trials, though many models, especially lower-capacity ones, generated fewer than the requested 100 uses per trial.

For both tasks, we recorded the raw responses, computed scores using the respective scoring systems, and performed statistical analyses to assess performance relative to human benchmarks.

[2] OpenScoring API at https://openscoring.du.edu/llm
[3] OpenScoring documentation at https://openscoring.du.edu/docs

3. Results

Below, we first reported the results for the DAT, followed by the AUT. Our main analyses focus on the DAT-awareness and the AUT-creative prompt conditions because they were more commonly used in former human studies (e.g., Acar, 2023; Reiter-Palmon et al., 2019). The DAT-unawareness condition and the practical and feasible prompt condition were included for exploratory purposes. For each creativity test, we focused on percentile ranks to facilitate comparisons with human performance measures. However, we report the comparisons between humans and LLMs using the untransformed scores in the online Supplementary Material on OSF, which replicate the results from the percentiles. Furthermore, we tested which LLM performed best and how stable its responses were (i.e., computed the variability within each model).

3.1. Divergent association task (DAT)

The results revealed widespread differences in DAT performance across LLMs. The average percentile was M = 60.16, *SD* = 26.15, significantly higher than the average human response (i.e., 50th percentile; for descriptive statistics, see Fig. 1 and Table 2 in the Supplementary Material), t = 14.48, p < .001. Several LLMs performed, on average, poorer than the average human (i.e., a percentile rank of <50). Some models performed better than others: Llama 3.3, Claude 3.7, and Grok beta outperformed most other models and, on average, were better than 80 % of humans. In contrast, DeepSeek R1 Distill performed only better than 22.91 % of humans. A pairwise comparison of models is shown in Figure 3 in the Supplementary Material.

In the next step, we compared our findings with those reported in the literature to assess whether LLM performance had increased. For example, Cropley (2023) reported in 2023 that GPT-4 had a percentile rank in the DAT of 82.54, *SD* = 13.94 across 102 responses. Surprisingly, an independent-samples *t*-test revealed that GPT-4o performed worse at the end of February 2025 (i.e., in our data) than GPT-4 in Cropley's 2023 data, *t*(200) = 14.46, p < .001, d = 2.04. Further, Hubert et al. (2024) reported that GPT-4 had a score (no percentile ranks were reported) of *M* = 84.56, *SD* = 3.05, across 151 responses, also in 2023. This was again higher than the scores we found in February 2025 (M = 77.34, *SD* = 2.92; *t*(249) = 18.84, p < .001, d = 2.41).

However, the variability within the responses was substantial for almost all LLMs (cf. Fig. 1). Across all 14 LLMs, 495 tests were below the 50th percentile and 894 were above the 50th percentile, *range* = 0.16, 99.78. The only LLM that produced consistent responses above the average human (i.e., >50th percentile) was Llama. Even for LLMs that are considered powerful such as ChatGPT-4.5, 6 responses were below the 50th percentile, although 94 were above the 50th percentile (M = 74.58, *SD* = 13.45, Table 2 in the Supplementary Material). Since transforming scores into percentiles can affect the distance between scores, we also compared raw scores, which replicated the findings reported in this section (see Figures 4 and 5 in the Supplementary Material). Additionally, we tested whether mentioning the DAT in the prompt mattered. We compared the 100 responses per LLM in the DAT-aware condition with the 100 responses per LLM in the DAT-unaware condition using a linear mixed-effects model (R packages lme4 and lmerTest; Kuznetsova et al. 2017), specifying random intercepts for the 14 LLMs. On average, the LLMs in the aware condition scored higher (M = 60.16, *SD* = 26.15) than in the unaware condition (M = 52.56, *SD* = 25.38, B = 11.89, *SE* = 0.90, p < .001). Interestingly, exploratory follow-up analyses revealed that this effect differed between LLMs: Claude 3.5 and Grok 2 performed much better in the DAT-aware condition (Cohen's ds > 1.00), whereas DeepSeekR1 Distill Qwen 7B performed significantly worse, with most other models performing somewhat better in the aware condition (cf. Table 2 in the Supplementary Material). We only interpreted findings that are significant at $\alpha = 0.005$ to adjust for the 14 comparisons (for visualizations and pairwise comparisons between the

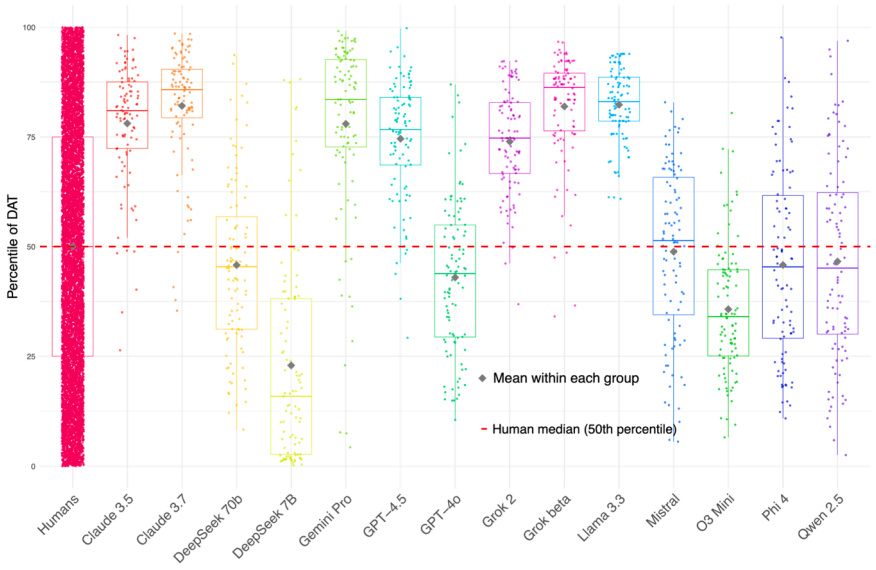

Fig. 1. Percentile scores of each large language model (LLM) in the DAT-awareness condition. The first group reflects the distribution of human percentile ranks.
Note. Percentiles for the 14 LLMs were computed by benchmarking their DAT scores against the human distribution, such that higher percentiles indicate better performance relative to humans Each boxplot displays the distribution of percentiles for a given LLM, with black diamonds indicating mean performance. The red dashed line represents the average human performance (50th percentile). The human responses are from Olson et al. Olson et al. (2021).

models, see Figures 6, 7, 8, and 9 in the Supplementary Material).

3.2. Alternate use task (AUT)

Again, we focused on the percentiles to facilitate comparisons with human responses. A one-way between-subjects ANOVA with 14 levels was significant, $F(13, 356.64) = 9.34$, $p < .001$. To control for multiple comparisons, we applied the Holm correction (Holm, 1979), a stepwise procedure that adjusts p-values to control the family-wise error rate while maintaining greater statistical power than the Bonferroni method. Holm-adjusted follow-up tests revealed that several of the models performed differently from each other, with GPT-4o performing overall best and Gemini Pro relatively worst (Fig. 2; for pairwise comparisons between the 14 LLMs and human responses, see Figure 10 in the Supplementary Material). Overall, the average model performance was more homogeneous than for the DAT, with all means ranging between 65.66 and 77.85 (Figure 3 in the Supplementary Material). A series of one-sample t-tests revealed that each model performed significantly better than the average human (i.e., percentile of 50), $ps < 0.001$.

In the next step, we compared our findings with those reported in the literature to test whether the performance of the LLMs had increased. For example, Haase and Hanel (2023) found that the average originality score generated by GPT-4 in March 2023 for the prompts *pants, ball, tire, fork,* and *toothbrush* was $M = 3.22$, $SD = 0.36$, as analyzed with OpenScoring (Organisciak et al., 2023). This was not significantly different from the average in our sample, $M = 3.67$, $SD = 0.27$, $t(7.44) = 2.23$, p =

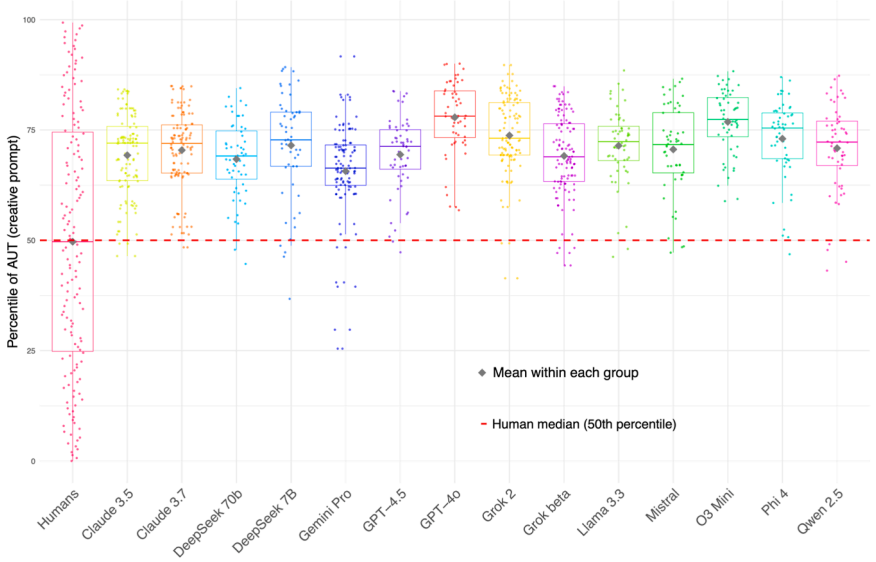

Fig. 2. Percentile scores of each large language model (LLM) in the creative-prompt condition (AUT). The first group reflects the distribution of human percentile ranks.
Note. Percentiles for the 14 LLMs were computed by benchmarking their AUT scores against the human distribution, such that higher percentiles indicate better performance relative to humans. Each boxplot displays the distribution of percentiles for a given LLM, with black diamonds indicating mean performance. The red dashed line represents the average human performance (50th percentile). The human responses are from Hubert et al. Hubert et al. (2024).

.059, d = 1.41.

We again found heterogeneity in the responses. The best response in each model was 25 percentile points (i.e., one quartile) higher than the model's worst response (Fig. 2). Somewhat surprisingly, most responses were below the top 10 % of human-generated responses. Only 3 out of 1061 responses, 0.28 %, were in the top 10 %. Additionally, we tested whether prompting the LLMs to be more creative (vs practical and feasible) would result in more original responses using again a linear mixed-effects model with random intercepts across the 14 LLMs. On average, LLMs in the creative condition scored higher (M = 70.85, *SD* = 10.00) than LLMs in the practical condition (M = 63.49, *SD* = 12.45), B = 7.89, *SE* = 0.51, p < .001. Exploratory follow-up analyses revealed that this effect was mostly consistent across LLMs: All LLMs performed better when instructed to be creative (vs practical), even though this effect was not significant for Mistral (Figures 11 and 12 for visualizations of the scores from the practical prompt condition in the Supplementary Material).

4. Discussion

In the present study, we tested whether LLMs have become more creative over time across two commonly used creativity tests, assessed the variability of their responses, and examined performance differences across 14 widely used models. Regarding the first research question, we found that GPT-4o previously benchmarked in 2023 as GPT-4—performed substantially worse on the Divergent Association Task (DAT) but retained its performance on the Alternative Uses Task (AUT). Even for the AUT, however, only 0.28 % of responses reached the 90th percentile. In other words, highly creative responses remain rare, and humans are still approximately 35.7 times more likely to produce such standout ideas.

This finding offers one possible explanation for the increasingly documented trend toward homogenization in LLM-assisted output (Anderson et al., 2024; Doshi & Hauser, 2024; Moon et al., 2024). While LLMs may generate text that appears individually novel, they often lack the type of originality required to break into the top decile of human creativity.

Interestingly, we found substantial differences between models in the DAT, whereas performance on the AUT was, on average, higher and more consistent across models. One explanation may be that the AUT, being widely used in GenAI research, is overrepresented in training data or has influenced model optimization, leading to inflated and stable performance. The DAT, by contrast, is less common and structurally more demanding: it requires generating a fixed number of semantically distant words, a task that combines lexical control with abstract association. This may be harder for some models to interpret, especially given sparse prompt cues and their tendency to favor locally coherent outputs. Further, such constraints may be less compatible with the default autoregressive generation mode of LLMs, particularly if the prompt does not provide sufficient context or examples. Recent evidence suggests that LLMs can struggle with tasks that require generation under sparse or underspecified instructions (Petrov et al., 2025).

We further observed substantial within-model variability, especially in DAT performance. Even the same model under the same conditions often produced responses ranging from below-average to exceptional. This intra-model instability complicates the evaluation of LLM creativity. While prior studies often relied on single-shot assessments or few-shot comparisons, our findings suggest that such designs may over- or underestimate the actual creative potential of these systems. The variability we observed exceeded that previously reported in LLM evaluations focused on structured or closed-ended tasks (e.g., logical reasoning, legal analysis, see Blair-Stanek & Van Durme, 2025; Liu et al., 2024), likely because DT tasks invite broader exploration and lower constraints by design. One plausible explanation for this trend may lie in recent optimization efforts to reduce so-called "AI hallucinations"—outputs that deviate from factual accuracy but often reflect greater associative freedom. While such constraints improve reliability, they may simultaneously curb generative flexibility and, thus, creative diversity across newer models.

This variability carries important implications for both creativity research and practical applications. It may help explain the inconsistent findings in studies exploring human-AI co-creativity, where collaboration with LLMs sometimes enhances and sometimes inhibits creative performance (cf. Table 4 in the Supplementary Material; Taheri et al., 2024; Vaccaro et al., 2024). As our results show, creativity outcomes are highly sensitive to model choice, prompt framing, and stochastic model behavior. Without accounting for these factors, conclusions about LLM effectiveness in creative domains may be premature or misleading.

Prompt design emerged as another significant modulator of performance. Merely disclosing the creativity test context (e.g., mentioning the DAT) influenced LLM performance in model-specific ways—improving results for some models (e.g., Claude 3.5 and Grok 2), while worsening performance for DeepSeek R1 Distill Qwen 7B This aligns with recent findings showing that LLMs are sensitive to goal framing and task specification (Memmert et al., 2024a), and echoing human creativity research on priming effects, where some individuals perform better when prompted to "be creative" (Acar et al., 2020; Sassenberg & Moskowitz, 2005). The implication is that creativity in LLMs may, in part, be prompt-contingent—a result of interaction dynamics rather than an inherent capacity.

Our findings also speak to the larger philosophical debate on *artificial creativity*. Critics argue that LLMs merely remix existing data, lacking the emotional depth, intentionality, or conceptual leaps characteristic of human creativity (Cropley & Cropley, 2023; Runco, 2023). Indeed, the absence of high-end originality in LLM output could be taken as support for this view. However, we caution against such binary thinking. While LLMs may not engage in creative processes in the human sense, their ability to generate outputs that score above the average human in both semantic divergence and usefulness indicates a form of functional or output-based creativity. Beyond all critical analysis discussed so far, 80 % of LLM's AUT-output is on average better than that of humans—when specifically picking a "creative LLM", like Claude, Gemini, GPT-4o, or Grok, one can expect original output.

What these results lead to, especially the variability, is the importance of actively working with and reflecting upon the output generated by LLMs. Thus, our results contribute to the growing literature on LLMs in human-AI co-creativity. While LLMs clearly offer utility as brainstorming aids or creative scaffolds, their performance is not reliably high nor uniformly distributed across contexts. Our findings resonate with broader research on human–AI collaboration, showing that GenAI augments ideation and productivity but tends to yield convergent, mid-novelty outputs. The most effective creative workflows are human-led and iterative, where the human defines the problem and constraints, and the LLM supplies fluent associative variations. When supported by thoughtful design—preserving human control, transparency, and exploration—these teaming setups can enhance both output quality and learning effects, enabling users to refine their problem framing and idea evaluation over time (Doshi & Hauser, 2024; Haase & Pokutta, 2024; Zhou & Lee, 2024).

The most promising use cases may lie not in full automation but in human-AI teaming, where the human provides context and framing, and the LLM contributes fluent, varied, or unexpected content. Especially the LLMs' capabilities to be extremely fast, easy to access, and very fluent in output length make them very efficient co-creators. Still, as our findings show, such systems may encourage mid-level novelty but rarely produce radically original ideas—thus reinforcing combinatorial rather than conceptual creativity (Orrù et al., 2023; Soroush et al., 2025). Without thoughtful human oversight and critical engagement, GenAI may unintentionally constrain creative diversity and reinforce existing patterns, rather than expanding the overall human-AI-enhanced creative process.

4.1. Limitations and ethical considerations

This study has several limitations. First, we used the default settings provided by each LLM API, including temperature and sampling parameters. While this reflects common usage patterns and hence enhances ecological validity, these parameters are known to influence output variability and originality (Peeperkorn et al., 2024). Future research should systematically vary such parameters to explore their effects on DT performance.

Second, we focused exclusively on English-language tasks and did not investigate multilingual capabilities. Given the global reach of GenAI tools and recent evidence suggesting that LLMs may behave differently across languages (Zhang et al., 2023), this remains an important area for future exploration. Third, while we evaluated a wide range of models on standardized DT tasks, we did not address more complex aspects of creativity such as problem finding, iterative elaboration, or social validation—dimensions often considered central in high-level creative work (Amabile, 2017; Simonton, 2013). As such, our findings should be interpreted as reflecting creative potential rather than holistic creative capacity.

Ethical considerations also arise from our findings. On the one hand, the ability of LLMs to generate original content that rivals or exceeds average human output suggests a democratizing potential: these tools can help individuals with limited experience engage in creative expression. On the other hand, the proliferation of AI-generated content raises concerns about authenticity, plagiarism, and the dilution of originality in digital culture. Repeated exposure to AI-generated ideas may subtly influence human ideation, narrowing the perceived range of what is creative or acceptable—potentially reinforcing the very homogenization effects observed in this study and elsewhere (Doshi & Hauser, 2024; Toma & Yánez-Pérez, 2024). Moreover, while exposure to others' ideas has been shown to benefit human creativity (Fink et al., 2009), the role of GenAI in this dynamic is more complex. Unlike human peers, LLMs generate content that is simultaneously vast in volume and narrow in conceptual distribution. This tension necessitates thoughtful integration of GenAI into educational, artistic, and professional workflows, ideally with mechanisms that encourage critical thinking and safeguard human agency.

As GenAI continues to reshape creative practice, the importance of maintaining meaningful human oversight cannot be overstated. Responsible use requires not only technological literacy but also ethical awareness and reflective practices that ensure AI augments rather than undermines human creative expression.

Acknowledgments

We would like to thank the Zuse Institute Berlin for hosting various LLM models for testing and Peter Organisciak for providing us with an API key for the OpenScoring API. We would also like to thank the authors of Olson et al. (2021) and Organisciak et al. (2023) for making their work and codes, including their scoring sites, publicly available.

Data availability

The datasets and R code supporting the findings of this study are available at the Open Science Framework: https://osf.io/e62fx/.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the Cluster of Excellence MATH+ (grant number EXC-2046/1, project ID 390685689) and by the Federal Ministry of Research, Technology and Space under the grant 16DII133, founded by the German Federal Ministry of Education and Research in 2017.

CRediT authorship contribution statement

Jennifer Haase: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Paul H.P. Hanel:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Sebastian Pokutta:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.yjoc.2025.100113.

References

Acar, S. (2023). Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research Journal, 0*(0), 1–7.

Acar, S., Runco, M. A., & Park, H. (2020). What should people be told when they take a divergent thinking test? A meta-analytic review of explicit instructions for divergent thinking. In *Psychology of aesthetics, creativity, and the arts, 14* pp. 39–49). Place: US Publisher: Educational Publishing Foundation.

Amabile, T. M. (2017). In pursuit of everyday creativity. *The Journal of Creative Behavior, 51*(4), 335–337. _eprint: Https://onlinelibrary.wiley.com/doi/pdf/10.1002/jocb.200.

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on Human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition, C&C '24* (pp. 413–425). Association for Computing Machinery.

Benedek, M., Bruckdorfer, R., & Jauk, E. (2020). Motives for creativity: Exploring the what and why of everyday creativity. *The Journal of Creative Behavior, 54*(3), 610–625.

Blair-Stanek, A. and Van Durme, B. (2025). Llms provide unstable answers to legal questions. *arXiv* preprint arXiv:2502.05196.

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence, 103*(1), 347–356.

Boden, M. A. (2009). Computer models of creativity. *AI Magazine, 30*(3), 23–23. Number: 3.

Botella, M., Zenasni, F., & Lubart, T. (2018). What are the stages of the creative process? What visual art students are saying. *Frontiers in Psychology*, 9.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol., 15*(3), 39:1–39:45.

Christensen, P., Guilford, J., Merrifield, R., & Wilson, R. (1960). Alternate uses test. *Beverly Hills, CA: Sheridan Psychological Service*.

Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M. J. (2022). A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications, 209*, Article 118-190.

Cromwell, J., Haase, J., & Vladova, G. (2022). The creative thinking Profile: Measuring internal preferences for multiple creative thinking styles. *Academy of Management Proceedings*, (1), 2022.

Cromwell, J., Harvey, J.-F., Haase, J., & Gardner, H. K. (2023). Discovering where ChatGPT can create value for your company. *Harvard Business Review*.

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal, 18*(3), 391–404.

Cropley, D. (2023). Is artificial intelligence more creative than humans?: ChatGPT and the Divergent Association Task. *Learning Letters, 2*, 13–13.

Cropley, D., & Cropley, A. (2023). Creativity and the Cyber Shock: The Ultimate Paradox. *The Journal of Creative Behavior*, 57(4), 1-3.

da Costa, S., Páez, D., Sánchez, F., Garaigordobil, M., & Gondim, S. (2015). Personal factors of creativity: A second order meta-analysis. *Revista de Psicología del Trabajo y de las Organizaciones, 31*(3), 165–173.

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., & Kohli, P. (2021). Advancing mathematics by guiding human intuition with ai. *Nature, 600*(7887), 70–74.

DiPaola, S. and McCaig, G. (2016). Using artificial intelligence techniques to emulate the creativity of a portrait painter. In *Electronic visualisation and the arts (EVA)*. BCS Learning & Development.

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), 1–9.

Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature, 613*(424).

Fink, A., Grabner, R. H., Benedek, M., Reishofer, G., Hauswirth, V., Fally, M., Neuper, C., Ebner, F., & Neubauer, A. C. (2009). The creative brain: Investigation of brain activity during creative problem solving by means of EEG and FMRI. *Human Brain Mapping, 30*(3), 734–748.

Gaessler, F., & Piezunka, H. (2023). Training with AI: Evidence from chess computers. *Strategic Manage- ment Journal, 44*(11), 2724–2750. _eprint: Https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.3512.

Gilhooly, K. (2023). AI vs humans in the AUT: Simulations to LLMs. *Journal of Creativity, 34*(1), 1–5.

Gosling, S. D., Ybarra, K., & Angulo, S. K. (2024). A widely used generative-ai detector yields zero false positives. *Aloma: Revista de Psicologia. Ciències de l'Educació i de l'Esport, 42*(2), 31–43.

Guzik, E. E., Byrge, C., & Gilde, C. (2023). The originality of machines: AI takes the Torrance test. *Journal of Creativity, 33*(3).

Gómez-Rodríguez, C., & Williams, P. (2023). *A Confederacy of Models: A Comprehensive Evaluation of LLMs on Creative Writing*. arXiv:2310.08433 [cs].

Haase, J., Djurica, D., & Mendling, J. (2023). The art of inspiring creativity: Exploring the unique impact of AI-generated images. In *AMCIS 2023 Proceedings*.

Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity, 33*(3), 1–7.

Haase, J., Hanel, P. H. P., & Pokutta, S. (2025). S-DAT: A Multilingual, GenAI-Driven Framework for Automated Divergent Thinking Assessment. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(2), 1194–1205.

Haase, J., & Pokutta, S. (2024). *Human-AI Co-Creativity: Exploring Synergies Across Levels of Creative Collaboration*. arXiv:2411.12527.

Hadas, E. (2025). Assessing creativity across multi-step intervention using generative AI models. *Journal of Learning Analytics, 12*(1), 91–109.

Holford, W. D. (2019). The future of human creative knowledge work within the digital economy. *Futures, 105*, 143–154.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports, 14*(1), 3440.

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). *The platonic representation hypothesis*. arXiv preprint arXiv:2405.07987.

Kaufman, J. C., Beghetto, R. A., & Watson, C. (2016). Creative metacognition and self-ratings of creative performance: A 4-C perspective. *Learning and Individual Differences, 51*, 394–399.

Kirkpatrick, K. (2023). Can AI demonstrate creativity? *Communications of the ACM, 66*(2), 21–23.

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences, 118*(22), Article E2018340118. Publisher: Proceedings of the National Academy of Sciences.

Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. In *Scientific Reports, 13*. Number: 1 Publisher: Nature Publishing Group, Article 13601.

Krakowski, S., Luger, J., & Raisch, S. (2023). Artificial intelligence and the changing sources of competitive advantage. *Strategic Management Journal, 44*(6), 1425–1452. _eprint: Https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.3387.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26.

Lan, M., Torr, P., Meek, A., Khakzar, A., Krueger, D., & Barez, F. (2025). *Sparse Autoencoders Reveal Universal Feature Spaces Across Large Language Models*. arXiv:2410.06981 [cs].

Liu, J., Liu, H., Xiao, L., Wang, Z., Liu, K., Gao, S., Zhang, W., Zhang, S., & Chen, K. (2024). *Are your llms capable of stable reasoning?*. arXiv preprint arXiv:2412.13147.

Marcus, G., Davis, E., & Aaronson, S. (2022). *A very preliminary analysis of DALL-E 2*. arXiv:2204.13807 [cs].

Memmert, L., Cvetkovic, I., and Bittner, E. (2024a). The more is not the merrier: Effects of prompt engineering on the quality of ideas generated by GPT-3. In *HICSS2024*, Honolulu, Hawaii.

Memmert, L., Mies, J., & Bittner, E. (2024b). Brainstorming with a generative language model: The role of creative ability and tool-support for Brainstorming performance. In *ICIS 2024 Proceedings*.

Miller, A. I. (2019). *The artist in the machine: The world of AI-powered creativity*. MIT Press.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of what art? A call for Multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics, 12*, 933–949.

Moon, K., Green, A., & Kushlev, K. (2024). Homogenizing Effect of a Large Language Model (LLM) on Creative Diversity: An Empirical Comparison of Human and ChatGPT Writing. 10.31234/osf.io/8p9wu.

Mundinger, K., Pokutta, S., Spiegel, C., & Zimmer, M. (2024). Extending the continuum of six- colorings. Geombinatorics Quarterly, arXiv preprint arXiv:2404.05509.

Mundinger, K., Zimmer, M., Kiem, A., Spiegel, C., & Pokutta, S. (2025). *Neural Discovery in Mathematics: Do Machines Dream of Colored Planes?*. arXiv preprint arXiv:2501.18527.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences, 118*(25). E2022340118.

OpenAI (2025). Introducing GPT-4.5.

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*, Article 101356.

Organisciak, P., Dumas, D., Acar, S., & de Chantal, P. L. (2025). *Open Creativity Scoring Computer software*.

Orrù, G., Piarulli, A., Conversano, C., & Gemignani, A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence, 6*.

Pandya, V. (2024). The age of generative AI: Over half of Americans have used generative AI and most believe it will help them be more creative | Adobe Blog.

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. (2024). Is Temperature the Creativity Parameter of Large Language Models? arXiv:2405.00492 [cs].

Petrov, I., Dekoninck, J., Baltadzhiev, L., Drencheva, M., Minchev, K., Balunović, M., Jovanović, N., & Vechev, M. (2025). *Proof or bluff? evaluating llms on 2025 usa math olympiad*. arXiv preprint arXiv:2503.21934.

Plucker, J. A. (2004). Generalization of creativity across domains: Examination of the method effect hypothesis. *The Journal of Creative Behavior, 38*(1), 1–12.

Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour, 7*(11), 1836–1838.

Raz, T., Reiter-Palmon, R., & Kenett, Y. N. (2023). The role of asking more complex questions in creative thinking. *Psychology of aesthetics, creativity, and the arts*. Educational Publishing Foundation. pages No Pagination Specified–No Pagination Specified. Place: US Publisher:.

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.

Renze, M. (2024). The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: Emnlp 2024*, pages 7346–7356.

Runco, & Jaeger, G. J (2012). The Standard Definition of creativity. *Creativity Research Journal, 24*(1), 92–96.

Runco, M. A. (2023). AI can only produce artificial creativity. *Journal of Creativity, 33*(3), 1–7.

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66–75.

Runco, M. A., Turkman, B., Acar, S., & Abdulla Alabbasi, A. M. (2025). Examining the idea density and semantic distance of responses given by AI to tests of divergent thinking. *The Journal of Creative Behavior, 59*(3), e1528. https://doi.org/10.1002/jocb.1528

Salinas, A. and Morstatter, F. (2024). The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. arXiv preprint arXiv:2401.03729.

Sassenberg, K., & Moskowitz, G. B. (2005). Don't stereotype, think different! overcoming automatic stereotype activation by mindset priming. *Journal of Experimental Social Psychology, 41*(5), 506–514.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*(2), 68–85.

Simonton, D. K. (2013). What is a creative idea? Little-c versus big-C creativity. *Handbook of research on creativity, 2*, 69–83.

Sinha, R., Song, Z., and Zhou, T. (2023). A Mathematical Abstraction for Balancing the Trade-off Between Creativity and Reality in Large Language Models. arXiv:2306.02295 [cs].

Soroush, M. Z., Sourav, M. S. U., & Zeng, Y. (2025). Creativity and AI. editor. In S. C. Suh (Ed.), *Artificial intelligence for design and process science* (pp. 29–43). Cham: Springer Nature Switzerland. pages.

Steele, L. M., Johnson, G., & Medeiros, K. E. (2018). Looking beyond the generation of creative ideas: Confidence in evaluating ideas predicts creative outcomes. *Personality and Individual Differences, 125*, 21–29.

Swirszcz, G., Wagner, A.Z., Williamson, G., Blackwell, S., Georgiev, B., Davies, A., Eslami, A., Racaniere, S., Weber, T., and Kohli, P. (2025). Advancing geometry with ai: Multi-agent generation of polytopes. *preprint*.

Taheri, A., Khatiri, S., Seyyedzadeh, A., Ghorbandaei Pour, A., Siamy, A., & Meghdari, A. F. (2024). Investigating the impact of Human-robot collaboration on creativity and team efficiency: A case study on brainstorming in presence of robots. In A. A. Ali, J.-J. Cabibihan, N. Meskin, S. Rossi, W. Jiang, H. He, & S. S. Ge (Eds.), *Social robotics* (pp. 94–103). Singapore: Springer Nature. Lecture Notes in Computer Science, pages.

Toma, R. B., & Yánez-Pérez, I. (2024). Effects of ChatGPT use on undergraduate students' creativity: A threat to creative thinking? *Discover Artificial Intelligence, 4*(1), 74.

Torrance, E. P. (1972). Can we teach children to think creatively? *The Journal of Creative Behavior, 6*(2), 114–143.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*(12), 2293–2303.

Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B., & Lu, Z. (2024). "It felt like having a second mind": Investigating Human-AI Co-creativity in prewriting with large language models. *Proc. ACM Hum.-Comput. Interact, 8*(CSCW1), 84:1–84:26.

Wenger, E. and Kenett, Y. (2025). We're Different, We're the Same: Creative Homogeneity Across LLMs. arXiv:2501.19361 [cs].

White, C. (2023). Opinion: Artificial intelligence can't reproduce the wonders of original human creativity. *The Star*.

Williams, R., Runco, M., & Berlow, E. (2016). Mapping the themes, impact, and cohesion of creativity research over the last 25 years. *Creativity Research Journal, 28*, 385–394.

xAI (2025). Grok 3 beta — The age of reasoning agents.

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., and Carroll, J.M. (2024). Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis.
Zhang, X., Li, S., Hauer, B., Shi, N., and Kondrak, G. (2023). Don't Trust ChatGPT when Your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. arXiv:2305.16339 [cs].
Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus, 3*(3), 1–8.