AI is Misled by GenAI: Stylistic Bias in Automated Assessment of Creativity in Large Language Models

Marek Urban[*]
Institute of Psychology
Czech Academy of Sciences
urban@praha.psu.cas.cz

Petra Kmoníčková
Institute of Psychology
Czech Academy of Sciences

Kamila Urban
Institute for Research in Social Communication
Slovak Academy of Sciences

Abstract

Outputs from large language models (LLMs) are often rated as highly original yet show low variability (or greater homogeneity) compared to human responses, a pattern we refer to as the *LLM creativity paradox*. Yet, prior work suggests that assessments of originality and variability may reflect stylistic features of LLM outputs rather than underlying conceptual novelty. The goal of the present study was to investigate this issue using outputs from seven distinct LLMs on a modified Alternative Uses Task. We scored verbatim and "humanized" LLM responses—reworded to reduce verbosity but maintain core ideas—using four automated metrics (supervised OCSAI and CLAUS models, and two unsupervised semantic-distance tools) and compared them with responses from 30 human participants. As expected, verbatim LLM responses were rated as substantially more original than human responses (median $d = 1.46$) but showed markedly lower variability (median $d = 0.85$). Humanizing the responses strongly decreased originality and weakly increased variability, indicating that part of the LLM creativity paradox is driven by stylistic cues. Nevertheless, even after humanization, originality scores of LLM responses remained higher (median $d = 0.80$) and their variability lower ($d = 0.57$) than those of human responses. These findings suggest that automated assessment tools can be partially misled by the style of LLM outputs, highlighting the need for caution when using automated methods to evaluate machine-generated ideas, particularly in real-world applications such as providing feedback or guiding creative workflows.

1 Introduction

Large language models (LLMs) have shown remarkable performance on various creativity tasks, often matching or even surpassing average human originality scores (Haase and Hanel 2023, Haase et al. 2025, Hubert et al. 2024). However, a paradoxical trend has emerged in which individual LLM-generated ideas may seem novel, yet as a group they are much more alike than ideas produced by humans (Wenger and Kenett 2025). Recent studies across domains from story writing to scientific research found that LLM-aided outputs converge on a narrower range of ideas than the diverse solutions generated by humans (Anderson et al. 2024, Doshi and Hauser 2024, Moon et al. 2025, Si et al. 2024, Wenger and Kenett 2025). These findings have raised concerns about "generative

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

monocultures" (Wu et al. 2024), where widespread reliance on LLMs could lead users to produce variations of the same few ideas, slowing scientific progress or limiting innovation (Messeri and Crockett 2024). However, Hubert et al. (2024) suggest that part of this paradox may arise from the style of LLMs outputs rather than the underlying ideas. Specifically, LLMs tend to use more abstract and non-tangible words whereas humans often focus on concrete and observable ideas. This raises the question of whether the high originality and low variability observed in LLM outputs reflect truly essence of the ideas or merely the way these ideas are expressed. This issue is particularly important when creativity is assessed using automated AI-based tools trained on human responses.

1.1 Automated Creativity Assessment Methods

In recent years, several automated methods have been developed to evaluate originality of answers, partly to overcome the subjective and labor-intensive nature of human scoring (Beaty and Johnson 2021). The unsupervised approach estimates originality of the answers through semantic distance, based on the assumption that ideas are more original when they are semantically farther from common responses or a prompt. Using word embedding models (e.g., GloVe), these methods quantify how remote a response is (e.g., the distance between *paperclip* and *lockpick* is smaller than that between *paperclip* and *tiny catapult for sunflower seeds*). The semantic-distance scores moderately to strongly correlate with human originality ratings (Dumas et al. 2021, Patterson et al. 2023).

More recently, supervised models have been developed that are trained on large sets of human-rated responses to learn how humans score creativity. Open Creativity Scoring with AI (OCSAI; Organisciak et al. 2023) uses a fine tuned transformer (v1.6 built on GPT 4o mini) trained on over 27,000 human rated responses to the Alternate Uses Task. Cross Lingual Alternate Uses Scoring (CLAUS; Patterson et al. 2025) applies a multilingual XLM-RoBERTa model trained on creativity responses in multiple languages. Both approaches achieve high agreement with human judges, with correlations up to $r = .81$ with human originality ratings. Organisciak et al. (2023) even reported that fine-tuned LLM scorers can match or exceed the reliability of human raters, and Saretzki and Benedek (2025) showed that their scores can outperform human ratings in predictive validity. Yet, when these methods are applied to evaluate creativity in LLM-generated responses, important questions arise about what exactly they are measuring.

1.2 Present study

The present study examines whether automated creativity assessments capture the underlying originality of ideas or are biased by stylistic features of responses. Researchers noted that LLM-generated answers often have a distinctive style, tending to be longer and more fluent (Wenger and Kenett 2025). Verbose responses suggest an *elaboration bias*, i.e., an elaborate response may receive a higher originality rating even if the core idea is fairly ordinary. Conversely, the longer responses may use more similar words and therefore exhibit higher internal similarity (Wenger and Kenett 2025). Furthermore, Hubert et al. (2024) observed that LLM responses include more abstract or non-tangible words (e.g., freedom or philosophy), whereas human responses are often more concrete. This could inflate originality scores in favor of AI simply because the chosen words are rarer.

In light of these gaps, the present study examines whether automated creativity scoring reflects the novelty of ideas or simply the style in which LLMs present them. We first aim to replicate what we call the "LLM creativity paradox," i.e., the tendency of LLM responses to score as highly original yet overly uniform. We then directly test the influence of style by "translating" LLM-generated answers into a more human-like form, preserving their core meaning but reducing verbosity and polished phrasing. Comparing automated scores for the verbatim and "humanized" versions allows us to assess whether AI scoring models respond to conceptual originality or are swayed by surface-level features. In doing so, we evaluate widely used creativity metrics—including semantic distance measures (Dumas et al. 2021, Patterson et al. 2023), supervised scoring models (OCSAI, Organisciak et al. 2023; CLAUS, Patterson et al. 2025), and cosine distance as a measure of variability (Wenger and Kenett 2025)—to determine whether they capture the essence of ideas in LLM outputs or are biased by the eloquence of LLM-generated language.

2 Methods

2.1 Sample

Human participants. A priori sample size calculation was based on Haase et al. (2025), who found differences between LLM-generated and human outputs ranging from $d = 1.33$ to 3.67. A power analysis assuming the smallest of these effects ($d = 1.33$), $\alpha = .05$, and desired power $(1 - \beta) = .80$ indicated that only 10 participants per group were required. We therefore recruited 30 participants to provide a conservative buffer and to ensure that the sample size would also satisfy normality assumptions for the planned statistical analyses. Thirty U.S. university students (16 men, 12 women, 2 other; $M_{\text{age}} = 22.6$ years, $SD = 2.6$) were recruited on Prolific. Prescreen filters required that volunteers are at least 18 years old, that they are enrolled in a college or university program, are native English speakers, and report no learning disability. There were 6 first-year, 7 second-year, 4 third-year, 4 fourth-year bachelor's students, and 9 MA students. Race/ethnicity was 22 White/Caucasian, 5 Asian/Pacific Islander, and 3 Black or African American.

Participants completed the anonymous study online via SurveyMonkey as part of a larger project. The session lasted on average 13.6 min, and participants were compensated £4.00. Two attention checks were embedded in the survey; all participants passed and were retained. The study was approved by the Institutional Review Board of the third author's institution, and informed consent was obtained electronically before participation.

Large language models (LLMs). Outputs were generated from seven contemporary large language models (LLMs), each accessed through the interfaces provided by their respective developers. Three models were obtained from OpenAI via the Playground interface: ChatGPT-4o-latest (version 20241120), GPT-o1 (version 20241120), and GPT-o3 (version 20250703). Three models were obtained from Google through its AI Studio platform: Gemini 2.0 "Flash-Thinking" (version 20241120), Gemini 2.5 Pro (version 20250703), and Gemini 2.5 "Flash-Thinking" (version 20250703). One additional model, DeepSeek r1 (version 20250703), was accessed through the standard chat interface.

All models were queried thirty times using identical prompt. Default generation settings were used for all responses, no parameters were manually altered. This approach ensured that the outputs reflected the standard behavior of each model rather than being influenced by user-defined tuning. For reasoning models, the chain of thought was enabled.

2.2 Stimuli

Both human participants and LLMs completed a modified version of the Alternative Uses Task (AUT). Participants were instructed to generate only one use for an ordinary paperclip, focusing on producing the most original idea. The prompt was as follows:

> Imagine an ordinary paperclip. Paperclips are commonly used to hold papers together, but they can have many other uses as well. Some of these uses are typical, but some uses can be completely original—uses that are novel, surprising, and different from anything we usually associate with a paperclip. Your task is to think of only one use for a paperclip. However, this use must be as original as possible. Use your imagination to come up with the most original use for a paperclip that you can think of.

The single response design emphasizes not only divergent (the ability to generate large number of ideas) but also a convergent thinking (selecting the single most original idea). This hybrid format was chosen because LLMs can produce practically unlimited number of ideas in one session (Hubert et al. 2024, Wenger and Kenett 2025). This approach makes differences between humans and LLMs easier to interpret because it aligns human inter-individual variation with the inter-session variation of LLMs.

In addition to the verbatim responses, we created a set of "humanized" responses. These were produced by the second author, who rewrote each LLM-generated response in randomized order. At the time of this task, the second author was unaware of the specific purpose of the procedure. The goal was to preserve the core idea expressed in each original response while rephrasing it in a more human-like manner, using more concrete, conversational language and reducing unnecessary verbosity or stylistic embellishment typical of LLM outputs.

2.3 Analytical procedure

Automated originality scoring. All responses were scored for originality using four automated tools accessed on July 14, 2025: CAP SemDis (Patterson et al. 2023; https://cap.ist.psu.edu/), CAP CLAUS (Patterson et al. 2025; https://cap.ist.psu.edu/claus), Semantic Distance Scoring (Dumas et al. 2021; https://openscoring.du.edu/scoring), and OCSAI (Organisciak et al. 2023; https://openscoring.du.edu/scoringllm). SemDis and Semantic Distance Scoring estimate originality from semantic distance and return scores between 0 and 1, whereas CLAUS and OCSAI are supervised models trained on large sets of human originality ratings and return scores from 0 to 1 (CLAUS) or 1 to 5 (OCSAI). In all cases, lower values indicate lower originality.

Variability of responses. To quantify how much the ideas within each population (humans versus the various LLMs) differed, we reproduced Wenger and Kenett (2025) cosine-distance procedure in R. For every participant or LLM session we first concatenated the single paper-clip use into one character string, then lower-cased, removed punctuation and English stop-words (package *tm*; Feinerer and Hornik 2025) and applied Snowball stemming (package *SnowballC*; Bouchet-Valat 2023). Each cleaned string was embedded with the pre-trained all-MiniLM-L6-v2 sentence-transformer accessed through the *text* package (Kjell et al. 2025). Cosine similarity was computed with *proxy* package (Meyer and Buchta 2025) and converted to cosine distance (1 – similarity). Values near 0 indicate that population contains similar ideas, whereas values approaching 1 indicate maximally distinct ideas.

The data are available on the Open Science Framework (OSF) at https://osf.io/qax9w.

3 Results

The analyses addressed two main questions: (a) whether LLM-generated responses are assessed as more original but show lower variability compared to human responses, and (b) whether this pattern changes when LLM responses are "humanized" to reduce stylistic differences.

3.1 LLM creativity paradox: high originality, low variability

We first compared the originality scores and response variability of verbatim LLM responses with those of human participants to examine whether automated scoring methods rated LLM outputs as more original yet less diverse. Originality scores are summarized in Table 1, and variability (cosine distance) values are reported in Table 2.

Table 1: Originality scores for long (verbatim) versus humanized responses

Model	OCSAI (LLM)				OCSAI (Sem. Dis.)				CAP CLAUS				CAP SemDis			
	Long		Hum.		Long		Hum.		Long		Hum.		Long		Hum.	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
human	2.98	0.68	—	—	0.85	0.06	—	—	0.57	0.08	—	—	0.43	0.06	—	—
gpt-4o	4.25	0.41	3.99	0.60	0.88	0.04	0.86	0.05	0.71	0.06	0.67	0.06	0.46	0.05	0.43	0.06
gpt-o1	4.12	0.34	3.77	0.49	0.89	0.04	0.86	0.06	0.71	0.06	0.69	0.07	0.47	0.06	0.45	0.05
gpt-o3	4.25	0.25	3.90	0.46	0.90	0.04	0.86	0.03	0.65	0.07	0.63	0.07	0.47	0.05	0.46	0.05
2.0 Flash thinking	3.92	0.42	3.58	0.60	0.88	0.07	0.88	0.09	0.65	0.07	0.62	0.07	0.48	0.06	0.44	0.05
2.5 Flash thinking	4.35	0.35	4.10	0.46	0.92	0.07	0.90	0.03	0.72	0.06	0.66	0.07	0.51	0.06	0.44	0.05
2.5 Pro	4.40	0.33	3.72	0.63	0.88	0.05	0.88	0.03	0.74	0.06	0.65	0.09	0.53	0.04	0.46	0.05
r1	4.27	0.31	3.87	0.51	0.90	0.03	0.88	0.04	0.70	0.07	0.65	0.07	0.51	0.06	0.44	0.05

Independent samples t-tests were conducted to compare human responses against LLM-generated responses, with a Holm-Bonferroni correction applied for multiple comparisons. The results indicate that the LLMs consistently outperformed human responses across all conditions. The supervised models, OCSAI and CLAUS, yielded substantially higher scores for LLM-generated answers, with a median Cohen's $d = 2.59$ $[1.92,\ 2.90]$ and 1.97 $[1.18,\ 2.45]$, respectively. Two semantic-distance models also yielded significantly higher scores for LLM-generated responses, with a median $d = 0.61$ $[0.50,\ 1.19]$ and 0.84 $[0.47,\ 1.89]$. The moderate to large effect sizes show that verbatim LLM-generated answers are scored substantially higher than human responses.

The variability analysis mirrored this pattern. Variability was moderately greater for human responses compared with those from LLMs, with a median $d = 0.85$ $[0.76,\ 1.46]$. This result indicates that although LLM answers tended to receive higher originality scores, they were semantically more clustered than human responses. All analyses can be found in Table 3.

Table 2: Variability of long (verbatim) versus humanized responses

Model	Distance			
	Long		Hum.	
	M	*SD*	*M*	*SD*
human	0.43	0.07	—	—
gpt-4o	0.34	0.10	0.38	0.09
gpt-o1	0.36	0.08	0.37	0.08
gpt-o3	0.36	0.08	0.38	0.07
2.0 Flash thinking	0.36	0.08	0.40	0.07
2.5 Flash thinking	0.34	0.07	0.39	0.08
2.5 Pro	0.31	0.09	0.34	0.09
r1	0.36	0.08	0.40	0.08

3.2 Effect of style on automated assessment of originality and variability

To examine whether humanizing the LLM-generated answers affected their originality scores, we conducted separate 2 (verbatim vs. humanized) × 7 (model) mixed-design ANOVAs for each of four originality measures: OCSAI and CLAUS, and two semantic-distance tools. Across all measures, humanized answers were judged as less original than the verbatim versions. The decrease in originality as well as comparison between LLM and human answers can be seen in Figures 1, 3, 4, and 5.

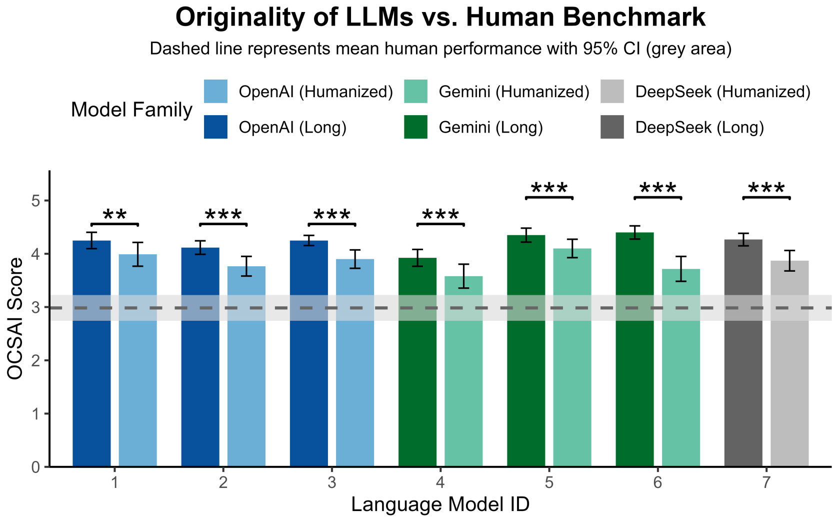

Figure 1: Originality of long and humanized responses rated by OCSAI. *Note:* Model IDs correspond to: (1) gpt-4o, (2) gpt-o1, (3) gpt-o3, (4) 2.0 Flash thinking, (5) 2.5 Flash thinking, (6) 2.5 Pro, and (7) r1.

For OCSAI, there was a strong main effect of humanization, $F(1, 203) = 132.41$, $p < .001$, partial $\eta^2 = .40$, with a significant humanization × model interaction, $F(6, 203) = 2.82$, $p = .01$, partial $\eta^2 = .08$, indicating that the drop in originality varied across models. The OCSAI semantic distance measure showed the same pattern, $F(1, 203) = 24.04$, $p < .001$, partial $\eta^2 = .11$, with no significant interaction, $F(6, 203) = 1.67$, $p = .13$, partial $\eta^2 = .05$, suggesting a fairly uniform effect across models. For the CLAUS, originality scores were again lower for humanized answers, $F(1, 203) = 79.90$, $p < .001$, partial $\eta^2 = .28$, accompanied by a significant interaction, $F(6, 203) = 3.53$, $p = .004$, partial $\eta^2 = .10$. Finally, the CAP SemDis metric showed the strong reduction, $F(1, 203) = 103.29$, $p < .001$, partial $\eta^2 = .34$, with a robust interaction, $F(6, 203) = 5.05$, $p < .001$, partial $\eta^2 = .13$. Overall, humanizing the outputs consistently lowered originality across all metrics, with some differences in effect sizes between models and measures. Yet, it is important to note that originality scores of humanized responses were still substantially higher than scores of human responses (median $d = 0.80$ $[-0.01, 2.29]$; see Table 3).

Furthermore, to examine whether humanizing the LLM-generated answers influenced their variability, we fitted a linear mixed-effects model with humanization as a fixed factor and random intercepts for the seven language models. Variability was significantly higher for humanized answers ($M =$

0.38, $SE = 0.01$, 95% CI $[0.36,\ 0.40]$) compared to verbatim answers ($M = 0.35$, $SE = 0.01$, 95% CI $[0.33,\ 0.36]$), with an estimated difference of $\Delta = 0.04$ ($b = -0.04$, $SE = 0.00$, 95% CI $[-0.04,\ -0.03]$, $t(6.08) = -16.91$, $p < .001$). The fixed effects alone explained marginal $R^2 = 4.3\%$ of the variance, while the full model that also included random effects explained conditional $R^2 = 8.9\%$. These results indicate that humanizing the responses reliably increased their variability across all models. However, as can be seen in Figure 2 and Table 3, the variability of human-generated responses was still higher (median $d = 0.57$ $[0.29,\ 1.05]$).

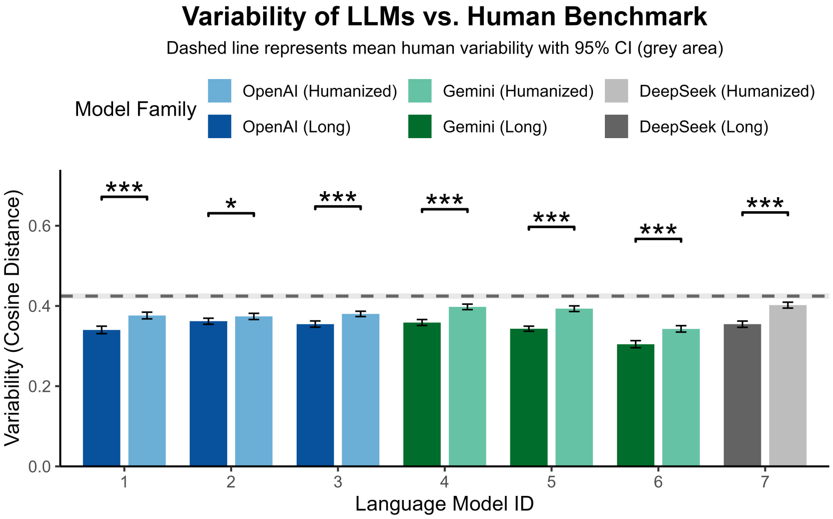

Figure 2: Variability of long and humanized responses.

4 Discussion

Our findings replicate and extend what we call the *LLM creativity paradox*: outputs from large language models (LLMs) are often scored as more original than human responses while simultaneously exhibiting lower semantic variability. Automated scorers, particularly supervised models such as OCSAI and CLAUS, consistently rated verbatim LLM outputs as more original, but when we humanized these outputs—rephrasing them to resemble typical human wording while preserving the core ideas—the measured originality decreased and variability increased. This indicates that part of the originality advantage attributed to LLMs is driven by stylistic features such as verbosity and abstract language rather than by fundamentally novel concepts (Hubert et al. 2024, Wenger and Kenett 2025). Automated scoring tools thus need to incorporate style-insensitive metrics or explicitly control for features such as concreteness or abstractness (compare Muraki et al. 2023).

Yet, our results further show that even after controlling for style, LLM outputs remain more original but still less diverse than human outputs. Given the growing reliance on LLMs for ideation, there is a risk of narrowing the collective creative search space (Darbellay et al. 2017, Messeri and Crockett 2024, Wu et al. 2024). If users adopt highly similar AI-generated solutions, innovation could be inadvertently constrained. As generative AI tools increase the self-confidence in one's abilities (compare Urban et al. 2024), inflated originality scores may give individuals false confidence that their AI supported pipelines foster creativity when, in fact, they promote uniformity.

Limitations and conclusion. The present study has several limitations. Although a priori sample size calculation was performed for statistical testing, the number of responses may not have been sufficient to fully capture the true level of semantic similarity. Furthermore, only a single task was used, which may limit the generalizability of the findings. Finally, the humanization of LLM responses was performed by a single person, which may introduce subjective bias. Future research should therefore examine whether these patterns generalize across diverse tasks and prompts and develop automated metrics that more effectively distinguish genuine conceptual novelty from stylistic cues. Such refinements will be essential for integrating LLMs responsibly into creative workflows and educational or feedback systems (compare de Chantal et al. 2025).

Acknowledgments and Disclosure of Funding

The authors declare that they have no competing interests. The research was approved by the Ethical Committee of the Institute for Research in Social Communication, Slovak Academy of Sciences, in conformity with the APA ethical code. Participants provided their own written consent. The data are available on the Open Science Framework (OSF) at https://osf.io/qax9w. M.U. was funded by the project „Mediated Society (MEDIS:ON) CZ.02.01.01/00/23_025/0008713“ which is co-financed by the European Union. K. U. was funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00423.

References

Anderson, B. R., Shah, J. H. and Kreminski, M.: 2024, Homogenization effects of large language models on human creative ideation, *Proceedings of the 16th Conference on Creativity & Cognition (C&C '24)*, Association for Computing Machinery, pp. 413–425.
URL: *https://doi.org/10.1145/3635636.3656204*

Beaty, R. E. and Johnson, D. R.: 2021, Automating creativity assessment with semdis: An open platform for computing semantic distance, *Behavior Research Methods* **53**(2), 757–780.
URL: *https://doi.org/10.3758/s13428-020-01453-w*

Bouchet-Valat, M.: 2023, Snowballc: Snowball stemmer based on the c 'libstemmer' utf-8 library (version 0.7.1) [r package], `https://github.com/nalimilan/R.TeMiS`.

Darbellay, F., Moody, Z. and Lubart, T.: 2017, Introduction: Thinking creativity, design and interdisciplinarity in a changing world, *in* F. Darbellay, Z. Moody and T. Lubart (eds), *Creativity, design thinking and interdisciplinarity*, Springer, pp. xi–xviii.

de Chantal, P., Beaty, R., Laverghetta, A., J., Pronchick, J., Patterson, J., Organisciak, P. and Karwowski, M.: 2025, Artificial intelligence enhances human creativity through real-time evaluative feedback [preprint].
URL: *https://doi.org/10.31219/osf.io/qrgbn_v1*

Doshi, A. R. and Hauser, O. P.: 2024, Generative ai enhances individual creativity but reduces the collective diversity of novel content, *Science Advances* **10**(28).
URL: *https://doi.org/10.1126/sciadv.adn5290*

Dumas, D., Organisciak, P. and Doherty, M.: 2021, Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods, *Psychology of Aesthetics, Creativity, and the Arts* **15**(4), 645–663.
URL: *https://doi.org/10.1037/aca0000319*

Feinerer, I. and Hornik, K.: 2025, tm: Text mining package (version 0.7-16) [r package], `https://tm.r-forge.r-project.org/`.

Haase, J. and Hanel, P. H. P.: 2023, Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity, *Journal of Creativity* **33**(3), 100066.
URL: *https://doi.org/10.1016/j.yjoc.2023.100066*

Haase, J., Hanel, P. H. P. and Pokutta, S.: 2025, Has the creativity of large-language models peaked?—an analysis of inter- and intra-llm variability [preprint].
URL: *https://doi.org/10.48550/arXiv.2504.12320*

Hubert, K. F., Awa, K. N. and Zabelina, D. L.: 2024, The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks, *Scientific Reports* **14**(1), 3440.
URL: *https://doi.org/10.1038/s41598-024-53303-w*

Kjell, O., Giorgi, S. and Schwartz, A.: 2025, text: Analyses of text using transformers models from huggingface, natural language processing and machine learning (version 1.6) [r package], `https://r-text.org/`.

Messeri, L. and Crockett, M. J.: 2024, Artificial intelligence and illusions of understanding in scientific research, *Nature* **617**, 49–58.
URL: *https://doi.org/10.1038/s41586-024-07146-0*

Meyer, D. and Buchta, C.: 2025, proxy: Distance and similarity measures (version 0.4-27) [r package], `https://CRAN.R-project.org/package=proxy`.

Moon, K., Green, A. E. and Kushlev, K.: 2025, Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing, *Computers in Human Behavior: Artificial Humans* **6**, 100207.
URL: *https://doi.org/10.1016/j.chbah.2025.100207*

Muraki, E. J., Abdalla, S., Brysbaert, M. and Pexman, P. M.: 2023, Concreteness ratings for 62,000 english multiword expressions, *Behavior Research Methods* **55**(5), 2522–2531.
URL: *https://doi.org/10.3758/s13428-022-01912-6*

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

Patterson, J. D., Merseal, H. M., Johnson, D. R., Agnoli, S., Baas, M., Baker, B. S., Barbot, B., Benedek, M., Borhani, K., Chen, Q., Christensen, J. F., Corazza, G. E., Forthmann, B., Karwowski, M., Kazemian, N., Kreisberg-Nitzav, A., Kenett, Y. N., Link, A., Lubart, T., Mercier, M., Miroshnik, K., Ovando-Tellez, M., Primi, R., Puente-Díaz, R., Said-Metwaly, S., Stevenson, C., Vartanian, M., Volle, E., van Hell, J. G. and Beaty, R. E.: 2023, Multilingual semantic distance: Automatic verbal creativity assessment in many languages, *Psychology of Aesthetics, Creativity, and the Arts* **17**(4), 495–507.
URL: *https://doi.org/10.1037/aca0000618*

Patterson, J. D., Pronchick, J., Panchanadikar, R., Fuge, M., van Hell, J. G., Miller, S. R., Johnson, D. R. and Beaty, R. E.: 2025, Cap: The creativity assessment platform for online testing and automated scoring, *Behavior Research Methods* **57**, Article 264.
URL: *https://doi.org/10.3758/s13428-025-02761-9*

Saretzki, J. and Benedek, M.: 2025, Investigating the validity evidence of automated scoring methods for divergent thinking assessments [preprint].
URL: *https://doi.org/10.31219/osf.io/8n2cb_v1*

Si, C., Yang, D. and Hashimoto, T.: 2024, Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers [preprint].
URL: *https://arxiv.org/abs/2409.04109*

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C. and Urban, K.: 2024, Chatgpt improves creative problem-solving performance in university students: An experimental study, *Computers Education* **215**, 105031.
URL: *https://www.sciencedirect.com/science/article/pii/S0360131524000459*

Wenger, E. and Kenett, Y. N.: 2025, We're different, we're the same: Creative homogeneity across llms [preprint].
URL: *https://doi.org/10.48550/arXiv.2501.19361*

Wu, F., Black, E. and Chandrasekaran, V.: 2024, Generative monoculture in large language models [preprint].
URL: *https://arxiv.org/abs/2407.02209*

A Technical Appendices and Supplementary Material

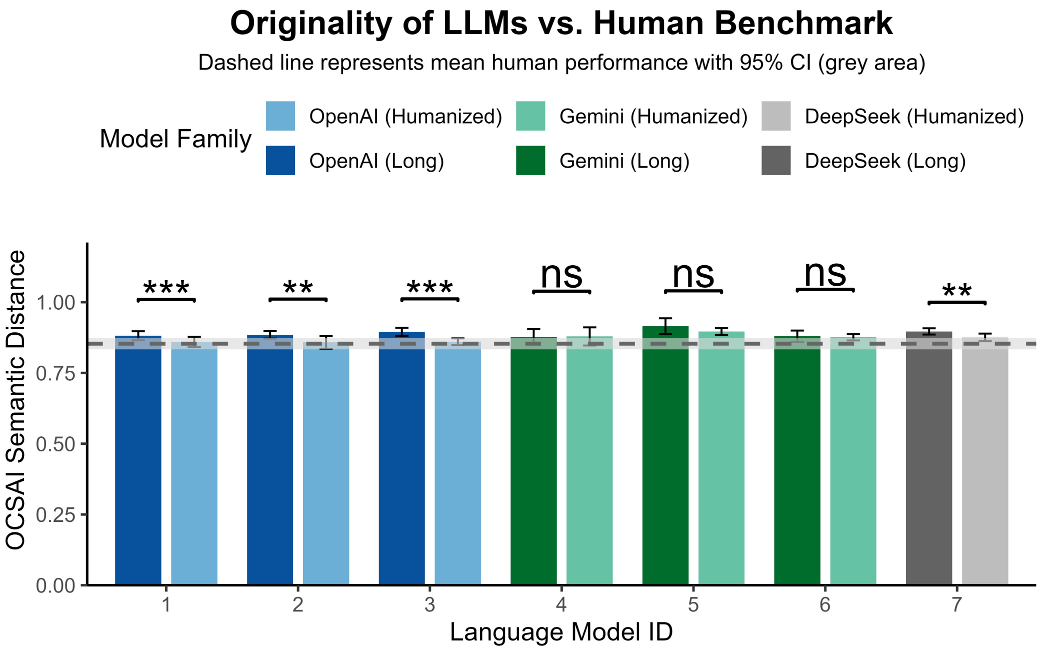

Figure 3: Originality of long and humanized responses rated by OCSAI Semantic Distance. *Note:* Model IDs correspond to: (1) gpt-4o, (2) gpt-o1, (3) gpt-o3, (4) 2.0 Flash thinking, (5) 2.5 Flash thinking, (6) 2.5 Pro, and (7) r1.

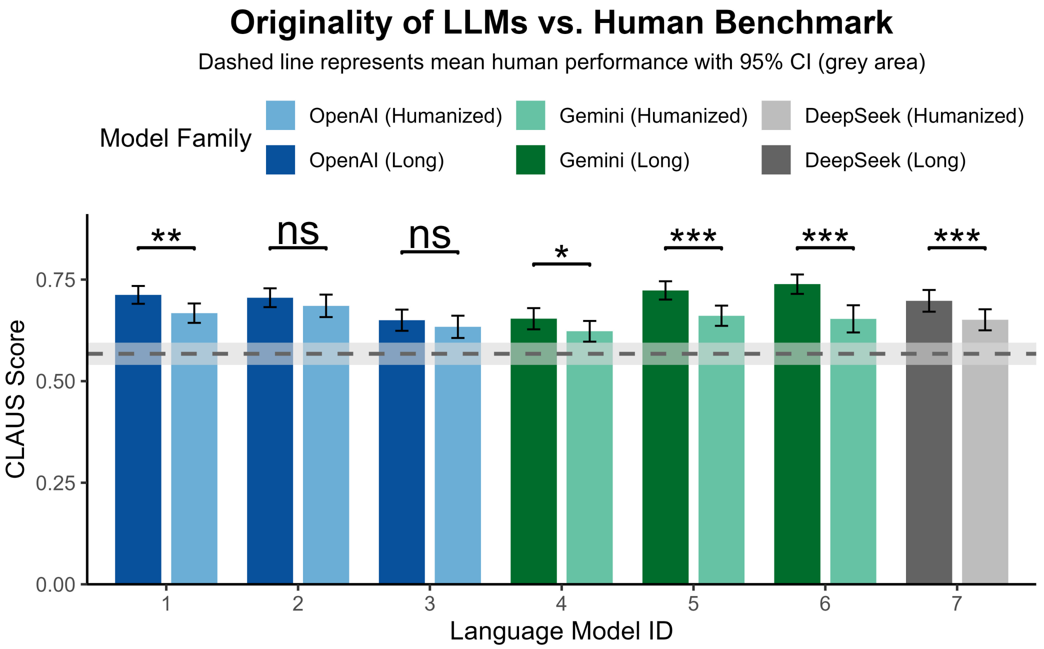

Figure 4: Originality of long and humanized responses rated by CLAUS. *Note:* Model IDs correspond to: (1) gpt-4o, (2) gpt-o1, (3) gpt-o3, (4) 2.0 Flash thinking, (5) 2.5 Flash thinking, (6) 2.5 Pro, and (7) r1.

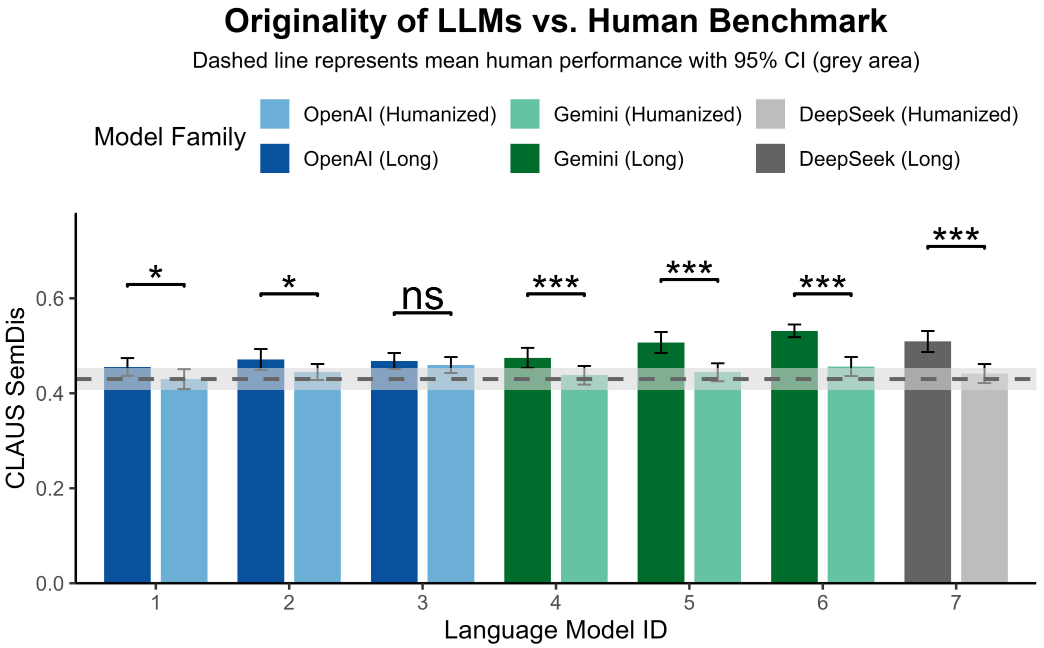

Figure 5: Originality of long and humanized responses rated by CAP SemDis. *Note:* Model IDs correspond to: (1) gpt-4o, (2) gpt-o1, (3) gpt-o3, (4) 2.0 Flash thinking, (5) 2.5 Flash thinking, (6) 2.5 Pro, and (7) r1.

Table 3: Comparisons of originality scores and variability of answers between LLM-generated and human responses

			OCSAI			OCSAI (semantic distance)			CLAUS			CAP SemDis			Variability (cosine distance)		
		Comparison	t(232)	d	p_{Holm}	t(232)	d	p_{Holm}	t(232)	d	p_{Holm}	t(232)	d	p_{Holm}	t(3470)	d	p_{Holm}
human	vs.	gpt-4o (long)	-12.10	-2.59	<.001	-2.02	-0.54	1.00	-8.40	-2.07	<.001	-1.81	-0.47	1.00	15.11	1.03	<.001
human	vs.	gpt-o1 (long)	-10.82	-2.32	<.001	-2.28	-0.61	1.00	-8.00	-1.97	<.001	-2.93	-0.77	.30	11.24	0.76	<.001
human	vs.	gpt-o3 (long)	-12.10	-2.59	<.001	-3.03	-0.80	.30	-4.79	-1.18	<.001	-2.72	-0.71	.54	12.51	0.85	<.001
human	vs.	2.0 Flash Thinking (long)	-8.98	-1.92	<.001	-1.78	-0.47	1.00	-5.00	-1.23	<.001	-3.20	-0.84	.13	11.81	0.80	<.001
human	vs.	2.5 Flash Thinking (long)	-13.05	-2.80	<.001	-4.48	-1.19	<.001	-9.04	-2.23	<.001	-5.51	-1.44	<.001	14.58	0.99	<.001
human	vs.	2.5 Pro (long)	-13.53	-2.90	<.001	-1.90	-0.50	1.00	-9.93	-2.45	<.001	-7.25	-1.89	<.001	21.49	1.46	<.001
human	vs.	r1 (long)	-12.25	-2.63	<.001	-3.13	-0.83	.22	-7.55	-1.86	<.001	-5.65	-1.48	<.001	12.56	0.85	<.001
human	vs.	gpt-4o (humanized)	-7.98	-2.06	<.001	-0.46	-0.12	1.00	-5.53	-1.43	<.001	0.02	0.01	1.00	9.12	0.62	<.001
human	vs.	gpt-o1 (humanized)	-6.21	-1.60	<.001	-0.34	-0.09	1.00	-6.53	-1.69	<.001	-1.09	-0.28	1.00	9.55	0.65	<.001
human	vs.	gpt-o3 (humanized)	-7.27	-1.88	<.001	-0.53	-0.14	1.00	-3.67	-0.95	.02	-2.12	-0.55	1.00	8.36	0.57	<.001
human	vs.	2.0 Flash Thinking (humanized)	-4.73	-1.22	<.001	-1.88	-0.49	1.00	-3.06	-0.79	.16	-0.58	-0.15	1.00	5.04	0.34	<.001
human	vs.	2.5 Flash Thinking (humanized)	-8.85	-2.29	<.001	-3.14	-0.82	.22	-5.18	-1.34	<.001	-1.01	-0.26	1.00	5.90	0.40	<.001
human	vs.	2.5 Pro (humanized)	-5.81	-1.50	<.001	-1.66	-0.43	1.00	-4.76	-1.23	<.001	-1.91	-0.49	1.00	15.42	1.05	<.001
human	vs.	r1 (humanized)	-7.03	-1.82	<.001	-1.64	-0.43	1.00	-4.63	-1.20	<.001	-0.82	-0.21	1.00	4.23	0.29	<.001