

Prompt	Individual creativity		Population-level variability	
	$\mu(O_{AUT}(\mathcal{P}))$	$t(df) = X, p$ (vs. humans)	$\mu(V_{AUT}(\mathcal{P}))$	$t(df) = X, p$ (vs. humans)
Humans	0.695	-	0.738	-
Baseline	0.711	$t(2094) = -3.4, 0.001$	0.459	$t(10078) = 19.1, 3.9e^{-80}$
More creative	0.733	$t(5020) = -9.8, 1.0e^{-22}$	0.503	$t(10078) = 16.1, 3.5e^{-58}$
Very creative	0.754	$t(5206) = -15.9, 3.2e^{-56}$	0.576	$t(10078) = 11.1, 5.6e^{-29}$
Not creative	0.702	$t(2507) = -1.28, 0.1$	0.492	$t(10078) = 16.8, 1.1e^{-62}$

Table 7. Varying the system prompt slightly increases LLM individual creativity and response variability, but variability remains far lower than that of humans.

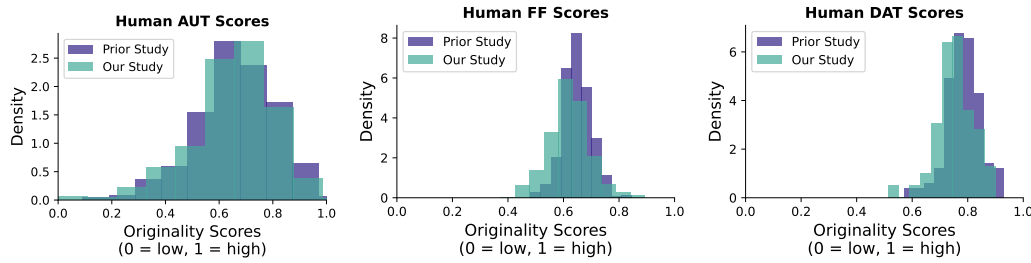


Fig. 8. Humans in prior studies have higher individual originality scores than humans in our study for all three tests. For the AUT and DAT tests, a t-test for a difference in means (alternative hypothesis is that prior study has higher mean than ours) is significant at the 0.01 level but not the 0.001 level: AUT has $t(5064) = 3.21, p = 0.001$ and DAT has $t(206) = 3.32, p = 0.001$. For FF, the difference more significant: $t(892) = 6.91, p < 0.0001$.

We evaluate the same subset of LLMs from §4 on the AUT using these system prompts and report summary statistics in Table 7. As the Table shows, using more creative system prompts slightly increases individual creativity for LLMs (and vice versa for the less creative prompt). However, the system prompt does not substantially improve LLM response variability—across all prompts, LLM variability remains much lower than that of humans.

5.4 Validation with preexisting survey data

Finally, we compare responses in our user study to prior user studies to ensure that our human subject pool is reliable and not unduly skewed by possible use of LLMs. We test both the individual originality of our human responses and population-level variability and find that while *respondents in prior studies score better individually on the tests*, *respondents to our study exhibit equal or greater population-level variability* (the more important metric for our study) on the more-informative AUT and FF tests.

Figure 8 compares individual creativity results for our study ($n = 102$) to that of prior studies ($n = 141$ for DAT, $n = 92$ for AUT, $n = 146$ for FF). T-tests for differences of individual performance (see caption of Figure 8) find that the mean score is higher for prior studies on all tests at a significant level of $p \leq 0.001$. Figure 9 compares the response variability of our study to that of the prior study. Using a t-test for difference in population means, we find that responses in our study have slightly higher variability on the FF and DAT ($p < 0.0001$), and lower on the AUT ($p < 0.001$). From this, we conclude that, our results roughly mirror those of prior studies, making them a reasonable baseline for our analysis.

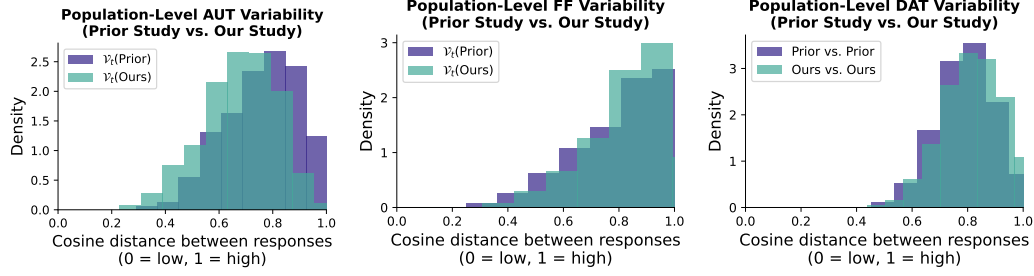


Fig. 9. Our study responses have higher population-level variability than the prior study on the DAT and FF tests, but slightly lower variability on the AUT. We use t -tests to compare means of the two population-level originality distributions. Null hypothesis is that means are equal, and alternative is that they are not. For AUT, the prior study has higher mean variability, $t(1046) = 11.0, p < 0.001$. For FF, our study has a higher mean, $t(366689) = -28.9, p < 0.001$. For the DAT, our study also has higher mean, $t(19832) = -19.6, p < 0.001$.

6 DISCUSSION

Motivated by measured homogeneity in creative outputs produced by specific LLMs and observed feature space overlap in LLMs, we study whether responses to creative prompts produced by a group of LLMs exhibit more, less, or equal variance as a set of human responses to the same creative prompts. We find that LLMs exhibit *much* lower population-level output variability than humans, even after controlling for potential model similarities and structural differences between LLM and human responses. Our work upholds prior work showing that LLMs perform well on tests of divergent thinking but adds the nuance that such performance is homogeneous—LLMs return a narrower range of responses to creative prompts than humans. This result enhances prior observations of LLM-induced homogeneity, which only considered the effect of specific LLMs on creative outputs, and suggests that the use of LLMs *in general* may homogenize creative outputs.

Implications. These results have significant implications if LLMs are widely adopted as creativity support tools for writing, idea generation, or similar tasks. If all LLMs respond similarly to specific creative requests, then the population of users leveraging to LLMs to aid creativity will converge towards a limited set of creative outputs. In other words, LLM users may be self-limited from being exhibiting the divergent creativity that defined well-recognized artistic geniuses like Tolkein, Mozart, and Picasso because their LLM “creative” partners may collectively drive them towards a mean.

Limitations. Our work has several limitations. First, while we have demonstrated LLM homogeneity in response to certain creativity tests, this does not prove that LLMs in general produce homogeneous outputs when asked to behave creatively. It merely provides an indication that future work should explore this subject. Additionally, we measure a single metric of divergent thinking or creativity—originality, as measured by semantic similarity between responses—and finds that LLMs are homogeneous along this dimension. However, there are other well-known metrics of divergent thinking, such as flexibility, fluency, and elaboration (see §3.1), and LLMs may demonstrate more or less homogeneity along these dimensions. Future work should consider these alternatives.

Acknowledgments. We thank Austin Liu for helping us design the system prompts of §5.3.

7 ETHICAL CONSIDERATIONS

We took care to ensure the user study in this paper was conducted in accordance with ethical standards. IRB approval for the study was obtained, and participants signed a clearly written consent form before completing our survey. To ensure privacy, participant data was anonymized and stored on secure servers. Other ethical risks from this paper are minimal, as our LLM experiments do not involve sensitive data and elicit only benign model responses.

REFERENCES

- [1] 2018. APA Dictionary of Psychology - Creativity. <https://dictionary.apa.org/creativity>.
- [2] 2024. Apple Intelligence | Writing Tools | iPhone 16. <https://www.youtube.com/watch?v=3m0MoYKwVTM>.
- [3] 2024. Command R and Command R Plus Model Card. <https://docs.cohere.com/docs/responsible-use>.
- [4] 2024. Use Notion AI to write better, more efficient notes and docs. <https://www.notion.com/help/guides/notion-ai-for-docs>.
- [5] Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).
- [6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [7] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*. 413–425.
- [8] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. ‘Revisiting model stitching to compare neural representations. *Proc. of NeurIPS* (2021).
- [9] Baptiste Barbot. 2018. The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in psychology* (2018).
- [10] Roger E Beaty, Paul J Silvia, Emily C Nusbaum, Emanuel Jauk, and Mathias Benedek. 2014. The roles of associative and executive processes in creative cognition. *Memory & cognition* (2014).
- [11] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.
- [12] Honghua Chen and Nai Ding. 2023. Probing the Creativity of Large Language Models: Can models produce divergent semantic association? (Oct. 2023). <http://arxiv.org/abs/2310.11158>
- [13] Jacob Cohen. 2016. A power primer. (2016).
- [14] David Cropley. 2023. Is artificial intelligence more creative than humans?: ChatGPT and the divergent association task. *Learning Letters* 2 (2023), 13–13.
- [15] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. 2019. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security)* 19. 321–338.
- [16] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10, 28 (July 2024). <https://doi.org/10.1126/sciadv.adn5290>
- [17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [18] Denis Dumas and Kevin N Dunbar. 2014. Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity* (2014).
- [19] Denis Dumas, Peter Organisciak, and Michael Doherty. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts* (2021).
- [20] Matt Ellis. 2024. How to Use AI to Enhance Your Storytelling Process. <https://www.grammarly.com/blog/writing-with-ai/ai-story-writing/>.
- [21] Google. 2024. Google + Team USA - Dear Sydney. <https://www.youtube.com/watch?v=NgtHJKn0Mck>.
- [22] Kurt Gray, Stephen Anderson, Eric Evan Chen, John Michael Kelly, Michael S Christian, John Patrick, Laura Huang, Yoed N Kenett, and Kevin Lewis. 2019. ‘Forward flow’: A new measure to quantify free thought and predict creativity. *American Psychologist* 74, 5 (2019), 539.
- [23] Joy Paul Guilford, Paul R Christensen, Philip R Merrifield, and Robert C Wilson. 1978. Alternate uses. (1978).
- [24] Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 301–314.
- [25] Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14, 1 (2024), 3440.
- [26] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987* (2024).
- [27] Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. 2024. Bias Similarity Across Large Language Models. *arXiv preprint arXiv:2410.12010* (2024).
- [28] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7B (2023). *arXiv preprint arXiv:2310.06825* (2023).
- [29] Max Klabunde, Mehdi Ben Amor, Michael Granitzer, and Florian Lemmerich. 2023. Towards Measuring Representational Similarity of Large Language Models. In *UniReps: the First Workshop on Unifying Representations in Neural Models*.
- [30] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118.
- [31] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*. PMLR, 3519–3529.
- [32] Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981* (2024).
- [33] Karel Lenc and Andrea Vedaldi. 2015. Understanding image representations by measuring their equivariance and equivalence. In *Proc. of CVPR*.