Characterising the Creative Process in Humans and Large Language Models

Surabhi S. Nath[1,2,3], **Peter Dayan**[1,2], **Claire Stevenson**[4]
[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]University of Tübingen, Tübingen, Germany
[3]Max Planck School of Cognition, Leipzig, Germany
[4]University of Amsterdam, Department of Psychological Methods, Amsterdam, Netherlands

Abstract

Large language models appear quite creative, often performing on par with the average human on creative tasks. However, research on LLM creativity has focused solely on *products*, with little attention on the creative *process*. Process analyses of human creativity often require hand-coded categories or exploit response times, which do not apply to LLMs. We provide an automated method to characterise how humans and LLMs explore semantic spaces on the Alternate Uses Task, and contrast with behaviour in a Verbal Fluency Task. We use sentence embeddings to identify response categories and compute semantic similarities, which we use to generate jump profiles. Our results corroborate earlier work in humans reporting both persistent (deep search in few semantic spaces) and flexible (broad search across multiple semantic spaces) pathways to creativity, where both pathways lead to similar creativity scores. LLMs were found to be biased towards either persistent or flexible paths, that varied across tasks. Though LLMs as a population match human profiles, their relationship with creativity is different, where the more flexible models score higher on creativity. Our dataset and scripts are available on GitHub: `https://github.com/surabhisnath/Creative_Process`

Introduction

Much recent work has benchmarked and quantified the generative creative aptitudes of large language models (LLMs) (Chakrabarty et al. 2023; Gilhooly 2023; Franceschelli and Musolesi 2023; Tian et al. 2023; Wang et al. 2024; Hubert, Awa, and Zabelina 2024). LLMs often perform as well as the average human on creative thinking tasks such as the Alternate Uses Task (AUT) (Orwig et al. 2024; Koivisto and Grassini 2023; Stevenson et al. 2022; Góes et al. 2023; Guzik, Byrge, and Gilde 2023). However, these works largely analysed creativity from a *Product* perspective (Rhodes 1961), assessing how original and useful model responses are to determine "what makes them creative (or not)". An equally important component of creativity, less studied in the field of Artificial Creativity, is the *Process* perspective (Rhodes 1961), addressing the question of "how creativity arises". This paper aims to fill this gap and characterise human and LLM creativity by looking at the creative process (Stevenson et al. 2022), comparing the way humans and LLMs explore semantic spaces while generating sequences of creative ideas. This comparison makes sense given that the tasks involve creative response-sequence generation, in line with LLM objective of next-word prediction.

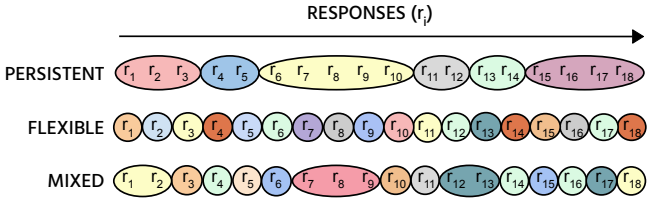


Figure 1: Example persistent, flexible and mixed response sequences. r_i denotes the i^{th} response, coloured regions denote the semantic spaces/concepts/categories. Note, in practice, most sequences will be mixed, containing different patterns of persistence and flexibility.

When humans generate creative ideas, for example, alternate uses for a "brick", two types of response pathways are observed (Baas et al. 2013; Nijstad et al. 2010). In the *persistent* pathway, responses stem from deeper search within limited conceptual spaces, exhibiting high clustering and similarities in responses (e.g., using a brick to break a window, break a lock, and as a nutcracker; i.e., for breaking things). In the *flexible* pathway, responses arise from broader search across multiple conceptual spaces, exhibiting frequent jumps between categories and dissimilarities in responses (e.g., using a brick to build a dollhouse, as an exercise weight, and as a coaster) (Figure 1).

There are two complementary ways of quantifying response clustering borrowed from the literature on memory search and semantic fluency. The first is to categorise responses temporally using inter-item retrieval times, *i.e.* responses that occur shortly after each other are expected to belong to the same category and longer pauses are expected to signal jumps from one category to another. The second method is to group successive responses semantically using a set of pre-defined categories (e.g., into "building" or "breaking" for uses of a brick). The number of categories divided by the number of responses provides a flexibility index (Hills, Jones, and Todd 2012). Hass (2017) compared clustering in creative thinking tasks like AUT to that in a verbal fluency task (VFT) of naming animals and reported less evident clustering and higher flexibility in AUT than VFT (where responses were highly clustered, for example

naming zoo animals followed by sea animals).

However, the methods used in these works are either based on handcrafted lists of categories or on response-time profiles which do not apply to responses from LLMs. In addition, these works show that semantic similarity is related to jumps in response sequences, but semantic similarity has not been used to code for jumps directly until now. In this paper, we propose a fully automated, data-driven method to signal jumps in response sequences using response categorisation and semantic similarites and apply it to characterise the creative process in both humans and LLMs.

In the next sections, we first introduce our method and investigate its reliability and validity. We then apply it to characterise human and LLM flexibility on the AUT and VFT. We find that LLMs as a population match the variability in human response sequences on AUTs, but unlike humans, their relationship to creativity differs. We also discuss how to use these insights to use LLMs as artificial participants or co-creators.

Method

Data Collection: We collected data from humans and LLMs on the AUT for "brick" and "paperclip", and the VFT of naming animals (Figure 2A).

Human data were collected from undergraduate participants at University of Amsterdam using a within-subjects design. For AUT, participants listed as many creative uses for "brick" and "paperclip" as possible in a fixed time of 10 minutes. For VFT, participants named as many animals as possible in a fixed time of 2 minutes. Participants not adhering to instructions were removed, resulting in a total of 220 participants. The responses, originally in Dutch were translated to English for analysis using the `deep-translator` Python package. Translations were manually inspected to correct for errors due to spelling mistakes.

LLM data were collected in English by prompting several recent open and closed source models. For open source models, we used the Together AI API. The prompt matched instructions given to humans, but with specific response number and length requirements. We tested multiple prompt versions to achieve the best quality LLM responses. The final prompt for the AUT instructed LLMs to generate n_{AUT} creative uses for "brick" or "paperclip", and to answer in short phrases of maximum m_{AUT} words. For the VFT, the final prompt instructed LLMs to name n_{VFT} animals, and to answer in short phrases of maximum m_{VFT} words. n_{AUT}, n_{VFT} were set to the mean number of human responses (N) in AUT (=ceil[max[N_{brick}, $N_{\text{paperclip}}$]]) and VFT tasks. m_{AUT}, m_{VFT} were set to the maximum mean human response word length (M) in AUT (=floor[max[M_{brick},$M_{\text{paperclip}}$]]) and VFT.

In pilots, only $\sim$20 models gave valid responses for the AUT tasks, of which we selected the 4 that followed the prompt instructions for length and number of responses, namely, `Meta 70B Llama 3 chat HF (Llama)` model, `Mistral AI 7B Instruct (Mistral)` model, `NousResearch 7B Nous-Hermes Mistral DPO (NousResearch)` model and `Upstage 10.7B SOLAR Instruct (Upstage)` model.

We experimented with temperature and repetition penalty parameters. However, varying the repetition penalty did not produce higher quality responses, so we only varied the temperature, through 11 levels (0-1, inclusive, at every 0.1).

We also tested the latest versions of 4 closed source models: `OpenAI GPT-4 turbo (GPT)`, `Google Palm bison (Palm)`, `Google Gemini 1.0 pro (Gemini)` and `Anthropic Claude 3 (Claude)`, with the same prompt and parameters as for the open models. All 4 models generated valid responses and adhered to the response number and length instructions.

We generated 5 samples per model $\times$ temperature combination, and therefore our LLM data set consisted of 440 (8 $\times$ 11 $\times$ 5) LLM response sequences in all.

The 220 human and 440 LLM response sequences were cleaned by removing stopwords, punctuations and common words such as "use" or "brick"/"paperclip". They were also manually inspected for correctness and validity. Invalid responses (verbatim repeats/junk responses) were removed [1].

Response Categorisation, Semantic Similarities and Jump Signal: First, we encoded all responses using sentence-transformers, using the `gte-large` model given its encodings' suitability for clustering. Each response was encoded as a 1024 dimensional normalised embedding vector. Next, all responses were aggregated, dropping duplicates, resulting in 2770 unique alternate uses for brick, 3512 unique alternate uses for paperclip and 482 unique animals. The vector embeddings of these response sets were categorised using the `scipy linkage, fcluster` hierarchical clustering functions with the `ward` distance metric and a distance threshold chosen such that the mean minimum pairwise semantic similarity (vector dot product) per category was just above 0.7. This resulted in 26 brick, 28 paperclip, and 15 animal categories.

Using these categories, we defined a binary variable $jump_{cat}$ for each response in a response sequence (except for the first response) as 1 if it marked a change in category compared to the previous response, and 0 otherwise.

$jump_{cat}$ provided us course-grained similarities (for example 'elevation' and 'table leg' belonged to the same category as did 'keep scarf together' and 'hang clothes'). To address finer-grained differences, we evaluated the semantic similarity (SS) between successive embeddings of responses in a response sequence (Hass 2017; Camenzind et al. 2024). Using SS, we defined a second binary variable $jump_{SS}$ and set it to 0 if SS was above a threshold and 1 otherwise. $jump_{SS}$ signaled finer-grained similarities (for example'piercing' and 'ring').

A combined jump signal was defined as their logical AND: $jump = jump_{cat} \wedge jump_{SS}$. We set the threshold for $jump_{SS}$ such that *jump* has at least 0.8 True Positive and True

[1]For AUT brick, we removed low temperature responses in `Mistral` and `NousResearch` models as these were verbatim repeats. For VFT, we excluded `NousResearch` and `Palm` models fully, as they only listed animals in alphabetical order.

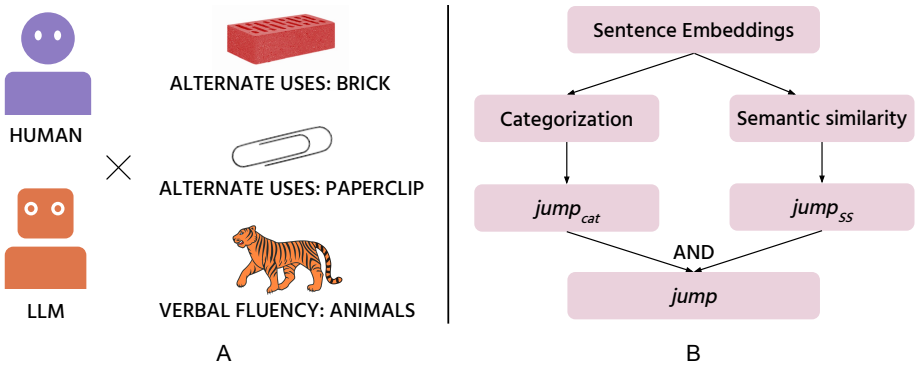


Figure 2: (A) Humans and LLMs perform 3 tasks—Alternate Uses Task (AUT) for brick and paperclip, and a Verbal Fluency Task (VFT) of naming animals. (B) Our method for obtaining jumps in the response sequence. Sentence embeddings are used for assigning response categories and evaluating semantic similarities, which respectively give $jump_{cat}$ and $jump_{SS}$. Their logical AND gives *jump*.

Negative Rates on hand-coded[2] jump signals for AUT brick. Our entire procedure is illustrated in Figure 2B. We conduct psychometric analyses to investigate the reliability and validity of the method.

Jump Profiles and Participant Clustering: Using the jump signals, we determined a jump profile for each response sequence as the cumulative count of jumps at each response (for example, a response sequence of length 4 with jumps $[1, 0, 1]$ will have a jump profile $[1, 1, 2]$). Different human participants produced different numbers of responses, so we considered just the first 18 responses from each sequence (the median human sequence length), excluding shorter sequences. The remaining profiles (AUT brick: 97; AUT paperclip: 103; VFT: 195) were clustered using K-Means (`sklearn KMeans`) with K-Means++ initialization (Arthur, Vassilvitskii, and others 2007) per task. LLM jump profiles were assigned to the closest human cluster.

Evaluating Response Creativity: We used Open Creativity Scoring `ocsai-chatgpt` (Organisciak et al. 2023) to score response originality in AUT brick and paperclip.

Results

Jump Signal Reliability and Validity: We first test the reliability and validity of the *jump* signal. For reliability, we measured the test-retest correlation of the number of jumps for AUT brick and paperclip response sequences from 81 participants (who had $>=18$ responses in both). We found a positive Pearson correlation of r=0.42 (p<0.001, CI=[0.22, 0.58]), which is high considering the test-retest and alternate-form reliability of AUT *product* creativity seldom exceeds r=0.5 (Barbot 2019).

For validity, we test for agreement with past findings in humans. In keeping with Hass 2017, who showed more jumping in AUT than VFT, we found significantly more jumps in AUT brick and paperclip than in VFT (both p<0.001). Moreover, in line with Hass (2017), Hills *et al.* (2012), we also found greater mean response times for $jump = 1$ than $jump = 0$ (p<0.001).

Participant Clusters: Based on the literature and clustering elbow plots, we assigned human jump profiles to 3 clusters for each task (Figure 3A). These map to different levels of flexibility in the response sequences—cluster 1: *persistent* profiles (7-12 jumps for AUT and 1-6 jumps for VFT); cluster 2: *flexible* profiles (15-18 jumps for AUT and 6-11 jumps for VFT); and cluster 3: *mixed* profiles (12-16 jumps for AUT and 4- jumps for VFT). The different numbers of jumps in AUT and VFT are clear, where the *flexible* cluster in VFT closely resembles the *persistent* cluster in AUTs. Thus the classifications are task-relative. The proportion of participants assigned to each cluster further reinforces that people are more flexible in AUT and more persistent in VFT.

LLM Assignments: The LLM jump profiles were assigned to one of the 3 human clusters with proportions of assignment shown in Figure 3B.

Different models exhibited different biases towards persistence or flexibility in the AUTs. For example, in AUT brick, `Upstage`, `GPT`, `Claude` and `Palm` are mostly flexible while `Llama` and `Gemini` are mostly persistent. However, models were less consistent across the two AUTs. In AUT paperclip, while `Upstage` and `GPT` remained mostly consistent in their assignments, but `Llama` and `Gemini` switched from *persistent* to *flexible*. This is also evident in the test-retest correlation, which was lower than for humans (r=0.22, p<0.001, CI=[0.12, 0.31]). Taken together, we find that LLMs are not significantly different than humans in number of jumps on AUTs (p>0.05 in both). However, on the VFT, LLMs were overwhelmingly persistent, and significantly more persistent than humans (p<0.001).

Comparing the human and model cluster assignment percentages, we observe that `Mistral` and `NousResearch` models closely resemble the human distribution in AUT brick; `Gemini` model does so in AUT paperclip; but no model resembles humans for VFT.

Temperature neither influenced cluster assignments nor number of jumps in AUTs (p>0.05). In VFT, temperature did influence jumping (p<0.001), but did not influence cluster assignment. This is consistent with previous research suggesting no role of temperature in flexibility (Stevenson et al. 2022) and suggests that model responses cannot be easily manipulated parametrically.

Relationship to Creativity: We calculated the mean originality ratings in each response sequence. For humans, mean originality was similar for *persistent* and *flexible* clusters in both AUTs (both p>0.05). Mean originality did not predict the number of jumps in AUT brick (p>0.05), and weakly predicted jumps in AUT paperclip (0.01<p<0.05). This is

[2]The jump signals were hand-coded by the first author.

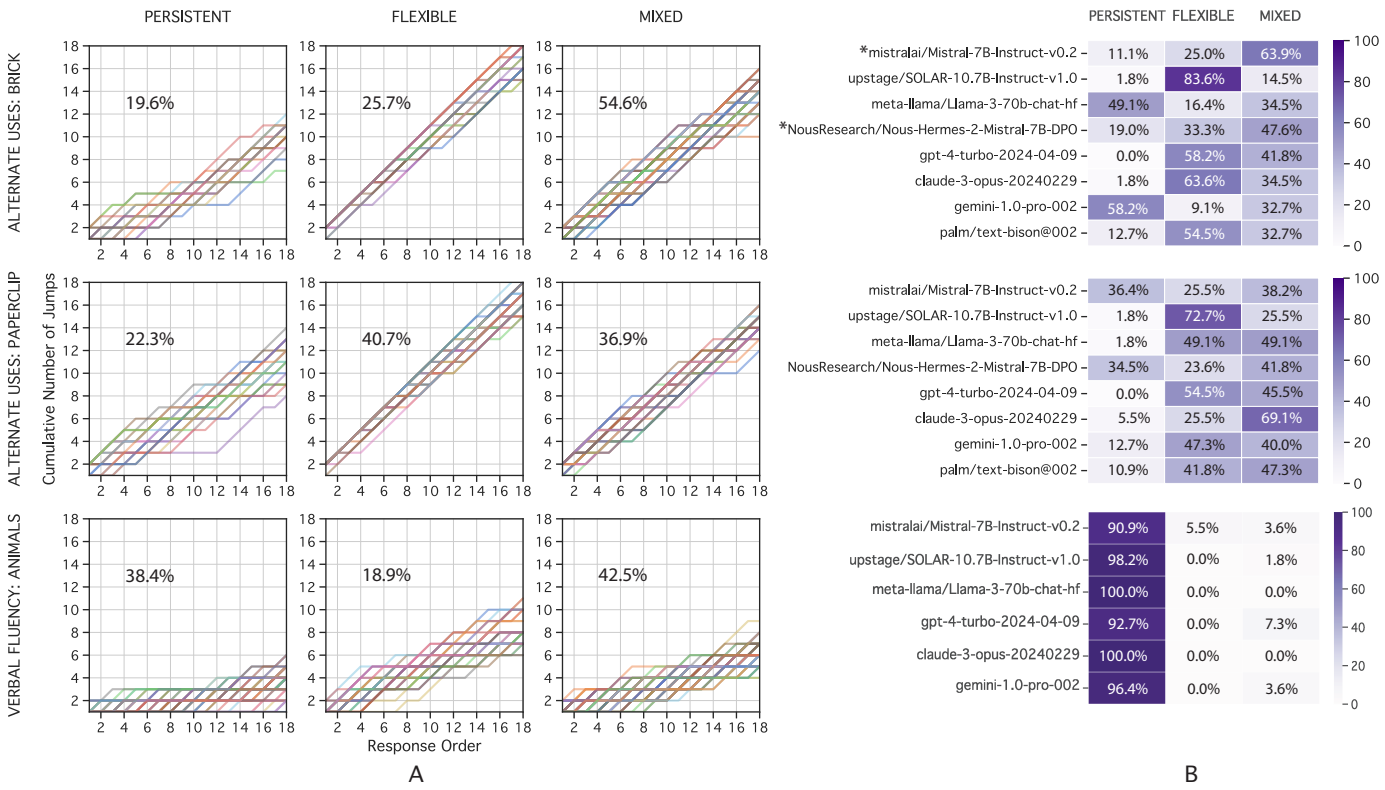


Figure 3: (A) 3 human clusters for each task–persistent, flexible and mixed. Each coloured trajectory represents 1 participant. Percentages in each row indicate the percentage of participants assigned to that cluster. (B) Percentages of each LLM response sequences assigned to each cluster. * indicates not all temperatures for that model were included (0.4-1 for `Mistral`, and 0.7-1 for `NousResearch` were used).

in line with the literature suggesting that creativity can arise both from deeper and broader search of semantic spaces.

In contrast, for LLMs, on both AUTs, mean originality was higher in the *flexible* cluster compared to the *persistent* cluster (both $p<0.01$), and mean originality predicted the number of jumps(both $p<=0.01$). Therefore, even though the number of jumps in AUT tasks for humans and LLMs do not differ, their relationship with originality differs. Further, LLMs also scored higher on overall mean response sequence originality compared to humans on AUTs (both $p<0.001$).

Discussion

We introduce an automated, data-driven method to study the creative *process* in humans and LLMs. We defined an algorithmic *jump signal* to indicate persistance or flexibility while solving divergent thinking tasks such as the AUT and VFT. Our jump signal proved reliable and valid for human responses and replicated findings using the traditional methods (Hills, Jones, and Todd 2012; Hass 2017). We used this signal to investigate human and LLM jump profiles. For AUT, we found that both human and LLM jump profiles spanned from *persistent* to *flexible*. As in previous literature, human creativity was not correlated with flexibility profile (De Dreu, Baas, and Nijstad 2008). However, in LLMs more flexible models had higher originality scores.

Our work has limitations to be addressed with further research. First, we used the same embedding model (with different distance metrics) for response categorisation and semantic similarities. Second, each response was categorised into a single category—however, a response such as "throw a brick to produce sound" should include concepts of both "throw" and "produce sound". Multiclass classification based on predefined categories could tackle this issue. Third, we used a very basic prompt for LLMs in order to match human instructions. Future work can experiment with more advanced prompt engineering. Fourth, we only scored response originality ignoring utility, which could inflate creativity comparisons as inappropriate responses (e.g., "using a brick as a ribbon") have high originality scores but are not considered creative. Lastly, the AUT being a popular creativity task (especially AUT brick[3]) could be a source of LLM data contamination (Gilhooly 2023).

The implications of our work are many. There is an emerging trend in cognitive science to use LLMs as artificial participants (Argyle et al. 2023; Frank 2023; Dillion et al. 2023; Binz and Schulz 2023). Our results suggest that often LLMs are biased towards either persistence or flexibility, regardless of parameter settings such as temperature, and may be inconsistent across tasks. Therefore, we suggest using a host of models to approximate the human distribution and draw valid inferences.

The creative collaboration literature suggests that more diverse teams yield more creative ideas (Hoever et al. 2012).

[3]checked with WIMBD tool (Elazar et al. 2023)

An implication of our work for human-AI co-creativity, is to use an LLM to complement one's own brainstorming pathway. For example, more *persistent* participants could collaborate with a *flexible* model such as `Upstage`, which could help them diversify their ideas.

Through our work, we offer a first step to study human and LLM creative processes under the same metric. We provide some directions that are worth exploring in the future to further our understanding of human and artificial verbal creativity processes.

Acknowledgements

Authors thank Fleur Doorman and Emma Schreurs for their Masters theses work focussing on aspects of this project. Nath thanks Vishaal Udandarao, Andrew Webb, Sahiti Chebolu for useful suggestions and code review. The data collection and previous work on this project was funded by the Amsterdam Brain & Cognition (ABC) Talent Grant (University of Amsterdam) 2016–2018 awarded to Stevenson and the Jacobs Foundation Fellowship 2019–2022 awarded to Stevenson (2018 1288 12).

References

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3):337–351.

Arthur, D.; Vassilvitskii, S.; et al. 2007. k-means++: The advantages of careful seeding. In *Soda*, volume 7, 1027–1035.

Baas, M.; Roskes, M.; Sligte, D.; Nijstad, B. A.; and De Dreu, C. K. 2013. Personality and creativity: The dual pathway to creativity model and a research agenda. *Social and Personality Psychology Compass* 7(10):732–748.

Barbot, B. 2019. Measuring creativity change and development. *Psychology of Aesthetics, Creativity, and the Arts* 13(2):203.

Binz, M., and Schulz, E. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Camenzind, M.; Single, M.; Gerber, S. M.; Nef, T.; Bassetti, C. L.; Müri, R. M.; and Eberhard-Moscicka, A. K. 2024. Applying german word vectors to assess flexibility performance in the associative fluency task. *Psychology of Aesthetics, Creativity, and the Arts*.

Chakrabarty, T.; Laban, P.; Agarwal, D.; Muresan, S.; and Wu, C.-S. 2023. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

De Dreu, C. K.; Baas, M.; and Nijstad, B. A. 2008. Hedonic tone and activation level in the mood-creativity link: toward a dual pathway to creativity model. *Journal of personality and social psychology* 94(5):739.

Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Elazar, Y.; Bhagia, A.; Magnusson, I.; Ravichander, A.; Schwenk, D.; Suhr, A.; Walsh, P.; Groeneveld, D.; Soldaini, L.; Singh, S.; et al. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Franceschelli, G., and Musolesi, M. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Frank, M. C. 2023. Openly accessible llms can help us to understand human cognition. *Nature Human Behaviour* 7(11):1825–1827.

Gilhooly, K. 2023. Ai vs humans in the aut: simulations to llms. *Journal of Creativity* 100071.

Góes, L. F.; Volpe, M.; Sawicki, P.; Grses, M.; and Watson, J. 2023. Pushing gpt's creativity to its limits: Alternative uses and torrance tests.

Guzik, E. E.; Byrge, C.; and Gilde, C. 2023. The originality of machines: Ai takes the torrance test. *Journal of Creativity* 33(3):100065.

Hass, R. W. 2017. Semantic search during divergent thinking. *Cognition* 166:344–357.

Hills, T. T.; Jones, M. N.; and Todd, P. M. 2012. Optimal foraging in semantic memory. *Psychological review* 119(2):431.

Hoever, I. J.; Van Knippenberg, D.; Van Ginkel, W. P.; and Barkema, H. G. 2012. Fostering team creativity: perspective taking as key to unlocking diversity's potential. *Journal of applied psychology* 97(5):982.

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.

Koivisto, M., and Grassini, S. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific reports* 13(1):13601.

Nijstad, B. A.; De Dreu, C. K.; Rietzschel, E. F.; and Baas, M. 2010. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European review of social psychology* 21(1):34–77.

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.

Orwig, W.; Edenbaum, E. R.; Greene, J. D.; and Schacter, D. L. 2024. The language of creativity: Evidence from humans and large language models. *The Journal of Creative Behavior*.

Rhodes, M. 1961. An analysis of creativity. *The Phi delta kappan* 42(7):305–310.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting gpt-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*. Association for Computational Creativity (ACC).

Tian, Y.; Ravichander, A.; Qin, L.; Bras, R. L.; Marjieh, R.; Peng, N.; Choi, Y.; Griffiths, T. L.; and Brahman, F.

2023. Macgyver: Are large language models creative problem solvers? *arXiv preprint arXiv:2311.09682*.

Wang, H.; Zou, J.; Mozer, M.; Zhang, L.; Goyal, A.; Lamb, A.; Deng, Z.; Xie, M. Q.; Brown, H.; and Kawaguchi, K. 2024. Can ai be as creative as humans? *arXiv preprint arXiv:2401.01623*.