How do Humans and Language Models Reason About Creativity? A Comparative Analysis

Antonio Laverghetta Jr.[1]* **Tuhin Chakrabarty**[2] **Tom Hope**[3]
Jimmy Pronchick[1] **Krupa Bhawsar**[1] **Roger E. Beaty**[1]
[1]Pennsylvania State University
[2]Stony Brook University
[3]Hebrew University of Jerusalem

Abstract

Creativity assessment in science and engineering is increasingly based on both human and AI judgment, but the cognitive processes and biases behind these evaluations remain poorly understood. We conducted two experiments examining how including example solutions with ratings impact creativity evaluation, using a finegrained annotation protocol where raters were tasked with explaining their originality scores and rating for the facets of remoteness (whether the response is "far" from everyday ideas), uncommonness (whether the response is rare), and cleverness. In Study 1, we analyzed creativity ratings from 72 experts with formal science or engineering training, comparing those who received example solutions with ratings (example) to those who did not (no example). Computational text analysis revealed that, compared to experts with examples, no-example experts used more comparative language (e.g., "better/worse") and emphasized solution uncommonness, suggesting they may have relied more on memory retrieval for comparisons. In Study 2, parallel analyses with state-of-the-art LLMs revealed that models prioritized uncommonness and remoteness of ideas when rating originality, suggesting an evaluative process rooted around the semantic similarity of ideas. In the example condition, while LLM accuracy in predicting the true originality scores improved, the correlations of remoteness, uncommonness, and cleverness with originality also increased substantially — to upwards of 0.99 — suggesting a homogenization in the LLMs evaluation of the individual facets. These findings highlight important implications for how humans and AI reason about creativity and suggest diverging preferences for what different populations prioritize when rating.

Keywords: creativity; large language models; text analysis; STEM

Introduction

How do people evaluate and reason about creativity? In science and engineering, creativity assessment has traditionally relied on human experts to evaluate everything from grant proposals and scientific manuscripts to new technologies and engineering designs. Although experts routinely make these high-stakes decisions by weighing factors such as novelty and technical feasibility, the cognitive processes and biases that shape their evaluations are poorly understood. This challenge takes on new importance as artificial intelligence (AI) systems, particularly large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022), increasingly assume advanced roles in scientific research and innovation, from idea generation to peer review (Boiko et al., 2023; D'Arcy et al., 2024; Lu et al., 2024; Si et al., 2024; Q. Wang et al., 2024). LLMs can achieve impressive accuracy in predicting human creativity assessments (Organisciak et al., 2023), yet we know little about how they arrive at their judgments, what features they prioritize, or whether their evaluation strategies align with those of human experts. Understanding these cognitive and computational processes is crucial not only for advancing creativity research but also for developing AI systems that can better aid creative evaluation in STEM.

Modern creativity assessment is based on the Consensual Assessment Technique (CAT) (Amabile, 1982; Silvia et al., 2008), which uses human judgments from experts to reach a consensus on the creativity of a product/idea, often by assessing its originality or quality. In the standard implementation of this method, expert raters independently evaluate creative products without specific scoring criteria, relying on their implicit domain understanding. While this approach has demonstrated remarkable reliability and predictive validity — CAT ratings predict real-world creative achievements across multiple domains (Silvia et al., 2008) — it typically relies on global originality scores that may mask underlying evaluation processes (Cseh & Jeffries, 2019). Notably, originality itself can be understood as an aggregation of three distinct facets: uncommonness, remoteness, and cleverness (Silvia et al., 2008), though their relative contributions to expert judgments remain unclear. Understanding these evaluation patterns is especially crucial for STEM domains, where creative solutions must carefully balance novelty with real-world technical constraints. Even studies of simpler ideation tasks like the Alternate Uses Task (AUT) suggest this complexity, showing how different factors like novelty and appropriateness contribute distinctly to creativity judgments (Diedrich et al., 2015).

The cognitive processes underlying creativity evaluation are becoming clearer through recent empirical work J. Wang et al. (2025). Early think-aloud studies like Gilhooly et al. (2007) focused on idea-generation processes in the AUT, showing how participants move from memory retrieval to more abstract strategies. More recently, Orwig et al. (2024) used linguistic inquiry and word count (LIWC) analysis, a computerized text analysis method that quantifies psychological dimensions of language (Tausczik & Pennebaker, 2010), to analyze how participants explain their originality ratings on the AUT. Results revealed that even when judges agree on creativity scores, they often employ different cognitive processes in their evaluations, as evidenced by variations in their use of memory-related terms, temporal focus (past vs. future orientation), and analytical language.

*Corresponding author: <aml7990@psu.edu>

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

STEM creativity represents a crucial yet understudied domain for creativity assessment. Although extensive research has examined scientific hypothesis generation and engineering design thinking, we know surprisingly little about how experts evaluate creative merit in STEM contexts, where ideas must balance novelty with technical feasibility. To assess creative thinking in STEM, Patterson et al. (2025) developed a novel Design Problems Task (DPT) that measures the ability to generate solutions to real-world STEM challenges, capturing the dual constraints that characterize expert-level scientific and engineering creativity without the need for expert level knowledge to solve the items. The DPT spans three domains: ability difference and limitations (e.g., assisting people with learning impairments), transportation and mobility (e.g., reducing traffic congestion), and social environments and systems (e.g., improving access to clean water), with participants generating multiple solutions that are then rated for both originality and effectiveness.

The demanding nature of creativity evaluation, requiring expert raters to assess thousands of open-ended responses while maintaining consistent judgment standards, has sparked interest in automating assessment using AI. Recent work has demonstrated that LLMs trained on human creativity ratings can achieve impressive accuracy in evaluating novel responses (Organisciak et al., 2023; Patterson et al., 2025). Like human experts, these models may be influenced by contextual information and examples provided during evaluation. However, these AI evaluators present their own interpretability challenges. While AI may agree with human ratings, we know little about how they arrive at their judgments, what features of responses they attend to during evaluation, or whether those features are similar to those attended to by human experts. Understanding AI creativity evaluation mechanisms for STEM-related tasks has become increasingly urgent as LLMs take on expanded roles in scientific research, from peer review (Huang et al., 2025; Lin et al., 2024) to idea generation (Gu & Krenn, 2024; Si et al., 2024). If these models are to serve as reliable creativity judges, we must better understand their evaluation processes and potential biases. This knowledge could inform both model interpretability efforts and attempts to align AI creativity assessment with human judgment. Moreover, improving evaluation capabilities may enhance idea generation abilities, consistent with cognitive models that link these processes (Smith et al., 1995).

The present research examines how human experts and LLMs evaluate STEM creativity through real-world design problems. In two studies, we conduct a fine-grained analysis of the factors influencing creativity assessment in STEM domains. First, we examine how human experts evaluate originality in STEM solutions, analyzing both their numerical ratings and their written explanations to understand the cognitive processes involved. We also investigate how providing examples impacts expert judgment, testing whether contextual information (example design solutions and originality ratings) changes how experts weigh different aspects of creativity. Second, we perform a parallel analysis of LLMs, examining whether these models show similar patterns in their creativity evaluations and exploring potential differences between human and AI assessment strategies.[1]

Study 1: Human Creativity Evaluations

Our first study sought to understand the key factors underlying human expert evaluation of the creativity of solutions to design problems (DPT) items. A participant in this task is given a scientific or engineering problem (e.g., increasing the use of renewable energy) and is instructed to come up with as many novel solutions to the problem as they can think of. Similar to expert-level science, the best solutions are both original and feasible, though unlike other STEM assessments the DPT benefits from but is not contingent on expertise to come up with creative ideas. The greater complexity of DPT responses compared to those from other creativity tests and its relationship to scientific creativity more broadly makes it a strong choice for our analysis. Unlike prior studies, which often have experts rate only the originality or quality of products, we instead ask our raters to provide fine-grained assessments of cleverness (whether the solution is insightful or witty), remoteness (whether the solution is "far" from everyday ideas), and uncommonness (whether the solution is rare, given by few people) in addition to originality, each of which is thought to influence ratings of creativity (Silvia et al., 2008). These assessments are performed both with and without the presence of example creativity ratings to DPT items, enabling us to examine how added context affects the evaluation process. Finally, we ask experts to briefly explain their originality scores, enabling us to employ methods from computational text analysis to probe the cognitive processes experts employ when rating and how such processes may be modulated by added context.

Methods

We use the data from Patterson et al. (2025), who obtained more than 7000 responses to DPT items from undergraduate STEM majors. Each response was rated for originality using a five-point Likert scale by at least three expert raters with formal training in engineering. We drop items that did not obtain at least one rating from every point of the scale (certain items never had a response that received a five). We convert Likert scores into factor scores, as this has been shown to provide more accurate creativity ratings (Silvia, 2008), and we treat these factor scores as the ground truth originality scores of each response.

We recruit 80 participants on Prolific to provide finegrained creativity ratings to DPT responses, requiring that they have a bachelor's degree or higher in a STEM field and are fluent in English. We split participants into two conditions: a *no example* condition where participants are given responses to rate without any additional context, and an *example* condition where participants are first shown example solutions

[1]Code, data, and supplementary materials are available at: https://github.com/Beaty-Lab/CogSci-2025-Scientific-Creativity

with originality scores for responses to the same prompt being rated. We pull three example solutions from the same dataset while ensuring that participants never rate them. We include a solution with a score of one, one with a score of three, and one with a score of five, to avoid biasing participants towards either end of the scale. We first have each participant rate for originality following the same procedure, instructions, and facet definitions as Patterson et al. (2025). After rating originality, participants in both groups then provide 1-2 sentences explaining their rating process (Orwig et al., 2024), and they finish by rating the uncommonness, remoteness, and cleverness of the response using a five-point Likert scale for each. We instruct participants to be specific in their explanations, to draw on their domain expertise as holders of a STEM degree, and to avoid overly simplistic explanations (e.g., "it's not original" or "it's an obvious answer"). We define a good explanation as being at least one sentence long and including specific details from the participant's prior experience, the response, or the examples (if applicable). We also provide definitions of uncommonness, remoteness, and cleverness for the final rating task, emphasizing that each facet is related while being distinct from originality. We include educational background and AI use checks at the end of the survey.

We administer each participant 15 DPT responses at random. To encourage high-quality explanations, we offer $20 per hour to complete a 30-minute study. We exclude participants with an approval rating of less than 90%, who report using AI to complete the task, or who report an education level lower than the minimum specified on Prolific. We also exclude participants who were exceptionally slow or fast (with a completion time further than three standard deviations from the mean), who gave the same rating for every response, or who did not follow our instructions for formatting explanations (as checked by a research assistant). This resulted in a final sample size of 37 participants and 481 ratings in the example condition and 35 participants and 455 responses for the no example.

When examining the participants' explanations, we employ an analysis plan similar to Orwig et al. (2024), who used LIWC to analyze explanations of originality scores for AUTs. However, recent work has found that LLMs can predict psycholinguistic features of text more strongly than LIWC, even zero shot (Rathje et al., 2024). Therefore, we use LLMs to automatically rate linguistic markers in the explanations. We instruct LLMs to rate for the following variables:

- *Past/future expressions*: Is the explanation past-focused or future-focused in its evaluation of the response?
- *Perceptual details*: Does the explanation focus on the process of perceiving ("observe", "seen", "heard", "feel", etc.)?
- *Causal/analytical*: Does the explanation involve a structured evaluation of the response, evidencing an analytical process, or is the explanation more intuitive in its justifications?
- *Comparative*: Does the explanation make explicit references to standards or examples or compare the response to other ideas?
- *Cleverness*: Does the explanation refer to the cleverness, wittiness, shrewdness, or ingenuity (or lack thereof) of the response?

Both past/future language use and perceptual details have been explored to assess cognitive strategies employed on other creativity tests (Orwig et al., 2024). We elect to use causal/analytical, comparative, and cleverness linguistic markers to aid in assessing whether participants employed a more structured process — which might be evidenced by causal/analytical or comparative language use — or a more intuitive process, as evidenced by language indicating sensory experiences or other "gut reactions" (e.g, "it feels like a clever idea"). These linguistic markers also map onto the finegrained facets participants were asked to rate, with cleverness language mapping onto cleverness and comparative language mapping onto remoteness and uncommonness (as both remoteness and uncommonness often require making references to prior solutions). We use both CLAUDE-3.5-SONNET[2] and GPT-4O[3] to check for reliability in ratings and avoid biases specific to a single LLM, though due to space constraints we mainly report results from GPT-4O as this is the model Rathje et al. (2024) validated. To encourage deterministic output, we set the temperature for both models to 0 and top P to 1. We instruct LLMs to rate each facet and provide a binary evaluation of whether the explanation does or does not contain the feature. Prompts are provided in the supplementary materials.

Results

We begin by examining inter-correlations among all facets (cleverness, remoteness, uncommonness) and correlations between each facet and originality for both conditions. Results are in Figure 1. As expected, each facet is moderately correlated with originality as well as each other, with Pearson r in the range 0.45–0.67 (all correlations are significant).[4] Comparing the example to no example conditions, we see an increase in correlation between originality and cleverness and a decrease in correlation between originality and both remoteness and uncommonness. Changes in correlation across conditions were significant for cleverness-remoteness (Fisher's z = 2.83, $p < 0.01$), remoteness-uncommonness (z = -4.61, $p < 0.001$), and remoteness-originality (z = -2.96, $p < 0.01$), but were insignificant for all other comparisons. Notably, the presence of the examples did not make experts significantly more accurate in their evaluations of originality, with correlations in the moderate range for both conditions (no example r = 0.44, example r = 0.47).

[2]https://www.anthropic.com/news/claude-3-5-sonnet

[3]https://openai.com/index/hello-gpt-4o/

[4]Results from all correlational analysis in both studies were similar using Spearman ρ.

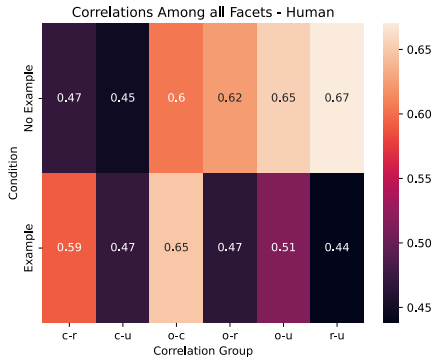

Figure 1: Pearson correlations among pairwise Likert ratings for both conditions. o = originality, c = cleverness, u = uncommonness, r = remoteness.

Turning to participant explanations, GPT-4O's ratings did not reveal significant differences per condition for perceptual details, past/future language use, or cleverness, but differences are significant for both causal/analytical language (Mann-Whitney U = 78039.5, $p < 0.05$) and comparative language (U = 75627.5, $p < 0.01$) with the example condition using less comparative and causal/analytical language than the no examples. Distributions for linguistic markers are shown in Figure 4. CLAUDE-3.5-SONNET's ratings generally agreed with GPT-4O (Cramer's V in the range 0.549–0.798) with the only notable departure being that CLAUDE-3.5-SONNET found no significant difference in causal/analytical language between the conditions (U = 75076.5, $p < 0.5$). We report additional linguistic marker analysis in the supplementary materials.

Discussion

As expected, the facet ratings did not correlate above 0.67 for any pair, implying that participants at least partially distinguished among each facet when assessing originality. Further, correlations changed by a significant degree when including example ratings, with both remoteness and uncommonness becoming weaker predictors of originality and cleverness becoming a stronger one. Given that participants in the no example condition needed to actively retrieve example solutions from memory when evaluating, a possible explanation is that this retrieval process biased them towards placing stronger emphasis on the remoteness and uncommonness of the response in relation to solutions they had seen in the past, while example participants would instead have these cognitive resources free for other aspects of evaluation. Notably, participants in both groups did not differ significantly in terms of education, making it unlikely this effect could be explained as a skill confound. The idea that participants in the example condition were biased toward cleverness rather than the other facets was also partially supported by their explanations, as no example participants used significantly more comparative language than example participants. Given that assessing remoteness or uncommonness often requires making direct comparisons to prior solutions, it makes sense that an evaluation rooted around these facets would contain more comparisons than an evaluation rooted around cleverness, which is more readily evaluated in isolation (e.g., whether the idea is resource efficient, not immediately obvious, etc.).

Study 2: LLM Creativity Evaluations

Our analysis thus far has shed light on how experts reason about originality when taking on the role of creativity evaluators and how this is affected by the inclusion of context. But do these same effects hold for LLMs when they are used to evaluate creativity? Conceptually, the examples given to humans serve a similar role as few-shot learning, a well-established method of improving the accuracy of LLMs on classification tasks (Brown et al., 2020). Yet exemplars may also bias models in other ways, especially in technically complex domains like science and engineering where the nature of truly creative ideas can be complex and difficult to discern a priori (Schmidt et al., 2011; Simonton, 2004). Scientific evaluations given by LLMs also tend to be markedly different from humans, with LLMs often overestimating the quality of scientific research (Schmidgall et al., 2025). Given the increasing role generative AI is having both in expert-level science and in STEM education, it is crucial to perform a similar finegrained analysis of LLM creativity evaluations to enable a head-to-head comparison between human experts and LLMs as evaluators. Our second experiment sought to perform this comparison, using the same methods as in the first experiment but using ratings from LLMs.

Methods

We use CLAUDE-3.5-HAIKU and GPT-4O-MINI, as these LLMs tend to achieve competitive performance on AI benchmarks while also being cost-effective (Chiang et al., 2024). Further, because we use the larger variants of OpenAI and Anthropic models to rate explanations (CLAUDE-3.5-SONNET and GPT-4O), we chose to use these smaller variants for this study to avoid possible biases from LLMs recognizing their own output (Panickssery et al., 2024). We set the temperature to 0 and top P to 1 for all trials while leaving other hyperparameters at their defaults. We structure our prompt similar to the instructions given to the human participants. We instruct the LLM to rate originality, cleverness, remoteness, and uncommonness and to explain its originality score. LLMs are given no exemplars in the no example condition and are given the same examples as humans in the example condition. We administer the same datasets for both conditions as we used in the first experiment, including duplicate archival responses, to make results from humans and LLMs as comparable as possible.[5] We include our prompts in the supplementary materials.

[5]Note that, even with temperature set to zero, these LLMs may generate different ratings for the same response.

Results

LLM originality predictions correlated strongly with the ground truth, and examples significantly boosted this correlation (no example: r = 0.6, 0.67; example: r = 0.74, 0.76; all correlations significant). CLAUDE-3.5-HAIKU and GPT-4O-MINI exhibited strong agreement in their ratings, with correlations between their facet scores in the range 0.73–0.88. Figure 2 summarizes facet correlations for GPT-4O-MINI. Cleverness was the weakest predictor of originality scores in the no example condition, with remoteness and uncommonness being much more strongly correlated with originality. However, this effect dissipated in the example condition, with the strength in correlation between originality and each facet increasing significantly (all changes in correlations were found significant using Fisher's z test). To account for the possibility of these correlations being influenced by chance agreement among the facets, we also performed the same comparisons using Cohen's Kappa and obtained similar findings, detailed results are in the supplementary materials. We compare the distributions of human and GPT-4O-MINI originality scores across both conditions in Figure 3. We find that LLMs rarely use a five and rate a majority of responses as a two.

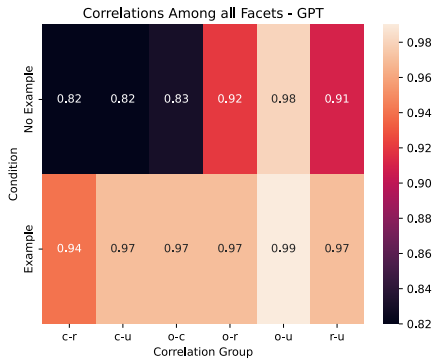

Figure 2: Pearson correlations among pairwise Likert ratings for GPT-4O-MINI in both conditions. o = originality, c = cleverness, u = uncommonness, r = remoteness.

Figure 3: Human and GPT-4O-MINI originality scores.

Figure 4 compares human and GPT-4O-MINI explanations. For GPT-4O-MINI, both GPT-4O and CLAUDE-3.5-SONNET found no significant difference in perceptual details per condition, but did find significant differences in use of cleverness language (Claude U = 106495, $p < 0.001$; GPT U = 1031445.5, $p < 0.001$), with models producing conflicting ratings for the remaining markers (results were similar for CLAUDE-3.5-HAIKU). Comparing LLMs to humans, we find more variability in the presence or absence of all linguistic markers in humans as opposed to LLMs, with LLMs having much more heavily skewed rating distributions and sometimes having a marker completely absent across all explanations, which never occurred in humans. LLM explanations also tended to follow a more rigidly analytical structure, implying a more structured evaluation compared to humans, who had more instances of perceptual details that implied a more intuitive process drawing on memory. Results for CLAUDE-3.5-HAIKU are similar and are included in the supplementary materials.

Discussion

Our findings highlight key differences in how LLMs reason about creativity as compared to humans. While human judges appeared to more strongly associate originality with cleverness, the opposite pattern emerged in LLMs, where originality was more strongly associated with remoteness and uncommonness. Though one might be inclined to trust the LLM originality evaluations more, given their stronger correlation with the true ratings, this should be balanced against the LLMs' apparent inability to distinguish among cleverness, remoteness, or uncommonness while rating. Human facet correlations were consistently much weaker than they were for LLMs, which reflects the conceptualization of these facets as impacting originality while not being the same construct. This lies in stark contrast to the LLMs' much stronger facet correlations — in some cases nearing perfect correlation with originality in the example condition. It is especially noteworthy that differences in facet correlations appeared washed out by the examples, even as model predictions became more accurate. While LLM originality scores had stronger predictive validity, they also had weaker construct validity due to a homogenization of cleverness, remoteness, and uncommonness, which unlike for humans was only strengthened by the presence of the examples.

Our analysis of LLM explanations pointed to similar discrepancies between humans and AI. LLMs exhibited less diversity in explanation styles, with much more heavily skewed distributions of linguistic markers as compared to humans. This mirrors both the distributions of Likert originality scores, and similar findings in other areas of creativity and the social sciences, where generative AI has been found to suffer from less diversity in generations (Park et al., 2024). This trend may have partially been driven by our use of a low temperature value, though we note that exact duplicate explanations were uncommon even for the same problem (especially for GPT-4O-MINI), making it unlikely this is the sole

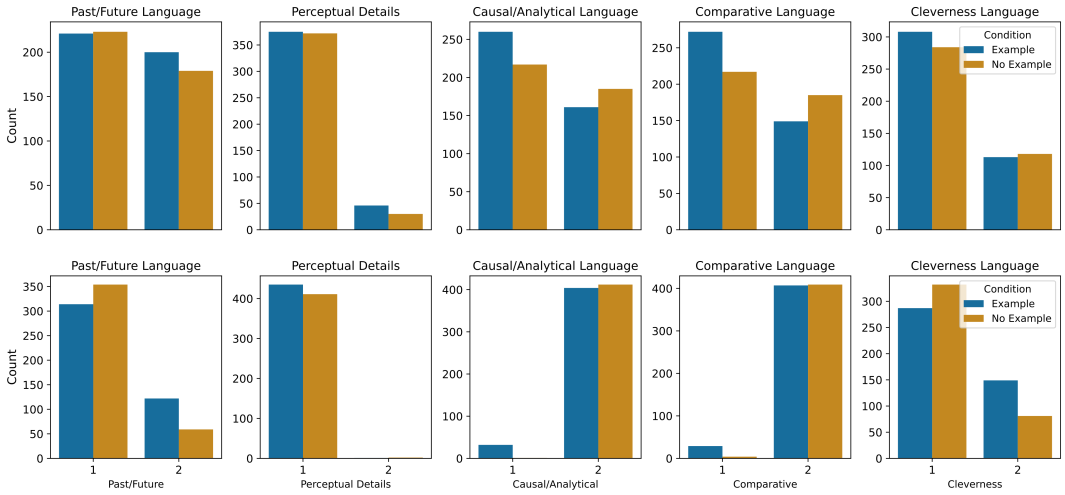

Figure 4: Comparison between linguistic marker use from humans (top) and GPT-4O-MINI (bottom), as assessed by GPT-4O. A rating of 1 indicates the feature is absent in the response, 2 indicates it is present.

reason for redundancy. LLMs did align with humans along several linguistic markers, with both tending to not employ perceptual details and using less future-focused language in the no example condition as compared to the example condition, though not all such trends were statistically significant and LLMs did not closely align with humans for the majority of the linguistic markers explored. It appears that, like with humans, the examples qualitatively impacted LLM explanations, leading to shifts in the content of the explanations as opposed to the no example, which is an important consideration when LLMs are used to evaluate creativity in the real world.

Conclusion

Understanding how AI reasons about the creativity of products and whether that reasoning process aligns with human experts becomes ever more paramount as AI assumes the role of creativity evaluators. Although many works have studied LLMs' ability to accurately rate originality across various tasks (Lin et al., 2024; Schmidgall et al., 2025), the coarse-grained nature of such evaluations makes drawing conclusions about their rating process difficult. We contribute to this literature by collecting finegrained originality evaluations of responses to science and engineering prompts from both human experts and LLMs. Our analysis reveals substantial differences in how these populations both rate for originality in relation to other facets and in how they structure explanations of their rating process. It appears that how LLMs are affected by context and how they rate individual facets of originality is markedly different from human judges, carrying important implications for deploying LLMs as evaluators. Notably, these differences point to potential validity issues with the LLM scores, with AI producing vastly different facet correlations despite more accurately predicting ground truth originality scores. While it is possible that such differences may fade as new LLMs are developed, identifying and understanding these discrepancies remains crucial for advancing work in automated creativity scoring, given that current LLMs already excel at creativity scoring at a coarse-grained level (Organisciak et al., 2023; Luchini et al., 2025).

Future work can expand on our contributions in several ways. Due to budget constraints, we were unable to run multiple pairwise comparisons between humans and LLMs under multiple instruction sets and prompt variations, which is an important step for quantifying the sensitivity of LLM ratings to the input prompt. Though we mitigated this by using multiple LLM evaluators, it remains possible that our results may have been driven in part by the structure of the prompt. Similarly, due to time constraints in human studies, we selected only a small number of examples and were unable to run studies that varied either the examples themselves or their order of presentation. It is possible that this could have biased LLM judgments of cleverness, uncommonness, and remoteness, given that we were unable to include multiple examples at each originality rating. Though we do not believe this could fully explain the homogenization effect, since correlations among the facets were already stronger than in humans zero shot, it remains an important analysis to help us further understand this effect.

Creativity is often considered one of the most critical skills to master in modern economies (Illéssy & Makó, 2020; Tsegaye et al., 2019), and AI must be developed to evaluate it both accurately and fairly. Achieving this goal requires understanding how AI arrives at creativity judgments and whether it rates individual facets of creativity in the same way as humans. We hope our work provides insights into how AI reasons about creativity and serves as a call to action to perform similar finegrained assessments of creativity in other domains.

Acknowledgments

R.E.B. is supported by grants from the National Science Foundation [DRL-1920653; DRL-240078; DUE-2155070].

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, *43*(5), 997.

Boiko, D. A., MacKnight, R., & Gomes, G. (2023). Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... others (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Cseh, G. M., & Jeffries, K. K. (2019). A scattered cat: A critical evaluation of the consensual assessment technique for creativity research. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 159–166. Retrieved from https://doi.org/10.1037/aca0000220 doi: 10.1037/aca0000220

D'Arcy, M., Hope, T., Birnbaum, L., & Downey, D. (2024). Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, *9*(1), 35–40. Retrieved from https://doi.org/10.1037/a0038688 doi: 10.1037/a0038688

Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, *98*(4), 611–625.

Gu, X., & Krenn, M. (2024). Interesting scientific idea generation using knowledge graphs and llms: Evaluations with 100 research group leaders. *arXiv preprint arXiv:2405.17044*.

Huang, S., Huang, Y., Liu, Y., Luo, Z., & Lu, W. (2025). Are large language models qualified reviewers in originality evaluation? *Information Processing & Management*, *62*(3), 103973.

Illéssy, M., & Makó, C. (2020). Automation and creativity in work. *Intersections*, *6*(2), 112–129.

Lin, E., Peng, Z., & Fang, Y. (2024). Evaluating and enhancing large language models for novelty assessment in scholarly publications. *arXiv preprint arXiv:2409.16605*.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Luchini, S. A., Maliakkal, N. T., DiStefano, P. V., Laverghetta Jr, A., Patterson, J. D., Beaty, R. E., & Reiter-Palmon, R. (2025). Automated scoring of creative problem solving with large language models: A comparison of originality and quality ratings. *Psychology of Aesthetics, Creativity, and the Arts*.

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

Orwig, W., Beaty, R. E., Benedek, M., & Schacter, D. L. (2024). Creative evaluation: The role of memory in novelty & effectiveness judgements. *Creativity Research Journal*, 1–9.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744.

Panickssery, A., Bowman, S. R., & Feng, S. (2024). Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 1–17.

Patterson, J. D., Pronchick, J., Panchanadikar, R., Fuge, M., van Hell, J. G., Miller, S. R., ... Beaty, R. E. (2025). *Cap: The creativity assessment platform for open tests and automated scoring.* (Manuscript under review)

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., ... Barsoum, E. (2025). Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Schmidt, A. L., et al. (2011). Creativity in science: Tensions between perception and practice. *Creative Education*, *2*(05), 435.

Si, C., Yang, D., & Hashimoto, T. (2024). Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Silvia, P. J. (2008). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. *Personality and Individual differences*, *44*(4), 1012–1021.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., ... 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.

Simonton, D. K. (2004). Creativity in science: Chance, logic, genius, and zeitgeist. *Cambridge Univ Pr*.

Smith, S. M., Ward, T. B., & Finke, R. A. (1995). *The creative cognition approach*. MIT press.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Tsegaye, W., Su, Q., & Malik, M. (2019). The antecedent impact of culture and economic growth on nationscreativity and innovation capability. *Creativity Research Journal*, *31*(2), 215–222.

Wang, J., Yang, S., & Long, H. (2025). Behind the scores: Unraveling rater judgment in subjective creativity assessments. *Psychology of Aesthetics, Creativity, and the Arts*.

Wang, Q., Downey, D., Ji, H., & Hope, T. (2024, August). SciMON: Scientific inspiration machines optimized for novelty. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.18/