

Evaluating LLMs' Divergent Thinking Capabilities for Scientific Idea Generation with Minimal Context

Corresponding Author: Professor Hao Sun

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

- What are the noteworthy results?

This paper presents a new evaluation framework to test the divergent capabilities of LLMs in the context of scientific discovery. It addresses a huge gap in evaluation, which is primarily focused on intelligence and “convergent” abilities. The framework is based on theoretical discussions on creativity and divergent thinking in science, but implemented in such a way that the entire process is fully automatized. Its significance is demonstrated by running this evaluation framework on a great variety of state-of-the-art LLMs, showing how divergent capabilities are not directly correlated with convergent ones and deserve proper treatment.

- Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.

The work will be of significance to the field. It provides a new framework to evaluate LLMs from an underrepresented perspective; while its focus is on scientific ideas, I believe it can also be of inspiration for other domains soon. While based on well-known theories of creativity and similar in spirit to other divergent tasks, it provides an original and valuable contribution to the literature.

- Does the work support the conclusions and claims, or is additional evidence needed?

The work mainly supports the conclusions and claims, even if some issues should be addressed. Additional evidence should be provided both in terms of complementary experiments (see the methodology question) and presentation of current results (see the data analysis question); I will expand this discussion below.

- Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?

In my opinion, the data analysis and presentation lack some important statistical details. In particular, all the figures but Fig. 6 report average performances without including standard errors or confidence intervals. Moreover, I suggest reporting the number of keywords per category (which is available only in the code); Fig. 3 discusses scores across various disciplines, but we cannot know whether the number of keywords per discipline is sufficiently high to be statistically significant. Finally, the results in Fig. 4 should deserve a greater discussion, as they are somehow unexpected. There exists a huge literature arguing that current LLMs are good in general intelligence tasks but not in creative ones; here, results from LiveBench and LiveIdeaBench suggest the opposite. In addition, it seems that the variance of scores is very limited: while we can observe scores between 21 and 82 (and the averages between 39 and 66) for LiveBench, for LiveIdeaBench scores are between 56 and 91, and the averages between 66 and 77.

- Is the methodology sound? Does the work meet the expected standards in your field?

The methodology is generally sound; however, I believe two main issues should be addressed before publication.

The first is about the theoretical aspects behind the article: some important literature has not been discussed, and certain claims are too strong and should be more nuanced. Firstly, it is important to remark that scientific creativity is not the same as

divergent thinking. Of course, divergent thinking is a crucial aspect of (scientific) creativity, but it is not the only one; creativity also requires domain-specific skills and a creative result should also be valuable and not only original. While this is somehow considered in the assessment through the "feasibility" component, I believe it is important to avoid using them as synonyms and clarify that what this framework aims to evaluate is only a portion of aspects of creativity. Similarly, the discussion about "human creativity research" lacks some fundamental theories, such as Boden's (e.g., "The Creative Mind: Myths and Mechanisms"); Amabile's (e.g., "The Social Psychology of Creativity: A Componential Conceptualization"); or Rhodes' (see An Analysis of Creativity). I understand that a complete treatment of creativity would require separate work, but I still believe this paper should correctly mention and discuss them in relation to the proposed framework.

The second issue is about the practical consequences of using LLMs as evaluators. While using LLM-as-a-Judge techniques help automatize evaluation, some issues have to be carefully considered, e.g., biases induced by the evaluating model and correlation between human and AI-based judges. As correctly pointed out in the article, we still have to demonstrate the soundness of LLM-based evaluation approaches applied to creative tasks; I was surprised not to see any study about whether the proposed evaluation approach correlates with judgment from humans (and especially experts in scientific domains). In addition, asking a model trained to "converge" on training data to evaluate "divergent" properties seems to me counterintuitive: intuitively, what the model will consider mostly original is what is similar to what has been judged as original in the past, that is, something that is not original anymore. I believe that the article should expand on this expected tension, and should also provide more empirical data in defense of using LLM-based evaluation for divergent thinking.

As an additional point, the use of LLM-as-a-Judge for fluency is not sufficiently motivated in my opinion. There are various diversity metrics considering either the syntax or semantics of text that can be used to evaluate potential overlapping without relying on LLMs (which requires more computing resources and presents the issues of LLM-as-a-Judge discussed above). Moreover, the fluency scores are obtained in a 0-3 scale, and then translated into a 0-10 scale: how is that exactly done? Does this "discretization" (as apparent from Fig. S.3) affect the overall scores?

- Is there enough detail provided in the methods for the work to be reproduced?

There are enough details in the Methods and Supplementary Information for the work to be roughly, but not exactly, reproduced (some fine-grained implementation details, e.g., about how refused responses are managed and text is generated, are not reported).

(Remarks on code availability)

Unfortunately, the code does not provide enough instructions in its README for installing and running the application. The repository does not contain any requirements.txt file or any instruction to get a functioning Docker container, so it is not intuitive how to reproduce the results. Moreover, the code correctly contains a .py file to instantiate and run LLMs and process their outputs, as well as a notebook to analyze the results, but no main .py file is available, thus impeding any correct reproduction of the results.

Reviewer #3

(Remarks to the Author)

The manuscript introduces LiveldeaBench, a benchmark for evaluating the scientific creativity of large language models (LLMs) under minimal contextual conditions. The study builds upon Guilford's creativity theory and assesses models across originality, feasibility, fluency, and flexibility. The authors analyze 20 models using a dataset of 1,180 scientific keywords, highlighting the independence of creativity from general intelligence metrics.

While the research addresses an important and underexplored aspect of LLM capabilities, there are key concerns regarding the theoretical grounding, methodological robustness, and clarity of presentation that should be addressed.

Major Comments

1. Lack of References in Introduction (p. 1-3)

The introduction is missing references to support claims, especially regarding the importance of creativity in scientific discovery and the limitations of existing LLM evaluation frameworks. Authors should cite relevant work in AI creativity assessment (i.e., work by Janet Rafner as a starting point, but there are multiple papers on AI creativity from theoretical and empirical research that must be addressed here) and cognitive theories of creativity (i.e., work by Roger Beaty, Mathias Benedek, Adam Green, Yoed Kenett, etc.).

2. Motivation for the Research (p.2)

The justification for LiveldeaBench is not entirely convincing. The manuscript argues that general intelligence and scientific creativity are independent, but the need for a new benchmark should be clarified.

3. Divergent vs. Convergent Thinking

The study positions LiveldeaBench as assessing divergent thinking but does not convincingly differentiate it from convergent thinking. Please explain why the chosen methodology isolates divergent thinking. Perhaps reference studies that contrast divergent and convergent thinking and also in AI.

4. Claim that Convergent thinking is secondary in creativity (p.3)

This point has since been challenged, for example Cortes et al. (2019) claim that creative measures involve mixtures of the two constructs together.

Cortes, R. A., Weinberger, A. B., Daker, R. J., & Green, A. E. (2019). Re-examining prominent measures of divergent and

convergent creativity. *Current Opinion in Behavioral Sciences*, 27, 90-93. <https://doi.org/10.1016/j.cobeha.2018.09.017>

5. IQ and LLM Intelligence (p.3)

The paper compares LLM intelligence to human IQ, which is at the very least problematic since IQ is not designed for LLMs. The authors should remove direct comparisons to human IQ and instead compare LLMs based on creativity-relevant cognitive benchmarks.

6. Anthropomorphic conclusions

The authors should avoid claims such as LLM's are creative or intelligent, rather they may exhibit signs of human-like creativity or intelligence. Also please reference current literature on whether LLM's are truly intelligent or creative and how that influences our creativity, as well as the dangers of LLM hallucinations.

Some relevant literature:

Marrone, R., Cropley, D., & Medeiros, K. (2024). How Does Narrow AI Impact Human Creativity?. *Creativity Research Journal*, 1-11.

Wenger, E., & Kenett, Y. (2025). We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint arXiv:2501.19361*.

Cropley, D. (2023). Is artificial intelligence more creative than humans? : ChatGPT and the Divergent Association Task. *Learning Letters*, 2, 13. <https://doi.org/10.59453/ll.v2.13>

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is Chatgpt a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning*, 40(2), 919–930. <https://doi.org/10.1111/jcal.12929>

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1-55.

7. Confounding Factors in Creativity Measurement (p.11)

The study does not sufficiently address potential biases in creativity scoring, particularly in relation to model training data. For example the multiple evaluator methods, may involve models with similar training data, such that they are not as independent as we may think.

8. Lack of Human Evaluation

The reliability of the automated creativity ratings is unclear without comparison to human judges. Perhaps include a validation study where human raters assess a subset of generated ideas to compare against LiveIdeaBench scores or examples of the reliability of these models from other research.

Minor Comments

1. The manuscript refers to LiveIdeaBench both as a "framework" and a "benchmark," leading to potential confusion. Please define and consistently use one term throughout the paper.

2. Figures are not always referenced in the text at appropriate locations (e.g., Fig. 1 lacks in-text reference). Ensure each figure is clearly cited where relevant.

3. The claim that idea length is independent of quality is supported with an R^2 value (0.003) as correlation strength, but this is a measure of explained variance rather than correlation.

4. Several claims in the related work and discussion sections lack citations, making it difficult to verify assertions.

5. Some LLMs evaluated are proprietary and not widely accessible, which could affect reproducibility. Discuss limitations regarding access to closed-source models.

6. Some figures (e.g., Fig. 7) may be redundant or better integrated into supplementary materials.

Recommendation

The study presents an interesting and novel contribution to evaluating LLM creativity, but major revisions are needed before it can be considered for publication. Specifically:

- Strengthen the theoretical justification for LiveIdeaBench, in the field of creativity and in general.
- Clarify the theoretical distinction between divergent and convergent thinking, and ground claims in current literature regarding AI abilities in these fields.
- Address concerns regarding the absence of human evaluation as a benchmark comparison in this study as well as potential biases in LLM scoring and the multiple evaluator approach.
- Improve and add referencing and clarity in data interpretation and terms

With these improvements, the manuscript could make a valuable contribution to field of AI-driven creativity assessment and scientific idea generation.

(Remarks on code availability)

This review process was conducted together with my PhD student, Tuval Raz. I do not see an option below but I would be greatly interested in having him also officially recognized for his review of this ms, officially by the journal and via ORCID which he has. Please be in touch to verify that this will happen.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)

Firstly I want to thank the authors for their timely and relevant paper. It represents original research which will contribute to the field and I will eventually be published. However, I note some significant aspects which I feel require additional work before publication.

Overall, I think that the article addresses an important topic and the methodology seems sound (apart from a small comment I present later on), however I think that the article need to work on sharpening the framing. The article presents that models have achieved super human performance across disciplines, and presents the fundamental question “does the creative potential of LLMs grow in tandem with their analytical capabilities, particularly in the realm of scientific discovery, however it does not address if or why we want our LLMs to make such discoveries alone or rather if they should be made as part of a tight human in the loop approach, as is recommended by the fields of hybrid intelligence and human centered AI. How often do we expect to use LLMs alone for scientific discovery? The authors allude to this in the second to last paragraph of the discussion when they state that their findings extend beyond model evaluation to the broader landscape of AI assisted scientific discovery (a section that is completely un referenced). However, the article would greatly benefit from a position that perhaps the benchmarking is needed as an initial step in developing such human-AI systems, or if the authors take the stance that AI scientific discovery alone (AGI with no human involvement) is beneficial then to preface that.

For more references on human centered AI and hybrid intelligence and hybrid intelligence and generative AI in creativity tasks I suggest the following references:

Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., ... & Welling, M. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8), 18-28.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637-643.

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

Additionally Falk et al., 2025 propose the idea of instead of LLM as a judge, it is LLM as an (additional judge), leveraging humans and AI in synergy.

<https://arxiv.org/pdf/2503.04290>

Now I will go through the sections one by one with my comments

Abstract

The language in the abstract should be improved for clarity. In the first sentence I would appreciate examples of the scientific tasks as well as what is meant by performance. The meaning behind the sentence “we reveal that scientific creativity shows distinct patterns from general intelligence metrics” is not clear to me. Additionally “suggest that the development of creative capabilities in LLMs may follow different trajectories than traditional problem solving” is too vague.

Introduction

Overall the introduction seems quite light on references, and of the references in this manuscript, half of them are arxiv preprints. I am acutely aware of the issue of publishing work with the rapidly moving field that is genAI and LLMS, and I myself often face this dilemma of needing to reference the most up to date work which is usually only on arxiv. However, I encourage the authors to review some of my suggested literature and to incorporate more peer reviewed articles into the manuscript.

Figure 1 is confusing to me. I understand the issues with page limitations and trying to put multiple figures together into one figures, however I am confused if the entire figure should be seen as a flow chart, following a logical order or if they are distinct separate branches. Additionally it is confusing that in part b it shows 3 llms but in the text it says the evaluation panel was done by 10 llms

Related work

The authors mention the threshold theory, but they do not acknowledge that there is quite a debate about whether this theory actually holds. The article they cite (Jauk et al 2013) only shows part of the story and has been addressed in more recent work. If the article is situated in a comparison of intelligence vs. creativity, i believe it is important to present a more accurate and nuanced discussion on the topic. Additionally the authors would benefit from citing specific literature on intelligence and creativity in scientific creativity.

Excerpt from Benedek, M. (2024). On the relationship between creative potential and creative achievement: Challenges and future directions. *Learning and Individual Differences*, 110, 102424.

Specifically, the threshold hypothesis assumes that high intelligence (e.g., IQ above a threshold of 120) represents a necessary condition for high creativity and that the influence of intelligence should become lower or vanish above this intelligence threshold (Getzels & Jackson, 1962; Kim, 2005). Research has reported mixed evidence for the threshold effect so far, with some studies finding no support (Kim, 2005; Preckel et al., 2006; Weiss et al., 2020), and others finding support under certain conditions (Gralewski et al., 2012; Jauk et al., 2013). But these works mainly considered creativity in terms of DT ability. Studies focusing on creative achievement actually showed that intelligence is relevant for creative achievement across the entire spectrum of intelligence (Jauk et al., 2013; Park et al., 2008) or becomes even more important beyond the threshold (Runco & Albert, 1986). These empirical findings are generally consistent with the notion of interactive effects, by emphasizing the importance of high creative potential without assuming reduced relevance beyond a given threshold.

Getzels, J. W., & Jackson, P. W. (1962). Creativity and intelligence: Explorations with gifted students.

Kim, K. H. (2005). Can only intelligent people be creative? A meta-analysis. *Journal of Secondary Gifted Education*, 16(2-3), 57-66.

Kim, K. H. (2005). Can only intelligent people be creative? A meta-analysis. *Journal of Secondary Gifted Education*, 16(2-3), 57-66.

Preckel, F., Holling, H., & Wiese, M. (2006). Relationship of intelligence and creativity in gifted and non-gifted students: An investigation of threshold theory. *Personality and individual differences*, 40(1), 159-170.

Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). A reappraisal of the threshold hypothesis of creativity and intelligence. *Journal of Intelligence*, 8(4), 38.

Gralewski, J., Weremczuk, E., & Karwowski, M. (2012). Intelligence and creativity of Polish middle-school students: Looking for the threshold hypothesis. *The New Educational Review*, 29, 328-338.

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, 41(4), 212-221.

Park, G., Lubinski, D., & Benbow, C. P. (2008). Ability differences among people who have commensurate degrees matter for scientific creativity. *Psychological science*, 19(10), 957-961.

Runco, M. A., & Albert, R. S. (1986). The threshold theory regarding creativity and intelligence: An empirical test with gifted and nongifted children. *Creative Child and Adult Quarterly*, 11(4), 212-218.

In the section on Evaluating Creative Capabilities of LLMs the second paragraph is not clear to me. I do not fully understand what framework has been developed and how that relates to this study.

On Page 8 and 9 the authors address the topic of length of responses, however I would also very much like to see something on clarity of responses. Lee and Chung (2024) found that chatgpt assisted humans in producing creative problem solving solutions that were more well articulated than with out ai assistance. This could be see as clarity of the idea, which would also be very interesting in the context of the authors work.

Lee, B.C., Chung, J.(. An empirical investigation of the impact of ChatGPT on creativity. *Nat Hum Behav* 8, 1906–1914 (2024). <https://doi.org/10.1038/s41562-024-01953-1>

The axis on figure 6 and 7 need to be better described. For example the x axis of on figure 6 A just reads "score" and should state score of what.. Similarly 7B just has "average" on the y axis but does not state an average of what..

Lastly there is a growing movement to report the carbon impact of studies which use LLMs. I strongly suggest that the authors do so. I recommend they use the following tool for guidelines: <https://mlco2.github.io/impact/#about>

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Here are some additional references that I recommend the author review to better contextualize their work.

Peer reviewed and published:

M. Krenn, L. Buffoni, B. Coutinho, S. Eppel, J. G. Foster, A. Gritsevskiy, H. Lee, Y. Lu, J. P. Moutinho, N. Sanjabi, et al., Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network, *Nature Machine Intelligence* 5, 1326 (2023).

F. Shi and J. Evans, Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines, *Nature Communications* 14, 1641 (2023).

Q. Wang, D. Downey, H. Ji, and T. Hope, Scimon: Scientific inspiration machines optimized for novelty, Annual Meeting of the Association for Computational Linguistics (2024).

H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al., Scientific discovery in the age of artificial intelligence, *Nature* 620, 47 (2023).

J. Sourati and J. A. Evans, Accelerating science with human-aware artificial intelligence, *Nature Human Behaviour* 7, 1682 (2023).

Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. <https://doi.org/10.18653/v1/2023.acl-long.870>

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. In Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings), Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 361–374. <https://doi.org/10.18653/v1/2023.findings-ijcnlp.32>

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT Rates Natural Language Explanation Quality like Humans: But on Which Scales?. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 3111–3132. <https://aclanthology.org/2024.lrec-main.277>

Arxiv papers:

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.

X. Gu and M. Krenn, Forecasting high-impact research topics via machine learning on evolving knowledge graphs, arXiv:2402.08640 (2024).

J. Baek, S. K. Jauhar, S. Cucerzan, and S. J. Hwang, Researchagent: Iterative research idea generation over scientific literature with large language models, arXiv:2404.07738 (2024).

M. R. AlScience and M. A. Quantum, The impact of large language models on scientific discovery: a preliminary study using gpt-4, arXiv:2311.07361 (2023).

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. arXiv preprint arXiv:2405.01535 (2024).

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

I sincerely thank you for your thorough responses and careful revisions. I am really impressed by the quality of the additional work you have conducted to satisfy all my concerns.

The article has been greatly improved in all of its aspects. From a theoretical standpoint, it now includes relevant additional references, and the scope of the research has been made clearer. The method has been expanded and better clarified (especially considering the new supplementary materials), and the potential limitations have been correctly discussed. Finally, the evaluation is now statistically sound, and I really appreciate the additional study comparing LLM and human evaluation, which, in my opinion, is crucial to demonstrate the validity of the benchmark.

All considered, I have no further requests, and I believe the article is ready for acceptance.

(Remarks on code availability)

The code is now appropriate, with all necessary information to reproduce the results.

As a final suggestion, I believe it would be useful to include all the instructions to add new models in the README file. This would simplify the adoption of the benchmark by developers of future LLMs.

Reviewer #5

(Remarks to the Author)

The authors present a comprehensive evaluation benchmark (LiveIdeaBench) to assess large language models' (LLMs) capacity for divergent thinking in the context of scientific idea generation. The revised manuscript reflects substantial effort to address initial review comments, including strengthening theoretical grounding, expanding model comparisons, incorporating human expert validations, and enhancing the clarity of both methodology and presentation. I am also very happy to see that the authors have included a statement on carbon impact of the work.

I think that at this point to polish the article it would be nice to see if the discussion could be (slightly) expanded to highlight practical applications and next steps:

How can LiveIdeaBench guide future LLM development (e.g., training for idea generation vs problem-solving)?
Could hybrid human-AI frameworks benefit from the insights provided by LiveIdeaBench?

The authors include this statement: Our benchmark provides a foundation for understanding and improving these capabilities, potentially enabling more effective human-AI collaboration [11–13] and informing the design of human-aware AI systems for science [68] in pushing the boundaries of scientific knowledge.

But could benefit from elaborating on how the benchmark could inform design of human-aware AI systems for science.

Additionally this sentence could be supplemented by a short example: The creative strengths we observed in different model architectures suggest that a diverse ecosystem of AI tools, each with complementary capabilities, could support different aspects of the scientific process.

Lastly, this sentence could use a reference, or a specific link back to findings: As LLMs demonstrate increasingly sophisticated idea generation capabilities, they hold promise as powerful tools for accelerating scientific innovation - from hypothesis generation to experimental design.

(optional) I recommend the authors look at these new articles,

DiStefano et al. 2025

Evaluating AI's Ideas: The Role of Individual Creativity and Expertise in Human-AI Co-Creativity.

This article indicates the changing nature of creativity, shifting from generation to judgement.

This new article discusses the adapting role of humans and AI to work together for evaluation of creativity.

Falk et al. 2025. How Do Hackathons Foster Creativity? Towards Automated Evaluation of Creativity at Scale. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-23).

Specifically relevant for the LLMs-as-a-judge section

Overall I recommend this article is accepted after minor revisions

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)

I have been invited to step in as Reviewer 3 for the revised version of your manuscript “LiveIdeaBench: Evaluating LLMs’ Divergent Thinking for Scientific Idea Generation with Minimal Context.” My role is therefore twofold: to evaluate the soundness of the research in its current form, and to assess how well the revisions address the concerns raised by the previous Reviewer 3. I want to acknowledge at the outset that the manuscript has been substantially improved since the initial submission, and I appreciate the considerable effort the authors have put into strengthening both the theoretical foundation and the methodological clarity.

Noteworthy results

The manuscript introduces LiveIdeaBench, a benchmark for evaluating LLMs’ divergent scientific idea generation under minimal context. This represents a valuable and original contribution, filling an important gap in existing evaluation frameworks that focus predominantly on convergent abilities. Expanding the benchmark to 41 models across 22 scientific domains provides impressive breadth. The key empirical finding, that divergent scientific idea generation is largely independent from general intelligence performance, is important for the AI evaluation literature. The addition of human expert validation in the Partial Differential Equations domain, showing strong alignment between human and LLM originality ratings, further supports the robustness of the benchmark.

Significance and relation to literature

This work is significant not only within AI evaluation but also in computational creativity and the study of human–AI scientific collaboration. The revised manuscript now situates the contribution more clearly in relation to established creativity research, including Guilford, Boden, Amabile, and Rhodes, as well as recent work by Beaty, Benedek, Kenett, and Rafner. This broader theoretical grounding was a key concern of the earlier Reviewer 3 and is now adequately addressed. Compared to prior creativity benchmarks such as DAT and CREA, LiveIdeaBench provides a distinct and complementary approach by isolating divergent thinking under minimal input conditions.

Support for conclusions

The main conclusions are supported by the data and analysis. The authors are careful not to overclaim: they now emphasize that LiveIdeaBench assesses a specific aspect of scientific creativity (divergent idea generation), not creativity in its entirety. The discussion of sycophancy, compressed score ranges, and feasibility evaluation challenges adds necessary nuance.

Data analysis and interpretation

The revisions have strengthened the statistical reporting: confidence intervals are now consistently included, keyword counts per discipline are shown, and the fluency scoring has been clarified. The discussion has been expanded to explain unexpected findings (e.g., some LLMs excelling in divergent tasks despite weaker performance on general benchmarks). While human validation is limited to one scientific domain, the authors are transparent about this constraint, and it does not undermine the validity of the broader findings.

Methodology and reproducibility

The methodology is sound and well detailed. Important refinements include the introduction of a “clarity” dimension, a transparent refusal-handling protocol, and clearer explanation of scaling decisions. Reproducibility has been substantially improved through a public repository with full code, requirements, and instructions. Some reliance on proprietary APIs remains, but this is unavoidable and is appropriately acknowledged.

Overall assessment

From my perspective as a new Reviewer, the manuscript has effectively addressed the prior concerns. The theoretical framing is now richer, the statistical analysis more rigorous, the limitations more explicitly acknowledged, and the reproducibility better supported.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

LiveIdeaBench: Evaluating LLMs’ **Divergent Thinking** for **Scientific Idea Generation** with Minimal Context

References: NCOMMS-25-09187-T

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, Hao Sun

We thank the reviewers for their constructive comments and suggestions, which have helped significantly enhance the quality of our manuscript. We have carefully considered and addressed the points raised. Here is a summary of our major revisions:

- **Modified and improved our benchmark:** We have modified our benchmark evaluation mechanism by adding a “clarity” dimension to better align with Guilford’s creativity framework; further refined the Judge LLM selection process to ensure more rigorous selection criteria that minimize bias; and improved the procedure for handling LLM refusals to generate ideas to ensure that the scientific creativity scores of LLMs with stronger safety guardrails are not underestimated.
- **Strengthened the theoretical justification:** We have carefully investigated, cited, and discussed the references recommended by the reviewers in our revised manuscript, which has significantly enhanced the theoretical foundation and conceptual significance of our work.
- **Validated our findings via human evaluation:** We designed and distributed questionnaires to domain experts to investigate the alignment between human judgments and LLM evaluations, providing empirical validation of our benchmark.
- **Expanded the comparison with more LLMs:** We have comprehensively updated our benchmark testing results, expanding our evaluation from the initial 20 models to 41 state-of-the-art LLMs based on LiveBench’s March 2025 results. This represents a significant expansion, incorporating 26 models non-included in our initial submission: `claude-3.7-sonnet:thinking`, `claude-3.7-sonnet`, `claude-3-opus`, `claude-3.5-haiku-20241022`, `o3-mini-high`, `gpt-4.5-preview`, `o1`, `o3-mini`, `gpt-4-turbo`, `gemini-2.0-flash-thinking-exp`, `gemini-2.0-pro-exp-02-05`, `gemini-2.0-flash-exp`, `gemini-2.0-flash-lite-001`, `gemma-2-27b-it`, `qwq-32b`, `qwen-max`, `qwen-2.5-7b-instruct`, `deepseek-r1`, `deepseek-chat (v3)`, `deepseek-r1-distill-llama-70b`, `deepseek-r1-distill-qwen-32b`, `llama-3.1-70b-instruct`, `mistral-small-24b-instruct-2501`, `mistral-small (v2409)`, `nova-lite-v1`, `phi-4`, and `step-2-16k-202411`. This expansion provides more comprehensive data and allows for more detailed analysis across a wider spectrum of current LLMs.
- **Improved the clarity:** The manuscript has been carefully proofread to correct typos and enhance overall readability and flow.

- **Refined the framing and title:** Based on feedback regarding the scope of “scientific creativity”, we have adjusted the terminology throughout the manuscript and revised the title to more accurately reflect the benchmark’s focus on *divergent thinking for scientific idea generation*, aka: “*Evaluating LLMs’ Divergent Thinking for Scientific Idea Generation with Minimal Context*”.

Our responses marked in blue are provided following each individual comment. Changes made in the revised manuscript are indicated in red. The updated leaderboard, codes, and datasets (v2) are found in the following repositories: <https://liveideabench.com/>, <https://github.com/x66ccff/liveideabench>, and <https://huggingface.co/datasets/6cf/liveideabench-v2>.

Reviewer 2

We sincerely appreciate the time and effort you dedicated to evaluating our work and providing constructive comments. Following your suggestions, we have conducted extensive experiments and made substantial revisions to further strengthen our manuscript. Please find our point-to-point responses below.

Comment 2.1 — *What are the noteworthy results?* This paper presents a new evaluation framework to test the divergent capabilities of LLMs in the context of scientific discovery. It addresses a huge gap in evaluation, which is primarily focused on intelligence and “convergent” abilities. The framework is based on theoretical discussions on creativity and divergent thinking in science, but implemented in such a way that the entire process is fully automatized. Its significance is demonstrated by running this evaluation framework on a great variety of state-of-the-art LLMs, showing how divergent capabilities are not directly correlated with convergent ones and deserve proper treatment.

Reply: Thanks for your positive and encouraging feedback!

Comment 2.2 — *Will the work be of significance to the field and related fields? How does it compare to the established literature? If the work is not original, please provide relevant references.* The work will be of significance to the field. It provides a new framework to evaluate LLMs from an underrepresented perspective; while its focus is on scientific ideas, I believe it can also be of inspiration for other domains soon. While based on well-known theories of creativity and similar in spirit to other divergent tasks, it provides an original and valuable contribution to the literature.

Reply: We sincerely value your recognition of our work’s significance and its potential to inspire developments in other domains!

Comment 2.3 — *Does the work support the conclusions and claims, or is additional evidence needed?* The work mainly supports the conclusions and claims, even if some issues should be addressed. Additional evidence should be provided both in terms of complementary experiments (see the methodology question) and presentation of current results (see the data analysis question); I will expand this discussion below.

Reply: Thanks for this excellent remark. Please see our responses below.

Comment 2.4 — *Are there any flaws in the data analysis, interpretation and conclusions? Do these prohibit publication or require revision?* In my opinion, the data analysis and presentation lack some important statistical details. In particular, all the figures but Figure 6 report average performances without including standard errors or confidence intervals.

Reply: Thank you for this excellent suggestion. We have now incorporated 95% confidence intervals throughout our results presentation in the manuscript. This addition provides a clearer

representation of the precision and robustness of our findings. Please see the updated Figures 2, 3, 4, and 6 in the Main text as well as Supplementary Figure S.4 and Supplementary Table S.3.

Comment 2.5 — Moreover, I suggest reporting the number of keywords per category (which is available only in the code); Figure 3 discusses scores across various disciplines, but we cannot know whether the number of keywords per discipline is sufficiently high to be statistically significant.

Reply: Thank you for this suggestion. We have added annotations indicating the number of keywords for each discipline directly in Figure 3 in the Main text. Following your previous recommendation, we have also incorporated 95% confidence intervals for all disciplines. These modifications enhance the transparency of our analysis’s statistical basis and aid readers in assessing the robustness of the observed disciplinary patterns.

Comment 2.6 — Finally, the results in Figure 4 should deserve a greater discussion, as they are somehow unexpected. There exists a huge literature arguing that current LLMs are good in general intelligence tasks but not in creative ones; here, results from LiveBench and LiveIdeaBench suggest the opposite.

Reply: Thank you for your insightful observation. We have expanded our discussion in the Discussion section (page 12 in the Main text) to provide a more comprehensive analysis of these interesting findings.

Comment 2.7 — In addition, it seems that the variance of scores is very limited: while we can observe scores between 21 and 82 (and the averages between 39 and 66) for LiveBench, for LiveIdeaBench scores are between 56 and 91, and the averages between 66 and 77.

Reply: Thank you for pointing out the limited score variance in LiveIdeaBench compared to LiveBench. We agree this is an interesting observation. We believe this is primarily due to the LLM-as-a-judge evaluation mechanism itself, as aligned models tend to exhibit sycophancy—a behavior potentially stemming from alignment training focused on harmlessness, helpfulness, and honesty. This may lead to inflated scores and a compressed score range. This is also evident in Supplementary Figure S.3, where the center of the distribution for various scoring dimensions is not around 5, but rather in the 6-7 range.

However, we believe this compressed range does not undermine the benchmark’s utility. Since LiveIdeaBench functions as a dynamic benchmark, its main purpose is to differentiate between LLMs with stronger scientific idea generation capabilities, thus focusing on relative rather than absolute scores. While the ensemble approach (using multiple judges) aims to mitigate individual model biases, we acknowledge that varying degrees of sycophancy across different judge models could theoretically influence relative rankings to some extent. This reinforces the importance of interpreting these scores primarily in terms of relative model capabilities rather than absolute values, acknowledging this known challenge in LLM-as-a-judge evaluations.

Furthermore, the addition of confidence intervals, as you suggested, reveals narrow 95% CIs (often ± 0.5 points) even when the mean scores are close (see revised Figure 4 in the Main text).

This indicates that LiveIdeaBench scores maintain good discriminative power for ranking models, despite the compressed range.

Nevertheless, we acknowledge that the upper half of Figure 4 in the original manuscript does not show comparisons clearly enough. Therefore, we have improved Figure 4 by implementing dual colorbars. The heatmap for LiveIdeaBench in the upper half and LiveBench in the lower half now uses two separate color schemes. We believe this will make it easier for readers to distinguish the relative high and low scores.

Crucially, we have also incorporated this point into our revised manuscript: the potential for score inflation and the nuances of sycophancy are now explicitly discussed as a limitation within the Discussion section (page 12-16 in the Main text), citing relevant literature (Wei et al., 2024; Gu et al., 2025; Fanous et al., 2025) provided below.

References: (the following number coincides with that in the Main text)

[64] Wei, J., Huang, D., Lu, Y., Zhou, D., & Le, Q. V. (2024). Simple synthetic data reduces sycophancy in large language models. *arXiv*. <https://arxiv.org/abs/2308.03958>

[65] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2025). A Survey on LLM-as-a-Judge. *arXiv*. <https://arxiv.org/abs/2411.15594>

[66] Fanous, A., Goldberg, J., Agarwal, A. A., Lin, J., Zhou, A., Daneshjou, R., & Koyejo, S. (2025). SycEval: Evaluating LLM Sycophancy. *arXiv*. <https://arxiv.org/abs/2502.08177>

Comment 2.8 — *Is the methodology sound? Does the work meet the expected standards in your field?* The methodology is generally sound; however, I believe two main issues should be addressed before publication. The first is about the theoretical aspects behind the article: some important literature has not been discussed, and certain claims are too strong and should be more nuanced. Firstly, it is important to remark that scientific creativity is not the same as divergent thinking. Of course, divergent thinking is a crucial aspect of (scientific) creativity, but it is not the only one; creativity also requires domain-specific skills and a creative result should also be valuable and not only original. While this is somehow considered in the assessment through the “feasibility” component, I believe it is important to avoid using them as synonyms and clarify that what this framework aims to evaluate is only a portion of aspects of creativity.

Reply: Thank you for your insightful comment! You raise a very valid point. We agree that scientific creativity and divergent thinking are not synonymous concepts, and we have revised our manuscript to clarify this important distinction. Based on your suggestions, we have made the following changes:

- In the **Abstract**, we revised our description to: “We introduce LiveIdeaBench, a comprehensive benchmark evaluating LLMs’ scientific idea generation by assessing divergent thinking capabilities using single-keyword prompts” instead of implying our benchmark evaluates scientific creativity comprehensively.

- In the [Introduction section \(page 3 in the Main text\)](#), we clarified: *“While acknowledging that comprehensive scientific creativity involves an interplay of both divergent and convergent processes, LiveIdeaBench focuses on and evaluates the foundational, generative phase: the ability to produce diverse scientific concepts from sparse cues. This capacity is critical for sparking innovation but is often overlooked by conventional benchmarks. Thus, our focus is on assessing this specific aspect of creative potential relevant to scientific discovery, rather than evaluating the entire scientific process, which also involves deep domain expertise and subsequent validation.”*
- In the [Introduction section](#), we revised the wording of our first contribution point from “... evaluating scientific creativity in LLMs ...” to “... evaluating scientific idea generation in LLMs ...”, better reflecting this distinction.
- At the beginning of [Methods \(page 16 in the Main text\)](#), we explicitly emphasized: *“It is important to note that our methodology evaluates an essential but not exhaustive aspect of scientific creativity, focusing primarily on divergent thinking capabilities.”*

We believe these revisions clarify that our benchmark aims to evaluate a specific dimension of creativity rather than claiming to assess scientific creativity in its entirety.

Comment 2.9 — Similarly, the discussion about “human creativity research” lacks some fundamental theories, such as Boden’s (e.g., “The Creative Mind: Myths and Mechanisms”); Amabile’s (e.g., “The Social Psychology of Creativity: A Componential Conceptualization”); or Rhodes’ (see An Analysis of Creativity). I understand that a complete treatment of creativity would require separate work, but I still believe this paper should correctly mention and discuss them in relation to the proposed framework.

Reply: Thank you for this valuable feedback! We appreciate your suggestion to include key theoretical frameworks. We have added a new paragraph to the [Related Works section \(Human Creativity Research, paragraph 4 in the Main text, page 5\)](#) discussing the contributions of Margaret Boden (P-creativity vs. H-creativity), Teresa Amabile (componential theory: domain skills, creativity processes, motivation), and Melvin Rhodes (4 P’s: Person, Process, Press, Product). This discussion now explicitly contextualizes LiveIdeaBench within these broader frameworks, noting its primary focus on the “Process” aspect (specifically idea generation under constraints) and acknowledging the importance of the other components (like Product value) for overall scientific creativity.

We believe these revisions clarify that our benchmark aims to evaluate a specific dimension of creativity rather than claiming to assess scientific creativity in its entirety. Furthermore, reflecting this clarified focus, we have also revised the manuscript’s title to *“Evaluating LLMs’ Divergent Thinking for Scientific Idea Generation with Minimal Context.”*

Comment 2.10 — The second issue is about the practical consequences of using LLMs as evaluators. While using LLM-as-a-Judge techniques help automatize evaluation, some issues

have to be carefully considered, e.g., biases induced by the evaluating model and correlation between human and AI-based judges. As correctly pointed out in the article, we still have to demonstrate the soundness of LLM-based evaluation approaches applied to creative tasks; I was surprised not to see any study about whether the proposed evaluation approach correlates with judgment from humans (and especially experts in scientific domains).

Reply: Thank you for this thoughtful and critical suggestion. We agree that comparing automated evaluations with human expert judgments is indeed essential for validating the soundness of our LLM-as-a-judge approach, especially for subjective tasks like assessing scientific creativity.

Given the diverse range of scientific domains covered by our benchmark keywords, recruiting human experts across all fields posed a significant practical challenge. Therefore, we focused our human evaluation study on a domain where we could readily access domain expertise: Partial Differential Equations (PDEs). We designed a questionnaire (a screenshot of the interface is shown in [Supplementary Figure S.8](#); the online questionnaire can be accessed at <https://wj.qq.com/s2/20424422/j0kj/>). For this validation study, we utilized the set of 22 scientific ideas generated by 11 LLMs (i.e. `claude-3.7-sonnet:thinking`, `qwq-32b`, `claude-3.5-sonnet`, `gpt-4.5-preview`, `deepseek-r1-distill-llama-70b`, `qwen2.5-dracarys2-72b`, `o1`, `qwen-2.5-coder-32b-instruct`, `qwen-2.5-7b-instruct`, `claude-3.5-haiku-20241022`, and `nova-lite-v1`, sampled at equal intervals based on rankings (see [Table Supplementary Table S.3](#)) from the 41 LLMs included in our benchmark, two ideas per model) specifically for the keyword “partial differential equations”. We invited 6 experts (including 4 Ph.D. candidates, a postdoc, and a faculty member) in the PDE field to rate these 22 ideas on the same three dimensions used by our LLM judges: originality, feasibility, and clarity (on a 1-10 scale). Importantly, the evaluation was conducted blindly, with idea order randomized, to minimize potential bias (details found in [Supplementary Note 7](#)).

To assess the reliability of the human judgments themselves, we calculated the Inter-Rater Reliability (IRR) using the Intraclass Correlation Coefficient, aka, $ICC(A, k)$. The results showed a good agreement among experts for Originality ($ICC = 0.823$) and Clarity ($ICC = 0.782$), but Poor agreement for Feasibility ($ICC = 0.453$). This indicates that while experts consistently judged novelty and articulation, they significantly disagreed on the feasibility of the ideas (see details in [Supplementary Note 7](#)).

The detailed statistical results comparing average human ratings to LLM ratings are presented in [Supplementary Note 7](#). Our correlation analysis ([Supplementary Figure S.9](#)) demonstrates statistically significant positive correlations between human and LLM ratings across the assessed dimensions. The correlation is strong for originality ($r = 0.820^{***}$), moderate for feasibility ($r = 0.572^{**}$), and weaker, yet still significant, for clarity ($r = 0.420$). The asterisks indicate the statistical significance level (p-value), representing the likelihood that the observed correlation occurred purely by chance. Following standard conventions: ** signifies $p < 0.01$ (less than 1% chance probability), and *** signifies $p < 0.001$ (less than 0.1% chance probability). A lower p-value indicates a more statistically meaningful correlation.

This alignment, particularly the strong correlation for originality which is supported by high human IRR, provides substantial empirical support for the validity of our LLM-as-a-judge methodology for assessing novelty. The less satisfactory human agreement on feasibility, however, pro-

vides crucial context for the moderate human-LLM correlation on that dimension. It suggests the challenge lies not only potentially with the LLM judges but also with the inherent difficulty and subjectivity of assessing feasibility from brief descriptions, even for experts. This finding highlights a limitation in reliably benchmarking feasibility using this setup. Despite this specific challenge with feasibility, the overall results suggest that, at least within the PDE domain studied, the LLM panel is capable of discerning relative quality differences in scientific ideas in a manner largely consistent with human expert judgment, especially for originality. This strengthens the rationale for using this automated approach for the LiveIdeaBench benchmark, while acknowledging the nuances revealed by the validation.

We sincerely thank you for prompting us to conduct this validation study and the reliability analysis, which has significantly strengthened the empirical grounding and credibility of our evaluation approach by providing a more nuanced understanding of its strengths and limitations.

Comment 2.11 — In addition, asking a model trained to “converge” on training data to evaluate “divergent” properties seems to me counterintuitive: intuitively, what the model will consider mostly original is what is similar to what has been judged as original in the past, that is, something that is not original anymore. I believe that the article should expand on this expected tension, and should also provide more empirical data in defense of using LLM-based evaluation for divergent thinking.

Reply: Thank you for raising this thoughtful point! While we acknowledge there is a theoretical tension here, recent empirical evidence strongly suggests LLMs can effectively evaluate creative scientific ideas.

Notably, Google’s recent “AI co-scientist” framework (Gottweis et al. 2025) successfully employed LLM-as-a-judge approaches within its multi-agent system built with Gemini 2.0. Their system includes specialized “Ranking” and “Meta-review” agents that evaluate scientific hypotheses. Most significantly, their framework not only identified promising novel ideas but also produced hypotheses that were subsequently validated through wet-lab experiments across multiple domains. In one compelling example, their system proposed candidate drugs for acute myeloid leukemia that demonstrated tumor inhibition in vitro at clinically applicable concentrations. Similarly, the AI co-scientist identified novel epigenetic targets for liver fibrosis that were validated through anti-fibrotic activity and liver cell regeneration in human hepatic organoids. These findings underwent laboratory validation, confirming the system’s ability to generate genuinely valuable scientific insights. Expert assessments of novelty and impact also aligned with the system’s automated evaluation metrics, further validating the approach. This real-world validation provides compelling evidence that LLMs can effectively recognize and evaluate genuinely innovative scientific concepts. These empirical results suggest that while the theoretical concern has merit, the practical limitations may be less significant than one might initially anticipate.

We appreciate your insightful comment about this important issue, and we have expanded our Discussion section (page 12-16 in the Main text) to address the “Reliability of LLM-as-Judge” as an important consideration. As you correctly point out, when ideas contain concepts outside the judge model’s knowledge, the model might misunderstand them and consequently misjudge

their originality. Our revised discussion addresses this inherent challenge, acknowledging that while our dynamic panel utilizing state-of-the-art judge models likely provides broader and more current knowledge coverage than static approaches, the fundamental limitation persists. We further propose in the discussion that this limitation could be more comprehensively addressed through retrieval-augmented generation (RAG) approaches that incorporate up-to-date scientific literature. By combining retrieval of the latest scientific papers with LLM generation capabilities, such approaches could significantly mitigate this issue, although implementing this solution falls outside the scope of the current study. We look forward to exploring this direction in future work.

References: (the following number coincides with that in the Main text)

[47] Gottweis, J., Weng, W., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., ... Natarajan, V. (2025). Towards an AI co-scientist. arXiv. <https://arxiv.org/abs/2502.18864>

Comment 2.12 — As an additional point, the use of LLM-as-a-Judge for fluency is not sufficiently motivated in my opinion. There are various diversity metrics considering either the syntax or semantics of text that can be used to evaluate potential overlapping without relying on LLMs (which requires more computing resources and presents the issues of LLM-as-a-Judge discussed above).

Reply: Thank you for this insightful comment. To clarify the motivation for using LLM-as-a-Judge for fluency, we evaluated various non-LLM diversity metrics (semantic, syntactic), transforming scores to a common scale where higher scores indicate greater dissimilarity. As shown in Figure 1 below (see next page), their correlation with LLM-as-a-Judge assessments was generally moderate (Pearson $r < 0.7$), indicating limitations in non-LLM metrics for capturing the specific conceptual divergence relevant to our benchmark.

Crucially, metrics based on surface similarity (lexical/embedding) often failed to capture nuanced conceptual differences. A concrete example illustrates this: consider these two research ideas, where we highlight the identical introduction, problem/goal, and method/solution:

Idea A:

Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies the most important features in a dataset. Proposed idea: Apply PCA to large-scale single-cell RNA sequencing data to identify the key genes driving cellular heterogeneity. This reduced set of principal component genes could then be used to develop targeted therapies for specific cell subpopulations in complex diseases like cancer. Focusing on the most informative genes allows for more precise and effective treatments, potentially improving patient outcomes. This approach leverages the power of PCA to distill high-dimensional data into actionable insights, advancing precision medicine efforts.

Idea B:

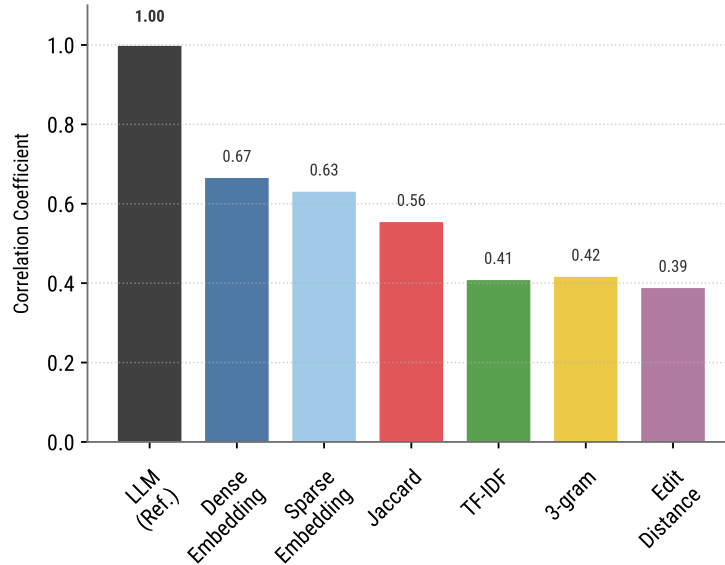


Figure 1: Correlation analysis (Pearson r) between LLM-as-a-Judge dissimilarity scores and various non-LLM similarity/distance metrics (all transformed to a scale where higher scores indicate greater dissimilarity) across the benchmark dataset.

Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies the most important features in a dataset. Idea: Develop an adaptive PCA algorithm that automatically adjusts the number of principal components based on the explained variance ratio. This dynamic approach would optimize the balance between dimensionality reduction and information retention, making PCA more efficient and applicable to a wider range of datasets. The adaptive nature of the algorithm would enable it to handle datasets with varying complexity and structure, ultimately facilitating more accurate data analysis and pattern recognition across diverse scientific fields.

Despite sharing the introductory sentence (orange), these ideas diverge fundamentally in their core contributions (purple and teal): Idea A proposes a specific **application**, while Idea B proposes a **methodological improvement**.

Our LLM-as-a-Judge correctly identified this divergence, assigning a high dissimilarity score of **10** (well above the benchmark mean of 6.03, see Figure 2 on next page). In contrast, the bge-m3 dense embedding yielded a score of only **0.73**, indicating very low dissimilarity (far below its benchmark mean of 2.35). This metric clearly failed, likely overemphasizing shared introductory sentence (orange) while missing the distinct goals (purple) and methods (teal).

This direct comparison underscores the limitations of relying solely on embedding-based metrics for nuanced scientific idea evaluation.

Therefore, despite its acknowledged challenges, we maintain that the LLM-as-a-Judge approach is currently the most suitable for capturing the specific type of conceptual divergence targeted by LiveIdeaBench. The PCA example clearly demonstrates its superior ability compared

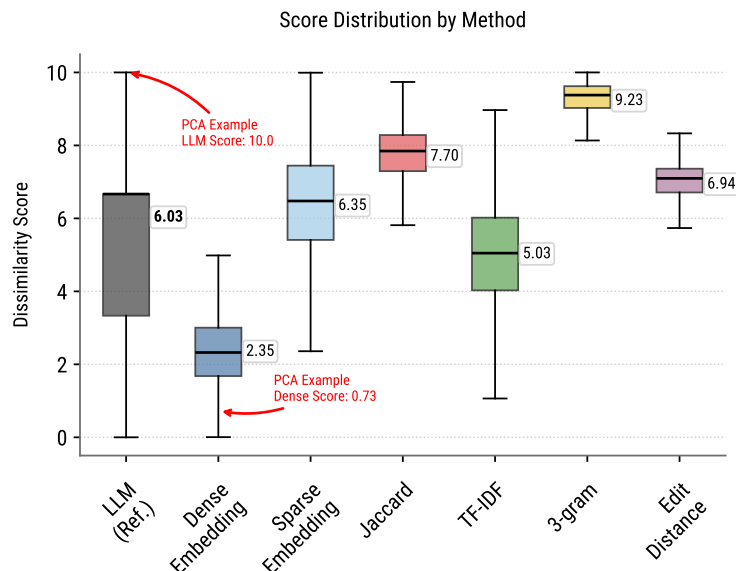


Figure 2: Distribution of dissimilarity scores from LLM-as-a-Judge and transformed Dense Embeddings (bge-m3). Both metrics are scaled such that higher scores indicate greater dissimilarity. Vertical lines mark the scores for the example PCA ideas: LLM (10, high dissimilarity) vs. Dense Embedding (0.73, low dissimilarity), illustrating the LLM’s superior ability to capture conceptual divergence in this case.

to non-LLM metrics (see Figure 1) in capturing the nuanced conceptual distinctions vital for evaluating scientific idea divergence. This assessment precision is essential for reliably measuring the divergent thinking capabilities targeted by our benchmark. We hope this clarifies our rationale.

Comment 2.13 — Moreover, the fluency scores are obtained in a 0-3 scale, and then translated into a 0-10 scale: how is that exactly done? Does this “discretization” (as apparent from Figure S.3) affect the overall scores?

Reply: Thank you for pointing out the need to clarify the fluency score scaling. To address this and ensure consistency, we have revised the scaling. These four qualitative grades are now directly mapped linearly to the integer scores 1 (for D), 4 (for C), 7 (for B), and 10 (for A), respectively. This ensures:

- Fluency aligns with the 1-10 score (not 0-10) range used for originality, feasibility, and clarity, enhancing consistency.
- Equal intervals represent the steps between the four distinct qualitative levels assessed.

This mapping translates the discrete assessment outcomes (A–D) to the target 1–10 scale, rather than discretizing a continuous measure. This revised mapping improves internal consistency, and all reported results now reflect this updated scaling. This is clarified in the revised Methods section (Fluency subsection, page 17 in the Main text).

Comment 2.14 — *Is there enough detail provided in the methods for the work to be reproduced?* There are enough details in the Methods and Supplementary Information for the work to be roughly, but not exactly, reproduced (some fine-grained implementation details, e.g., about how refused responses are managed and text is generated, are not reported).

Reply: Thank you for pointing out the need for further details to ensure reproducibility, specifically regarding how we handle instances where models refuse to generate an idea (e.g., for sensitive keywords). We appreciate this constructive feedback, as robust handling of refusals is crucial for fair evaluation.

As part of our benchmark refinement, we have implemented and now detail the specific procedure for handling model refusals in the revised manuscript and supplementary information. This procedure addresses the challenge that directly penalizing safety-related refusals could underestimate the capabilities of models with strong safety alignments.

The implemented protocol involves a two-step process:

- **First Attempt & Check:** An initial request is made using the standard idea generation prompt (detailed in [Supplementary Section 1](#)). The model's response is then automatically checked for common refusal-indicating phrases (the full list is provided in [Supplementary Section 5](#)).
- **Fallback Attempt (if refusal detected):** If a refusal is detected in the initial response, a specific **fallback prompt** is triggered for a second attempt. This fallback prompt (also detailed in [Supplementary Section 5](#)) explicitly frames the task within an academic research context to encourage a valid response despite potential keyword sensitivity.

This two-step process aims to elicit a valid scientific idea even for sensitive keywords, ensuring their idea generation capability is assessed more fairly.

Furthermore, to enhance transparency and reproducibility, the revised [Supplementary Note 5](#) now provides: (1) visualization of this refusal handling workflow ([Supplementary Figure S.6](#)), and (2) data on the initial refusal rates per model ([Supplementary Figure S.7](#)).

We believe detailing this refined procedure and providing the supporting visualizations and data in the Supplementary Information directly addresses your concern and significantly enhances the clarity and reproducibility of our evaluation methodology regarding response handling. This improvement is part of our broader effort to ensure LiveIdeaBench robustly and fairly assesses scientific idea generation capabilities.

Comment 2.15 — Unfortunately, the code does not provide enough instructions in its README for installing and running the application. The repository does not contain any requirements.txt file or any instruction to get a functioning Docker container, so it is not intuitive how to reproduce the results. Moreover, the code correctly contains a .py file to instantiate and run LLMs and process their outputs, as well as a notebook to analyze the results, but no main .py file is available, thus impeding any correct reproduction of the results.

Reply: Thank you for highlighting the issues with the code repository's reproducibility. To address this, we have now:

- Uploaded the **complete codebase** required to run the experiments and reproduce our results to the GitHub repository: [<https://github.com/x66ccff/liveideabench>]
- Added comprehensive, step-by-step instructions to the README.md file, detailing the setup process and the workflow for executing the experiments and analysis.
- Included requirements.txt file specifying all necessary dependencies, allowing for straightforward recreation of the required environment.

These updates, particularly the detailed README guidance and the environment file, should now provide the necessary information and structure to reproduce our findings effectively.

Overall Remark: In summary, we sincerely appreciate your constructive comments and suggestions, which have been greatly helpful for improving the quality of our manuscript!

Reviewer 3

We sincerely appreciate the time and effort you dedicated to evaluating our work and providing constructive comments. Following your suggestions, we have conducted extensive experiments and made substantial revisions to further strengthen our manuscript. Please find our point-to-point responses below.

Comment 3.1 — The manuscript introduces LiveIdeaBench, a benchmark for evaluating the scientific creativity of large language models (LLMs) under minimal contextual conditions. The study builds upon Guilford’s creativity theory and assesses models across originality, feasibility, fluency, and flexibility. The authors analyze 20 models using a dataset of 1,180 scientific keywords, highlighting the independence of creativity from general intelligence metrics.

While the research addresses an important and underexplored aspect of LLM capabilities, there are key concerns regarding the theoretical grounding, methodological robustness, and clarity of presentation that should be addressed.

Reply: Thanks for this remark. Please see our responses below.

Comment 3.2 — *Major Comment 1. Lack of References in Introduction (p. 1-3)* The introduction is missing references to support claims, especially regarding the importance of creativity in scientific discovery and the limitations of existing LLM evaluation frameworks. Authors should cite relevant work in AI creativity assessment (i.e., work by Janet Rafner as a starting point, but there are multiple papers on AI creativity from theoretical and empirical research that must be addressed here) and cognitive theories of creativity (i.e., work by Roger Beaty, Mathias Benedek, Adam Green, Yoed Kenett, etc.).

Reply: Thank you for this helpful remark. We have revised the [Introduction](#) and [Related Works](#) sections extensively to address this concern. Specifically, we added citations supporting the importance of creativity in science (e.g., Wang *et al.*, 2023; Shi & Evans, 2023), acknowledging the role of diverse perspectives (Shi & Evans, 2023). We also cited the key work on AI creativity assessment, including Rafner *et al.* (2023), and expanded the discussion in the [Related Works](#) section. Furthermore, we followed your suggestion and incorporated discussions and citations related to cognitive theories by Beaty (cited in Rafner *et al.*, 2023), Benedek (e.g., Jauk *et al.*, 2013; Benedek, 2024), and Green (e.g., Cortes *et al.*, 2019), particularly within the [Human Creativity Research](#) subsection (page 4 in the Main text). We also integrated relevant computational creativity work, including from Kenett (e.g., Wenger & Kenett, 2025), into the [Related Works](#) section, providing a richer overall theoretical context.

References: (the following number coincides with that in the Main text)

[1] Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., ... & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972), 47-60.

- [2] Shi, F., & Evans, J. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1), 1641.
- [5] Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11), 1836–1838.
- [7] Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, 41(4), 212–221.
- [19] Benedek, M. (2024). On the relationship between creative potential and creative achievement: Challenges and future directions. *Learning and Individual Differences*, 110, 102424.
- [21] Cortes, R. A., Weinberger, A. B., Daker, R. J., & Green, A. E. (2019). Re-examining prominent measures of divergent and convergent creativity. *Current Opinion in Behavioral Sciences*, 27, 90–93.
- [32] Wenger, E., & Kenett, Y. (2025). *We're different, we're the same: Creative homogeneity across LLMs*. arXiv preprint arXiv:2501.19361.

Comment 3.3 — *Major Comment 2. Motivation for the Research (p.2)* The justification for LiveIdeaBench is not entirely convincing. The manuscript argues that general intelligence and scientific creativity are independent, but the need for a new benchmark should be clarified.

Reply: Thank you for this helpful remark regarding the motivation for the benchmark. We acknowledge our initial presentation may not have fully clarified the benchmark’s necessity beyond the parallel drawn between human cognition and LLM capabilities.

To address this concern, we have substantially revised the [Introduction \(page 1-3\)](#), [Related Works \(pages 4-8\)](#), and [Discussion \(pages 12-16\)](#) sections in the [Main text](#) to provide a multi-faceted and robust motivation for LiveIdeaBench.

In essence, the need for LiveIdeaBench arises from two primary, interconnected factors, discussed as follows:

1. **Methodological Gap & Focus on Divergent Thinking:** Existing benchmarks excel at evaluating *convergent thinking* using rich contextual inputs. However, they fail to assess the crucial capability of *divergent scientific idea generation* from minimal context. LiveIdeaBench is specifically designed to fill this gap by assessing this distinct skill relevant to early-stage scientific exploration. The inspiration from cognitive science regarding the potential distinction between general intelligence and creativity prompted us to hypothesize that performance here might differ from standard metrics.
2. **The “Live” Dynamic Evaluation:** Static benchmarks quickly become outdated or susceptible to overfitting. A core motivation for LiveIdeaBench is its *dynamic* nature. As described in the [Discussion and Methods](#), it employs a continuously updated panel of

state-of-the-art LLMs (from LiveBench) for both idea generation and judging. This “Live” approach ensures ongoing relevance, resists data contamination, and provides a more robust, evolving measure of current creative capabilities, distinct from static tests.

Crucially, our empirical findings (e.g., [Figure 4 in the Main text](#) and [Supplementary Figure S.4](#)) demonstrate a notable divergence between performance on LiveIdeaBench and general intelligence metrics, validating that we are measuring a distinct capability and underscoring the need for this specialized, dynamic benchmark.

We believe these revisions demonstrate that the justification for LiveIdeaBench is fundamentally grounded in addressing critical gaps within existing LLM evaluation. By focusing on minimal-context divergent thinking—a capability our results confirm is distinct from general intelligence—and employing an essential dynamic, “Live” methodology, our benchmark provides necessary insights into scientific idea generation capabilities that existing evaluations overlook.

Comment 3.4 — *Major Comment 3. Divergent vs. Convergent Thinking.* The study positions LiveIdeaBench as assessing divergent thinking but does not convincingly differentiate it from convergent thinking. Please explain why the chosen methodology isolates divergent thinking. Perhaps reference studies that contrast divergent and convergent thinking and also in AI.

Reply: Thank you for highlighting this important distinction and prompting us to clarify the theoretical underpinnings of our approach. We have revised the [Introduction \(page 3\) in the Main text](#) to provide a much stronger justification for LiveIdeaBench as an assessment of divergent thinking, explicitly grounding its methodology in [Guilford \(1967\)](#) seminal theory of divergent production.

As detailed in the revised [Main text](#), our methodology directly operationalizes key principles from Guilford’s framework for assessing divergent thinking. The use of single-keyword prompts provides the minimal context and broad stimulus characteristic of classic divergent thinking tasks, designed to elicit the generation of multiple and varied potential ideas and foster a “*ready flow of ideas*” ([page 138 in Guilford \(1967\)](#)), rather than convergence on a single solution. This approach necessitates that models “*produce their own answers, not to choose them from alternatives given to them*” ([page 138 in Guilford \(1967\)](#)), a hallmark of divergent production assessment, distinguishing it from selection-based or highly constrained convergent thinking tasks. Furthermore, our evaluation directly measures core divergent production dimensions like fluency (reflecting the capacity to generate a “*number of varied responses*”, [page 138 in Guilford \(1967\)](#)) and originality, as conceptualized by Guilford.

We further clarify in the [Main text](#) (see [paragraph 2, page 3](#)) how this approach contrasts sharply with standard LLM benchmarks primarily focused on convergent thinking. While acknowledging that comprehensive scientific creativity involves an interplay of both divergent and convergent processes, LiveIdeaBench specifically focuses on and evaluates the foundational, generative phase crucial for sparking innovation: the ability to produce diverse scientific concepts from sparse cues. We believe this revised explanation, now explicitly linked to Guilford’s foundational work, robustly addresses the distinction and clarifies our benchmark’s targeted assessment of divergent thinking capabilities in LLMs.

References: (the following numbers coincide with those in the Main text)

[10] Guilford, J. P. (1967). *The Nature of Human Intelligence*. McGraw-Hill.

Comment 3.5 — *Major Comment 4. Claim that Convergent thinking is secondary in creativity (p.3)*. This point has since been challenged, for example Cortes et al. (2019) claim that creative measures involve mixtures of the two constructs together.

Cortes, R. A., Weinberger, A. B., Daker, R. J., & Green, A. E. (2019). Re-examining prominent measures of divergent and convergent creativity. *Current Opinion in Behavioral Sciences*, 27, 90-93. <https://doi.org/10.1016/j.cobeha.2018.09.017>

Reply: Thank you for pointing out this important nuance and providing the relevant reference. We agree that the original statement oversimplified the relationship. We have revised the manuscript, specifically in the [Introduction \(page 2-3 in the Main text\)](#) and [Related Works \(Human Creativity Research, page 4 in the Main text\)](#) sections, to address this concern. We removed the claim that convergent thinking is merely “secondary”. Instead, we now acknowledge the complex interplay between divergent and convergent thinking in creativity, explicitly citing Cortes et al. (2019) regarding the finding that creative tasks often involve mixtures of both. Our revision clarifies the role of both processes in overall creativity.

References: (the following number coincides with that in the Main text)

[21] Cortes, R. A., Weinberger, A. B., Daker, R. J., & Green, A. E. (2019). Re-examining prominent measures of divergent and convergent creativity. *Current Opinion in Behavioral Sciences*, 27, 90-93.

Comment 3.6 — *Major Comment 5. IQ and LLM Intelligence (p.3)*. The paper compares LLM intelligence to human IQ, which is at the very least problematic since IQ is not designed for LLMs. The authors should remove direct comparisons to human IQ and instead compare LLMs based on creativity-relevant cognitive benchmarks.

Reply: Thank you for this comment. We agree that directly applying human IQ metrics to LLMs is problematic and potentially misleading.

In response to this concern, we have revised the manuscript, particularly the [Introduction \(page 2-3 in the Main text\)](#) and [Related Works \(page 4-5 in the Main text\)](#) sections, to remove all direct comparisons and analogies to human IQ. We believe this revision provides a more accurate and appropriate context for our study.

Comment 3.7 — *Major Comment 6. Anthropomorphic conclusions*. The authors should avoid claims such as LLM’s are creative or intelligent, rather they may exhibit signs of human-like creativity or intelligence. Also please reference current literature on whether LLM’s are truly intelligent or creative and how that influences our creativity, as well as the dangers of LLM hallucinations.

Some relevant literature:

- Marrone, R., Cropley, D., & Medeiros, K. (2024). How Does Narrow AI Impact Human Creativity?. *Creativity Research Journal*, 1-11.
- Wenger, E., & Kenett, Y. (2025). We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint arXiv:2501.19361*.
- Cropley, D. (2023). Is artificial intelligence more creative than humans? : ChatGPT and the Divergent Association Task. *Learning Letters*, 2, 13. <https://doi.org/10.59453/11.v2.13>
- Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is Chatgpt a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning*, 40(2), 919–930. <https://doi.org/10.1111/jcal.12929>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1-55.

Reply: This is a crucial point – we appreciate your suggestion along with the helpful references. We have carefully revised the manuscript to use more precise and cautious language, avoiding direct anthropomorphic claims. Instead of stating LLMs “are creative,” the text consistently refers to their “capabilities for initial idea generation,” their “performance on idea generation tasks,” or their “potential for tasks analogous to human creative processes” (as noted in the [Related Works](#) section). This focuses language on observable abilities and performance rather than attributing inherent creative states.

The literature you suggested provided valuable context regarding current discussions on AI creativity, including its potential impact, e.g., [Marrone et al. \(2024\)](#), limitations like homogeneity (e.g., [Wenger & Kenett, Cropley](#)), and the importance of considering issues like hallucinations. We found these perspectives very helpful in refining our manuscript. Accordingly, our revised [Related Works \(Evaluating Idea Generation in LLMs, page 5-6\)](#) section now incorporates discussion of several relevant studies, including those by [Marrone et al. \(2024\)](#), [Wenger & Kenett \(2025\)](#), and [Cropley \(2023\)](#), to better situate our work within the current landscape. Thank you. In line with using more precise language and accurately representing the benchmark’s scope, we have also updated the manuscript’s title to “*Evaluating LLMs’ Divergent Thinking for Scientific Idea Generation with Minimal Context*”.

References: (the following number coincides with that in the Main text)

- [29] Cropley, D. (2023). Is artificial intelligence more creative than humans?: ChatGPT and the divergent association task. *Learning Letters*, 2, 13.
- [30] Marrone, R., Cropley, D., & Medeiros, K. (2024). How Does Narrow AI Impact Human Creativity?. *Creativity Research Journal*, 1-11.

[32] Wenger, E., & Kenett, Y. (2025). *We're different, we're the same: Creative homogeneity across LLMs*. arXiv preprint arXiv:2501.19361.

Comment 3.8 — *Major Comment 7. Confounding Factors in Creativity Measurement (p.11)*. The study does not sufficiently address potential biases in creativity scoring, particularly in relation to model training data. For example the multiple evaluator methods, may involve models with similar training data, such that they are not as independent as we may think.

Reply: Thank you for this remark regarding potential biases. This is indeed a valid point; even with multiple evaluators, correlated biases could potentially influence the assessment if the judge models lack sufficient independence due to overlapping training histories or methodologies.

To mitigate this risk and strengthen the evaluation’s robustness, we have refined our judge selection protocol for the 10-model panel drawn from top LiveBench performers. As now detailed in the revised [Experimental Protocol subsection, page 17-18 in the Main text](#)), we incorporate specific diversity constraints during panel formation:

- (a) **Limiting Organizational Representation:** We ensure no single organization (e.g., OpenAI, Google) contributes more than 20% (i.e., a maximum of 2 models) to the judge panel. If necessary, models are replaced following LiveBench rank order to meet this cap. This promotes a broader representation of different training philosophies and potential data sources.
- (b) **Handling Identical Base Models:** For model pairs where the underlying base architecture is identical but they differ primarily in inference-time configuration or “reasoning effort” (e.g., o3-mini vs. o3-mini-high), we select only one representative (prioritizing the higher-performing variant on general intelligence benchmarks). Including both would be redundant and could disproportionately amplify inherent biases from that specific base model, without adding genuinely diverse evaluative perspectives.

We believe these refined selection criteria help ensure a more diverse and independent judge panel, reducing the likelihood of systematic biases related to shared training data dominating the evaluation process. We appreciate you prompting us to clarify and strengthen this aspect of our methodology, which is now reflected in the manuscript.

Comment 3.9 — *Major Comment 8. Lack of Human Evaluation*. The reliability of the automated creativity ratings is unclear without comparison to human judges. Perhaps include a validation study where human raters assess a subset of generated ideas to compare against LiveIdeaBench scores or examples of the reliability of these models from other research.

Reply: Thank you for this thoughtful and critical suggestion. We agree that comparing automated evaluations with human expert judgments is indeed essential for validating the soundness of our LLM-as-a-judge approach, especially for subjective tasks like assessing scientific creativity.

Given the diverse range of scientific domains covered by our benchmark keywords, recruiting human experts across all fields posed a significant practical challenge. Therefore, we focused our

human evaluation study on a domain where we could readily access domain expertise: Partial Differential Equations (PDEs). We designed a questionnaire (a screenshot of the interface is shown in [Supplementary Figure S.8](#); the online questionnaire can be accessed at <https://wj.qq.com/s2/20424422/j0kj/>). For this validation study, we utilized the set of 22 scientific ideas generated by 11 LLMs (i.e. `claude-3.7-sonnet:thinking`, `qwq-32b`, `claude-3.5-sonnet`, `gpt-4.5-preview`, `deepseek-r1-distill-llama-70b`, `qwen2.5-dracarys2-72b`, `o1`, `qwen-2.5-coder-32b-instruct`, `qwen-2.5-7b-instruct`, `claude-3.5-haiku-20241022`, and `nova-lite-v1`, sampled at equal intervals based on rankings (see [Table Supplementary Table S.3](#)) from the 41 LLMs included in our benchmark, two ideas per model) specifically for the keyword “partial differential equations”. We invited 6 experts (including 4 Ph.D. candidates, a postdoc, and a faculty member) in the PDE field to rate these 22 ideas on the same three dimensions used by our LLM judges: originality, feasibility, and clarity (on a 1-10 scale). Importantly, the evaluation was conducted blindly, with idea order randomized, to minimize potential bias (details found in [Supplementary Note 7](#)).

To assess the reliability of the human judgments themselves, we calculated the Inter-Rater Reliability (IRR) using the Intraclass Correlation Coefficient, aka, $ICC(A,k)$. The results showed a good agreement among experts for Originality ($ICC = 0.823$) and Clarity ($ICC = 0.782$), but Poor agreement for Feasibility ($ICC = 0.453$). This indicates that while experts consistently judged novelty and articulation, they significantly disagreed on the feasibility of the ideas (see details in [Supplementary Note 7](#)).

The detailed statistical results comparing average human ratings to LLM ratings are presented in [Supplementary Note 7](#). Our correlation analysis ([Supplementary Figure S.9](#)) demonstrates statistically significant positive correlations between human and LLM ratings across the assessed dimensions. The correlation is strong for originality ($r = 0.820^{***}$), moderate for feasibility ($r = 0.572^{**}$), and weaker, yet still significant, for clarity ($r = 0.420$). The asterisks indicate the statistical significance level (p-value), representing the likelihood that the observed correlation occurred purely by chance. Following standard conventions: ** signifies $p < 0.01$ (less than 1% chance probability), and *** signifies $p < 0.001$ (less than 0.1% chance probability). A lower p-value indicates a more statistically meaningful correlation.

This alignment, particularly the strong correlation for originality which is supported by high human IRR, provides substantial empirical support for the validity of our LLM-as-a-judge methodology for assessing novelty. The poor human agreement on feasibility, however, provides crucial context for the moderate human-LLM correlation on that dimension; it suggests the challenge lies not only potentially with the LLM judges but also with the inherent difficulty and subjectivity of assessing feasibility from brief descriptions, even for experts. This finding highlights a limitation in reliably benchmarking feasibility using this setup. Despite this specific challenge with feasibility, the overall results suggest that, at least within the PDE domain studied, the LLM panel is capable of discerning relative quality differences in scientific ideas in a manner largely consistent with human expert judgment, especially for originality. This strengthens the rationale for using this automated approach for the LiveIdeaBench benchmark, while acknowledging the nuances revealed by the validation.

We sincerely thank you for prompting us to conduct this validation study and the reliability

analysis; it has significantly strengthened the empirical grounding and credibility of our evaluation approach by providing a more nuanced understanding of its strengths and limitations.

Comment 3.10 — *Minor Comment 1.* The manuscript refers to LiveIdeaBench both as a “framework” and a “benchmark,” leading to potential confusion. Please define and consistently use one term throughout the paper.

Reply: We appreciate your careful attention to terminology consistency. After careful consideration, we have decided to consistently use the term “benchmark” when referring to LiveIdeaBench throughout the manuscript, as this more accurately reflects its primary purpose as an evaluation tool for comparing model performance.

We have conducted a thorough review of the manuscript and replaced instances of “framework” with “benchmark” when referring to LiveIdeaBench itself. In contexts where we discuss the methodological aspects of our work, we have used more specific terms such as “evaluation methodology” to clearly distinguish between the overall benchmark and its methodological components.

These revisions ensure terminological consistency while maintaining precision in describing the overall benchmark, its methodological components, and referenced theories. We believe these changes enhance clarity and better align with how LiveIdeaBench is presented in our external resources (e.g., website, GitHub repository, and Huggingface dataset).

Comment 3.11 — *Minor Comment 2.* Figures are not always referenced in the text at appropriate locations (e.g., Figure 1 lacks in-text reference). Ensure each figure is clearly cited where relevant.

Reply: We sincerely thank you for this valuable observation. We have addressed the missing in-text reference for Figure 1 by adding a reference on Page 3 in the Main text. We have also verified that all other figures and tables are appropriately referenced throughout the manuscript.

Comment 3.12 — *Minor Comment 3.* The claim that idea length is independent of quality is supported with an R^2 value (0.003) as correlation strength, but this is a measure of explained variance rather than correlation.

Reply: Thanks for this comment. We have revised our presentation accordingly. We now report the standard pairing of the Pearson correlation coefficient (r) and its p -value to indicate the strength and significance of the relationship. This revised reporting can be found in the Results section (page 11, paragraph 1 in the Main text) and Supplementary Figure S.2, which continues to show a weak relationship between idea length and quality.

Comment 3.13 — *Minor Comment 4.* Several claims in the related work and discussion sections lack citations, making it difficult to verify assertions.

Reply: Thank you for your valuable suggestion! Regarding the revisions to the Related Works section, please see our previous detailed responses to Comment 3.2. For the Discussion section,

we have revised potentially problematic claims to ensure rigor and accuracy, including adding supporting citations where appropriate as discussed above (e.g., the broader implications for AI-assisted scientific discovery). We believe these changes address your concerns.

Comment 3.14 — *Minor Comment 5.* Some LLMs evaluated are proprietary and not widely accessible, which could affect reproducibility. Discuss limitations regarding access to closed-source models.

Reply: Thanks for this important point regarding reproducibility and the use of proprietary models. We agree this is a relevant limitation inherent in evaluating the current LLM landscape.

To address this concern, we have added a statement in the [Discussion section \(page 12-16 in the Main text\)](#) of the revised manuscript acknowledging that the inclusion of closed-source models (e.g., GPT-4o) accessed via APIs presents challenges for exact reproducibility due to potential changes in access or underlying model updates. We clarify that while their inclusion is necessary for comprehensive coverage of the state-of-the-art, this impacts direct reproducibility compared to using only open-weight models.

Comment 3.15 — *Minor Comment 6.* Some figures (e.g., Figure 7) may be redundant or better integrated into supplementary materials.

Reply: Thank you for this helpful suggestion! We agree that the content of the original Figure 7 is better suited for the supplementary materials. Accordingly, we have moved this figure to the Supplementary Information, where it is now designated as [Supplementary Figure S.4](#).

Comment 3.16 — *Recommendation:* The study presents an interesting and novel contribution to evaluating LLM creativity, but major revisions are needed before it can be considered for publication. Specifically:

- Strengthen the theoretical justification for LiveIdeaBench, in the field of creativity and in general.
- Clarify the theoretical distinction between divergent and convergent thinking, and ground claims in current literature regarding AI abilities in these fields.
- Address concerns regarding the absence of human evaluation as a benchmark comparison in this study as well as potential biases in LLM scoring and the multiple evaluator approach.
- Improve and add referencing and clarity in data interpretation and terms

With these improvements, the manuscript could make a valuable contribution to field of AI-driven creativity assessment and scientific idea generation.

Reply: Thank you for this constructive summary and for recognizing the potential contribution of our work! We appreciate you highlighting these key areas for revision.

As detailed in our specific responses to your previous comments throughout this document, we believe we have addressed each of the points you raised: strengthening the theoretical background and justification, clarifying the distinction between thinking types and grounding our claims, addressing concerns about human evaluation and potential biases in LLM scoring (including refining our methodology), and improving referencing and overall clarity.

We believe the manuscript has been significantly improved as a result of these revisions, and we thank you again for your constructive and thoughtful comments/suggestions.

Overall Remark: In summary, we sincerely appreciate your constructive comments and suggestions, which have been greatly helpful for improving the quality of our manuscript!

Reviewer 4

Comment 4.1 — I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reply: We appreciate you noting your role as a co-reviewer for our manuscript. We understand this is part of the Nature Communications initiative supporting early career researchers in peer review, and we thank you for your valuable time and effort in contributing to the evaluation of our work.

Overall Remark: In summary, we sincerely appreciate your constructive comments and suggestions along with the co-reviewer, which have been greatly helpful for improving the quality of our manuscript!

Reviewer 5

We sincerely appreciate the time and effort you dedicated to evaluating our work and providing constructive comments. Following your suggestions, we have conducted extensive experiments and made substantial revisions to further strengthen our manuscript. Please find our point-to-point responses below.

Comment 5.1 — Firstly I want to thank the authors for their timely and relevant paper. It represents original research which will contribute to the field and will eventually be published. However, I note some significant aspects which I feel require additional work before publication.

Reply: Thanks for your positive and encouraging feedback. Please see our responses below.

Comment 5.2 — Overall, I think that the article addresses an important topic and the methodology seems sound (apart from a small comment I present later on), however I think that the article need to work on sharpening the framing. The article presents that models have achieved super human performance across disciplines, and presents the fundamental question “does the creative potential of LLMs grow in tandem with their analytical capabilities, particularly in the realm of scientific discovery, however it does not address if or why we want our LLMs to make such discoveries alone or rather if they should be made as part of a tight human in the loop approach, as is recommended by the fields of hybrid intelligence and human centered AI. How often do we expect to use LLMs alone for scientific discovery? The authors allude to this in the second to last paragraph of the discussion when they state that their findings extend beyond model evaluation to the broader landscape of AI assisted scientific discovery (a section that is completely un referenced). However, the article would greatly benefit from a position that perhaps the benchmarking is needed as an initial step in developing such human-AI systems, or if the authors take the stance that AI scientific discovery alone (AGI with no human involvement) is beneficial then to preface that.

For more references on human centered AI and hybrid intelligence and hybrid intelligence and generative AI in creativity tasks I suggest the following references:

- Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., ... & Welling, M. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8), 18-28.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637-643.
- Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11), 1836-1838.

Additionally Falk et al., 2025 propose the idea of instead of LLM as a judge, it is LLM as an (additional judge), leveraging humans and AI in synergy. <https://arxiv.org/pdf/2503.04290>

Reply: Thank you for raising this crucial point about the framing of our work and its position relative to human-AI collaboration paradigms like hybrid intelligence and human-centered AI. We sincerely appreciate your thoughtful feedback and the highly relevant references you provided.

We strongly agree with your perspective. The human-AI collaborative paradigm, as emphasized in the fields of hybrid intelligence (Akata *et al.*, 2020; Dellermann *et al.*, 2019) and human-centered AI, is indeed the approach we envision and advocate for.

Maintaining a central role for human researchers in the scientific discovery process is paramount. As highlighted in the literature (e.g., Rafner *et al.*, 2023), human oversight ensures that AI contributions align with human values and ethical considerations. Furthermore, human involvement is crucial for validating AI-generated hypotheses and mitigating the risks associated with potential LLM hallucinations or biases (Huang *et al.*, 2025). Moreover, the human element is vital for maintaining the crucial diversity of scientific exploration, as studies like Doshi & Hauser (2023) indicate that AI assistance, while beneficial individually, can lead to a convergence and reduction in collective novelty. Ultimately, fostering effective human-AI synergy is essential for maximizing benefits while safeguarding human well-being and preventing potential AI harms.

Following your valuable suggestion, we have refined the framing in the revised manuscript to make this aspect explicit. Specifically:

- In the **Introduction** (page 3, paragraph 1 in the Main text), we now explicitly state: “We posit that such benchmarking serves as a crucial initial step towards developing effective human-AI hybrid intelligence systems (Rafner *et al.*, 2022; Akata *et al.*, 2020; Dellermann *et al.*, 2019) for scientific discovery. By systematically assessing LLMs’ performance on idea generation tasks, we aim to provide insights into how these models can best function as collaborative tools or creative sparks for human researchers within the scientific process.”
- We have integrated the references you kindly provided on hybrid intelligence (Akata *et al.*, 2020; Dellermann *et al.*, 2019) and AI creativity (Rafner *et al.*, 2023) into the **Introduction** section and relevant parts of the **Related Works** section to better contextualize our work within these important research areas.
- The **Discussion** section now also reinforces this perspective, emphasizing the potential for human-AI collaboration (page 14 of the revised manuscript) and explicitly referencing AI-assisted scientific discovery.

We believe these revisions clearly articulate our stance and position LiveIdeaBench appropriately as a tool aimed at understanding and ultimately facilitating, rather than replacing, human scientific endeavors through effective AI assistance. Thanks again for your great suggestion, which is extremely helpful to strengthen the critical aspect of our paper’s framing.

References: (the following number coincides with that in the Main text)

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

- [11] Rafner, J., Gajdacz, M., Kragh, G., Hjorth, A., Gander, A., Palfi, B., Berditchevskiaia, A., Grey, F., Gal, K., Segal, A., et al. (2022). Mapping citizen science through the lens of human-centered AI. *Human Computation*, 9(1), 66–95.
- [12] Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8), 18–28.
- [13] Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637–643.
- [33] Doshi, A. R., & Hauser, O. P. (2023). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(1), ead0137.
- [67] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55.

Comment 5.3 — Now I will go through the sections one by one with my comments. *Abstract*. The language in the abstract should be improved for clarity. In the first sentence I would appreciate examples of the scientific tasks as well as what is meant by performance. The meaning behind the sentence “we reveal that scientific creativity shows distinct patterns from general intelligence metrics” is not clear to me. Additionally “suggest that the development of creative capabilities in LLMs may follow different trajectories than traditional problem solving” is too vague.

Reply: Thanks for this valuable comment. We have revised the [Abstract](#) accordingly. We clarified the first sentence by adding examples of scientific tasks (“*literature analysis and experimental design*”) and explicitly defined what is meant by performance via providing examples of the specific capabilities involved, such as “*accurately extracting key findings from papers or generating coherent experimental procedures*”. We also replaced the unclear phrase “*shows distinct patterns from*” by “*are poorly predicted by standard general intelligence metrics*”, with the subsequent sentence giving a concrete example of this divergence. Finally, instead of the vague “*different trajectories*”, we now state more concretely that enhancing these creative capabilities “*may require different training strategies than those used for improving general problem-solving abilities*”. We believe these revisions directly address your points and improve the abstract’s clarity.

Comment 5.4 — *Introduction*. Overall the introduction seems quite light on references, and of the references in this manuscript, half of them are arxiv preprints. I am acutely aware of the issue of publishing work with the rapidly moving field that is genAI and LLMs, and I myself often face this dilemma of needing to reference the most up to date work which is usually only on arxiv. However, I encourage the authors to review some of my suggested literature and to incorporate more peer reviewed articles into the manuscript.

Reply: We appreciate you highlighting the important balance between timely preprints and peer-reviewed literature. Following your valuable suggestion, as well as related feedback from other reviewers, we have undertaken a comprehensive review of the relevant literature. We have diligently worked to revise the manuscript, particularly the [Introduction](#) and [Related Works](#) sections, to incorporate a significantly greater number of peer-reviewed articles. This includes several of the insightful references you and others kindly provided, which have helped us to better contextualize our work and strengthen its theoretical underpinnings.

While the rapid evolution of the field necessitates citing some recent preprints to capture the absolute state-of-the-art, we have made a concerted effort to ensure our core arguments and background discussions are now more thoroughly grounded in established, peer-reviewed research. We believe this blend now provides both timeliness and robust academic grounding.

Thank you once again for this constructive feedback. It has been instrumental in improving the overall quality and scholarly rigor of our paper.

Comment 5.5 — Figure 1 is confusing to me. I understand the issues with page limitations and trying to put multiple figures together into one figures, however I am confused if the entire figure should be seen as a flow chart, following a logical order or if they are distinct separate branches. Additionally it is confusing that in part b it shows 3 llms but in the text it says the evaluation panel was done by 10 LLMs.

Reply: Thank you for this feedback regarding Figure 1. We appreciate you identifying these points. Figure 1 is intended as a conceptual overview illustrating the distinct stages (**a-f**) of the LiveldeaBench benchmark, representing key components rather than a strict sequential flowchart. Regarding the judge LLMs, your observation accurately identified a key aspect of our methodology requiring clarification.

Specifically, as now detailed within “*Judge LLM Panel Formation, Independence, and Application*” in the [Methods section \(Experimental Protocol subsection, page 18, Main text\)](#), the panel is applied as follows: for the dimensions of originality, feasibility, and clarity, each generated idea is assessed by a randomly selected subset of 3 judges drawn from the eligible panel members (i.e., the top 10, excluding the model being evaluated). The depiction of three LLMs in [Figure 1b \(page 2 in the Main text\)](#) illustrates this sampling process for evaluating a single idea. For the fluency dimension, which assesses the diversity among ideas generated for the same keyword, a single judge is randomly sampled from the eligible panel to perform the comparison for that keyword-model pair.

We recognize that these crucial sampling steps were not sufficiently detailed either within that specific paragraph in the original manuscript or explicitly visualized in the figure. To address this concern and enhance clarity, we have made the following revisions:

- We have expanded the single paragraph in the [Methods section](#) (under “*Judge LLM Panel Formation, Independence, and Application*”) to explicitly describe these sampling procedures for evaluating originality, feasibility, clarity, and fluency ([page 18 in the Main text](#)).
- We have revised [Figure 1 \(in the Main text\)](#) and its caption. The updated figure now includes a clearer visual flow and explicit labels (e.g., “*Random Sampling 3 Judges*” in part **b**) to better illustrate how the judge panel is utilized via sampling for the evaluation steps.

We believe these revisions to both the textual description within the [Methods](#) section and the visual representation in [Figure 1 \(in the Main text\)](#) now provide a much clearer and more accurate picture of our evaluation methodology, directly addressing the points of confusion you raised. We hope this clarifies our approach, and we thank you again for helping us improve the manuscript's clarity.

Comment 5.6 — *Related work.* The authors mention the threshold theory, but they do not acknowledge that there is quite a debate about whether this theory actually holds. The article they cite (Jauk et al 2013) only shows part of the story and has been addressed in more recent work. If the article is situated in a comparison of intelligence vs. creativity, i believe it is important to present a more accurate and nuanced discussion on the topic. Additionally the authors would benefit from citing specific literature on intelligence and creativity in scientific creativity.

Excerpt from Benedek, M. (2024). On the relationship between creative potential and creative achievement: Challenges and future directions. *Learning and Individual Differences*, 110, 102424.

Specifically, the threshold hypothesis assumes that high intelligence (e.g., IQ above a threshold of 120) represents a necessary condition for high creativity and that the influence of intelligence should become lower or vanish above this intelligence threshold (Getzels & Jackson, 1962; Kim, 2005). Research has reported mixed evidence for the threshold effect so far, with some studies finding no support (Kim, 2005; Preckel et al., 2006; Weiss et al., 2020), and others finding support under certain conditions (Gralewski et al., 2012; Jauk et al., 2013). But these works mainly considered creativity in terms of DT ability. Studies focusing on creative achievement actually showed that intelligence is relevant for creative achievement across the entire spectrum of intelligence (Jauk et al., 2013; Park et al., 2008) or becomes even more important beyond the threshold (Runco & Albert, 1986). These empirical findings are generally consistent with the notion of interactive effects, by emphasizing the importance of high creative potential without assuming reduced relevance beyond a given threshold.

- Getzels, J. W., & Jackson, P. W. (1962). Creativity and intelligence: Explorations with gifted students.
- Kim, K. H. (2005). Can only intelligent people be creative? A meta-analysis. *Journal of Secondary Gifted Education*, 16(2-3), 57-66.
- Preckel, F., Holling, H., & Wiese, M. (2006). Relationship of intelligence and creativity in gifted and non-gifted students: An investigation of threshold theory. *Personality and individual differences*, 40(1), 159-170.
- Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). A reappraisal of the threshold hypothesis of creativity and intelligence. *Journal of Intelligence*, 8(4), 38.
- Gralewski, J., Weremczuk, E., & Karwowski, M. (2012). Intelligence and creativity of Polish middle-school students: Looking for the threshold hypothesis. *The New Educational Review*, 29, 328-338.

- Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, 41(4), 212-221.
- Park, G., Lubinski, D., & Benbow, C. P. (2008). Ability differences among people who have commensurate degrees matter for scientific creativity. *Psychological science*, 19(10), 957-961.
- Runco, M. A., & Albert, R. S. (1986). The threshold theory regarding creativity and intelligence: An empirical test with gifted and nongifted children. *Creative Child and Adult Quarterly*, 11(4), 212-218.

Reply: Thank you for this comment regarding the threshold theory discussion. We agree our initial discussion could be more nuanced. To address this concern, we have substantially revised the discussion in both the [Introduction \(page 2, paragraphs 4-5\)](#) and the [Related Works \(Human Creativity Research, page 4-5\)](#) sections. Specifically, we now:

- Explicitly acknowledge the mixed evidence surrounding the threshold hypothesis in the [Related Works](#) section, citing key studies suggested by you that found limited support or proposed alternative relationships (e.g., [Getzels & Jackson, 1962](#); [Kim, 2005](#); [Park et al., 2008](#); [Gralewski et al., 2012](#); [Jauk et al., 2013](#); [Weiss et al., 2020](#)).
- Incorporate the important distinction between creative potential (often linked to divergent thinking ability) and creative achievement, referencing literature suggesting that intelligence may remain linearly related or even become more critical for achievement across the entire ability spectrum ([Jauk et al., 2013](#); [Runco & Albert, 1986](#); [Benedek, 2024](#)).
- In the [Introduction](#) section, we now frame the question about LLMs' capabilities by referencing these complexities in human cognition, suggesting that idea generation and analytical abilities might develop differently in AI, mirroring the observed distinctions between creativity and intelligence in humans.
- We have also added a sentence in the [Related Works](#) section acknowledging that while much of the debate centers on general creativity, the specific link between intelligence and scientific creativity requires nuanced investigation, positioning our work within this context.

We believe these revisions provide a more accurate, balanced, and nuanced discussion of the intelligence-creativity relationship, grounding our research question more effectively and incorporating the valuable literature highlighted by you. Thank you for your great comment, which is very helpful to improve the clarity of our manuscript.

References: (the following number coincides with that in the Main text)

[7] Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, 41(4), 212-221.

- [8] Kim, K. H. (2005). Can only intelligent people be creative? A meta-analysis. *Journal of Secondary Gifted Education*, 16(2-3), 57-66.
- [9] Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). A reappraisal of the threshold hypothesis of creativity and intelligence. *Journal of Intelligence*, 8(4), 38.
- [14] Getzels, J. W., & Jackson, P. W. (1962). Creativity and intelligence: Explorations with gifted students.
- [18] Runco, M. A., & Albert, R. S. (1986). The threshold theory regarding creativity and intelligence: An empirical test with gifted and nongifted children. *Creative Child and Adult Quarterly*, 11(4), 212-218.
- [19] Benedek, M. (2024). On the relationship between creative potential and creative achievement: Challenges and future directions. *Learning and Individual Differences*, 110, 102424.

Comment 5.7 — In the section on Evaluating Creative Capabilities of LLMs the second paragraph is not clear to me. I do not fully understand what framework has been developed and how that relates to this study.

Reply: Thank you for pointing out the lack of clarity in our discussion of related assessment frameworks. We apologize for the confusion and have revised the relevant paragraphs in the manuscript to better explain the context and relationship between prior work and LiveIdeaBench. Our clarification of the key works cited is described as follows:

- **Bellemare-Pépin et al. (2024):** This large-scale study provides crucial empirical evidence demonstrating LLMs' strong performance, sometimes exceeding human levels, on *general* creative tasks like the DAT. While validating the feasibility of assessing divergent capabilities, it focuses on these general tasks and does not address the specific context of *scientific* idea generation.
- **Lu et al. (2024):** This work focused specifically on assessing creativity in the *text-to-code* domain. Their methodology involved using programming challenges (using Codeforces dataset) combined with specific prompting techniques designed to encourage novel solutions by restricting previously used methods. This differs from our broader scientific scope, and potential reliance on *static* problem sets was noted as a limitation.
- **Zhao et al. (2024):** Drawing from foundational creativity theories (like Guilford's), this work proposed a comprehensive, though *static*, framework. Their approach involved adapting tasks derived from established psychological tests to evaluate *general* creativity across multiple dimensions.

These studies represent important advancements in assessing different facets of LLM creativity. LiveIdeaBench complements these efforts by addressing a specific and previously underexplored niche. In particular:

- **Focus on Scientific Idea Generation from Minimal Context:** Unlike the general tasks (Bellemare-Pépin *et al.*, Zhao *et al.*) or the code-specific domain (Lu *et al.*), LiveIdeaBench uniquely targets the initial generation of *scientific* ideas prompted only by single keywords.
- **Dynamic Evaluation Methodology:** Our benchmark employs a *dynamic, contamination-limited* approach, continuously updating keywords, models, and judges, contrasting with the static frameworks used in the cited studies and mitigating potential associated issues.
- **Large-Scale Assessment:** LiveIdeaBench operates at a significant scale (41 LLM models, 1180 keywords, 22 domains).
- **Investigating Creativity vs. Intelligence in Science:** A key contribution is our unique investigation into how these specific divergent scientific thinking capabilities relate to general intelligence metrics, revealing distinct patterns not explored by the prior works focused primarily on creativity assessment itself.

In essence, while these works provide valuable assessments for their respective areas (general creative tasks, coding creativity), LiveIdeaBench provides a necessary, complementary benchmark specifically designed for evaluating the foundational capability for divergent *scientific* thinking from minimal cues, using a robust dynamic methodology, and yielding unique insights into its relationship with general intelligence.

We have revised the relevant paragraphs in the [Related Works](#) section (subsection: [Evaluating Idea Generation in LLMs](#), page 6, paragraphs 4-5 in the Main text) to reflect these specific descriptions and distinctions. We hope this provides a much clearer explanation of the relationship between our work and the existing literature.

References: (the following number coincides with that in the Main text)

- [31] Bellemare-Pépin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., Bengio, Y., & Jerbi, K. (2024). *Divergent creativity in humans and large language models*. arXiv preprint arXiv:2405.13012. <https://arxiv.org/abs/2405.13012>
- [34] Lu, Y., Wang, D., Li, T., Jiang, D., & Khashabi, D. (2024). *Benchmarking language model creativity: A case study on code generation*. arXiv preprint arXiv:2407.09007. <https://arxiv.org/abs/2407.09007>
- [35] Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X., & Du, Z. (2024). *Assessing and understanding creativity in large language models*. arXiv preprint arXiv:2401.12491. <https://arxiv.org/abs/2401.12491>

Comment 5.8 — On Page 8 and 9 the authors address the topic of length of responses, however I would also very much like to see something on clarity of responses. Lee and Chung (2024) found that chatgpt assisted humans in producing creative problem solving solutions that were more well articulated than with out ai assistance. This could be see as clarity of

the idea, which would also be very interesting in the context of the authors work. Lee, B.C., Chung, J.J. An empirical investigation of the impact of ChatGPT on creativity. Nat Hum Behav 8, 1906–1914 (2024). <https://doi.org/10.1038/s41562-024-01953-1>

Reply: Thank you for this insightful suggestion and for bringing the relevant findings from Lee & Chung (2024) to our attention! We found this point particularly valuable and agree that the clarity or articulation of a generated scientific idea is indeed a critical aspect to evaluate.

Your comment prompted us to carefully reconsider our initial adaptation of Guilford’s creativity framework. In our original submission, we replaced Guilford’s “elaboration” dimension with “feasibility”, arguing the latter was more pertinent in a scientific context. However, upon reflection, spurred by your feedback and the Lee & Chung (2024) study highlighting articulation, we realize that “elaboration” aligns more closely with the concept of “clarity” than “feasibility” does. Furthermore, we now believe that removing the “elaboration/clarity” aspect entirely in favor of “feasibility” might have been an oversimplification. Both the clarity of presentation and the scientific viability are crucial for evaluating generated scientific ideas.

Therefore, we decided to act upon your excellent suggestion by incorporating “clarity” as an additional, distinct dimension in our evaluation benchmark. LiveIdeaBench now assesses LLMs across five key dimensions: originality, feasibility, fluency, flexibility, and the newly added clarity. This addition not only addresses your point but also brings our evaluation closer to Guilford’s original multifaceted framework (by including a dimension akin to elaboration) while retaining feasibility, which we still consider essential for the scientific domain.

This significant change is now reflected throughout the entire revised manuscript, specifically:

- Figure 1 in the Main text (Overall design) now depicts the five dimensions being assessed.
- Figure 2 in the Main text (Bar plot & Radar plot) and Figure 4 (Heatmap comparison) include clarity scores alongside the other metrics.
- In Figure 5 in the Main text (Bubble plot), we now use the *edge width* of the bubbles to represent the clarity score for each model.
- We have added Figure 6e in the Main text, which specifically shows the ranked clarity scores for all evaluated models.
- The Methods section (page 16 in the Main text) detailing the evaluation dimensions has been updated accordingly, explaining how clarity is assessed by the judge LLMs (similar to originality and feasibility).

Furthermore, to explore the interplay between the newly added clarity and the other dimensions, we analyzed the relationship between clarity scores and the existing originality and feasibility scores across all generated ideas. The results are visualized in Figure 3 below (see next page).

The analysis reveals statistically significant positive correlations between clarity and the other core dimensions evaluated by the LLM judges. This suggests that ideas perceived as clearer also tend to score higher on both originality and feasibility, reinforcing the importance of articulation alongside novelty and viability in the context of scientific idea generation.

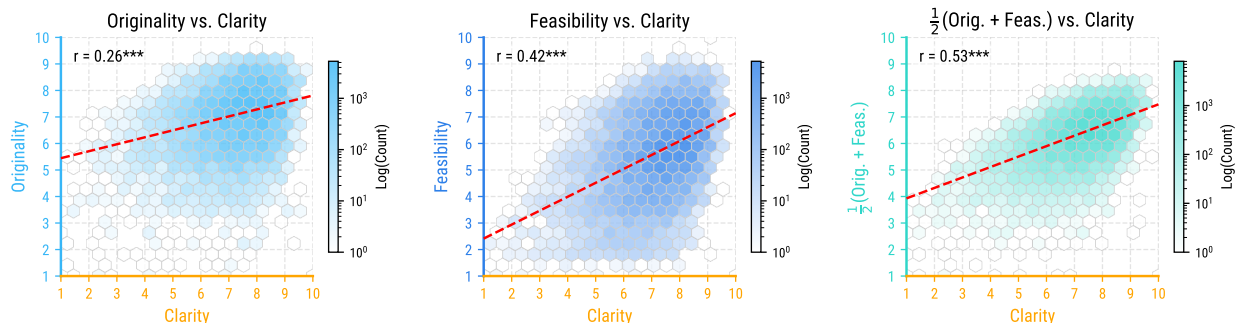


Figure 3: Relationship between clarity and other core evaluation dimensions across all generated ideas. Hexbins show data density ($\text{Log}(\text{Count})$), red dashed lines show linear regressions. Clarity demonstrates positive correlations with originality ($r = 0.26^{***}$), feasibility ($r = 0.42^{***}$), and their average ($r = 0.53^{***}$), indicating clearer ideas are generally rated higher overall (the superscript *** denotes $p < 0.001$).

We believe incorporating this “clarity” dimension makes our benchmark more comprehensive and better captures the nuances of effective scientific idea generation. We are grateful for your thoughtful suggestion, which has helped us significantly refine and strengthen our evaluation methodology, aligning it more robustly with both foundational creativity theory and practical considerations for scientific communication. Thank you again for prompting this insightful improvement to our work.

References: (the following number coincides with that in the Main text)

[28] Lee, B.C., Chung, J.J. An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour* 8, 1906–1914 (2024). <https://doi.org/10.1038/s41562-024-01953-1>

Comment 5.9 — The axis on figure 6 and 7 need to be better described. For example the x axis of on figure 6 A just reads “score” and should state score of what.. Similarly 7B just has “average” on the y axis but does not state an average of what..

Reply: Thank you for noting the need for more descriptive axis labels in Figures 6 and 7 in the Main text. We have revised these figures to improve clarity:

Regarding the original Figure 6a where the axis simply read “score”, we have revised this figure significantly. The new Figure 6 now presents a multi-part analysis (Fig 6a–e). Specifically, Figure 6a displays the “Idea Score Avg. by Model: $\frac{1}{3} (\text{Orig.} + \text{Feas.} + \text{Clar.})$ ”. The caption clearly states this is the average of originality, feasibility, and clarity scores for each model, addressing the previous ambiguity about what the score represented.

Similarly, for the figure originally labeled 7B (which has been moved and is now located as Supplementary Figure S.4), we have ensured both axes are clearly labeled. The x-axis now reads “Average LiveBench Score” and the y-axis reads “Average LiveIdeaBench Score”, removing the vague term “average”.

We believe these more descriptive labels improve the clarity of these figures and directly address the points you raised. Thanks again for helping enhance the presentation of our results.

Comment 5.10 — Lastly there is a growing movement to report the carbon impact of studies which use LLMs. I strongly suggest that the authors do so. I recommend they use the following tool for guidelines: <https://mlco2.github.io/impact/#about>

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700.

Reply: This is an excellent remark! We strongly agree that reporting carbon impact is a crucial aspect for the field to consider. We thank you for recommending resources, including the “mlco2” calculator (<https://mlco2.github.io/impact/>) and the foundational paper by Lacoste *et al.* (2019), which underscores the importance of quantifying these emissions. We have added this valuable reference to our manuscript.

We investigated the “mlco2” calculator tool itself. However, we found it requires inputs like specific GPU models and runtimes, which are difficult to ascertain accurately for the proprietary models accessed via APIs, as the underlying hardware infrastructure used by service providers is typically opaque to the end-user.

Therefore, to provide an estimate suitable for our API-centric evaluation context, we adopted an alternative methodology using the “EcoLogits Calculator” (<https://huggingface.co/spaces/genai-impact/ecologits-calculator>; <https://doi.org/10.5281/zenodo.15132099>). This approach allows for estimating emissions based on accessible factors like model parameter counts (which we estimated for proprietary models based on public information) and the number of generated tokens, making it more applicable to our specific scenario.

Following this methodology, we have calculated the estimated energy consumption and CO₂ emissions for each model’s usage in our benchmark. These results, broken down by model and role (idea generator vs. judge), are now presented in detail in [Supplementary Table S.4](#) in the revised manuscript. We have also added a brief summary of the total estimated environmental impact (approximately 3074 kgCO₂eq.) in the [Discussion section \(page 12-16 in the Main text\)](#)

We appreciate you encouraging this reporting and highlighting the importance of environmental awareness in AI research, supported by work like the reference you provided. We agree it’s a crucial consideration for the field moving forward, and we commend your focus on promoting sustainable research practices.

Comment 5.11 — Here are some additional references that I recommend the author review to better contextualize their work.

Peer reviewed and published:

- M. Krenn, L. Buffoni, B. Coutinho, S. Eppel, J. G. Foster, A. Gritsevskiy, H. Lee, Y. Lu, J. P. Moutinho, N. Sanjabi, et al., Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network, *Nature Machine Intelligence* 5, 1326 (2023).
- F. Shi and J. Evans, Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines, *Nature Communications* 14, 1641 (2023).

- Q. Wang, D. Downey, H. Ji, and T. Hope, Scimon: Scientific inspiration machines optimized for novelty, *Annual Meeting of the Association for Computational Linguistics* (2024).
- H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al., Scientific discovery in the age of artificial intelligence, *Nature* 620, 47 (2023).
- J. Sourati and J. A. Evans, Accelerating science with human-aware artificial intelligence, *Nature Human Behaviour* 7, 1682 (2023).
- Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15607–15631. <https://doi.org/10.18653/v1/2023.acl-long.870>
- Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 361–374. <https://doi.org/10.18653/v1/2023.findings-ijcnlp.32>
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.
- Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT Rates Natural Language Explanation Quality like Humans: But on Which Scales?. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 3111–3132. <https://aclanthology.org/2024.lrec-main.277>

ArXiv papers:

- 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*.
- X. Gu and M. Krenn, Forecasting high-impact research topics via machine learning on evolving knowledge graphs, *arXiv:2402.08640* (2024).

- J. Baek, S. K. Jauhar, S. Cucerzan, and S. J. Hwang, Researchagent: Iterative research idea generation over scientific literature with large language models, *arXiv:2404.07738* (2024).
- M. R. AI4Science and M. A. Quantum, The impact of large language models on scientific discovery: a preliminary study using gpt-4, *arXiv:2311.07361* (2023).
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535* (2024).

Reply: Thank you for providing this extensive list of relevant references. We have carefully reviewed and incorporated these references into the revised manuscript. These additions are particularly prominent in the updated **Introduction** and **Related Works** sections, where they help to better situate LiveIdeaBench within the current landscape of research on AI for scientific discovery, LLM evaluation, and creativity assessment.

We truly appreciate you taking the time to compile and share these resources. Your comments and suggestions have been instrumental in improving the depth and rigor of our literature review.

Overall Remark: In summary, we sincerely appreciate your constructive comments and suggestions, which have been greatly helpful for improving the quality of our manuscript!

Evaluating LLMs’ Divergent Thinking for Scientific Idea Generation with Minimal Context

References: NCOMMS-25-09187-T

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, Hao Sun

We sincerely appreciate the positive feedback from the reviewers. The remaining minor comments have been addressed in our point-to-point response below as well as the revised paper.

Reviewer 2

Comment 2.1 — I sincerely thank you for your thorough responses and careful revisions. I am really impressed by the quality of the additional work you have conducted to satisfy all my concerns.

The article has been greatly improved in all of its aspects. From a theoretical standpoint, it now includes relevant additional references, and the scope of the research has been made clearer. The method has been expanded and better clarified (especially considering the new supplementary materials), and the potential limitations have been correctly discussed. Finally, the evaluation is now statistically sound, and I really appreciate the additional study comparing LLM and human evaluation, which, in my opinion, is crucial to demonstrate the validity of the benchmark.

All considered, I have no further requests, and I believe the article is ready for acceptance.

Reply: Thanks for your positive and encouraging feedback!

Comment 2.2 — *Remarks on code availability.* The code is now appropriate, with all necessary information to reproduce the results.

As a final suggestion, I believe it would be useful to include all the instructions to add new models in the README file. This would simplify the adoption of the benchmark by developers of future LLMs.

Reply: Thank you for this excellent suggestion. The README file contains an "Evaluation Instruction" section, which would help developers test other LLMs.

Reviewer 5

Comment 5.1 — The authors present a comprehensive evaluation benchmark (LiveIdeaBench) to assess large language models’ (LLMs) capacity for divergent thinking in the context of scientific idea generation. The revised manuscript reflects substantial effort to address initial review comments, including strengthening theoretical grounding, expanding model comparisons, incorporating human expert validations, and enhancing the clarity of both methodology and presentation. I am also very happy to see that the authors have included a statement on carbon impact of the work.

Reply: Thanks for your positive and encouraging feedback!

Comment 5.2 — I think that at this point to polish the article it would be nice to see if the discussion could be (slightly) expanded to highlight practical applications and next steps: How can LiveIdeaBench guide future LLM development (e.g., training for idea generation vs problem-solving)? Could hybrid human-AI frameworks benefit from the insights provided by LiveIdeaBench?

Reply: Thanks for this suggestion. We have now slightly expanded the discussion to highlight these practical applications. Specifically, we added that future development might benefit from “prioritizing training objectives that reward diverse conceptual associations, thereby laying the groundwork for hybrid human-AI frameworks tailored to specific creative stages.” (in Main text, page 5)

Comment 5.3 — The authors include this statement: Our benchmark provides a foundation for understanding and improving these capabilities, potentially enabling more effective human-AI collaboration [11–13] and informing the design of human-aware AI systems for science [68] in pushing the boundaries of scientific knowledge. But could benefit from elaborating on *how* the benchmark could inform design of human-aware AI systems for science.

Reply: Thanks for this suggestion. We have now elaborated on this in the text to specify that the benchmark informs design “by leveraging specific model profiles to complement human cognitive blind spots,” thereby clarifying the mechanism for improving human-aware AI systems. (in Main text, page 7)

Comment 5.4 — Additionally this sentence could be supplemented by a short example: The creative strengths we observed in different model architectures suggest that a diverse ecosystem of AI tools, each with complementary capabilities, could support different aspects of the scientific process.

Reply: Thank you for the suggestion. We have added a short example to this sentence to illustrate the complementary roles. It now reads: “...support different aspects of the scientific process, for instance, by pairing high-originality models for initial hypothesis generation with high-feasibility models for experimental implementation.” (in Main text, page 7)

Comment 5.5 — Lastly, this sentence could use a reference, or a specific link back to findings: As LLMs demonstrate increasingly sophisticated idea generation capabilities, they hold promise as powerful tools for accelerating scientific innovation - from hypothesis generation to experimental design.

(optional) I recommend the authors look at these new articles, DiStefano et al. 2025 Evaluating AI's Ideas: The Role of Individual Creativity and Expertise in Human-AI Co-Creativity. This article indicates the changing nature of creativity, shifting from generation to judgement.

This new article discusses the adapting role of humans and AI to work together for evaluation of creativity. Falk et al. 2025. How Do Hackathons Foster Creativity? Towards Automated Evaluation of Creativity at Scale. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (pp. 1-23). Specifically relevant for the LLMs-as-a-judge section

Reply: Thanks for this excellent suggestion. We have now include these articles in our references.

Comment 5.6 — Overall I recommend this article is accepted after minor revisions

Reply: Thanks for your positive and encouraging feedback!

Reviewer 6

Comment 6.1 — I have been invited to step in as Reviewer 3 for the revised version of your manuscript “LiveIdeaBench: Evaluating LLMs’ Divergent Thinking for Scientific Idea Generation with Minimal Context.” My role is therefore twofold: to evaluate the soundness of the research in its current form, and to assess how well the revisions address the concerns raised by the previous Reviewer 3. I want to acknowledge at the outset that the manuscript has been substantially improved since the initial submission, and I appreciate the considerable effort the authors have put into strengthening both the theoretical foundation and the methodological clarity.

Noteworthy results

The manuscript introduces LiveIdeaBench, a benchmark for evaluating LLMs’ divergent scientific idea generation under minimal context. This represents a valuable and original contribution, filling an important gap in existing evaluation frameworks that focus predominantly on convergent abilities. Expanding the benchmark to 41 models across 22 scientific domains provides impressive breadth. The key empirical finding, that divergent scientific idea generation is largely independent from general intelligence performance, is important for the AI evaluation literature. The addition of human expert validation in the Partial Differential Equations domain, showing strong alignment between human and LLM originality ratings, further supports the robustness of the benchmark.

Significance and relation to literature. This work is significant not only within AI evaluation but also in computational creativity and the study of human–AI scientific collaboration. The revised manuscript now situates the contribution more clearly in relation to established creativity research, including Guilford, Boden, Amabile, and Rhodes, as well as recent work by Beaty, Benedek, Kenett, and Rafner. This broader theoretical grounding was a key concern of the earlier Reviewer 3 and is now adequately addressed. Compared to prior creativity benchmarks such as DAT and CREA, LiveIdeaBench provides a distinct and complementary approach by isolating divergent thinking under minimal input conditions.

Support for conclusions. The main conclusions are supported by the data and analysis. The authors are careful not to overclaim: they now emphasize that LiveIdeaBench assesses a specific aspect of scientific creativity (divergent idea generation), not creativity in its entirety. The discussion of sycophancy, compressed score ranges, and feasibility evaluation challenges adds necessary nuance.

Data analysis and interpretation. The revisions have strengthened the statistical reporting: confidence intervals are now consistently included, keyword counts per discipline are shown, and the fluency scoring has been clarified. The discussion has been expanded to explain unexpected findings (e.g., some LLMs excelling in divergent tasks despite weaker performance on general benchmarks). While human validation is limited to one scientific domain, the authors are transparent about this constraint, and it does not undermine the validity of the broader findings.

Methodology and reproducibility. The methodology is sound and well detailed. Important refinements include the introduction of a “clarity” dimension, a transparent refusal-handling

protocol, and clearer explanation of scaling decisions. Reproducibility has been substantially improved through a public repository with full code, requirements, and instructions. Some reliance on proprietary APIs remains, but this is unavoidable and is appropriately acknowledged.

Overall assessment. From my perspective as a new Reviewer, the manuscript has effectively addressed the prior concerns. The theoretical framing is now richer, the statistical analysis more rigorous, the limitations more explicitly acknowledged, and the reproducibility better supported.

Reply: Thanks for your positive and encouraging feedback!