MinervAI: Using Generative AI to Assist, Not Replace Humans in Peer Review

Imogen Hüsing and **Tim Petersen** and **Cornelius Wolff**[1]
Osnabrück University {ihuesing,tpetersen,cowolff}@uos.de
Moritz-André Weiher
German Research Center for Artificial Intelligence moritz-andre.weiher@dfki.de
Sebastian Musslick
Osnabrück University, Brown University sebastian.musslick@uos.de

Abstract

The exponential growth of submitted scientific articles is straining an already overburdened peer review system. This trend is fueled by the increasing use of large language models (LLMs) for scientific writing, which enables higher article throughput per author without a corresponding increase in available reviewers. While LLMs can support specific review tasks, such as citation verification and argumentation mapping, their limitations in domain expertise, consistency, and bias caution against their use as autonomous reviewers. Based on structured expert interviews with scientists, we argue that LLMs should complement rather than replace human judgment in peer review. We present a tool that leverages LLM strengths to address key pain points in the scientific review process, while preserving the responsibility of human experts for critical evaluations. This approach seeks to improve both the integrity and efficiency of peer review through transparent, constrained AI support in scientific practice.

Keywords: Peer Review, Large Language Models, Human-AI Interaction

1 Introduction

The peer review system is facing unprecedented challenges as the volume of scientific submissions continues to grow exponentially Bornmann et al. [2021]. This trend is accelerated by advances in large language models (LLMs), which enable researchers to produce manuscripts at a faster rate [Naddaf, 2025, Kim et al., 2025]. For example, submissions for the machine learning conference NeurIPS have grown 10.4× over the past decade, with an approximate compound annual growth rate of 26.4% [Lawrence, 2022, Neurips, 2024], with similar trends observed in other disciplines Bornmann et al. [2021]. The expanding pool of submissions is outpacing the growth of qualified reviewers [Walker and Rocha Da Silva, 2015, Kim et al., 2025], threatening the timeliness and quality of reviews. The widespread adoption of LLM-based writing tools further amplifies submission rates [Liu and Shah, 2023, Wei et al., 2025], raising concerns about maintaining quality control in the review process [Aczel et al., 2025, Musslick

[1]Now Centrum Wiskunde & Informatica, Amsterdam

et al., 2025]. Commercial and open-source tools now offer to assist overburdened reviewers by generating complete paper reviews, but LLM-generated reviewers introduce new risks and uncertainties for the integrity of peer review [Idahl and Ahmadi, 2025, D'Arcy et al., 2024]. At the same time, the need for faster review processes is undeniable, raising the question: How—and how far—should generative artificial intelligence (AI) be incorporated into peer review?

In response to the challenges posed by the increased use of LLMs in peer review and the growing demands on reviewers, our contributions are twofold: (1) By analyzing both the strengths and weaknesses of LLMs and synthesizing insights from expert interviews, we articulate a position that *AI should support, not replace human reviewers*; (2) We introduce an open-source tool that directly addresses key reviewer pain points, aiming to assist reviewers in specific tasks while preserving human judgment.

2 Expert Interviews

We conducted pilot expert interviews with eight cognitive science and AI researchers (one doctoral student, three postdoctoral researchers, and four professors) to investigate challenges in peer review and potential roles for LLMs. The first half of the interview was dedicated to individual review practices and pain-points encountered by the interviewees, while in the second half reflected on potential opportunities of LLMs in peer review as well as their limitations.

Most interviewees expressed openness to using LLMs for specific tasks, such as compiling full-text reviews from notes (4/8), checking citations (4/8), and clarifying unfamiliar concepts (3/8). Nonetheless, all participants expressed concerns to some degree, for example the risk of hallucinated content (8/8), potential amplification of biases (5/8), issues of accuracy and reproducibility (3/8), and data privacy risks (2/8).

3 Position

The concerns voiced in the expert interviews mirror the current state of research on using LLMs in the context of peer reviews. First, LLMs are unreliable in their quality assessment of texts. While they perform well at predicting scores from human-written reviews or approximating human rating behavior, they are unreliable at ranking papers, providing consistent review scores or reliable and constructive critique [Zhou et al., 2025, Song et al., 2025, Muehlhoff and Henningsen, 2024], making them ill suited for evaluative tasks, such as assigning scores to papers. Second, pre-trained LLMs do not reflect the latest state of knowledge. They rely on static datasets which often lack most recent research or the latest domain-specific knowledge [Nassiri and Akhloufi, 2024]. The latter is crucial for assessing novelty and existing work related to the contents of a manuscript under review.

Retrieval Augmented Generation (RAG) [Lewis et al., 2020] addresses the

second problem by enabling LLMs to interface with a knowledge base. While RAG systems still face challenges in verifying information and demand careful implementation and human oversight[Ahn, 2025, Chen et al., 2024], grounding LLMs in textual context can improve their robustness and avoid hallucinations [Béchard and Ayala, 2024]. Critically, by grounding them in text, statements produced by LLMs can be verified. Overall, the strengths of LLMs lie in their capabilities to process [Aly et al., 2025], summarize [Asgari et al., 2025] and reason over provided text [Wei et al., 2022, Wang and Shen, 2024].

Given these limitations and the influence of LLMs on human decision making [Krügel et al., 2023, Choi et al., 2024, Ikeda, 2024] LLMs should be limited to tasks where outcomes are objectively verifiable. Such tasks are typically narrow in scope but often tedious and repetitive for reviewers, such as verifying the correctness of citations against the cited publications or distilling a paper's core arguments for easier navigation. By contrast, evaluative judgments (e.g., assessing novelty or significance) must remain with human reviewers, as they require expertise, contextual knowledge, and value-based reasoning that cannot be mechanistically verified. In short, we argue that *LLMs should be used in scientific peer review only to support reviewers on well-defined, verifiable tasks, not to substitute human judgment or evaluation.*

4 MinervAI

Building on expert interviews and the issues reviewed above, we are developing an open-source tool[2] to support reviewers in the tedious and draining aspects of the review process. It is intended to address pain points in the review process while minimizing risks associated with LLM-based peer review. Accordingly, we focus on two pain points of the review process: Verifying citations and understanding each argument throughout the paper.

Citation Checker Verifying citations is crucial for maintaining scientific integrity, yet the process is often too time-consuming to perform manually. Indeed, one of the pain points we identified for human peer reviewers is judging whether a manuscript's citations are accurate and used appropriately (4/8). LLMs are well suited for this task as they can efficiently process large text collections.

To address this, MinervAI includes a citation checker which combines automated citation extraction with LLM-based verification, as suggested in prior work [Alvarez et al., 2024]. The UI for this feature can be seen in Figure 1. Citations are first identified using regular expressions or, if necessary, extracted by an LLM. For each extracted citation, the local claim and the referenced paper are provided to another LLM, which evaluates whether the claim is supported. The system outputs a structured *Yes, No, or Maybe* label, along with source-grounded rationales and quoted spans. This approach frames the task as one of textual entailment [Bar-Haim et al., 2025, Sanyal et al., 2024]. Critically, to

[2]https://study-project-ai-for-science.github.io/landing-page/

ensure transparency, reviewers can inspect the relevant source passages and the generated rationales for flagged citations.

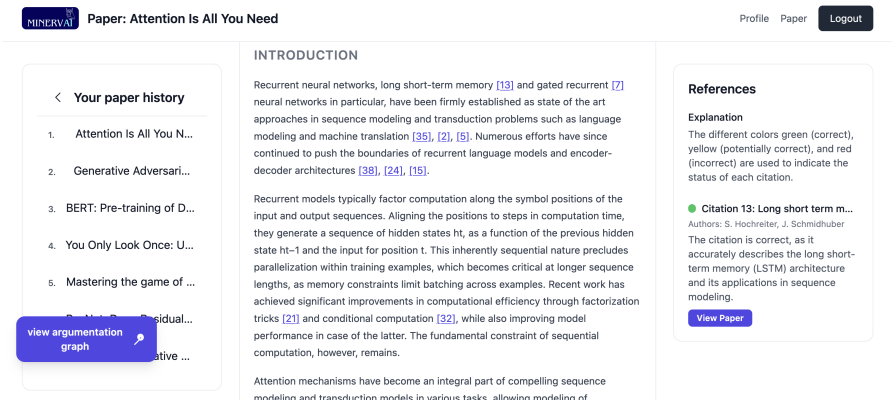

Figure 1: The MinervAI citation checker interface.

Argumentation Graphs The expert interviews revealed that navigating papers and going back and forth between different sections is tedious (5/8). Understanding the logic and argument of the paper was also mentioned as time consuming and mentally straining (5/8).

To address these two issues, MinervAI generates an argumentation graph which visualizes the argument structure of the manuscript. The nodes of the graph consist of different classes of argumentation units (AUs). Examples for these classes of AUs are *Claim, Hypothesis, Result,* or *Conclusion*. The links of the graph represent the relationships between the different AUs, for example *motivates, challenges, supports,* or *is part of*. The chosen classes are a subset of classes defined in Song et al. [2022].

The argumentation graph is constructed in a three-step process: First, the LLM is given the manuscript chunks and is tasked with identifying all AUs. In the second iteration, the LLM is provided with all the identified AUs and tasked with classifying them. In the final iteration, the LLM is provided with the entire manuscript along with the extracted AUs and is tasked with identifying the relations between the AUs [Lenz and Bergmann, 2025]. Finally, the reviewer is presented with a graph representation of the argument structure. The nodes of this graph contain summaries of the AUs and link to the respective section of the paper. Edges indicate relations between AUs.

5 Discussion

Although the primary objective was to develop a tool supporting the review process, MinervAI also offers considerable utility for authors during manuscript

preparation. Specifically, its argumentation graph can optimize the structural coherence of the manuscript's logical flow, while its citation checker aids in maintaining fidelity to the original sources' findings.

A potential risk involves authors over-optimizing their manuscript based on the tool's output. Nonetheless, because the tool provides supportive visualizations rather than prescriptive quality metrics, any adverse effect is likely limited to subtle biases, such as favoring particular argumentation structures or citation practices.

Crucially, even when utilized extensively by authors, the tool's utility for reviewers remains intact. It is not designed to flag manuscript deficiencies but rather to assist in one of the most laborious aspects of peer review: gaining an initial overview and understanding of the paper. The citation checker assists reviewers in distinguishing between accurately references sources and those that may be misattributed or cited inappropriately. Furthermore, the argumentation graph provides a strategic overview of the core arguments and improves manuscript navigation, enabling reviewers to jump directly between propositions rather than relying on extensive scrolling through the document.

In summary, although the design prioritized the needs of the review process, the tool provides mutual support for both authors and reviewers, all while ensuring that potential biases on the resultant text are minimized.

6 Conclusions and Outlook

Based on the existing literature and the expert interviews, we argue that the use of LLMs for scientific peer review should be confined to individual, verifiable tasks. Key judgments (novelty, technical soundness etc.) must remain with humans, at least as long as LLMs are prone to errors and biases. Based on this, we introduce MinervAI; a system for automating individual parts of the review process.

In a continuation of the project, the expert interviews could be expanded to include the perspectives of journal editors, thereby incorporating the viewpoints of publishers alongside those of authors and reviewers. Future work could also explore an expansion of the tool's functionalities. For instance, the citation checker could not only verify cited references but also suggest relevant papers that have not yet been cited in the manuscript. Similarly, the argumentation graph could be extended to connect content across multiple papers, helping researchers identify conflicting findings or gaps in the literature.

Furthermore, researchers in the expert interviews noted that a major barrier to adoption would be the introduction of a tool that provides information through a separate interface, adding extra steps and friction to their workflow. By integrating additional features such as note-taking capabilities and assisted review writing, MinervAI could evolve into an all-in-one platform for reviewers.

References

B. Aczel, A.-S. Barwich, A. B. Diekman, A. Fishbach, R. L. Goldstone, P. Gomez, O. E. Gundersen, P. T. von Hippel, A. O. Holcombe, S. Lewandowsky, et al. The present and future of peer review: Ideas, interventions, and evidence. *Proceedings of the National Academy of Sciences*, 122(5):e2401232121, 2025.

S. Ahn. A guide to evade hallucinations and maintain reliability when using large language models for medical research: a narrative review. *Annals of Pediatric Endocrinology & Metabolism*, 30(3):115–118, June 2025. ISSN 2287-1012, 2287-1292. doi: 10.6065/apem.2448278.139. URL `http://e-apem.org/journal/view.php?doi=10.6065/apem.2448278.139`.

C. Alvarez, M. Bennett, and L. Wang. Zero-shot Scientific Claim Verification Using LLMs and Citation Text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 269–276, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sdp-1.25. URL `https://aclanthology.org/2024.sdp-1.25`.

W. M. Aly, T. H. A. Soliman, and A. M. AbdelAziz. An evaluation of large language models on text summarization tasks using prompt engineering techniques. *arXiv preprint arXiv:2507.05123*, 2025.

E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, and D. Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274, 2025.

R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor. The Second PASCAL Recognising Textual Entailment Challenge. 2025.

P. Béchard and O. M. Ayala. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv preprint arXiv:2404.08189*, 2024.

L. Bornmann, R. Haunschild, and R. Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15, 2021.

J. Chen, H. Lin, X. Han, and L. Sun. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, Mar. 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i16.29728. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29728`.

A. S. Choi, S. S. Akter, J. Singh, and A. Anastasopoulos. The llm effect: Are humans truly using llms, or are they being influenced by them instead? *arXiv preprint arXiv:2410.04699*, 2024.

M. D'Arcy, T. Hope, L. Birnbaum, and D. Downey. MARG: Multi-Agent Review Generation for Scientific Papers, Jan. 2024. URL http://arxiv.org/abs/2401.04259. arXiv:2401.04259 [cs].

M. Idahl and Z. Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, 2025.

S. Ikeda. Inconsistent advice by chatgpt influences decision making in various areas. *Scientific Reports*, 14(1):15876, 2024.

J. Kim, Y. Lee, and S. Lee. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

S. Krügel, A. Ostermaier, and M. Uhl. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569, 2023.

N. d. Lawrence. The NeurIPS Experiment, 2022. URL https://inverseprobability.com/talks/notes/the-neurips-experiment-snsf.html.

M. Lenz and R. Bergmann. ArgueMapper Assistant: Interactive Argument Mining Using Generative Language Models. In M. Bramer and F. Stahl, editors, *Artificial Intelligence XLI*, volume 15446, pages 189–203. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-77914-5 978-3-031-77915-2. doi: 10.1007/978-3-031-77915-2_14. URL https://link.springer.com/10.1007/978-3-031-77915-2_14. Series Title: Lecture Notes in Computer Science.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020.

R. Liu and N. B. Shah. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing, June 2023. URL http://arxiv.org/abs/2306.00622. arXiv:2306.00622 [cs].

R. Muehlhoff and M. Henningsen. Chatbots im Schulunterricht: Wir testen das Fobizz-Tool zur automatischen Bewertung von Hausaufgaben, 2024. URL https://arxiv.org/abs/2412.06651. Version Number: 5.

S. Musslick, L. K. Bartlett, S. H. Chandramouli, M. Dubova, F. Gobet, T. L. Griffiths, J. Hullman, R. D. King, J. N. Kutz, C. G. Lucas, et al. Automating the practice of science: Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences*, 122(5):e2401238121, 2025.

M. Naddaf. AI is transforming peer review — and many scientists are worried. *Nature*, 639(8056):852–854, Mar. 2025. ISSN 1476-4687. doi: 10.1038/d41586-025-00894-7. URL https://www.nature.com/articles/d41586-025-00894-7. Bandiera_abtest: a Cg_type: News Feature Publisher: Nature Publishing Group Subject_term: Machine learning, Publishing, Lab life.

K. Nassiri and M. A. Akhloufi. Recent Advances in Large Language Models for Healthcare. *BioMedInformatics*, 4(2):1097–1143, Apr. 2024. ISSN 2673-7426. doi: 10.3390/biomedinformatics4020062. URL https://www.mdpi.com/2673-7426/4/2/62.

Neurips. Neurips 2024 Fact Sheet, 2024.

S. Sanyal, T. Xiao, J. Liu, W. Wang, and X. Ren. Are machines better at complex reasoning? unveiling human-machine inference gaps in entailment verification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10361–10386, 2024.

D. Song, W.-C. Lee, and H. Jiao. Exploring llm autoscoring reliability in large-scale writing assessments using generalizability theory. *arXiv preprint arXiv:2507.19980*, 2025.

N. Song, H. Cheng, H. Zhou, and X. Wang. Linking Scholarly Contents: The Design and Construction of an Argumentation Graph. *KNOWLEDGE ORGANIZATION*, 49(4):213–235, 2022. ISSN 0943-7444. doi: 10.5771/0943-7444-2022-4-213. URL https://www.imrpress.com/journal/KO/49/4/10.5771/0943-7444-2022-4-213.

R. Walker and P. Rocha Da Silva. Emerging trends in peer reviewâ€"a survey. *Frontiers in Neuroscience*, 9, May 2015. ISSN 1662-453X. doi: 10.3389/fnins.2015.00169. URL http://www.frontiersin.org/Brain_Imaging_Methods/10.3389/fnins.2015.00169/abstract.

L. Wang and Y. Shen. Evaluating Causal Reasoning Capabilities of Large Language Models: A Systematic Analysis Across Three Scenarios. *Electronics*, 13(23):4584, Nov. 2024. ISSN 2079-9292. doi: 10.3390/electronics13234584. URL https://www.mdpi.com/2079-9292/13/23/4584.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Q. Wei, S. Holt, J. Yang, M. Wulfmeier, and M. v. d. Schaar. The AI Imperative: Scaling High-Quality Peer Review in Machine Learning, June 2025. URL http://arxiv.org/abs/2506.08134. arXiv:2506.08134 [cs].

L. Zhou, R. Zhang, X. Dai, D. Hershcovich, and H. Li. Large Language Models Penetration in Scholarly Writing and Peer Review, Feb. 2025. URL http://arxiv.org/abs/2502.11193. arXiv:2502.11193 [cs].