SCIENCE CHINA
Information Sciences

• PERSPECTIVE •

November 2025, Vol. 68, Iss. 11, 216101:1–216101:3
https://doi.org/10.1007/s11432-025-4588-2

The dual-edged sword: artificial intelligence's evolving role in academic peer review

Xuanjing HUANG*, Shihan DOU† & Zhangyue YIN†

College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433, China

Received 23 April 2025/Revised 23 June 2025/Accepted 11 September 2025/Published online 28 October 2025

Citation Huang X J, Dou S H, Yin Z Y. The dual-edged sword: artificial intelligence's evolving role in academic peer review. Sci China Inf Sci, 2025, 68(11): 216101, https://doi.org/10.1007/s11432-025-4588-2

Peer review is a fundamental mechanism in academic publishing that necessitates specialized expertise. The rigor and reliability of the peer review process determine the quality of conferences and journals. As artificial intelligence (AI) becomes widely applied across various domains, the academic peer review process stands at a critical crossroads. With the explosive growth of research publications, particularly in AI-related fields, traditional review mechanisms are experiencing unprecedented strain [1, 2]. This issue is evident in multiple concerning aspects: the number of qualified reviewers has not matched the growing demand, and the quality of reviews often deteriorates under escalating workloads [2]. Consequently, an increasing number of manuscripts face difficulties in being paired with suitably specialized reviewers for comprehensive and rigorous evaluation.

Statistical data present a compelling scenario. For instance, Figure 1(a) illustrates that the number of manuscript submissions at AI-related conferences has reached unprecedented levels[1)]. Indeed, this trend is generally positive as it signifies that AI is attracting an increasing number of researchers, including those aiming to leverage AI in other areas and early-career researchers. However, it also brings negative repercussions. Many of these new researchers may lack a solid background in AI. Despite the rapid increase in submissions, the number of reviewers with specialized AI expertise has not grown correspondingly. As illustrated in Figure 1(b), this results in professional reviewers being assigned increasingly heavy workloads, substantially overburdening the review systems. All data presented in this figure are sourced from the official conference reports [1,2]. Specifically, as the volume of work grows, the review process may become superficial, with reviewers giving cursory feedback such as "lacks novelty" or requesting additional "experimental baselines" without substantial justification. The degradation in review quality ultimately affects the quality of papers presented at conferences and even leads to outstanding work not getting the recognition it deserves.

The risks behind AI in peer review. With the advancement of AI, some reviewers have attempted to utilize AI systems to assist with the review process for the substantial number of submissions they receive [3]. For example, recent research [3] using the GPTZero LLM detector estimated that at least 15.8% of reviews were AI-assisted in the International Conference on Learning Representations (ICLR) 2024, with 49.4% of all submitted manuscripts receiving at least one review classified as AI-assisted. The effectiveness of LLM detectors, however, remains widely debated within the scientific community. Some researchers advocate for embracing AI tools to improve review efficiency, while others maintain strong opposition. The Computer Vision and Pattern Recognition Conference (CVPR), for instance, explicitly states that LLM-assisted reviewing is "highly irresponsible" and constitutes grounds for immediate desk rejection[2)]. This conservative stance stems from legitimate concerns regarding AI systems' current limitations in critical scholarly evaluation, including potential hallucinations and the inability to verify empirical claims.

There is ample evidence indicating that AI systems are prone to generating hallucinations [4], potentially using incorrect information to assess manuscripts, providing suggestions and opinions that are irrelevant or erroneous. Moreover, when reviewers challenge AI-generated content, AI systems may be inclined to accommodate the reviewers' perspectives, which can inadvertently reinforce the reviewers' confidence in their own views—even if those views are incorrect. Another reasonable concern is that AI-assisted reviewing might reduce scholarly evaluation to a mere technical assessment, focusing solely on the technical aspects of a manuscript to determine its quality. However, the review process also includes critical evaluations of ethical values and disciplinary significance. Whether AI systems can align with human judgment in these aspects still requires thorough investigation.

The consistency of review opinions provided by AI is an-

* Corresponding author (email: xjhuang@fudan.edu.cn)
† These authors contributed equally to this work.

1) Acceptance rates for the major top-tier AI-related conferences. 2025. https://github.com/lixin4ever/Conference-Acceptance-Rate.

2) CVPR 2025 Reviewer Guidelines. 2025. https://cvpr.thecvf.com/Conferences/2025/ReviewerGuide lines.

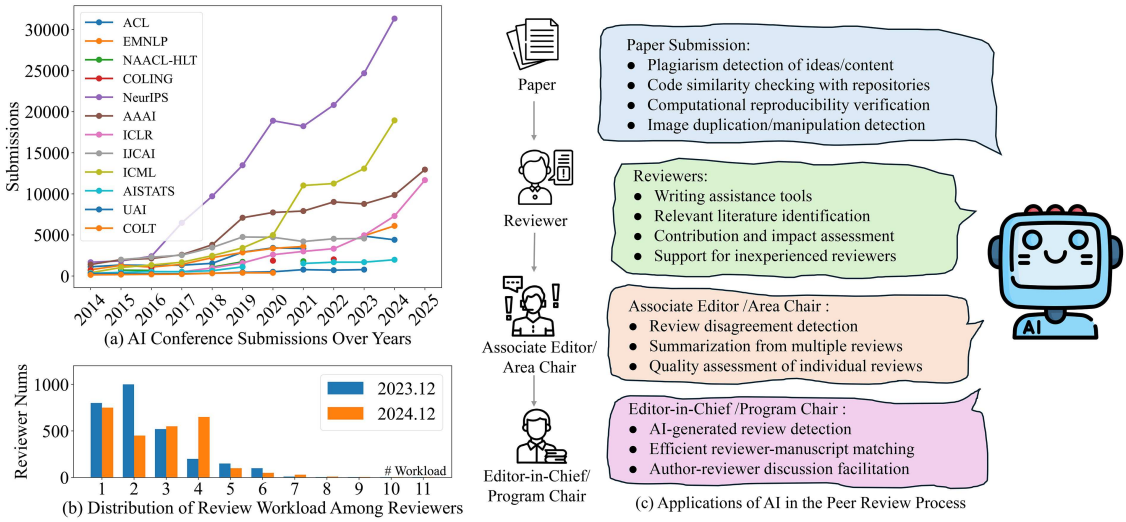

Figure 1 (Color online) The integration of AI in peer review systems. (a) Submissions to AI conferences have increased dramatically in recent years[1)]. (b) While submission volumes have surged, the reviewer pool has remained relatively stable, substantially increasing per-reviewer workload [1, 2]. (c) Potential applications of AI across the peer review ecosystem, including submission verification, reviewer assistance, associate editor decision support, and editorial process management.

other aspect to ponder. Different reviews from the same AI, or reviews from various AI systems, could yield differing opinions, creating uncertainty regarding which review to trust or whether to accept or reject all opinions. If reviewers blindly rely on AI-generated evaluations, it could significantly undermine the credibility of the review process and potentially interfere with the judgment of Area Chairs. Research also finds that AI-assisted reviewing could introduce unfairness in manuscript acceptance. These evidence suggest that AI-assisted reviewing can significantly influence the acceptance of a manuscript, which poses a considerable fairness issue for authors.

Moreover, AI-based review systems could be susceptible to jailbreak attacks. Malicious individuals or organizations might exploit vulnerabilities by submitting papers with specially crafted formats or embedded information to manipulate review outcomes in their favor. Authors could also attempt to identify and leverage weaknesses in AI review systems to secure advantageous results, thereby undermining the fairness and integrity of the peer review process. Fortunately, ICLR 2026 has explicitly recognized and prohibited such practices[3)]. The updated policy specifically bans the use of hidden "prompt injections", such as embedding invisible or misleading instructions in submissions to influence AI-generated reviews—and treats such acts as collusion. Violators may face severe consequences, including immediate desk rejection of their submissions. This regulatory stance further reinforces the commitment to research integrity in the era of AI-assisted peer review.

Beyond the concerns of conference publishers and reviewers, authors might also be apprehensive about AI-assisted review technology. Currently, many researchers prefer not to have their manuscripts reviewed by AI, expressing skepticism towards AI-generated suggestions. AI feedback could be too broad or ambiguous, making it challenging for authors to refine their manuscripts based on such feedback. This could create a negative cycle, leading to doubts about the overall quality of the conference. Consequently, some researchers might lose out due to such irresponsible reviewing, hindering advancements in the field. Most critically, the "black box" nature of many AI systems complicates accountability when errors occur. If an AI system's erroneous judgment leads to the rejection of valuable papers or the acceptance of flawed research, it raises the question of who should be held responsible.

The bright side of AI in peer review. Despite the aforementioned drawbacks of AI systems, some conferences are exploring the feasibility of AI-assisted review in limited ways. For instance, under the ARR rolling review policy, AI can be used responsibly[4)]. Some conferences have ventured into bolder experiments to explore the feasibility of AI in the review process. For example, in 2025, ICLR has implemented an AI Agent system designed to analyze reviewer comments, seek clarifications on ambiguous feedback, and encourage more detailed critiques [5]. The AI agent delivers focused and precise feedback specifically tailored to reviewer comments. In the study, 26.6% of reviewers chose to update their reviews based on the AI-generated feedback, incorporating a total of 12222 suggestions from the AI agent. The revised reviews were found to be more informative and clearer than the original versions, with an average increase of 80 words. This intervention also resulted in increased engagement between reviewers and authors during the rebuttal period. Overall, the use of AI feedback significantly extended the length of reviews and enhanced reviewer participation in author-reviewer discussions, thereby fostering more productive scholarly dialogue. Building on these experimental deployments, ICLR 2026 has further strengthened its policies to address the broader ethical and social implications of AI in peer review[3)], signaling a growing social consensus for the responsible governance of AI in scientific evaluation.

In Figure 1(c), we analyze the potential applications of AI across three key stages of the academic review process:

3) Policies on Large Language Model Usage at ICLR 2026. 2026. https://blog.iclr.cc/2025/08/26/policies-on-large-language-model-usage-at-iclr-2026.

4) ARR Reviewer Guidelines. 2025. https://aclrollingreview.org/reviewerguidelines.

paper submission verification, reviewer assistance, and administrative oversight. Modern AI systems offer substantial enhancements to submission verification through multiple sophisticated mechanisms. While traditional plagiarism detection focuses merely on textual overlap, advanced semantic models can now identify conceptual similarities between submissions and existing literature, detecting intellectual plagiarism even when ideas are expressed with different terminology. For computational research, AI can automatically execute submitted code against standardized datasets and compare the outputs with results reported in manuscripts, thus ensuring computational reproducibility without labor-intensive manual verification. In the visual domain, multimodal AI frameworks employ advanced computer vision techniques to detect subtle image duplication and manipulation across publications. Furthermore, specialized algorithms can analyze code similarity against open-source repositories and previously published works, efficiently identifying potential algorithmic plagiarism and improper attribution of computational contributions.

For reviewers, AI provides several practical tools to address common challenges in the peer review process. AI-assisted literature analysis platforms can efficiently retrieve, synthesize, and summarize relevant publications pertaining to a manuscript's subject matter, thereby offering reviewers a comprehensive overview of the current research landscape. This functionality is especially beneficial for reviewers whose expertise lies in related, but not identical, subfields, as it enables them to more accurately contextualize and assess the significance of a manuscript's contributions. Moreover, AI can facilitate more effective communication throughout the peer review process. For reviewers who are non-native English speakers, AI-powered tools can assist in refining their review comments, enabling them to produce well-structured, grammatically accurate assessments that clearly convey their evaluations. Similarly, authors can utilize AI to improve the clarity and coherence of their responses during the rebuttal stage, promoting more effective and constructive exchanges. AI systems can also support reviewers in contribution and impact assessment by analyzing citation networks, research trends, and methodological approaches across disciplines. For junior or less experienced reviewers, AI can serve as a training and support tool, highlighting standard evaluation criteria, suggesting relevant considerations specific to different research methodologies, and pointing out aspects that might merit particular attention. As noted by Thakkar et al. [5], AI assistants can help junior reviewers broaden their review perspective and provide more thorough assessments despite potential gaps in expertise or familiarity with the review process.

With the assistance of AI, Area Chairs or Associate Editors can efficiently synthesize the perspectives of all reviewers, identify points of consensus, and assess whether a manuscript exhibits significant flaws. Additionally, AI can help them detect disagreements among different reviews [6], enabling a more effective evaluation of manuscript quality and facilitating the production of more accurate meta-reviews. Given access to both the manuscript and reviewer comments, AI can also assist Area Chairs in analyzing whether reviewers' statements are factually consistent with the manuscript, allowing for the precise identification of a manuscript's weaknesses or potential errors in the reviews themselves. AI can significantly enhance the reliability of the peer review process.

At the administrative level, editors can employ AI tools to evaluate review quality systematically, identifying superficial or insufficient feedback that might require further clarification. These systems can automatically prompt reviewers to clarify ambiguous comments or address specific aspects of submissions that remain unevaluated. AI analysis can also detect potentially AI-generated reviews through linguistic pattern recognition and stylistic analysis. For reviewer-manuscript matching, AI algorithms can optimize the assignment process by analyzing the semantic content of submissions alongside reviewer expertise profiles, publication histories, and citation networks. During author-reviewer discussions, AI monitoring systems can facilitate more productive exchanges by identifying contentious language, suggesting constructive reformulations, and ensuring discussions remain focused on substantive scientific issues rather than personal disagreements.

Conclusion. The integration of AI into academic peer review presents a fundamental tension: while AI can substantially enhance efficiency in submission verification, literature analysis, and administrative tasks, it simultaneously raises critical concerns about hallucinations, bias, and the erosion of scholarly integrity. We argue that the authority to recognize and validate scientific knowledge must remain fundamentally in human hands, a responsibility that should never be delegated to machines. No human reviewers should be replaced by AI reviewing, and no human decision-making should be supplanted by AI decision-making. This principle extends beyond current AI limitations; even as models achieve considerable reliability, reviewers bear an unwavering professional and ethical obligation to thoroughly read each manuscript and provide independent, well-reasoned assessments rather than deferring to AI-generated outputs. This commitment to intellectual integrity, coupled with transparent disclosure of any AI assistance, forms the foundation of responsible AI integration. Moving forward, the academic community must establish clear guidelines that preserve human oversight at the core while leveraging AI for well-defined supportive roles. The goal is not to automate peer review but to augment human expertise in ways that address the growing imbalance between submission volumes and reviewer capacity, ultimately strengthening rather than replacing the human judgment that defines rigorous scientific evaluation.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62376061). The authors wish to thank Prof. Hong MEI for his valuable comments. We would also like to thank the anonymous reviewers for their helpful feedback.

References

1 Rogers A, Boyd-Graber J, Okazaki N. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: Association for Computational Linguistics, 2023
2 Ku L-W, Martins A, Srikumar V. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok: Association for Computational Linguistics, 2024
3 Latona G R, Ribeiro M H, Davidson T R, et al. The AI review lottery: widespread AI-assisted peer reviews boost paper scores and acceptance rates. ArXiv:2405.02150
4 Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inform Syst, 2025, 43: 1–55
5 Thakkar N, Yuksekgonul M, Silberg J, et al. Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025. ArXiv:2504.09737
6 Kumar S, Ghosal T, Ekbal A. When reviewers lock horns: finding disagreements in scientific peer reviews. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 16693–16704