Received: 28 Jul 2025, Revised: 29 Sep 2025, Accepted: 1 Jun 2026, Available online: 8 Sep 2026

Journal of Scholarly Communication
DOI: https://doi.org/10.62160/JSC44

RESEARCH ARTICLE

Generative AI in Peer Review: An Evaluation of Capabilities, Limitations, and Responsible Implementation Strategies

Zhongshi Wang, and Mengyue Gong*

Editorial Department of Systems Microbiology and Biomanufacturing, School of Biotechnology, Jiangnan University, Wuxi 214122, China

Corresponding author: Mengyue Gong (e-mail: mgong@jiangnan.edu.cn).

Abstract:

Generative artificial intelligence (GenAI) has been increasingly integrated in academic publishing process. This study aims to access the performance of GenAI in peer review using multi-disciplinary manuscript samples. Five GenAI models underwent peer review tests, systematically evaluated for completeness, accuracy, and rationality of responses using combined manual scoring and statistical analysis. Findings indicated high accuracy in summarizing manuscript content and identifying suitable journals. However, GenAI models exhibited significant limitations in identifying scientific errors and evaluating paper innovation. The results demonstrate clear competency boundaries for GenAI within the peer review, particularly concerning tasks requiring nuanced professional judgment. This study clarifies the potential application scope of GenAI in peer review and provides a basis for its responsible future implementation.

Keywords: Generative AI; Large language models; Peer review

1 | Introduction

Generative Artificial Intelligence (GenAI) refers to models and technologies capable of generating content such as text, images, and videos[1]. As GenAI tool, ChatGPT are employed in natural language processing[2], which have significantly impacted academic publishing. As the core of quality control in academic publishing, the peer review process is also undergoing a profound transformation[3]. The "Evaluate Manuscript" AI tool, developed by Elsevier and powered by the GenAI model Scopus AI, can automatically assess how well the manuscript fits the scope of the journal, demonstrating the technical feasibility of AI-assisted peer review[4]. Attitudes towards AI vary among different journals, publishers and academic publishing organizations. *The BMJ* has issued a statement permitting reviewers to use AI to enhance the language of their review reports, while emphasizing that reviewers must continue to uphold the confidentiality of the peer review process[5]. Taylor & Francis and SAGE have similar AI policies[6-7]. By contrast, Cell Press completely prohibit the use of AI tools in peer review[8]. Besides, Elsevier and Springer Nature are currently exploring and developing secure and controllable AI tools that meet confidentiality requirements[9-10]. The academic community generally holds a cautious yet positive stance towards GenAI tools. There is still controversy over how to use GenAI responsibly.

Researchers have conducted numerous studies on the applications of GenAI to peer review. Farber[11] employed ChatGPT-4.0 to assist in looking for suitable reviewers. The overlap rate between GenAI suggestions and manual selection results reached 42%, and the selection time was

© 2026 The Author(s). Published by Inforvidor Academic Service Group Ltd. This is an open access article under the terms of the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Received: 28 Jul 2025, Revised: 29 Sep 2025, Accepted: 1 Jun 2026, Available online: 8 Sep 2026

shortened by 73%. Liang et al.[12] conducted a large-scale study on GPT-4 feedback from over 5000 papers. The overlap rate between GPT-4 review feedback and the viewpoints of human reviewers exceeded 30%. Cox[13] tested the peer review of AI assistants in human-machine collaboration, showing that AI assistants could generate constructive feedback on research documents. The auxiliary performance of GenAI in peer review is limited by factors such as disciplinary fields, task types, and tested models. Although it has demonstrated certain capabilities in assisting peer review, its accuracy and review quality still require further verification, and there are also persistent issues such as bias and hallucination[14]. Exploring the boundaries of GenAI application in peer review has become a critical challenge that journals demand urgent attention to address.

Building on the technological advancements of GenAI, the study selected high-performance models such as ChatGPT-4o and DeepSeek R1 for evaluation. The tests centers on evaluating the capability of GenAI models in uploaded papers identification, journal scope alignment, scientific errors detection, and review comments drafting, aiming to assess their utility in assisting peer review and inform the formulation of GenAI usage guidelines for this process.

2.| Methodology

2.1 | Materials

A total of 160 papers were selected from the following sources: (1) Group A (Innovation Assessment): Papers submitted to *Systems Microbiology and Biomanufacturing* (SMAB) but rejected due to insufficient innovation; (2) Group B (Control Group): Published papers in SMAB; (3) Group C (Scientific Error Detection): Papers retracted due to scientific errors (Reported by https://retractionwatch.com); (4) Group D (Journal Scope Assessment): Papers published in journals outside the scope of SMAB. Each group comprised 40 papers.

To ensure the security of author information, for rejected papers (group A), only the abstracts were used for test. For papers in other groups, the entire text are uploaded to help the GenAI models fully understand the papers. The author information and publication information have been removed from all the texts and materials. The basic information of tested papers has been uploaded to Science Data Bank[15]. Among them, for the rejected papers, confidentiality measures were taken to protect the author information.

According to the GenAI model leaderboard of OpenCompass[16], 5 of the top 10 models on the list, namely ChatGPT-4o, Doubao-1.5, Qwen3, DeepSeek R1, and ERNIE-X1-Turbo, were selected. The ranking of this leaderboard is determined by OpenCompass based on the comprehensive performance of GAI models in language, knowledge, reasoning, math, and code. ChatGPT-4o is serving as an industry benchmark in terms of natural language understanding and generation. The other four GAI models represent the mainstream level of GAI models of China. The tested models and relevant websites are shown in Table 1.

TABLE 1 | Information of tested GAI models

	GAI model	Website	Company	Release
1	ChatGPT-4o	https://chatgpt.com	OpenAI	2024/5/13
2	DeepSeek R1	https://chat.deepseek.com	DeepSeek	2025/1/20
3	Qwen3	https://tongyi.aliyun.com/read	Alibaba	2025/4/29
4	ERNIE-X1-Turbo	https://yiyan.baidu.com	Baidu	2025/4/25
5	Doubao-1.5-pro	https://www.doubao.com/chat/	ByteDance	2025/4/17

2.2 | Methods

All GenAI models were accessed, registered and used according to the prompts through official website. In accordance with the peer review requirements of SMAB, programmatic prompt words were designed, and the GenAI models were required to summarize, evaluate and make judgments on the tested papers according to the prompt word. The prompt words were contentiously modified until the GenAI model generated clear and complete peer review comments. Prompt words were formed respectively for the original article and the review paper. Detailed prompt words, along with their various versions from the training process, have been uploaded to Science Data Bank[15].

GenAI model were asked according to the prompt words without using the online search function. Given that GenAI models are constantly being updated, to maintain the stability of the tests, all tests were completed within one week from May 19th to May 25th, 2025.

For the comments generated by the GenAI models, the accuracy rate of uploaded papers identification and the words of comments were statistically analyzed. The accuracy rate of GenAI models' comments were determined compared with editors' judgement. The rationality of generated comments were evaluated based on the number of consistent comments with human reviewers.

Two managing editors evaluated the test papers in accordance with the requirements of the SMAB independently. They summarized the manuscripts, assessed whether their themes aligned with the scope of SMAB and

evaluated their innovativeness. Based on this, they scored the comments of GenAI models. During the evaluation process, the managing editors compared each requirement listed in Table 2 with the corresponding comments generated by the GAI models, determined whether the GAI models' comments met the requirements, and then scored 1 or 0 point respectively. All managing editors hold PhDs in fields relevant to the journal. The scoring criteria are detailed in Table 2.

TABLE 2 | Scoring criteria for the comments generated by the GAI models

Dimensions	Score	Scoring criteria
Paper identification	0~1	One point for correctly identifying the uploaded paper.
Paper summary	0~1	One point for accurately summarizing the paper.
Scope identification	0~1	One point for accurately identifying the relationship between the topic of the paper and the scope of the journal.
Acceptance	0~1	If the judgment on whether the paper is accepted is consistent with the actual result, one point will be given.
Suggestions	0~1	One point will be given for providing targeted suggestions.

2.3 |Statistical processing and figure formatting

The significance of scores between different GAI models was tested with Wilcoxon signed-rank test, and the significance level *P* was 0.05 for significant differences. The abstract (group A) and the full text (group B-D) were evaluated separately. Spearman correlation tests were employed to analyze the relationships between the average word count of review comments and both judgment accuracy rates and average scores. The significance level *P* was 0.05 for significant differences, and Spearman's rank correlation coefficient ρ accessed the relationship between variables. Cohen's Kappa coefficient was utilized to assess inter-rater agreement between managing editors' scores, with separate analyses conducted for the 40 abstract-based test papers in Group A and the 120 full-text tested papers in Groups B-D. Scoring consistency was used as the evaluation indicator (score > 3 or ≤ 4).

Statistical analyses were carried out in R 4.2.2 using the packages ggplot2 and ggpubr. Final figure assembly and formatting were subsequently performed in Hiplot Pro (version 2.0, Hiplot, China), SigmaPlot 12.5 (Systat, USA), or Microsoft Excel 2021 (Microsoft, USA), as appropriate.

3 | Results

3.1 | GenAI models may misidentify papers

Several tested GenAI models misidentified the uploaded papers (Figure 1) and produced irrelevant review comments. Specifically, Qwen3 misidentified twenty-eight papers, accounting for 17.5%, whereas the misidentification rates of the remaining four GenAI models were all below 2%. The discrepancy of the misidentification rate may be attributed to differences in the file processing workflows of these models. For full-text evaluation, Qwen3 requires users to upload papers to its text repository. During the text extraction process, interference from other files in the repository may occur[17].

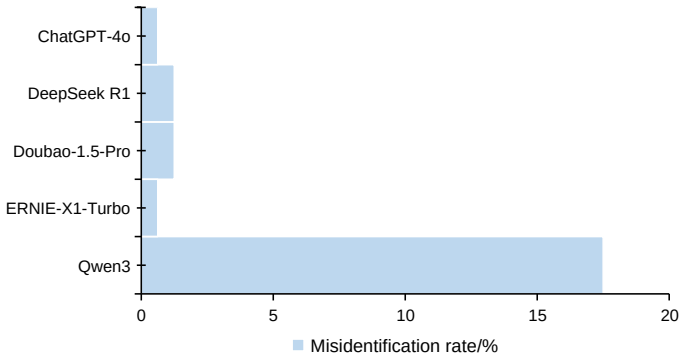

FIGURE 1 | Misidentification rate of the GAI models for uploaded papers.

GenAI models are subject to token limits[18]. Since the word count of the full text exceeds the parsing limit of ERNIE-X1-Turbo, some tested papers were uploaded in segments for evaluation. The segmented processing damaged the integrity of the papers, leading to inconsistent evaluation criteria of the GenAI models. Furthermore, when inputting an incomplete paper, ERNIE-X1-Turbo generated completions for the missing sections based on existing content. Although the completed content seemed reliable, it did not align with the original content of the paper. For the other four GAI models, the full text of tested papers was uploaded completely for testing.

3.2 | GenAI models were still unable to perform judgmental work

GenAI-generated comments were compared with editorial judgments to assess the accuracy of the GenAI models. The overall accuracy of tested models was between 30% and 50%, as shown in Figure 2.

Received: 28 Jul 2025, Revised: 29 Sep 2025, Accepted: 1 Jun 2026, Available online: 8 Sep 2026

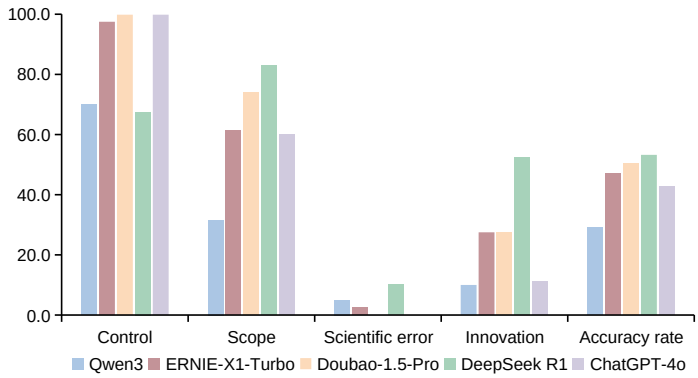

FIGURE 2 | The accuracy rate of GAI model comments.

Tested GAI models accepted most of the published papers. However, Qwen3 and DeepSeek R1 rejected some of the published papers in the control group. Specifically, DeepSeek R1 identified insufficient experimental data in some theoretical papers. For example, when evaluating tested paper B15 (means the 15th paper of Group B; the same applies below), DeepSeek R1 replied: "*Molecular docking is a well-established method; the novelty lies in the specific compounds tested rather than the methodology. No experimental validation (e.g., enzymatic degradation assays) to confirm computational predictions.*" Qwen3 lacked specific targeted comments. For example, when evaluating tested paper B24, Qwen3 replied: "*In terms of quality, the analysis of data and design of experiments seem reasonable, but lack detailed descriptions of experimental methods and statistical analyses.*" Qwen3 did not provide detailed explanations for this comment, nor did it offer suggestions for revision.

DeepSeek R1, Doubao-1.5-pro, and ENRIE-X1-Turbo performed well in distinguishing paper scopes. ChatGPT-4o correctly distinguished papers with obviously different scopes but misclassified those with similar topics. Qwen3 only correctly distinguished thirteen papers, achieving an accuracy rate of only 32.5%.

DeepSeek R1 also demonstrated strong performance in innovation assessment by evaluating the potential similarity between the reviewed papers and similar studies in the training dataset. When evaluating paper A25, DeepSeek R1 replied: "*Similar PEG-based purification methods are well-documented in prior literature*." ENRIE-X1-Turbo and Doubao-1.5 failed to clearly identify limitations of papers in innovation. Qwen3 and ChatGPT-4o only made judgments from the scope relevance of papers and did not provide assessments of innovation. Additionally, limited by the timeliness of training dataset updates, GenAI models inherently face constraints in assessing innovation. Furthermore, even for GenAI models equipped with network retrieval functions, the online data they can access remains quite limited.

All tested GenAI models failed to identify scientific errors in the evaluated papers. Notably, DeepSeek R1 sought to assess experimental findings, including the comprehensiveness of the experimental design and the reliability of the data. When evaluating paper A22, DeepSeek R1 replied: "*Small sensory panel (n=10), no statistical tests for significance (e.g., ANOVA), and minimal replication details.*" The comment provided references for reviewers and editors. However, when generating comments, GenAI models failed to provide clear evidence and were prone to forging references or wrongly citing literature[19]. Therefore, the comments of GAI models were unable to be directly adopted.

The above results indicated that the tested GenAI models failed to demonstrate judgment capabilities in peer review. Although the tested GenAI models summarized papers and accessed according to the review criteria, they still failed to draw accurate conclusions.

3.3 |The rationality of the GenAI models' comments was limited

ENRIE-X1-Turbo and ChatGPT-4o rarely provided targeted comments, primarily summarizing the content of the paper and making judgments. When evaluating multiple papers consecutively, Qwen3 generated templated responses. Constrained by the token limits of GAI models, the number of comments proposed by these models was relatively low, typically ranging from 2 to 4. Compared with the average consistency of human review comments (40 published papers in SMAB), Doubao-1.5-pro and DeepSeek R1 demonstrated relatively high consistency. Among them, Doubao-1.5-pro accumulated nearly 50 consistent comments (Figure 3). Overall, although some GenAI models offered revision comments, their ability to evaluate key aspects, such as the rationality of experimental design, data reliability, and writing logic, remained insufficient.

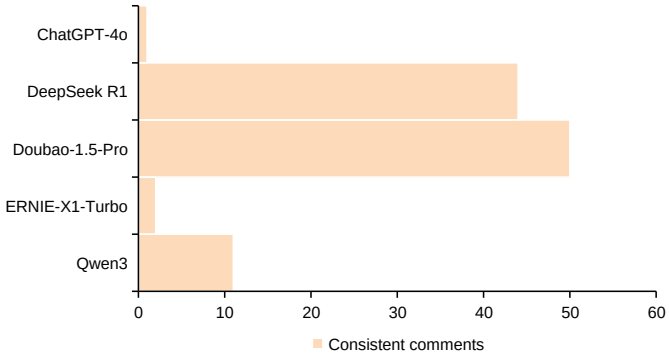

FIGURE 3 | Consistent comments between GAI models and human reviewers.

Received: 28 Jul 2025, Revised: 29 Sep 2025, Accepted: 1 Jun 2026, Available online: 8 Sep 2026

The managing editors evaluated the comments of the GenAI models across five dimensions: file identification, paper summarization, scope differentiation, innovation assessment, and targeted commenting. The evaluation scores were mainly clustered around 3 or 4 points (Table 3-4), which can be attributed to two main factors: First, the comments provided by the GenAI models lacked specificity and were predominantly template-based content. Second, the accuracy of judgment remained suboptimal, particularly in evaluating papers from similar fields and assessing innovation.

Using scoring consistency as the evaluation indicator (score > 3 or ≤ 4), the inter-editor agreement of group A was almost perfect agreement (κ=0.839), and the inter-editor agreement of group B-D was substantial agreement (κ=0.764). This result indicates that responsible editors exhibit high consistency in their scoring of the GAI model.

TABLE 3 | Manual scores of the comments generated by the GAI models in group A

GAI model	Scoring points of GAI models' comments					
	0 point	1 point	2 points	3 points	4 points	5 points
Qwen 3	0	0	18	42	14	6
ERNIE-X1-Turbo	0	0	10	50	20	0
Doubao-1.5-Pro	0	0	3	52	25	0
DeepSeek R1	0	2	4	12	31	31
ChatGPT-4o	0	0	18	56	6	0

TABLE 4 | Manual scores of the comments generated by the GAI models in group B-D

GAI model	Scoring points of GAI models' comments					
	0 point	1 point	2 points	3 points	4 points	5 points
Qwen 3	56	0	40	6	131	7
ERNIE-X1-Turbo	2	4	49	132	53	0
Doubao-1.5-Pro	6	0	15	67	115	37
DeepSeek R1	2	0	6	10	170	52
ChatGPT-4o	2	0	15	0	223	0

The average scores of the comments generated by the GAI models were calculated, as shown in Figure 4. DeepSeek R1, which accurately identified some papers lacking innovation and provided targeted revision comments, attained the highest average score (4.08 points). ChatGPT-4o, despite not offering targeted revision comments, demonstrated high accuracy in file identification and summarization (3.59 points). Doubao-1.5-pro also accurately identified files but produced overly brief paper summaries that failed to provide sufficient information (3.56 points). Other GenAI models, limited by issues such as low accuracy in paper summarization and errors in file identification, achieved comparatively low average scores.

3.4 |The completeness of the comments

Most comments generated by GenAI models ranged from 200 to 500 words in length. In the tests, ChatGPT-4o and DeepSeek R1 produced the longest comments (Figure 5). Specifically, longer comments typically included more comprehensive summaries of the papers.

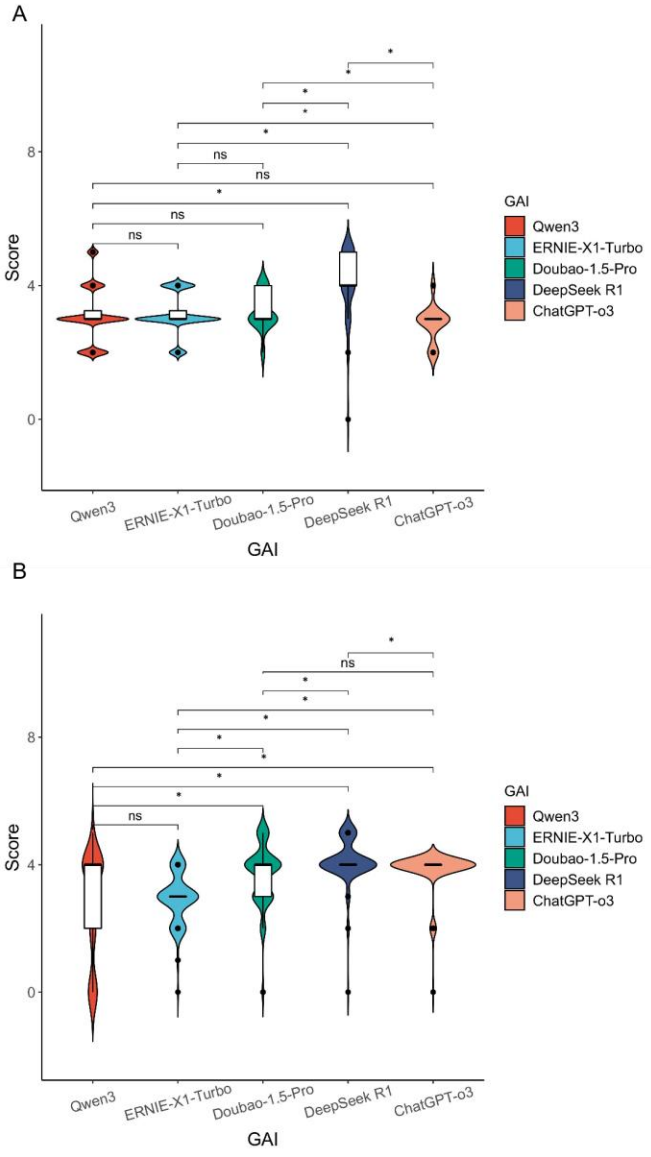

FIGURE 4 | Average score and distribution of comments generated by the GAI models.
A, A group, which used abstracts; B, B-D groups, which used full text. * indicates a significant correlation (P<0.05).

Received: 28 Jul 2025, Revised: 29 Sep 2025, Accepted: 1 Jun 2026, Available online: 8 Sep 2026

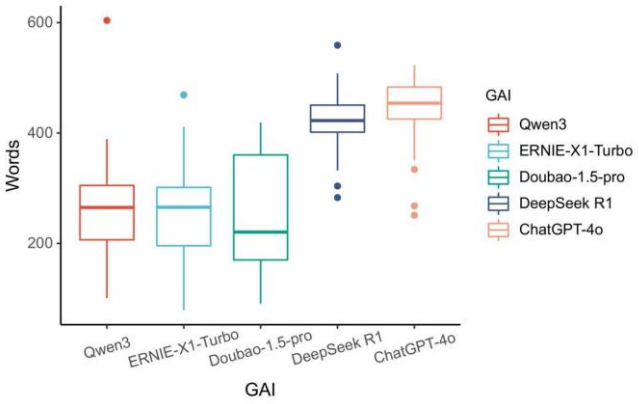

FIGURE 5 | The average number of words of the comments generated by the GAI models.

The relationships among the average word count of comments, judgment accuracy, and average score were analyzed (Table 5). Results from the Spearman correlation test showed no significant correlation between the average word count of comments and either judgment accuracy (P=0.660) or average score (P=0.491). Notably, a significant positive correlation was observed between judgment accuracy and average score (ρ=0.786, P=0.0459). This finding suggests that the length of evaluation comments does not necessarily translate to more accurate judgments.

TABLE 5 | Spearman correlation test of the comments word count, judgment accuracy rate and average scores

	Comments words	Judgment accuracy	Average scores
Comments words		0.660	0.491
Judgment accuracy	0.660		0.0251*
Average scores	0.491	0.0251*	

The figures in the table represent the P-values of the Spearman correlation test. * indicates a significant correlation (P<0.05).

3.5 |Application boundaries of GAI models in peer review

Tests indicate that GAI models can accurately summarize manuscripts, yet they still face challenges in manuscript assessment and exhibit clear boundaries in peer review. Mainstream GAI models were only capable of conducting basic evaluations based on journal scopes. Among Chinese GAI models, the average scores of DeepSeek R1 and Doubao-1.5-pro had already matched or even exceeded those of ChatGPT-4o. For Chinese journals, considering the convenience of use and localized optimizations of Chinese GAI models, they have gained practical advantages over ChatGPT-4o.

Among the GAI models tested in this study, the four Chinese GAI models are all reasoning models. However, Qwen3 was unable to invoke its "Deep Thinking" feature (which requires internet access) during use, while ENRIE-X1-Turbo encountered difficulties in identifying the test papers due to token limits. By contrast, DeepSeek R1 and Doubao-1.5-pro showed their reasoning abilities, and outperforming ChatGPT-4o in the "Scope" and "Innovation" tests. Nevertheless, ChatGPT-4o still excelled in document recognition and summary writing, highlighting its strengths in natural language processing.

Differences in capabilities ultimately influenced the performance of different GAI models in paper recognition, summary writing, and paper evaluation. GAI models with stronger natural language processing capabilities could more accurately identify the test papers and produce precise summaries of them. Meanwhile, those with stronger reasoning capabilities went beyond mere scope identification, and provided targeted revision suggestions tailored to the test papers.

4 |Discussion

4.1 |The application scenarios of GenAI in the peer review

With the rapid growth of academic papers, review resources have become relatively scarce[20]. During the peer review, issues like perfunctory reviewing, overdue reviews, and mid-review abandonment have emerged, which had a negative impact on publication efficiency[21]. GenAI is expected to alleviate the current shortage of review resources. However, the scope of GenAI application in peer review remains unclear, hindering the responsible use. This study has demonstrated that GenAI holds application potential in the following scenarios.

4.1.1 Summarize and analyze the manuscript

The performance of GenAI in summarizing and analyzing manuscripts is noteworthy. Tests showed that models such as DeepSeek R1 and ChatGPT-4o effectively summarized research content and generated accurate manuscript profiles, which helped reviewers quickly grasp the core points of papers. Given the time constraints of reviewers, it is challenging for them to conduct thorough and detailed evaluations of each manuscript.

Under human supervision, GenAI models can serve as supplementary tools to improve review thoroughness[22]. GenAI models can generate reports assessing experimental rationality and data credibility, providing actionable insights for reviewers and supplementing their feedback.

Based on these reports, reviewers can conduct manual checks on potential issues identified by the GenAI models, thereby enhancing the efficiency of peer review. Additionally, GenAI models significantly alleviate the language barriers on non-native English-speaking reviewers[23]. Given the relatively limited review resources, these applications are anticipated to enhance the overall efficiency of peer review.

4.1.2 Preliminary evaluation of the manuscript

During testing, GenAI models have demonstrated the capability to assess the alignment between the manuscript and the scope of journal. A recent study showed that GPT-4 Turbo has acceptable sensitivity and reasonably high specificity when screening citations[24], which is consistent with the results in this study. Publishers, academic societies, or journals can establish discipline-specific literature databases aligned with their scopes to train GenAI models, thereby enhancing the discipline specificity. If journals incorporate GenAI models into their review workflows, they can assist editors in rapidly screening manuscripts during the initial review phase.

The application of GenAI in academic integrity reviews also enhances efficiency for editors and reviewers. Well-known publishers such as Springer Nature and Taylor & Francis have launched a series of AI tools to detect text similarity and identify issues such as academic misconduct manuscripts from "paper mills"[25-26]. Despite this, GenAI models still face limitations in detecting scientific errors, given that the GenAI models failed to identify scientific errors in retracted papers (Figure 2). By correlating their training databases, GenAI models may tracked the retraction histories of similar studies, providing a reference for manual review. However, due to the outdated training databases, an online search function is necessary to obtain more timely information.

4.1.3 Language improvement

In language improvement, the capability of GenAI has been validated. Bai et al.[27] demonstrated that the GenAI models ChatGLM2 and Llama2-7B can correct grammatical errors in sentences and assist authors in polishing their papers. GenAI models tested in this study also identified some spelling and grammar errors, effectively supplementing the minor issues overlooked by reviewers and highlighting the strengths of GenAI models in simple auxiliary tasks.

The ethical guidelines of some publishers such as ACM, BMJ and Wiley permit reviewers and editors to leverage GenAI for enhancing language quality[28-30]. Given that reviewers hail from diverse global backgrounds, their English proficiency levels and writing styles inevitably vary. Using GenAI to enhance the language quality of peer review comments also helps reduce language barriers among authors, reviewers, and editors. Journals may consider allowing reviewers to use GenAI models for further refinement once the initial review report version has passed the AI similarity check.

4.2 | Suggestions for applying GenAI to the peer review

4.2.1 Ethical guidelines and transparent peer review

It is proposed that the GenAI-assisted peer review system be conditionally piloted. The practical deployment requires the establishment of safeguard mechanisms: Reviewers or editors must transparently declare the GenAI usage, with such declarations integrated into the transparent peer review process. Furthermore, it is necessary to develop and implement GenAI usage guidelines for peer review, which should establish clear boundaries and accountability frameworks for model application, along with standardized annotation guidelines for AI-generated content. Lin[31] noted that the issues such as biases and insufficient knowledge that people are concerned about in AI models are also problems existing in human peer reviewers. The concerns about AI peer review are actually structural issues in the peer review system. Crucially, the core distinction lies in human reviewers' ability to assume responsibility for their comments, whereas GenAI, as a black-box model, fundamentally lacks the capacity for accountability[32]. Establishing a human oversight and accountability framework for AI could therefore enhance the credibility of AI-assisted review systems.

For diverse use cases of GenAI models in peer review, tailored and specialized guidance should be provided. This ensures the stability and output consistency of GenAI models. Furthermore, it is equally critical to clarify supervisory accountabilities and validation criteria for human reviewers to safeguard the credibility of review outcomes. While reviewers may independently decide whether to use GenAI models, they are required to strictly adhere to relevant guidelines. Additionally, the standardized peer review templates should fully incorporate the application scenarios and requirements of GenAI models.

4.2.2 Data security and protection of decision-making process

In the application of GenAI models, the basic principles of privacy protection and data security must be strictly followed[33]. It is recommended to develop GenAI models under the control and supervision of publishers. When using the GenAI model to assist in peer review, encryption algorithms such as federated learning can be adopted to safeguard sensitive data, fully ensuring the security of personal information and unpublished research results[34]. For GenAI peer review systems involving academic integrity testing, a third-party oversight mechanism should

also be established to regularly verify the deviation rate and misjudgment probability of the algorithm. Furthermore, a complaint and redress mechanism should be implemented to ensure that GenAI in peer review remains supervisable, traceable, and correctable.

4.2.3 Construction of localized knowledge repositories

To mitigate the knowledge gaps in GenAI models, it is recommended to foster collaboration between publishers and GenAI developers, and open the abstract database to compliant GenAI platforms through structured data interfaces[35]. Given the current practical challenges arising from international technical barriers, such as NIH' s data blockade policy[36] and regional restrictions on GPT model access, it is recommended to prioritize cultivating and enhancing the academic service capabilities of domestic GAI platforms. Leveraging the open-source trend of the GenAI models[37] and with the support of publishers or academic societies, developing localized GenAI knowledge repositories would facilitate domain-specific training of GAI models. Journals in similar disciplines could jointly establish such knowledge repositories and locally deploy GenAI models for targeted training, which would mitigate the risk of unintended disclosure of author information or research findings.

4.2.4 Strengthen the GenAI application capabilities of editors and reviewers

GenAI is evolving rapidly, necessitating that editors and reviewers also keep pace with these advancements. To address this need, there is an urgent requirement to establish a standardized set of GenAI training guidelines. Specifically, editors and reviewers should be equipped with the capability to accurately evaluate GenAI-generated comments, with a particular focus on assessing their credibility and reliability. Appropriate tools needed to be available to editors for detecting content generated or altered by GenAI, not limited to the text content, but also covering the AI-generated images, AI data processing and other aspects, so as to comprehensively identify and prevent potential academic misconduct[38]. Editors and reviewers must also optimize the human-AI collaboration workflow, master prompt engineering techniques for GenAI models. Through the implementation of standardized and systematic GenAI training programs, their application proficiency in GenAI can be comprehensively elevated across both methodological and capability development dimensions.

4.3 | Limitations and bias-analysis

The main limitations of this study lie in the limited scale of tested papers and GAI models, insufficient disciplinary coverage, and potential subjective biases of editors. Although editors strictly adhered to the scoring criteria, and this study employed only binary scores (0 and 1) to mitigate subjective bias (Table 2), subjective judgment by editors was still required for sections like "Scope" and "Suggestions."

Given the rapid iterative updates of GenAI models, the judgment criteria and capability of these models may undergo changes during the study period, potentially affecting the robustness of the findings. The pre-training processes and historical interaction logs of prompt providers could influence the generated outputs of GenAI models, introducing additional variability in the results. Throughout this process, as researchers continuously queried the GAI models, the models inevitably compared the test papers, potentially leading to bias against papers tested later.

The process of modifying prompt language also impacts the GAI models' evaluation criteria. For instance, when testing the identification of journal's scope, using the term "*interdisciplinary*" caused the GAI model to exceed the journal's original scope, accepting manuscripts with high tolerance. Conversely, adding the phrase "*in the field of industrial microbiology and biomanufacturing*" led the model to strengthen its scrutiny of scope adherence. The specific prompts used in this study may still contain similar terminology that influences the GAI models' judgment standards.

Additionally, when testing papers, the output consistency of GAI models under repeated queries about the same paper was poor. For example, during testing of paper B1, although Qwen3 provided positive feedback in two consecutive queries, the second response was more comprehensive. In its first reply, Qwen3 stated: "*Based on the content and theme of the paper, this manuscript meets the scope and requirements of the journal and can be submitted for peer review. However, due to the lack of specific research cases and empirical evidence support, further enhancement and supplementation are needed.*" In the second response, Qwen3 replied: "*The manuscript meets the journal's requirements in terms of topic relevance and scientific significance, but it has some shortcomings in data support, depth of analysis, and writing quality. To enhance the quality of the manuscript and meet the journal's standards, it is recommended that the authors supplement more experimental data to support the paper's viewpoints and conclusions, further deepen the analysis of challenges faced by 'in silico' organisms and propose specific solutions, and optimize the structure and language expression of the paper to make it clearer and more understandable.*" It is inferred that the GAI model may have referenced its previous response when generating the answer. To ensure comparability across different GAI models, this study consistently used the first response from each GAI model.

The study prospectively evaluated the performance of GAI models in peer review through experiment, rather than relying solely on historical outcomes. Therefore, it is not classified as a retrospective study.

5 | Conclusion

Although the application of GenAI models in peer review faces significant challenges, their potential remains undeniable. This study demonstrated the applicability and scope constraints of the GenAI model in the peer review process by simulating peer review. Specifically, GenAI models can facilitate the summarization of manuscripts and assessment of their suitability for review, yet they lack the capability to undertake independent decision-making tasks.

In the future, the reliability and objectivity of test findings can be further enhanced through approaches such as longitudinal tracking tests, expanding the scope and scale of testing, and independent evaluations by more editors. It is important to note that the findings of this study only reflect the GenAI models' performance in peer review during the testing phase. With continuous optimizing, GenAI models are expected to improve assessment accuracy, mitigate issues such as misidentification and hallucinations, and play a more positive role in peer review.

Acknowledgement

This research is supported by the Science and Technology Journals Project by China Science Publishing & Media Ltd. (Science Press) and the Jiangsu Society of Science and Technology Periodicals and Science and Technology Journals of Yangtze River Delta & MPS/AJE English Journal Development Fund (No. JSKJQK-MPS/AJE-2025-003).

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Author Contribution Statement

Zhongshi Wang: Investigation, Formal analysis, Writing-Original Draft.

Mengyue Gong: Conceptualization, Investigation, Writing-Review and Editing.

Data Availability Statement

Data will be made available on request.

References

[1] Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33, 63.

[2] Open AI (2022). Introducing ChatGPT. accessed 30 November 2022, https://openai.com/index/chatgpt.

[3] Zou, J. (2024). ChatGPT is transforming peer review — how can we use it responsibly? *Nature*, 635, 10.

[4] Stoop, J. (2024). Announcing the new "Evaluate Manuscript". accessed 23 May 2024, https://www.elsevier.com/connect/announcing-the-new-evaluate-manuscript.

[5] The BMJ (2025). AI use. accessed 10 June 2025, https://www.bmj.com/content/ai-use.

[6] Taylor & Francis. AI Policy. https://taylorandfrancis.com/our-policies/ai-policy.

[7] SAGE. Artificial Intelligence Policy. https://uk.sagepub.com/en-gb/eur/artificial-intelligence-policy.

[8] Cell Press. Information for reviewers. https://www.cell.com/reviewers.

[9] Wodecki, B. (2023). Elsevier Debuts Generative AI Tool for Academic Researchers. AI Business, accessed 3 August 2023, https://aibusiness.com/nlp/elsevier-debuts-generative-ai-tool-for-academic-researchers#close-modal.

[10] Springer Nature (2023). AI use by peer reviewers. https://preview.springer.com/us/editorial-policies/artificial-intelligence--ai-/25428500.

[11] Farber, S. (2024). Enhancing peer review efficiency: A mixed-methods analysis of artificial intelligence-assisted reviewer selection across academic disciplines. *Learned Publishing*, 37(4), e1638.

[12] Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D.Y., Yang, X., Vodrahalli, K., He, S., Smith, D.S., Yin, Y., McFarland, D.A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1, 8.

[13] Cox Jr., L.A. (2024). An AI assistant to help review and improve causal reasoning in epidemiological documents. *Global Epidemiology*, 7, 100130.

[14] Hosseini, M., & Horbach, S.P.J.M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8, 4.

[15] Wang, Z., & Gong, M. (2025). Information of papers in an experiment of GenAI models peer review. accessed 14 April 2025, https://cstr.cn/31253.11.sciencedb.23681.

[16] OpenCompass (2025). CompassBench Large Language Model Leaderboard. accessed 13 May 2025, https://rank.opencompass.org.cn/leaderboard-llm/?m=25-04.

[17] Zhang, Z., Elfardy, H., Dreyer, M., Small, K., Ji, H., & Bansal, M. (2023). Enhancing multi-document summarization with cross-document graph-based information extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1696-1707, Dubrovnik, Croatia. Association for Computational Linguistics.

[18] He, X., Liu, J., & Duan, Y. (2025). 2-D Transformer: Extending large language models to long-context with few memory. *IEEE Transactions on Neural Networks and Learning Systems*, 1-15.

[19] Bengesi, S., El-Sayed, H., Sarker, M.K., Houkpati, Y., Irungni, J., & Oladunni, T. (2024), Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*, 12, 69812-69837.

[20] Kovanis, M., Procher, R., Ravaud, P., & Trinquart, L. (2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS One*, 11(11), e0166387.

[21] Vesper, I. (2018). Peer reviewers unmasked: Largest global survey reveals trends. *NATNEWS*, accessed 7 September 2018, https://doi.org/10.1038/d41586-018-06602-y.

[22] Krishnan, V. (2024). Artificial intelligence in scientific peer review. *Journal of the World Federation of Orthodontists*, 13(2), 55-56.

[23] Liu, S., Pan, L., Jian, D., & Xiong, D. (2025). Overcoming language barriers via machine translation with sparse Mixture-of-Experts fusion of large language models. *Information Processing and Management*, 62(3), 104078.

[24] Oami, T., Okada, Y., Nakada, T. (2024). Performance of a Large Language Model in Screening Citations. *JAMA Network Open*, 7(7), e2420496.

[25] Springer Nature (2024). Springer Nature unveils two new AI tools to protect research integrity. accessed 12 July 2024, https://group.springernature.com/gp/group/media/press-releases/new-research-integrity-tools-using-ai/27200740.

[26] Taylor & Francis (2024). Artificial intelligence (AI) for academic research at Taylor & Francis. accessed 27 May 2025, https://taylorandfrancis.com/about/ai.

[27] Bai, X., Xie, Y., Zhang, X., Han, H., & Li, J.-R. (2024). Evaluation of Open-Source Large Language Models for Metal–Organic Frameworks Research. *Journal of Chemical Information and Modeling*, 64(13), 4958-4965.

[28] ACM Publication Board (2023). ACM Peer Review Policy. accessed 30 November 2023, https://www.acm.org/publications/policies/peer-review.

[29] BMJ Journals. Peer Review Terms and Conditions. https://authors.bmj.com/policies/peer-review-terms-and-conditions/#ai%20use.

[30] Wiley. Wiley Peer Review Policy. https://authorservices.wiley.com/Reviewers/journal-reviewers/tools-and-resources/review-confidentiality-policy.html.

[31] Lin, Z. (2023). Why and how to embrace AI such as ChatGPT in your academic life. *Royal Society Open Sciences*, 10(8), 230658.

[32] Schwartz, I.S., Link, K.E., Daneshjou, R., & Cortés-Penfield, N. (2023). Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. *Clinical Infectious Diseases*, 78(4), 860-866.

[33] COPE council (2017). COPE Ethical guidelines for peer reviewers — English. https://doi.org/10.24318/cope.2019.1.9.

[34] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1-19.

[35] Wiley (2024). Wiley Launches New Partnership Innovation Program to Deliver AI Advantages for Researchers and Practitioners. accessed 17 October 2024, https://newsroom.wiley.com/press-releases/press-release-details/2024/Wiley-Launches-New-Partnership-Innovation-Program-to-Deliver-AI-Advantages-for-Researchers-and-Practitioners/default.aspx.

[36] National Institute of Health (2025). NIH security best practices for controlled-access data and repositories. accessed 4 April 2025, https://sharing.nih.gov/accessing-data/NIH-security-best-practices.

[37] Xiong, L., Wang, H., Chen, X., Sheng, L., Xiong, Y., & Liu, J. (2025). DeepSeek: Paradigm Shifts and Technical Evolution in Large AI Models. *IEEE/CAA Journal of Automatica Sinica*, 12(5), 841-858.

[38] Zielinski, C., Winker, M.A., Aggarwal, R., Ferris, L.E., Heinemann, M., Lapeña Jr, J.F., Pai, S.A., Ing, E., Citrome, L., Alam, M., Voight, M., & Habibzadeh, F. (2023). Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Colombia Medica*, 54(3), e1015868.