AI and the Future of Academic Peer Review

Sebastian Porsdam Mann,[1,2] Mateo Aboy,[3] Joel Jiehao Seah,[2] Zhicheng Lin,[4] Xufei Luo,[5,6] Daniel Rodger,[7] Hazem Zohny,[8] Timo Minssen,[1] Julian Savulescu,[2,8*] and Brian D. Earp[2]

1. Center for Advanced Studies in Bioscience Innovation Law (CeBIL), Faculty of Law, University of Copenhagen
2. Center for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore
3. Faculty of Law, University of Cambridge
4. Department of Psychology, Yonsei University, Seoul
5. Vincent V.C. Woo Chinese Medicine Clinical Research Institute, School of Chinese Medicine, Hong Kong Baptist University
6. Chinese EQUATOR Centre, Hong Kong
7. School of Allied and Community Health, London South Bank University, London
8. Uehiro Oxford Institute, University of Oxford, Oxford

*** Corresponding author.**

Abstract

Peer review remains the central quality-control mechanism of science, yet its ability to fulfill this role is increasingly strained. Empirical studies document serious shortcomings: long publication delays, escalating reviewer burden concentrated on a small minority of scholars, inconsistent quality and low inter-reviewer agreement, and systematic biases by gender, language, and institutional prestige. Decades of human-centered reforms have yielded only marginal improvements. Meanwhile, artificial intelligence—especially large language models (LLMs)—is being piloted across the peer-review pipeline by journals, funders, and individual reviewers. Early studies suggest that AI assistance can produce reviews comparable in quality to humans, accelerate reviewer selection and feedback, and reduce certain biases, but also raise distinctive concerns about hallucination, confidentiality, gaming, novelty recognition, and loss of trust.

In this paper, we map the aims and persistent failure modes of peer review to specific LLM applications and systematically analyze the objections they raise alongside safeguards that could make their use acceptable. Drawing on emerging evidence, we show that targeted, supervised LLM assistance can plausibly improve error detection, timeliness, and reviewer workload without displacing human judgment. We highlight advanced architectures, including fine-tuned, retrieval-augmented, and multi-agent systems, that may enable more reliable, auditable, and interdisciplinary review. We argue that ethical and practical considerations are not peripheral but constitutive: the legitimacy of AI-assisted peer review depends on governance choices as much as technical capacity. The path forward is neither uncritical adoption nor reflexive rejection, but carefully scoped pilots with explicit evaluation metrics, transparency, and accountability.

Introduction

Academic peer review serves as the primary quality control mechanism for scholarly research, tasked with validating methods, assessing significance, detecting errors, and ensuring fair evaluation of scientific contributions. Yet the system faces well-documented challenges: overwhelming submission volumes coupled with a limited pool of competent reviewers (who contribute voluntarily, and are often unpaid), reviewer burnout (Bravo et al. 2019), persistent biases against certain groups and institutions (Murray et al., 2019), inconsistent review quality (Bornmann et al., 2010), and lengthy delays that impede scholarly and scientific progress (Andersen et al. 2021). Previous attempts to address these problems through human-centered interventions, such as reviewer training programs, open review systems, multiple reviewer requirements, and quality checklists, have yielded disappointing results, suggesting that incremental reforms to the existing system may be insufficient (Bruce et al. 2016; Gaudino et al. 2021).

Artificial Intelligence (AI), including generative AI systems based on Large Language Models (LLMs), is now being piloted across the peer review pipeline. Major publishers, including Springer Nature, Elsevier, and Wiley, are testing AI for manuscript screening and compliance checking. Scholars are comparing human and AI assistance in review quality and in detecting AI-mediated writing (Shcherbiak et al. 2024; Liang et al. 2024), and funders like the NIH are starting to issue policies on the use of AI in grant research applications, and in peer review.

These developments are outpacing decision-relevant, task-level guidance on when and how to use LLMs, what safeguards are necessary, and where benefits plausibly outweigh risks. What is missing is a pragmatic, ethics-informed decision map: a linkage from (i) the aims and failure modes of peer review, to (ii) concrete LLM applications, to (iii) the principal objections they raise, and (iv) the safeguards and evaluation criteria that would render their use acceptable. This paper offers an overview of the evidence and arguments in favor and against the use of AI in peer review.

What is peer review supposed to accomplish?

Peer review serves multiple interconnected functions that collectively uphold the integrity of scholarly and scientific knowledge production and contributes to the wider public's trust in the credibility of published research. At its core, the process exists to ensure the quality, originality, and rigor of the findings, validate research methods and claims, ensuring that conclusions follow logically from evidence and that experimental designs meet disciplinary standards (Turner, 2003). The system is simultaneously supposed to act as a safeguard against error and misconduct; from catching statistical mistakes and logical fallacies to detecting plagiarism, data fabrication and manipulation, and undisclosed conflicts of interest (Mollaki, 2024). However, despite these aspirations, there remains a well-documented and widespread replication crisis, which has served to undermine the public's trust (Korbmacher et al. 2023); a problem that AI could, in the future, be used to help address (Straiton 2024). Beyond these evaluative dimensions, peer review embodies normative commitments to

fairness, timeliness, and constructive engagement that define legitimate scientific practice (Turner 2023).

The stakes of effective peer review extend far beyond individual manuscripts to encompass the credibility of entire research programs and policy decisions grounded in scientific evidence. Publication decisions also determine career trajectories through their impact on hiring, promotion, and grant funding. This gatekeeping function generates inherent tensions: reviewers must balance thoroughness against efficiency, maintain standards while encouraging innovation, and provide critical assessment without discouraging legitimate challenges to orthodoxy. The system thus operates under competing pressures that no single reform has successfully reconciled.

Problems with peer review

Unfortunately, empirical evaluations of the peer review process have found that the ability of peer review to fulfill the function described above is inhibited along several key dimensions.

First, the progress of scholarship depends on the availability of recent information, and thus on the rapid dissemination of scholarly works. This is true for all disciplines, but delays are especially consequential in certain fast-moving fields, such as AI or pandemic response. Yet, across fields, the journey from submission to publication often lasts many months. A cross-field analysis found typical end-to-end lags of roughly 9–18 months, varying by discipline and venue (Björk & Solomon, 2013). In line with these results, a more recent systematic review found large variations also within fields, with median submission to publication times for biomedical manuscripts ranging between 70 and 558 days (Andersen et al. 2021). These delays slow the dissemination of results. They can also disproportionately burden early-career researchers, often employed on time-limited grants rather than in tenured or tenure-track positions, and may encourage strategic behaviors (e.g. resubmissions, multiple submissions) that add little value in terms of epistemic progress. They can also negatively impact graduate students in terms of graduation requirements and/or career opportunities in "publish or perish" academia. Preprints might mitigate some, but not all, harms by allowing early access, since preprints are not always counted in evaluations, promotion decisions, and so on. Additionally, some journals still restrict or prohibit authors from posting submitted manuscripts onto a preprint repository.

Relatedly, the economic costs of peer review are substantial. Measured in time and foregone salary, reviewer effort amounts to what is essentially a very large subsidy (for journal publishers, other scholars, and wider society). A recent estimate placed global reviewer time in 2020 at well over 100 million hours, with the imputed wage value for the United States alone exceeding US$1.5 billion (Aczel et al., 2021). A separate survey estimated the annual global economic value of peer review at approximately US$6 billion when including reviewed-but-rejected manuscripts (LeBlanc et al., 2023). These burdens are not evenly distributed; multiple datasets show a highly skewed workload in which a small minority of scholars handles a large share of reviews. According to Publons (2018), which runs a service aimed at recognizing peer reviewers, approximately 10% of reviewers account for around

50% of all reviews; early-career researchers are disproportionately more likely than their senior colleagues to accept reviews (Bravo et al. 2019).

Since journal peer review is typically unpaid and anonymous, and academic advancement remains intensely competitive, individual incentives are weak: reviewers overwhelmingly cite time pressure as the main reason to decline; most want formal recognition from institutions; and very few attach their names to reviews (5.6% at one long-running ecology journal which allows for optional signing of reviews), which limits reputational returns (Publons, 2018; Fox, 2021). The result is a classic common-pool problem: conscientious (often junior) scholars absorb a disproportionate share of labor, resulting in opportunity costs, while the system moves more slowly than it might in the absence of these issues.

Partially as a result, and certainly consistent with this, refusals of peer review invitations have increased significantly over recent years. In a study of five Elsevier journals, the average percentage of accepted invitations to peer review declined from 43.6% to 30.9% between 2010 and 2017 (Bravo et al. 2019). As a result of both this and the fact that the number of submissions increases over time, one study estimated an increase of total peer review invitations at 9.8% per year (Publons, 2018). Fewer acceptances per invitation means longer handling times from having to search for other willing and suitable reviewers (who may be juniors in their fields), and heavier reliance on the same few names, perpetuating the skew in who bears the burden.

Once a peer review invitation is accepted, a different set of issues arise. Most journals require multiple reviewers; this is, in theory, a way to decrease the chance of bias, increase the odds of spotting errors, and help ensure the necessary expertise to review interdisciplinary submissions. However, inter-reviewer agreement for journal manuscripts is routinely low (mean $\varkappa \approx 0.17$) (Bornmann et al., 2010). In internal medicine, for example, reviewers' accept/revise/reject recommendations showed agreement "barely beyond chance" (Kravitz et al., 2010). Thus, the outcome of peer review may have as much to do with chance as with quality of the manuscript assessed.

Further, there is evidence that peer review only catches a small percentage of errors in manuscripts. Randomized "sting" experiments that deliberately seeded manuscripts with serious issues show that reviewers miss many of them. In a landmark BMJ/JAMA study, reviewers detected on average only a quarter (two of eight) of injected design/analysis weaknesses; only 10% of reviewers spotted four or more, and 16% detected none (Godlee et al., 1998). Subsequent studies have also found that reviewers catch less than a third of deliberately inserted major errors; and training programs targeted at improving these detection rates show only minor improvements not sustained past a six-month follow-up (Schroter et al. 2008).

These issues are compounded by multiple intersecting biases. For example, large-scale analyses show clear homophily: editors tend to choose reviewers of their own gender and from familiar networks. Because men occupy a larger share of editorial roles, male editors' homophily amplifies male over-representation among reviewers; female editors also show same-gender preference, though typically weaker (Murray et al., 2019). One such study that found a slightly but significantly greater acceptance

rate for male compared to female senior authors (53.5% versus 50%; Murray et al. 2019). Similarly, in a study looking at the single-blind review system (in which the identity and affiliations of submitting authors are made available to peer reviewers) at a high-end computer science conference, very large biases towards recommending acceptance for papers from famous authors, prestigious universities, and prestigious private enterprises were found (resulting in relative odds ratios of 1.63, 1.58, and 2.10, respectively; Tomkins et al. 2017). Experimental work further indicates that, consistent with linguistic bias, identical scientific content written in non-native-like English receives lower quality ratings (Politzer-Ahles et al., 2020).

None of this implies that peer review "doesn't work." It does many things tolerably well some of the time, and it is hard to replace its social functions outright. But the empirical record shows that the system, as practiced, often falls significantly short of its own ideals. It under-delivers on impartiality, it misallocates hidden labor, and it slows diffusion without always adding commensurate epistemic value. The upshot for this paper is straightforward; any proposal to deploy AI-assisted tools in peer review must be judged not against a romanticized ideal, but against the documented performance of the status quo.

How AI can assist peer review

The use of AI in peer review has been explored and employed for decades. Early applications focused on specific tasks; plagiarism detection through systems like Turnitin (founded 1998) and CrossRef's Similarity Check (introduced 2008), which created databases of millions of articles for manuscript screening. The Toronto Paper Matching System used topic modeling to match submissions with reviewers and was adopted by conferences including NIPS, ICML, and UAI (Charlin & Zemel, 2013). Statistical verification tools like Statcheck detected reporting errors (Nuijten et al., 2016), while the GRIM test found inconsistent means in approximately 50% of tested psychology articles (Brown & Heathers, 2017). While these systems reduced editorial workload, they remained limited to narrow tasks and could not replicate the evaluative functions of human review (Kousha & Thelwall, 2024).

On a theoretical level, several characteristics of LLMs suggest potential value for peer review. They can process large volumes of text quickly, potentially identifying errors or relevant prior work that reviewers might miss. They can provide immediate feedback over multiple rounds. Their multilingual capabilities could reduce bias against non-native English speakers. Several LLMs could be asked to review interdisciplinary work from each relevant disciplinary perspective. Perhaps most fundamentally, they could help address reviewer shortage.

Yet there are also documented concerns that temper the case for LLMs in peer review. Generative AI systems can 'hallucinate,' generating fabricated references or errors not present in manuscripts (Huang et al., 2025), especially if they have to rely on their parametric knowledge and are not given access to tools such as reference databases. They risk reproducing training data biases, potentially suppressing innovative approaches (Hosseini & Horbach, 2023). Moreover, there are concerns that manuscripts contain confidential research that should not be uploaded to commercial AI systems (NIH, 2023). In

addition, normative concerns about transparency have been raised, including how the AI system was developed, trained, and used, as well as the suggestion to disclose details such as the version, prompts, and date of use when AI is applied in the review process (Naddaf, 2025). Finally, LLMs are vulnerable to prompt injection and other adversarial attacks where hidden instructions manipulate outputs (Collu et al., 2025; Ye et al., 2024). We address these and other issues in the following section.

Evidence on general-purpose LLMs in peer review

The evidence on LLMs in peer review is early, heterogeneous, and context-specific (venues, tasks, prompts, and models vary). We therefore confine our claims to defined tasks in specific contexts, and treat any general conclusions as low-certainty signals that support pilot evaluation but not broad proof of superiority.

With that caveat in mind, there is mounting, indicative evidence that researchers already find value in AI for peer review assistance. Surveys indicate many peer reviewers already use LLMs to draft or polish reviews, particularly in computer science conferences (Latona et al., 2024). While self-reported uptake demonstrates willingness, the stronger question is whether such use improves review quality.

Emerging empirical work has begun to address this. One set of studies has looked at overall quality of peer review, as assessed by independent experts or by the authors receiving peer review. These generally find that LLM-assisted reviews, or even fully generated reviews, match or supersede the quality of human reviews. For example, the International Conference on Learning Representation (ICLR) found that offering LLM feedback to the authors of randomly selected reviews improved both the quality and length of reviews revised according to that feedback, as assessed by blinded raters (Thakkar et al. 2025). A large study of more than 5,000 papers from Nature journals, ICLR, and eLife (Liang et al. 2024) found that 57.4% of researchers in their sample found blinded reviews from GPT-4 (now a legacy model) to be helpful overall, and 82.4% rated these as more helpful than at least some human reviewers.

In addition to overall quality, studies suggest that LLM-assisted peer review can help address many of the specific concerns that plague the existing system. For one, the use of LLMs can reduce latency across several aspects of the peer review process. LLMs have been shown in some settings to (i) produce draft review text in minutes or hours rather than days (Thakkar et al. 2025), (ii) shorten reviewer-matching time (e.g., one study reports a 73% reduction in time-to-identify reviewers (Farber, 2024), and (iii) provide rapid triage input for desk screening, for example, by predicting a very low likelihood of positive review (Thelwall & Yaghi 2025). These are task-specific speedups; whether they change time-to-first-decision or submission-to-publication remains untested. However, even moderate improvements in these areas could substantially improve several related issues with existing reviews mentioned above, including the increasing number of peer review invitations, the substantial economic burden of peer review, as well as, of course, the substantial delays in publishing. Moreover, LLM assistance is easy to scale, thus partially addressing review issues related to the growth of scientific publishing.

The use of AI systems also addresses the disproportionate burden of peer review on volunteers and junior scholars. They do this indirectly, by reducing the overall burden; and may also do so directly, for example by helping to identify a broader pool of potential reviewers (cf. Stelmakh et al. 2021), better matching of these reviewers to papers, and (more speculatively) by helping to keep track of reviewer workloads and scheduling of requests (Zhang et al. 2025).

What about consistency, that is, agreement between reviewers? So far, there appears to be only one study that directly addresses this question. It found that GPT-4's feedback overlaps with that of an individual human reviewer by approximately 30.9% for Nature-journal submissions (versus 28.6% human-to-human overlap), and by 39.2% for the ICLR dataset (versus 35.3% human-to-human overlap) (Liang et al. 2024). This suggests that AI–human agreement is comparable to human–human agreement.

Similarly, there appears to be only two preliminary studies looking at the capacity of LLM systems to detect manuscript errors. Liu and Shah (2023) inserted deliberate mathematical and conceptual mistakes into 13 short computer science papers and found that GPT-4 was able to identify errors in 7 of them, suggesting some promise but far from comprehensive reliability. Another recent study (Zhang, 2025) used retracted arXiv papers with known critical errors (factual / methodological / logical flaws) to evaluate how well LLM-based quality checkers detect such errors, providing further evidence that LLMs can help flag serious issues, though with limitations in scope and sensitivity.

Advanced systems

These studies, while encouraging, were conducted using relatively simple implementations of now-legacy models. Recent technical advances suggest substantially greater potential for AI-assisted peer review. Modern LLMs incorporate chain-of-thought reasoning capabilities that allow them to break down complex evaluation tasks into sequential analytical steps, leading to improved accuracy and reduced hallucination rates (Wei et al., 2022; Yao et al., 2023). For example, OpenAI's o1 achieves 74% accuracy on Olympiad-level problems compared to 12% for GPT-4o (OpenAI, 2024). This is a level of mathematical ability exceeding that of most reviewers in most fields, with the exception of disciplines such as mathematics, physics, and engineering. Additionally, tool-use capabilities enable LLMs to access external databases for citation verification, statistical validation, and cross-referencing claims against published literature (Schick et al., 2023). Rather than relying solely on parametric knowledge (that is, the internal representation of knowledge stored in model weights), these systems can access external knowledge, for example via web searches or API requests. This can greatly reduce hallucinations and increase performance (Gao et al. 2023).

The discussion so far has focused on off-the-shelf, generic, commercially available generative AI systems based on LLMs like OpenAI's GPT/o and Anthropic's Claude series. However, it is possible to technically adapt general models to a specific function or area. For example, through further training on a specialized dataset ('fine-tuning') or by giving models access to an external knowledge base ('retrieval-augmented generation' or RAG; Lewis et al. 2020); and some of the most promising

applications of LLMs in peer review might come from such specialized, application-specific language models (cf. Porsdam Mann et al 2025; Liddicoat et al. 2025). For instance, the OpenReviewer system is an LLM fine-tuned on 79,000 expert reviews from machine learning conferences. The system's outputs match human review recommendations at more than twice the rate of the best LLM used in the comparison (GPT-4o), while maintaining rating distributions that closely match human reviewers (that is, shows less of a positive bias than general LLMs) (Idahl & Ahmadi, 2024).

RAG systems complement fine-tuning by grounding responses in external knowledge bases. Such custom knowledge bases might contain journal guidelines, relevant publications and submissions, and review precedents. This helps address key limitations of generic models by ensuring reviews reflect current disciplinary standards and maintain consistency with journal scope. Lin (2025) proposes restructuring scientific papers themselves to create machine-readable "structured appendices" that would transform manuscripts from static documents into queryable, executable research environments. These appendices would provide computational reproducibility, standardized stimulus metadata, ontologically-mapped constructs, and granular data formats collectively creating "networked knowledge" where papers become nodes in continuously updating evidence networks. Such infrastructure would fundamentally enhance AI-assisted peer review by grounding evaluations in verifiable, executable claims rather than ambiguous prose, enabling AI systems to directly validate computational reproducibility, check statistical analyses, and cross-reference findings against structured knowledge bases.

In theory, these approaches enable customization at multiple levels: journals could develop models aligned with their editorial voice and publication standards; fields could create specialized LLM-reviewer tools for methodological approaches unique to their discipline; and individual reviewers could fine-tune personal models on their previous reviews (cf. Porsdam Mann et al 2023), creating AI assistants that maintain their human-user's analytical style while handling routine evaluation tasks.

More sophisticated architectures promise further enhancements. Agentic systems are LLMs trained to break down overarching goals into multiple component tasks, with the ability to plan and interact with external environments to achieve these tasks autonomously (Park et al. 2023). Multi-agent systems extend this concept by enabling communication between multiple agents, with each agent carrying out one or more specialized roles (Han et al. 2024). The MARG-S (D'Arcy et al. 2024) is a demonstration of the potential of this approach in the peer review context. In this setup, different LLM agents are assigned specific functions corresponding to specific sections or features of a paper – i.e., methodology, clarity, or impact – and then enabled to communicate with each other. D'Arcy et al. (2024) report that this set-up produced an average of 3.7 'good' comments per paper, as compared to 1.7 'good' comments per paper from a single-LLM set up (they also report a decrease in generic comments from 60% to 29%).

Similarly, the Generative Agent Reviewers framework (Bougie & Watanabe 2024) combined agents with distinct reviewer personas and a meta-reviewer to synthesize their output. The reviews generated by this system were preferred both to those of other single-LLM review systems (Bradley–Terry score

of 0.684 v 0.242 for closest LLM system), as well as over human reviews (0.684 vs. 0.523). In addition, the system also predicted acceptance decisions more reliably, achieving F1 scores of 0.66–0.69 (on a 0–1 scale where higher values indicate better balance of precision and recall), compared with 0.49 for human reviewers (Bougie & Watanabe 2024). Beyond performance improvements, these multi-agent systems can be used to experiment with factors influencing peer review. Jin et al (2024) tested various social science theories (i.e., groupthink, authority effects) in their AgentReview framework, concluding that such social factors explain 37.1% of the variance in peer review recommendations.

These architectures leverage a unique advantage of AI systems: their capacity to maintain broad understanding across disciplines while specializing as needed. A multi-agent system reviewing an interdisciplinary manuscript could simultaneously evaluate it from multiple perspectives, helping to address the growing challenge of reviewing increasingly interdisciplinary research that often exceeds the expertise of individual human reviewers. Multi-agent/-perspective systems could also be valuable in disciplines where there might be a lack of consensus on higher-order matters or fundamental theoretical commitments – for example, moral philosophy (e.g. consequentialism vs. deontology), statistics (e.g. Bayesian vs. non-Bayesian), and so on. Such systems might mitigate biases that stem from reviewers' underlying empirical or normative commitments; for example, they could lessen the influence of a libertarian reviewer's predispositions when evaluating a manuscript defending the permissibility of paternalism in public health. The core idea of combining multiple LLMs can be taken even further, through workflow orchestration and ensemble systems (Dietterich 2000) which coordinate specialized models or aggregate the output of multiple models, respectively. While the details are beyond the scope of this discussion, it is important to note that the technical design space for multi-LLM peer review is still largely unexplored.

To bridge capabilities and governance, Table 1 summarizes persistent peer-review problems alongside AI-assisted solutions, foreseeable obstacles, and recommended mitigations.

Table 1. Human vs. AI-Assisted Academic Peer Review

Persistent Problem with Human Review	**LLM-based AI systems**		
	Solution	**Obstacle**	**Mitigation**
Lengthy delays & inefficiency	LLMs can find suitable reviewers much faster and can generate review reports in minutes (scalable)	Relying on AI could erode essential critical evaluation skills and lead to uncritical acceptance of AI outputs	Position AI as a tool to augment, not replace, human judgment
High cost & uneven burden	AI assistance reduces the overall time burden on volunteer reviewers and can	Uploading unpublished manuscripts to	Use models not trained on user input (e.g., with data-protection

	help identify a broader pool of potential reviewers	commercial AI systems risks data leaks	guarantees or open-source models on local, secure servers)
Inconsistent quality & low agreement	Advanced multi-agent systems can evaluate papers from multiple perspectives simultaneously, and AI–human agreement is comparable to, or slightly better than, human–human agreement	LLMs can generate plausible but false statements ("hallucinations") and amplify biases from training data	Grounding AI with external tools (like web search); compare AI to the biased status quo; AI biases are more auditable and correctable than human ones
Poor error detection	LLMs show promise in spotting mathematical and conceptual error; advanced systems can use tools to verify citations and check statistical analyses	Authors could embed hidden instructions (prompt injection) to manipulate AI systems	Use technical defenses like adversarial training; employing multi-agent or ensemble systems; human oversight remains the final and most important check
Systematic biases	AI biases can be systematically corrected through technical debiasing; multilingual capabilities can help reduce bias against non-native English speakers	LLMs may struggle to recognize paradigm-shifting research and may promote linguistic homogenization	Use specialized, fine-tuned models trained on diverse scholarly traditions; employ multi-agent systems designed to value disagreement and other potential signal of innovation

AI: Artificial Intelligence; LLM: Large Language model

These mappings are provisional and should be treated as pilot hypotheses rather than settled policy. Taken together, the current body of evidence highlighted above is still small, but the results are striking. At least across the limited domains studied to date, LLMs can approach human baselines on some proxies of review quality under constrained setups (e.g., blinded usefulness ratings, overlap metrics). Given the nascency and heterogeneity of the evidence, blanket claims are unwarranted. Consistent with Table 1, a more defensible reading is that targeted, supervised LLM assistance is plausible and worth tightly scoped pilots.

Objections and responses

Even if preliminary findings suggest that LLMs are not inferior in quality to human reviewers, some concerns go beyond technical performance. In what follows, we set out these objections in their strongest form and consider possible responses.

1. Hallucinations and Biases

The Objection.

LLMs are known to generate "hallucinations": plausible sounding but false statements, fabricated citations, and spurious critiques disconnected from manuscript content (Huang et al., 2025). Such errors are not incidental but arise especially when models operate without access to external tools or knowledge sources. In these cases, they are effectively working 'from memory,' drawing only on patterns encoded in their training data rather than verifying claims against evidence. Without the ability to double-check outputs, for example through citation databases, statistical packages, or fact-retrieval systems, models may fill gaps with "example" plausible but inaccurate content. In the context of peer review, where legitimacy depends on accurate representation of empirical findings and logical argument, this raises the prospect of introducing fabricated flaws, untrue assessments, or false endorsements into the scholarly and scientific record.

Also concerning is that LLMs, trained on text corpora, encode and may even amplify structural biases from those sources. This includes tendencies to equate quality with institutional prestige, established methodological norms, authors from specific demographic groups, or already well-established authors, potentially disadvantaging novel approaches or individuals from marginalized groups. Such biases are not hypothetical: one large-scale experiment in economics, covering nearly 9,000 submission evaluations, found that LLMs tended to favor prominent institutions, male authors, and renowned economists, despite controlling for manuscript quality (Pataranutaporn et al. 2025)

The Response.

The problems of accuracy, bias, and hallucinations go to the heart of what peer review is supposed to accomplish, that is, safeguarding the integrity of the scholarly record. A quality control system that itself introduces errors and biases is clearly a system that fails at its fundamental purpose. Any use of LLMs in the peer review process must therefore carefully consider both the probability and magnitude of these risks, and the available mitigation strategies.

Yet it is crucial to recognize that many so-called hallucinations occur when LLMs are used in isolation, without access to the very tools that human reviewers themselves take for granted. Requiring an LLM to review a manuscript without access to article databases to verify references, a web search to confirm current information, or an interpreter to run code and check mathematics is equivalent to asking a human reviewer to work entirely "from memory"—without a calculator, a browser, or a reference library. At a minimum, LLMs should be given access to the citations and sources they are expected to evaluate. When generic AI used without access to the necessary tools and knowledge corpus to

generate accurate answers, hallucinations are better understood not as intrinsic flaws but as a kind of *user error*: the model generating a plausible example or template when it cannot access the data needed to provide a definitive answer.

With appropriate tool integration and clear user practices, the evidence considered in the preceding section suggests these issues can be significantly mitigated, often more effectively than with human reviewers. Moreover, we must remember that the current human system already falls far short of its ideals. If technical and procedural safeguards reduce AI-generated errors and biases to levels comparable to or below those of human reviewers, then this objection loses much of its force.

Studies surveyed earlier demonstrate that even legacy models like GPT-4 achieve quality ratings comparable to human reviewers, with the majority of researchers finding AI feedback helpful and most rating it superior to at least some human reviewers (Liang et al., 2024). Current state-of-the-art models have advanced significantly beyond these systems. Agentic models equipped with web search capabilities, dedicated deep research features, and tool integrations that can digitally 'interface' with programs and websites can now: verify citations against scholarly databases; check calculations through computational engines like WolframAlpha; and access discipline-specific knowledge bases in real time through APIs and Model/Agent Context Protocol (MCP/ACP) connections. As already noted, more advanced technical solutions like fine-tuning, RAG, ensembles, and multi-agent systems can dramatically increase the quality and accuracy of LLM outputs (Idahl & Ahmadi, 2024; D'Arcy et al. 2024). While building and using such systems requires some degree of AI literacy, there are now no- and low-code solutions (such as n8n or Claude Code) that render them accessible to a much broader user base than was previously the case.

A fair evaluation of the ethics of LLM use in peer review should compare the current state of the art to the status quo, not yesterday's models to an idealized human reviewer. That status quo, as documented above, is deeply problematic: reviewers detect only a quarter of major errors (Godlee et al., 1998), show inter-reviewer agreement barely above chance (Bornmann et al., 2010), demonstrate systematic biases favoring prestigious institutions and male authors (Murray et al., 2019; Tomkins et al., 2017), and impose delays of months to years (Andersen et al., 2021).

The bias concern also admits of nuance Yes, AI systems trained on existing literature will reflect existing biases. But human reviewers also demonstrate strong biases (Tomkins et al. 2017). Critically, AI biases may prove more tractable than human ones. Once identified, they can be addressed systematically through retraining, debiasing techniques, or ensemble methods incorporating diverse perspectives. Human biases, rooted in decades of socialization, resist such direct intervention. Moreover, we can audit AI systems for bias in ways impossible with humans, running thousands of controlled experiments varying author characteristics while holding manuscript content constant (cf. the AgentReview study mentioned above, which found that social factors explained roughly a third of variance in peer review; Jin et al. 2024). Lastly, blinding reviewers, whether human or AI, to the identity of the submitting authors can at least partially address this issue for both types of reviewers.

Finally, and perhaps most importantly, we are not proposing that AI systems conduct peer review independently. Existing standards for LLM use in scholarship clearly mandate that humans using these systems must take full responsibility for the accuracy of any LLM-produced text (cf. Ganjavi et al 2024, Porsdam Mann et al. 2024, Luo et al 2025.). Similar standards can, and should, be insisted on for LLM use in peer review, a point we return to below (see Objection 3 and associated response).

2. Novelty Recognition and Homogenization

The Objection

The fundamental operation of LLMs is to predict the next most likely token (a fragment of a word or other symbols or characters) in a series of tokens, given patterns in their training data (Bender et al. 2021). Thus, even in an idealized case of "perfect" prediction, this mechanism constrains models to reproduce and recombine what is already known. The implication is clear: systems built to extend past distributions of language may struggle, in principle, to evaluate work that departs radically from established paradigms.

Empirical evidence reinforces this concern. Juzek and Ward (2025) found that 21 specific words, including "delve," "intricate," and "underscore", have increased dramatically in scientific abstracts since the widespread adoption of LLMs. Intriguingly, their investigations point towards reinforcement learning with human feedback (RLHF) as the causal mechanism. RLHF is a technique used to align LLMs with human preferences. When models are fine-tuned to maximize human approval, they apparently learn to favor words and phrases that evaluators find satisfying – typically those that sound authoritative, comprehensive, and familiar. It may be that this preference for the conventional extends beyond vocabulary; if RLHF teaches models to favor linguistic patterns that humans find immediately appealing, it may similarly train them to prefer conventional scientific ideas over radical departures from established thinking. In literature discovery, AI systems can act as powerful gatekeepers that amplify existing biases like the Matthew effect, in which those who already have advantages accumulate further advantages quicker than others (Lin, 2025a); indeed, one study found that over 60% of AI-generated paper recommendations are for articles in the top 1% of citation distributions (Algaba et al., 2025).

A further dimension to this issue relates to cultural and linguistic patterns. LLMs perform best with intellectual traditions that are well-represented in their training data (predominantly English-language and Chinese-language sources), reflecting where the most digital text is available and where most AI development occurs. Research grounded in less-represented philosophical frameworks, regional scholarly traditions, or languages with limited digital corpora may not be adequately evaluated. For instance, an AI system trained primarily on Western academic literature might lack sufficient exposure to evaluate research employing Ubuntu philosophy, Buddhist logic, or other non-Western analytical frameworks. This disparity in representation stems from practical constraints but nonetheless creates systematic disadvantages for scholarship from underrepresented traditions. Such technical limitations

risk amplifying existing disparities in whose knowledge gets recognized and validated in the global scientific community.

The Response

These concerns about innovation and diversity strike at the heart of scientific progress: Science advances in part through paradigm shifts (Kuhn 1962/1970) that violate existing norms, and underrepresented traditions are often a key source of alternative ways of engaging with a problem that might lead to such breakthroughs (Hofstra et al. 2020). Yet, as is the case with hallucination and biases, there are both technical and procedural methods that can at least mitigate these issues, if not solve them entirely.

The theoretical challenge appears fundamental: LLMs are trained to predict the next token based on patterns in existing data. How can systems built to reproduce past distributions recognize work that, by definition, breaks those patterns? The LLM predictive mechanism has often been taken to imply that such systems can only reproduce or recombine what is already known. In practice, however, contemporary AI systems have demonstrated the capacity to generate novel content, from new mathematical proofs and problem-solving strategies (Romera-Paredes et al. 2024) to original protein structures (Jumper et al. 2021) and chemical compounds (Bran et al. 2024). These outputs suggest that statistical prediction does not preclude innovation, particularly when models are scaled and equipped with reasoning strategies such as chain-of-thought prompting or tool integration. But here too, it must be borne in mind that human peer review also demonstrates conservative bias against innovative work (Guthrie et al 2019). The difference lies in how these limitations manifest and might be addressed.

Importantly, LLMs' broad training might sometimes enable them to recognize novelty that specialist reviewers miss. While a reviewer deeply embedded in one paradigm might reject work that violates their field's assumptions, an LLM trained across multiple disciplines (or a multi-agent setup with individual LLMs each singularly trained on a specific discipline) might recognize when concepts from one domain offer revolutionary potential in another. The system might identify that what appears as "error" from one paradigm's perspective represents valid methodology from another. This cross-pollination capacity could help identify certain forms of innovation, even if it turns out that the system cannot generate truly novel paradigms itself.

The homogenization concern, while related to this fundamental limitation, may be simpler to address because it reflects additional layers of design choice. The RLHF processes used in general-purpose LLMs optimize for broad human preferences, such as authority and clarity. While such optimization may serve commercial chat applications, it may also prove problematic for scientific peer review where breakthrough thinking often initially appears incomplete, unclear, or challengingly unconventional. When models are trained to maximize immediate human approval (i.e. their outputs are non-adversarial, and in some sense "sycophantic"), they learn to reproduce what feels satisfying and familiar rather than what might prove revolutionary. This suggests that for peer review applications, technical interventions are perhaps even more crucial than for other concerns.

Application-specific fine-tuning offers a direct counter to these homogenizing pressures. The OpenReviewer system, when trained specifically on peer review data rather than general human preferences, not only achieved superior accuracy but maintained rating distributions matching human reviewers' full range, avoiding the excessive positivity of general LLMs (Idahl & Ahmadi, 2024). More radically, multi-agent systems or ensembles involving separate models trained on different intellectual traditions and explicitly valuing disagreement between them as potentially signaling paradigm-challenging work. A manuscript that appears incoherent from one perspective might reveal methodological innovation from another.

Procedurally, editorial boards and funding agencies must make deliberate choices about what kind of diversity they wish to preserve. This means recognizing that the apparent trade-off between quality control and innovation may itself be an artifact of how we currently configure these systems. With appropriate design, AI assistance could potentially identify and champion the very innovations that human reviewers, constrained by their training within existing paradigms, might miss.

3. Human Oversight, Excessive Deference, and De-skilling

The Objection

The first two objections concerned the technical performance of LLM systems. However, there are other fundamental concerns which relate not to the models themselves, but rather to their effects on human agency and expertise. It is often said that AI systems should retain a human in the loop – that is, that human oversight is crucial to ethical and effective AI use. Yet, if AI systems require human validation, then the validators must possess sufficient expertise to critically evaluate AI outputs—and if they possess such expertise, why do they need AI assistance?

A related concern is the deference problem, otherwise known as automation bias. Multiple studies document that those charged with overseeing AI systems, for example in healthcare applications, gradually surrender their critical judgment, accepting AI recommendations even when their own expertise suggests otherwise (Abdelwanis et al. 2024; Nguyen 2024). In peer review, such deference could mean that subtle errors or biases in AI systems get amplified rather than corrected. Reviewers might accept AI-generated critiques without the careful scrutiny they would apply to their own thoughts, particularly when under time pressure or reviewing outside their core expertise.

Moreover, peer review serves a crucial pedagogical function that too much AI assistance threatens to erode. Reviewing others' work teaches essential skills: identifying methodological flaws, evaluating evidence quality, constructing constructive criticism, and appreciating diverse research approaches. These skills may then transfer to a researcher's own work, improving study design and presentation. The concern is not only about individual skill development but extends to the collective capacity of the scientific community to maintain critical standards. Uncritical delegation can foster AI complacency, a progressive disengagement from the evaluative struggle through which expert judgment is forged (Lin and Sohai, 2025).

The rise of AI in peer review also introduces new complexities for academic relationships, especially for those at the start of their careers. Traditionally, peer review has not only ensured the integrity of scholarly work but has also created moments of recognition, dialogue, and informal mentorship between early-career researchers and more established academics. If AI takes on a larger role in shaping evaluations, this relational dimension could be reduced. Automated systems are well-suited to checking format, consistency, or methodological precision, but they may struggle to fully capture the creativity, interdisciplinarity, or exploratory qualities often found in emerging scholarship. As a result, junior academics risk being assessed primarily through conformity to established norms rather than through the originality of their ideas. This shift could limit the visibility and recognition that early-career researchers need to build reputations and develop professional networks, reinforcing existing hierarchies in academic life.

The Response

Evidence from other fields such as radiology and aviation demonstrates that these concerns about oversight, deference, and de-skilling are indeed serious. In radiology, Dratsch et al. (2023) found that even highly experienced radiologists exhibited automation bias when using AI assistance, changing correct diagnoses to incorrect ones based on AI suggestions. Research in aviation has likewise shown that skilled pilots can lose critical competencies when over-relying on autopilot systems, leading to catastrophic failures when manual intervention becomes necessary (Mosier 2017). Yet this same body of evidence also helps us understand both the extent of these problems and which mitigation strategies prove effective.

What is framed as deskilling should instead be understood as a necessary recalibration of expertise. Effective AI integration demands the cultivation of essential AI meta-skills: the strategic direction of generative systems toward rigorous analysis, the critical discernment of their outputs against disciplinary standards, and the systematic calibration of human-AI collaboration through iterative refinement (Lin and Sohail, 2025). This follows established patterns of technological augmentation. When calculators became ubiquitous, the frequency of by-hand calculation diminished, but this freed up cognitive capacity for tackling more advanced mathematical and scientific problems. Similarly, fifth-generation fighter jets automate many lower-level piloting tasks, enabling pilots to focus on higher-level situational awareness, coordination, and tactical decision-making. The same logic applies to LLMs in peer review: if routine labor such as formatting checks, literature lookups, or drafting generic critiques is offloaded to AI, human reviewers can devote more attention to conceptual clarity, methodological innovation, and the broader significance of research findings.

First, however, let us consider the role and nature of oversight. It is true that the need for human oversight reduces the efficiency gains one might expect from a fully automated system, a feature of LLM use that Ohde et al. (2025) label a "new and tedious burden". Yet this shift from generation to oversight is neither new nor unique to AI. Academic careers already follow this trajectory: junior researchers write papers, senior researchers review them; PhD students conduct experiments, supervisors evaluate results, and so on (Hurshman et al., 2025). The progression from assistant to full

professor involves steadily less content creation and more quality control, a transition we accept as natural professional development rather than problematic burden. Indeed, the peer review system itself rests on the recognition that expert evaluation of others' work provides value even when the evaluator could theoretically have produced similar work themselves. The efficiency gain comes not from eliminating oversight but from the differential effort required: reviewing a manuscript takes hours, writing one takes months; similarly, fact-checking, editing, and generally improving an LLM-generated review draft might take a couple of hours, but this is likely less than the handful to dozens of hours needed to write one from scratch.

The deference problem (automation bias) represents perhaps the most serious of these concerns. However, precisely because it is a widespread issue across several important domains, it is also one that has been studied in detail. In general, three types of solutions have been studied, each of which helps to mitigate the problem: (1) training programs that emphasize user accountability for decisions, (2) design features that present confidence levels alongside automated recommendations, and (3) providing information rather than direct recommendations, requiring users to integrate and interpret data themselves (Goddard et al. 2012). Research has also shown that automation bias is strongly correlated with cognitive load (Lyell and Coiera 2017). These findings could help inform mitigation strategies in the peer review context. Journals, publishers, or universities could provide training courses in the effective use of LLMs in peer review. They could provide prompt libraries or even adapt models (via fine-tuning or custom instructions) to provide relevant information as opposed to direct accept/revise/reject decisions. They could instruct, advise, or require reviewers to use prompts or models designed to elicit confidence scores rather than direct requests. When an AI system flags that it has low confidence in a particular critique, reviewers know to examine that aspect more carefully.

Importantly, these same mitigation strategies also address the de-skilling concern. Training programs that reduce automation bias by teaching critical evaluation of AI outputs simultaneously preserve and develop the analytical skills that peer review is meant to cultivate. When reviewers must actively integrate information rather than passively accept recommendations, they maintain the cognitive engagement necessary for skill development. Requiring confidence scores forces reviewers to make independent judgments about areas of uncertainty, which requires precisely the kind of critical thinking that prevents atrophy of expertise. Furthermore, the de-skilling concern must be weighed against current realities. Peer review burden already falls disproportionately on junior researchers who accept invitations at higher rates than senior colleagues despite having less experience and facing greater career pressure. A harried postdoc reviewing their tenth manuscript while preparing job applications may gain little pedagogical benefit from rushed, mechanical review. Offloading part of this workload to AI would also help alleviate the significant opportunity costs faced by early career researchers.

Moreover, researchers develop critical evaluation skills through multiple activities beyond peer review: journal clubs, conference presentations, grant panels, and collaborative exchanges. If AI reduces

review burden, saved time could be invested in other activities that develop evaluation skills more effectively than the current review system.

Crucially, the appropriate response depends on the level of AI involvement, which spans a broad spectrum. For minimal involvement (formatting checks, citation verification, identification of reviewers), oversight requirements are modest and de-skilling risks negligible. For substantive involvement (AI-generated drafts that humans edit), reviewers maintain clear decisional authority while using AI as they might use reference software—as tools extending rather than replacing capability. The concern is strongest for areas in which LLMs replace, rather than supplement, human review. However, many of the benefits of LLM integration into peer review processes are likely to stem not from replacing, but from augmenting human efforts.

Similarly, to address relational concerns, risks can be mitigated if AI is positioned as an aid rather than a replacement for human judgment. AI can efficiently handle mechanical aspects of review—such as verifying references, checking data consistency, or identifying possible errors—while leaving space for reviewers to focus on interpretation, critique, and constructive advice. Journals and editorial boards can support this balance by encouraging reviewers to provide developmental feedback that actively helps junior scholars strengthen their work. Transparency about when and how AI tools are used is also essential to preserve trust and accountability, as discussed in more detail below. Beyond the review itself, additional initiatives such as editorial mentoring programs, feedback exchanges, or networking opportunities can help ensure that junior academics continue to benefit from the relational and developmental functions of peer review. In this way, AI can reduce administrative burdens while preserving the human connections that are essential for sustaining an inclusive and supportive academic community.

4. Confidentiality

The Objection

The use of AI systems for peer review raises fundamental concerns about the confidentiality of unpublished research. Many journals and funding agencies currently prohibit AI assistance on the grounds that uploading manuscripts to commercial AI providers could expose novel findings that represent years of intellectual investment (e.g., NIH 2023). Since priority determines credit in many fields, even minor leaks could prove professionally devastating. A researcher's career advancement, grant funding, and scientific legacy often depend on being first to publish key discoveries.

The concern applies at multiple levels. At the technical level, questions persist about data persistence: once a manuscript enters an AI system, does its content remain accessible in some form (Carlini 2021)? At the legal level, uncertainty exists about data ownership and usage rights when academic content is processed by commercial entities whose terms of service may claim broad licenses to user inputs (Chesterman 2025). At the ethical level, submitting authors' unpublished work to AI systems without explicit consent arguably violates the trust relationship fundamental to peer review (NIH 2023). At

the competitive level, if manuscripts enter training datasets, future users working on similar problems might gain unfair advantages through model behaviors influenced by that training, even without direct data extraction (Earp et al. 2025).

Unlike published papers, manuscripts under review contain ideas at their most vulnerable; polished enough to reveal their value but not yet protected by publication priority. The peer review process requires authors to share these ideas with journals and reviewers based on strict confidentiality assurances. Introducing LLM systems into this process without absolute guarantees of data security could be seen as betraying this foundational trust.

The Response

While confidentiality concerns deserve serious attention, these alone are not sufficient to justify a prohibition on LLM use in peer review. This is because the concerns only apply to a subset of LLM use, whereas current prohibitions citing these concerns tend to apply to all LLM use.

It is also worth recognizing that journals already rely heavily on cloud-based infrastructure. Manuscripts are routinely stored, transmitted, and processed using web-based submission and review platforms hosted on third-party servers. Editors and reviewers regularly access confidential content via Gmail, Outlook, Microsoft Word Online, Google Docs, and other cloud services. If it is considered acceptable to entrust manuscripts to Google Cloud or Microsoft Azure for email, document preparation, and journal management systems, then it is not clear why equivalent assurances of confidentiality from Google Gemini, Microsoft Copilot, or other AI services should be judged categorically differently. Provided that the same enterprise-level contractual protections and compliance standards apply, the risk profile is not fundamentally new.

Commercial AI providers have further responded to confidentiality concerns by offering dedicated APIs with explicit data governance provisions. OpenAI's enterprise API, Anthropic's Claude for Business, and Google's Vertex AI all contractually guarantee that content processed through these interfaces will not be retained for training or made accessible to other users (see e.g. OpenAI 2025). These are not merely terms of service but legally binding agreements with substantial financial penalties for breach. The major providers have compelling commercial incentives to honor these provisions, because their entire enterprise business, worth billions in revenue, depends on maintaining client trust. The suggestion that these companies would systematically violate contractual obligations and risk their core business requires evidence beyond theoretical possibility.

For institutions requiring absolute data control, technical alternatives eliminate external dependencies entirely. Open-source language models can be deployed on local infrastructure, ensuring complete data isolation. While early open-source models lagged commercial offerings significantly, this gap has narrowed dramatically. Models like DeepSeek-3,1, Kimi K2, Mistral, and Llama 4 rank highly on various metrics and benchmarks that are used to evaluate model performance (see, e.g. Guo et al. 2025; Moonshot AI 2025).

One must also keep in mind the comparison to human reviewers, who are known to have delayed manuscripts to publish competing work, appropriated ideas for grant proposals, or shared content with colleagues and students (e.g. Jackson et al., 2025). Human reviewers have career incentives, competitive pressures, and personal biases that can motivate confidentiality breaches. They might unconsciously incorporate ideas into their own thinking (Earp et al. 2025) or deliberately exploit their privileged access. LLM systems, whatever their limitations, have no papers to publish, no grants to win, and no careers to advance. They cannot engage in the deliberate copying or misuse that represents the most serious confidentiality threat in peer review.

This is not to dismiss confidentiality concerns but to evaluate them proportionally. The question is not whether AI systems present zero risk but whether available mitigations reduce risk to acceptable levels and whether the benefits justify residual risks. Properly configured AI systems may present no greater confidentiality risk, and potentially lesser, than existing practice. The key is recognizing that confidentiality protection is not binary but involves choosing appropriate tools for specific contexts, just as we currently do elsewhere throughout the research process.

5. The Efficiency Objection, or: Opening the Floodgates

The Objection

Using AI peer review risks exacerbating a longstanding problem in academic publishing: the sheer volume of research getting published. Between 2016 and 2022 the number of indexed articles increased from ~1.92 million to ~2.82 million (Hanson et al. 2024). Many scholars have warned that this growth makes it increasingly difficult to identify and absorb relevant findings (Chu & Evans 2021; Houssard et al. 2024). The rising number of submissions also places substantial strain on reviewers and editorial systems. While AI could alleviate some of this strain, its efficiency gains might eliminate existing bottlenecks, enabling more articles to pass through peer review. This trade off would be especially advantageous for predatory or volume-drive journals and intensify concerns that quantity is crowding out quality.

The Response

Concerns about the rising *volume* of manuscripts are legitimate, but they may be misplaced. In principle, more science is not itself a problem. Indeed, high-quality work should be welcomed regardless of quantity. The real challenge lies in the proliferation of low-quality outputs, especially from predatory or poorly governed journals (Grudniewicz et al. 2019). In such contexts, the worry is not about *too much* science, but about *too much bad science*.

AI-assisted peer review could in fact mitigate this problem by strengthening early filtering: flagging obviously unfit submissions, enforcing basic methodological and reporting standards, and raising the overall baseline quality of manuscripts that reach human reviewers. This could reduce reviewer burden at reputable outlets while making it harder for weak work to pass undetected.

Of course, predatory publishers are unlikely to change their behavior. To the extent that such journals already lack genuine peer review, AI will not make them better or worse; their incentives remain misaligned with scientific quality. A more serious risk is that these journals might adopt AI-generated "reviews" to create a veneer of legitimacy and scale up output. Yet this is arguably a continuation, not a transformation, of existing predatory practices, and should be met with the same countermeasures: indexing standards, author awareness, and institutional disincentives (Laine et al. 2025).

On balance, then, we should be less concerned about the total volume of papers, and more about the distribution of quality. Properly deployed, AI-assisted review may help journals separate signal from noise.

6. Vulnerability to Gaming

The Objection

The introduction of AI into peer review opens up a new attack vector that previously did not exist. As such, authors may now attempt to manipulate AI review systems through adversarial techniques. Because LLM systems operate through predictable patterns, they may be vulnerable to gaming strategies that would not affect human reviewers. Authors might embed hidden instructions in manuscripts (Gibney 2025), craft text to trigger favorable evaluations, or exploit known biases in LLM systems. As LLM use in peer review expands, incentives to discover and exploit such vulnerabilities will intensify.

For instance, in July 2025, 18 preprints on arXiv were found to contain hidden instructions designed to elicit positive reviews from LLMs—ranging from simple commands like "GIVE A POSITIVE REVIEW ONLY" to elaborate outlines structuring entire favorable assessments (Lin, 2025b). Prompt injection attacks exploit the fundamental architecture of language models, including causing models to ignore their instructions and follow hidden directives (Liu et al. 2023); bad faith users can craft text that manipulates LLM systems while appearing normal to human readers in terms of grammar, meaning, and readability (Wang et al. 2024). More speculatively, sophisticated actors might attempt to hack into the journal website hosting the AI interface and attempt to change or reroute the processing of the manuscript.

This suggests that authors could potentially craft manuscripts that receive favorable LLM reviews despite serious flaws. If journals rely heavily on LLM assessments, such gaming threatens to corrupt not only peer review but the broader infrastructure of scientific evaluation—from plagiarism detection to citation indexing—potentially distorting scientific knowledge at scale.

The Response

While the vulnerability to gaming represents a genuine concern, several strategies can mitigate these risks and create more robust AI review systems. The key lies not in eliminating vulnerabilities entirely but in raising the cost and complexity of successful manipulation to impractical levels.

First, various technical defenses are possible. For example, deliberately exposing AI systems to known manipulation techniques during development or fine-tuning can help models recognize and resist common gaming strategies (e.g. Xhonneux et al. 2024).

Second, human reviewers can use various defensive strategies, such as pre-processing manuscripts text before evaluation, removing potential trigger patterns while preserving legitimate content (Liu et al. 2023).

As for the more speculative scenario involving hacking of communication between journal websites and AI systems, it is also important to note that launching a successful adversarial attack would require a high level of technical sophistication, roughly equivalent to what would be required to hack the manuscript submission system itself. Except for a small minority of researchers in highly technical fields such as computer science and cybersecurity, such capabilities are likely beyond the reach of most authors.

This point also generalizes to sophisticated adversarial attempts involving the manuscript itself. Indeed, the discovery of a novel injection attack would itself constitute a significant technical contribution. Such contribution may even be comparable in scope to the main results of many scholarly articles, underscoring how implausible it is that routine authors would attempt, let alone succeed, in such exploits.

Nevertheless, this is a serious objection. As both adversarial and defensive strategies develop, we are likely to see a cat-and-mouse dynamic in which techniques are developed and mitigated, requiring some degree of technical vigilance and understanding on the part of journals and perhaps even at the level of individual reviewers.

The use of ensemble or multi-agent systems may be particularly effective in this context: An author attempting to game such a system would need to craft text that simultaneously exploits vulnerabilities across all models—a significantly more challenging task. Moreover, disagreement between models could trigger additional scrutiny, flagging potentially manipulated submissions for human review.

Importantly, it is worth noting that complete automation of peer review was never the goal. LLM systems are tools to augment human judgment, not replace it entirely. The combination of human and LLM review may be particularly effective at catching adversarial attempts, since few currently known strategies are effective against both humans and LLM systems.

7. Loss of Trust

The Objection

Finally, as with concerns raised in other domains where LLMs can be deployed to support or augment traditionally human-centric processes (e.g. Porsdam Mann et al 2025), the prospect of AI adoption in academic peer review raises central questions about the erosion of *trust* and *credibility* in science, both from within and among the general public.

Trust underpins the 'social contract' of human peer review. In the context of AI, it can be understood as operating on two levels: (i) trust from academic stakeholders (i.e. scholars and researchers) in the credibility of manuscripts peer reviewed with AI assistance or by an AI system, and (ii) broader societal trust in the legitimacy of science peer reviewed, at least in part, by AI agents or systems.

Empirical data suggest these concerns are salient. A 2025 *Nature* survey of 5,229 academics illustrates such tensions. Only 5% of respondents deemed it "appropriate" for a reviewer to use an AI's output (without 'AI usage' disclosure) as the basis of their peer-review report, while 52% judged such action "not appropriate under any circumstances". These numbers suggest that a significant portion of the academic community views AI involvement in peer review as unacceptable, at least without disclosure.

Although no studies (to our knowledge) have yet examined societal attitudes towards AI in peer review, it is plausible that public perspectives could be as, if not more, critical as those of researchers.

In sum, AI in peer review risks undermining the trust that is foundational to the scientific enterprise and which peer-reviewed is meant to maintain.

The Response

Concerns about AI's potential to undermine the trust and credibility of peer review are significant. However, a prudent response is not to categorically reject its use; instead, these legitimate concerns highlight the importance of safeguards and governance to ensure responsible adoption and trustworthy use. If journals choose to employ AI systems to assist or augment their peer review processes, they must take proactive and substantive actions. Beyond engaging relevant stakeholders to identify and address the reasons why integrating AI with peer review may be perceived or considered as inappropriate, there are a number of concrete methods of addressing concerns which journals should consider.

Transparency is perhaps the most fundamental requirement for maintaining trust in AI-augmented peer review processes. This necessitates the adoption of clear, publicly available policies and procedures that govern the use of AI in peer review, whether in initial desk screenings or as support for reviewers. Journals should provide technical details of any AI systems they propose to, or actually do, employ. Furthermore, journals must provide explicit disclosure regarding when and how AI was employed in the review process. This approach mirrors existing guidelines that require authors to disclose and acknowledge their use of AI in research and manuscript preparation, as similar principles

of transparency apply across all stages of the scientific publication process (cf. Porsdam Mann et al. 2024).

Accountability mechanisms must also be clearly established within journal policies to specify responsibility for AI use and potential misuse. When peer reviewers are granted autonomy to use journal-designated AI systems, they must assume full responsibility for the accuracy, integrity, and fairness of their reports. This means that regardless of AI assistance or contribution, the human reviewer remains ultimately accountable for the content and quality of their peer-review report. Such accountability structures ensure that AI remains a tool rather than a replacement for human judgment and expertise. Similarly, journals and publishers should explicitly assume responsibility for the accuracy and reliability of any in-house or licensed systems used for this purpose.

Journals should also promote literacy about AI capabilities and limitations among all stakeholders in the review process. Authors and reviewers need to be informed through submission guidelines and peer review policies about the journal's designated scope for AI use, including its specific roles and any restrictions on its application in the peer review process, as well as the rationale for the introduction of AI systems. This educational approach helps ensure that all participants understand both the potential benefits and inherent limitations of AI assistance.

Taken together, these measures can help mitigate concerns about trust and credibility, thereby enabling journals to adopt AI for peer review responsibly while maintaining confidence among both scholars and the broader public.

Conclusion

The current evidence is preliminary and heterogeneous; yet in several small-scale studies, even legacy models like GPT-4 have approached human baselines on specific review tasks. More recent technical advances, including chain-of-thought reasoning, tool use, fine-tuning, RAG systems, and multi-agent architectures, indicate that performance will likely improve dramatically as these methods diffuse into general use. While reasonable people may disagree, we find none of the stated objections fatal to the case for LLM-assisted peer review. The concerns reviewed above are genuine but admit of technical and procedural mitigations that can reduce risks to acceptable levels. Indeed, peer review may represent a particularly promising application for LLMs: the task is well-defined, quality metrics are established, and the potential benefits are substantial and might finally address documented failures of the current system that have resisted decades of reform attempts.

Nevertheless, realizing these benefits depends critically on implementation details. The difference between responsible augmentation and reckless automation lies in careful attention to technical design, governance structures, and user training. This means selecting appropriate tools (enterprise APIs over consumer chatbots, targeted use of local LLMs, specialized models over general-purpose systems), establishing clear oversight protocols (human verification requirements, confidence thresholds, audit

procedures), and developing new competencies within the research community. Most importantly, it requires recognizing that LLMs are tools for augmenting human judgment, not replacing it.

The path forward requires neither uncritical enthusiasm nor reflexive rejection, but careful empirical investigation of what works, for whom, and under what conditions. As LLM and general AI capabilities continue to evolve rapidly, so too must our frameworks for evaluating their appropriate use in scientific knowledge production. We invite the research community to engage with these questions through both theoretical analysis and practical experimentation. The future of peer review, and by extension, the validation of scientific knowledge, may depend on getting this balance right.

AI use acknowledgement

In the preparation of this manuscript, GPT-5 and Claude 4.1 Opus were used to edit and rephrase text supplied by the authors. Use of generative AI in this manuscript adheres to ethical guidelines for use and acknowledgment of generative AI in academic research (Hosseini et al. 2025; Porsdam Mann et al. 2024). Each author has made a substantial contribution to the work, which has been thoroughly vetted for accuracy, and assumes responsibility for the integrity of their contributions.

References

Abdelwanis, M., Alarafati, H.K., Tammam, M.M.S. and Simsekler, M.C.E.(2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. Journal of Safety Science and Resilience, 5(4), pp.460-469.

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: estimating the cost of researchers' time spent on peer review. Research Integrity and Peer Review, 6, 14.

Algaba, A., Holst, V., Tori, F., Mobini, M., Verbeken, B., Wenmackers, S., & Ginis, V. (2025). How deep do large language models internalize scientific literature and citation practices?. arXiv preprint arXiv:2504.02767.

Andersen, M. Z., Fonnes, S., & Rosenberg, J. (2021). Time from submission to publication varied widely for biomedical journals: a systematic review. Current medical research and opinion, 37(6), 985–993. https://doi.org/10.1080/03007995.2021.1905622.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

Björk, B.-C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. Journal of Informetrics, 7(4), 914–923.

Bornmann L, Mutz R, Daniel H-D (2010) A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. PLoS ONE 5(12): e14331. https://doi.org/10.1371/journal.pone.0014331

Bougie, N., & Watanabe, N. (2024). Generative adversarial reviews: When llms become the critic. arXiv preprint arXiv:2412.10415.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A.D. and Schwaller, P., 2024. Augmenting large language models with chemistry tools. Nature Machine Intelligence, 6(5), pp.525-535.

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., & Squazzoni, F. (2019). The effect of publishing peer review reports on referee behavior in five scholarly journals. Nature communications, 10(1), 322. https://doi.org/10.1038/s41467-018-08250-2.

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. Social Psychological and Personality Science, 8(4), 363–369. https://doi.org/10.1177/1948550616673876

Bruce, R., Chauvin, A., Trinquart, L., Ravaud, P. and Boutron, I., 2016. Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. BMC medicine, 14(1), p.85.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., et al.(2021). Extracting training data from large language models. In 30th USENIX security symposium (USENIX Security 21) (pp. 2633-2650).

Charlin, L. and Zemel, R.(2013). The Toronto paper matching system: an automated paper-reviewer assignment system. https://openreview.net/forum?id=caynafZAnBafx

Chesterman, S., (2025). Good models borrow, great models steal: intellectual property rights and generative AI. Policy and Society, 44(1), pp.23-37

Chu, J.S.G. and Evans, J.A. (2021). Slowed canonical progress in large fields of science. Proceedings of the National Academy of Sciences of the United States of America, 118(41), e2021636118

Collu, M. G., Salviati, U., Confalonieri, R., Conti, M., & Apruzzese, G. (2025). Publish to Perish: Prompt Injection Attacks on LLM-Assisted Peer Review. arXiv:2508.20863v1

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

Dietterich, T.G., (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., et al Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. Radiology, 307(4), e222176. https://doi.org/10.1148/radiol.222176.

Earp, B.D., Yuan, H., Koplin, J. , Porsdam Mann S. LLM use in scholarly writing poses a provenance problem. Nat Mach Intell 7, 1889–1890 (2025). https://doi.org/10.1038/s42256-025-01159-8

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

Fox C. W. (2021). Which peer reviewers voluntarily reveal their identity to authors? Insights into the consequences of open-identities peer review. Proceedings. Biological sciences, 288(1961), 20211399. https://doi.org/10.1098/rspb.2021.1399

Ganjavi, C., Eppler, M.B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G.S., Gill, I.S. and Cacciamani, G.E. (2024). Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. bmj, 384.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. and Wang, H., 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2(1).

Gaudino, M., Robinson, N.B., Di Franco, A., Hameed, I., Naik, A., et al., 2021. Effects of experimental interventions to improve the biomedical peer-review process: a systematic review and meta-analysis. Journal of the American Heart Association, 10(15), p.e019903.

Gibney E. (2025). Scientists hide messages in papers to game AI peer review. Nature, 643(8073), 887–888. https://doi.org/10.1038/d41586-025-02172-y

Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012 Jan-Feb;19(1):121-7. doi: 10.1136/amiajnl-2011-000089. Epub 2011 Jun 16. PMID: 21685142; PMCID: PMC3240751.

Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: A randomized controlled trial. JAMA, 280(3), 237–240.

Grudniewicz, A., Moher, D., Cobey, K.D., Bryson, G.L., Cukier, S., Allen, K., Ardern, C., Balcom, L., Barros, T., Berger, M. and Ciro, J.B., 2019. Predatory journals: no definition, no defence. Nature, 576(7786), pp.210-212.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Guthrie S, Rodriguez Rincon D, McInroy G et al (2019). Measuring bias, burden and conservatism in research funding processes [version 1; peer review: 1 approved, 1 approved with reservations]. F1000Research, 8:851 (https://doi.org/10.12688/f1000research.19156.1)

Han, S., Zhang, Q., Yao, Y., Jin, W. and Xu, Z.(2024). LLM multi-agent systems: Challenges and open problems. arXiv preprint arXiv:2402.03578.

Hanson, M.A., Barreiro, P.G., Crosetto, P., and Brockington, D. (2024). The strain on scientific publishing. Quantitative Science Studies, 5(4), pp. 823–843.

Hofstra, B., Kulkarni, V.V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., et al. (2020). The diversity–innovation paradox in science. Proceedings of the National Academy of Sciences, 117(17), pp.9284-9291.

Hosseini M, Horbach SPJM (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Res Integr Peer Rev. 18;8(1):4.

Hosseini, M., Gordijn, B., Kaebnick, G. E., & Holmes, K. (2025). Disclosing generative AI use for writing assistance should be voluntary. Research Ethics, 21(4), 728-735. https://doi.org/10.1177/17470161251345499

Houssard, A., Gargiulo, F., Venturini, T., Tubaro, P. and Di Bona, G. (2025). The Gerontocratization of Science: How hypergrowth reshapes knowledge circulation. arXiv. https://doi.org/10.48550/arXiv.2410.00788.

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

Hurshman, C., Porsdam Mann, S., Savulescu, J. and Earp, B.D., 2025. Authorship Without Writing: Large Language Models and the Senior Author Analogy. arXiv preprint arXiv:2509.05390.

Idahl, M. and Ahmadi, Z., 2024. Openreviewer: A specialized large language model for generating critical scientific paper reviews. arXiv preprint arXiv:2412.11948.

Jackson, D., Smith, G.D. and Cleary, M., 2025. The Privilege and Power of Peer Review: Advancing Science With Integrity, Vigilance and Fairness. Journal of Advanced Nursing. 82: 1-4

Jin, Y., Zhao, Q., Wang, Y., Chen, H., Zhu, K., et al. 2024). Agentreview: Exploring peer review dynamics with llm agents. arXiv preprint arXiv:2406.12708.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), pp.583-589.

Juzek, T. S., & Ward, Z. B. (2025). Word Overuse and Alignment in Large Language Models: The Influence of Learning from Human Feedback. arXiv preprint arXiv:2508.01930.

Kousha, K. and Thelwall, M. (2024), Artificial intelligence to support publishing and peer review: A summary and review. Learned Publishing, 37: 4-12. https://doi.org/10.1002/leap.1570

Kuhn, T. S. (1962/1970). The Structure of Scientific Revolutions. University of Chicago Press.

Kravitz, R. L., Franks, P., Feldman, M. D., Gerrity, M., Byrne, C., et al. (2010). Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? PLOS ONE 5(4): e10072. https://doi.org/10.1371/journal.pone.0010072

Korbmacher, M., Azevedo, F., Pennington, C.R. et al. (2023). The replication crisis has led to positive structural, procedural, and community changes. Communications Psychology, 1, 3.

Laine, C., Babski, D., Bachelet, V.C., Bärnighausen, T.W., Baethge, C., et al, 2025. Predatory journals: what can we do to protect their prey?. The Lancet, 405(10476), pp.362-364.

LeBlanc, A.G., Barnes, J.D., Saunders, T.J. et al. Scientific sinkhole: estimating the cost of peer review based on survey data with snowball sampling. Res Integr Peer Rev 8, 3 (2023). https://doi.org/10.1186/s41073-023-00128-2

Latona, G.R., Ribeiro, M.H., Davidson, T.R., Veselovsky, V. and West, R.(2024). The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. arXiv preprint arXiv:2405.02150.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D.Y., et al, 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI, 1(8), p.AIoa2400196.

Liddicoat, J.E., Lenarczyk, G., Aboy, M., Minssen, T. and Porsdam Mann, S.(2025). A policy framework for leveraging generative AI to address enduring challenges in clinical trials. npj Digital Medicine, 8(1), 33.

Lin, Z., & Sohail, A. (2026). Recalibrating Academic Expertise in the Age of Generative AI. Patterns, 7(1), 101473.

Lin, Z. 2025a. AI chatbots are already biasing research — we must establish guidelines for their use now. *Nature*, *645*(8080), p.285.

Lin, Z. (2025b). Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review. *arXiv preprint* arXiv:2507.06185.

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., et al, 2023. Prompt injection attack against llm-integrated applications. arXiv preprint arXiv:2306.05499

Liu, R. and Shah, N.B., 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. arXiv preprint arXiv:2306.00622.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 34, pp.9459-9474.

Luo, X., Tham, Y. C., Giuffrè, M., Ranisch, R., Daher, M., et al. (2025). Reporting guideline for the use of Generative Artificial intelligence tools in MEdical Research: the GAMER Statement. BMJ evidence-based medicine, bmjebm-2025-113825. Advance online publication. https://doi.org/10.1136/bmjebm-2025-113825

Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017 Mar 1;24(2):423-431. doi: 10.1093/jamia/ocw105. PMID: 27516495; PMCID: PMC7651899.

Mosier KL, Skitka LJ, Heers S, Burdick M. Automation bias: decision making and performance in high-tech cockpits. Int J Aviat Psychol. 1997;8(1):47-63. doi: 10.1207/s15327108ijap0801_3. PMID: 11540946.

Moonshot AI (2025). Kimi K2: Open Agentic Intelligence. Technical documentation. Available at: https://moonshotai.github.io/Kimi-K2/

Mollaki, V. (2024). Death of a reviewer or death of peer review integrity? the challenges of using AI tools in peer reviewing and the need to go beyond publishing policies. Research Ethics, 20(2), 239-250.

Murray, D., Siler, K., Larivière, V., Chan, W.M., Collings, A.M., Raymond, J. and Sugimoto, C.R., 2018. Author-reviewer homophily in peer review. BioRxiv, p.400515.

Naddaf M. (2025). AI is transforming peer review - and many scientists are worried. Nature, 639(8056), 852–854. https://doi.org/10.1038/d41586-025-00894-7

Nguyen, T., 2024. ChatGPT in medical education: a precursor for automation bias?. JMIR Medical Education, 10, p.e50174.

NIH, 2023. The Use of Generative Artificial Intelligence Technologies is Prohibited for the NIH Peer Review Process. NOT-OD-23-149. Available at: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html

Nuijten MB, Hartgerink CH, van Assen MA, Epskamp S, Wicherts JM. The prevalence of statistical reporting errors in psychology (1985-2013). Behav Res Methods. 2016 Dec;48(4):1205-1226. doi: 10.3758/s13428-015-0664-2. PMID: 26497820; PMCID: PMC5101263.

Ohde, J.W., Rost, L.M. and Overgaard, J.D., 2025. The burden of reviewing LLM-generated content. NEJM AI, 2(2), p.AIp2400979.

OpenAI. (2024). Learning to reason with LLMs. Retrieved from https://openai.com/index/learning-to-reason-with-llms/

OpenAI (2025). Data processing addendum. Retriev from https://openai.com/policies/data-processing-addendum/

Pataranutaporn, P., Powdthavee, N., Achiwaranguprok, C. and Maes, P., 2025. Can AI Solve the Peer Review Crisis? A Large Scale Cross Model Experiment of LLMs' Performance and Biases in Evaluating over 1000 Economics Papers. arXiv preprint arXiv:2502.00070.

Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P. Et al, 2023, October. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology (pp. 1-22).

Politzer-Ahles S, Girolamo T, Ghali S. Preliminary evidence of linguistic bias in academic reviewing. Journal of English for academic purposes. 2020 Sep 1;47:100895.

Porsdam Mann, S., Earp, B. D., Møller, N., Vynn, S., & Savulescu, J. (2023). AUTOGEN: A Personalized Large Language Model for Academic Enhancement—Ethics and Proof of Principle. The American Journal of Bioethics, 23(10), 28–41. https://doi.org/10.1080/15265161.2023.2233356

Porsdam Mann, S., Seah, J.J., Latham, S., Savulescu, J., Aboy, M. et al., 2025. Chat-IRB? How application-specific language models can enhance research ethics review. Journal of Medical Ethics.

Porsdam Mann, S., Vazirani, A.A., Aboy, M. et al. Guidelines for ethical use and acknowledgement of large language models in academic writing. Nat Mach Intell 6, 1272–1274 (2024). https://doi.org/10.1038/s42256-024-00922-7

Publons/Clarivate. (2018). Global State of Peer Review. https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf

Romera-Paredes B, Barekatain M, Novikov A, Balog M, Kumar MP, Det al. Mathematical discoveries from program search with large language models. Nature. 2024 Jan;625(7995):468-475. doi: 10.1038/s41586-023-06924-6. Epub 2023 Dec 14. PMID: 38096900; PMCID: PMC10794145.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., et al, 2023. Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 37, pp.68539-68551.

Schroter S, Black N, Evans S, Carpenter J, Godlee F, et al Effects of training on quality of peer review: randomised controlled trial. BMJ. 2004 Mar 20;328(7441):673. doi: 10.1136/bmj.38023.700775.AE. Epub 2004 Mar 2. PMID: 14996698; PMCID: PMC381220.

Shcherbiak A, Habibnia H, Böhm R, Fiedler S. Evaluating science: A comparison of human and AI reviewers. Judgment and Decision Making. 2024;19:e21. doi:10.1017/jdm.2024.24

Stelmakh, I., Shah, N.B., Singh, A. and Daumé III, H.(2021). A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 6, pp. 4785-4793).

Straiton, J. (2024). Artificial intelligence: help or hindrance in solving the reproducibility crisis? BioTechniques, 76(7), pp.291–294.

Thakkar, N., Yuksekgonul, M., Silberg, J., Garg, A., Peng, N., Set al. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. arXiv preprint arXiv:2504.09737

Thelwall, M., Yaghi, A. (2025). Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms. Scientometrics. https://doi.org/10.1007/s11192-025-05287-1

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single- versus double-blind peer review. Proceedings of the National Academy of Sciences, 114(48), 12708–12713.

Turner, L. Promoting F.A.I.T.H. in Peer Review: Five Core Attributes of Effective Peer Review. Journal of Academic Ethics 1, 181–188 (2003). https://doi.org/10.1023/B:JAET.0000006844.09724.98

Wang, J.L., Li, R., Yang, J. and Mao, C., 2024, November. RAFT: Realistic attacks to fool text detectors. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 16923-16936).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Ichter, B., et al (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 36, pp.24824-24837

Xhonneux, S., Sordoni, A., Günnemann, S., Gidel, G. and Schwinn, L., 2024. Efficient adversarial training in llms with continuous attacks. Advances in Neural Information Processing Systems, 37, pp.1502-1530.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., et al, 2022. React: Synergizing reasoning and acting in language models. In The eleventh international conference on learning representations.

Ye, R., Pang, X., Chai, J., Chen, J., Yin, Z., et al. (2024). Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. arXiv preprint arXiv:2412.01708

Zhang, Y., Shen, Y., Kang, S., Chen, X., Jin, B., et al (2025). Chain-of-factors paper-reviewer matching. In Proceedings of the ACM on Web Conference 2025 (pp. 1901-1910)