

Beyond Detection: Governing GenAI in Academic Peer Review as a Sociotechnical Challenge

Tatiana Chakravorti¹, Pranav Narayanan Venkit², Sourojit Ghosh³, Sarah Rajtmajer¹

¹Pennsylvania State University, USA

²Salesforce, USA

³University of Washington, USA

tfc5416@psu.edu, pnarayananvenkit@salesforce, ghosh100@uw.edu, smr48@psu.edu

Abstract

Generative AI (GenAI) tools are increasingly entering academic peer-review workflows, raising questions about fairness, accountability, and the legitimacy of evaluative judgment. While these systems promise efficiency gains amid growing reviewer overload, their use introduces new sociotechnical risks. This paper presents a convergent mixed-method study combining discourse analysis of 448 social media posts with interviews with 14 area chairs and program chairs from leading AI and HCI conferences to examine how GenAI is discussed and experienced in peer review. Across both datasets, we find broad agreement that GenAI may be acceptable for limited supportive tasks, such as improving clarity or structuring feedback, but that core evaluative judgments, assessing novelty, contribution, and acceptance, should remain human responsibilities. At the same time, participants highlight concerns about epistemic harm, over-standardization, unclear responsibility, and adversarial risks such as prompt injection. User interviews reveal how structural strain and institutional policy ambiguity shift interpretive and enforcement burdens onto individual scholars, disproportionately affecting junior authors and reviewers. By triangulating public governance discourse with lived review practices, this work reframes AI-mediated peer review as a sociotechnical governance challenge and offers recommendations for preserving accountability, trust, and meaningful human oversight. Overall, we argue that AI-assisted peer review is best governed not by blanket bans or detection alone, but by explicitly reserving evaluative judgment for humans while instituting enforceable, role-specific controls that preserve accountability. We conclude with role-specific recommendations that formalize the support judgment boundary.

Introduction

As generative AI (GenAI) is increasingly entering academic peer-review workflows, there is active discourse and a growing body of research around the opportunities for effective usage to reduce workloads on strained reviewers and increase academic equity (Biswas, Dobaria, and Cohen 2023; Chen, Brumby, and Cox 2025; Doslaliuk et al. 2025), as well as concerns around the impact of GenAI usage on the fairness, accountability, and legitimacy of scholarly evaluation. As debates continue, GenAI usage in peer review

is already prevalent; (Liang et al. 2024) estimates that between 6.5% and 17% of reviews submitted to prominent NLP conferences during the 2023-2024 cycles contained AI-generated text, even where conference policies explicitly prohibited such use. In this environment, we explore the following research questions:

- **RQ1:** How do peer-review stakeholders understand and negotiate appropriate use of GenAI in peer review?
- **RQ2:** What risks and harms do stakeholders observe or anticipate from GenAI?
- **RQ3:** How do governance practices and institutional factors shape how GenAI is adopted, contested, and regulated in peer review?

Our study centers the perspectives of area and program chairs (ACs/PCs) across AI and HCI venues, triangulated with public social media discourse, to examine how generative AI is negotiated within real-world peer review practices. By linking governance debates with the lived realities of reviewer labor and decision making, we move beyond abstract discussions of AI adoption to analyze peer review as a sociotechnical system under strain. We conceptualize AI-mediated peer review as a governance problem rather than a tool-use question, showing how institutional policy ambiguity operates as a deliberate yet consequential governance strategy that shifts interpretive and enforcement burdens onto individual scholars. Our findings further reveal how GenAI redistributes responsibility and risk across the review pipeline, while simultaneously reshaping peer review as a form of relational and care work, with implications for accountability, fairness, and epistemic legitimacy. This paper contributes:

1. An empirically grounded study to understand how AI and HCI stakeholders construct the “support versus judgment” boundary in AI-mediated peer review;
2. Identification of strategic policy ambiguity as an institutional governance strategy under technological uncertainty;
3. Analysis of how structural strain drives AI adoption and the ethical challenges that occur;
4. A reframing of peer review as a sociotechnical pipeline in which AI reshapes accountability, trust, and harm distribution

Finally, we provide one of the first triangulated empirical analyses linking public discourse to decision-making practices by ACs/PCs, demonstrating how AI-mediated peer review redistributes responsibility and risk across the peer review system.

Background

Peer Review as a Value-Laden Practice

Peer reviewing has been a cornerstone of the academic process for centuries, dating back to the editorial practices for medical journals in the 17th and 18th centuries, where editors would almost independently assess the quality of submitted research for publication (Drozdz and Ladomery 2024; Palmer 1908; Shiflett 1988). The modern model of academic peer review emerged in the 19th century, as editors began to acknowledge the limits of their own abilities and sought out others with some specialized knowledge required to evaluate a given submission (Burnham 1990; Kronick 1990). The practice of peer reviewing was adopted by scientific journals (including ACM journals such as the *Journal for the ACM*) in the mid-20th century, initially with the editor-only model before transitioning into the external-reviewers approach to account for growth and varying specializations.

Peer reviewing research is more than evaluating science to pass muster for some conference/journal. Like much of the scientific process, peer reviewing is not value-neutral; it reflects and reinforces (and sometimes contests) disciplinary values about what and whose knowledges are considered legitimate, rigorous, and worth disseminating, through judgments about quality, novelty, and impact entangled with social norms, institutional incentives, and ethical commitments. Peer reviewing is a performance of power (Berkenkotter 1995; Lipworth and Kerridge 2011) in which uncompensated reviewers can gatekeep the publication of knowledge to the scientific community. Authors are afforded little recourse, as resubmissions are more likely to face rejection (Stelmakh et al. 2021). This can be especially true for a conference with as broad a call for papers as FAccT, which can result in individuals being assigned reviews beyond their expertise (Laufer et al. 2022). While FAccT and similar conferences try to account for such subjectivity with reviewer guidelines requiring decisions to be strongly motivated and consensual decision-making among anonymous reviewers (Rastogi et al. 2024), it is still possible that a value alignment between reviewers unfairly affects authors.

Although prevalent practice in journals and conferences is double-blind peer review, there is also ample evidence of peer reviewers exhibiting subconscious bias based on reviewers' perceptions of authors' identities along the lines of gender (Borsuk et al. 2009; Helmer et al. 2017; Pontille and Torny 2014; Whittaker 2008), nationality and English proficiency (Marsh, Jayasinghe, and Bond 2008; Tregenza 2002), and prestige and institutional affiliation (Bornmann, Weymuth, and Daniel 2010; Lee et al. 2013). Anonymizing authorial (and sponsor/institution) identities during peer review can sometimes create other issues, such as potentially ethically-conflicted research appearing benign (Young,

Katell, and Krafft 2022) and reviewers consciously or subconsciously trying to identify authors (Chung et al. 2015; Haffar, Bazerbachi, and Murad 2019; Pontille and Torny 2014).

GenAI in Peer Review: Approaches, Risks, and Concerns

Peer reviewing as a practice has been radically impacted by the advent and popularity of GenAI tools, as their ability to generate natural language reviews for papers or respond to specific questions required during peer review means that reviewers can easily offload their tasks to these systems. There exist GenAI tools specifically for peer review, e.g., ReviewWriter (Su et al. 2023), which guides users through the writing process with intermittent nudges and instructions, without allowing them to upload a document and receive an AI-generated review¹. This (and similar approaches (Drori and Te'eni 2024; Neshaei et al. 2024; Thakkar et al. 2025)) attempts to strike a balance in AI-assisted peer review, whereby humans write their own reviews with augmentation and support from GenAI, including review outlining, suggested line completion, dropdown lists of suggestions, and more, addressing issues of reviewer fatigue while still maintaining academic objectivity and rigor (Checco et al. 2021; Chidambaram 2025; Feetham-Walker et al. 2025; Gruda 2025; Robertson 2023; Zhuang et al. 2025). Other GenAI tools produce entirely synthetic paper reviews, whose capabilities including writing broad reviews (Chitale et al. 2025; Gao et al. 2025; Taechoyotin and Acuna 2025; Wang et al. 2020; Zeng et al. 2025), adapting review text to the specific requirements of a conference (Du et al. 2024; Idahl and Ahmadi 2025; Zhou et al. 2024), or predicting peer review feedback and suggesting improvements (Bougie and Watanabe 2024; Liu and Shah 2023; Sukpanichnant, Rapberger, and Toni 2024; Tan et al. 2024; Weng et al. 2025).

In the field, one perspective views GenAI as a pathway for more equitable access into academia (Austin et al. 2023; Berdejo-Espinola and Amano 2023; Ghosh 2024; Kohnke and Zaugg 2025; Nixon, Lin, and Snow 2024; Roshanaei, Olivares, and Lopez 2023, e.g.), where students and researchers from communities historically marginalized in academia can leverage AI systems to participate in research activities at similar levels as their more privileged peers. This can take several forms in peer review, such as accurately understanding papers' contents by parsing the often-complicated lexicon of academic writing (e.g., Al Sawi and Alaa 2024; Ghosh 2024) and proofreading reviews written outside writers' first or native languages (e.g., Alalaq 2024; Sebastian and Baron 2024). On the other hand, using GenAI tools for peer review raises serious concerns about the erosion of academic integrity (Liang et al. 2024), over and above the fact that AI-generated reviews might contain hallucinations and otherwise shallow evaluations of rich scholarly content (Conroy 2023; Chawla 2024; Gao et al. 2025). The overuse of GenAI threatens an epistemic convergence as the low ceiling of originality in AI-generated text (Bender

¹For information about other LLM-assisted generative review agents, see (Sun 2025).

et al. 2021; Birhane et al. 2022) promises the reinforcement of traditional methods and established disciplinary norms. AI-generated reviews also tend to overinflate praise without substantive critique (Donker 2023; Zhou, Chen, and Yu 2024), miss potential errors made by researchers (Goldberg et al. 2024; Liu and Shah 2023), and take up different points of view from human reviewers (Du et al. 2024; Saad et al. 2024; Suleiman et al. 2024; Yu et al. 2024), thus contributing to higher rates of acceptance for poorer quality papers (Russo et al. 2025; Thelwall and Yaghi 2024).

A majority of research around the impact of GenAI on peer reviewing does not consider the direct perspectives of the area/program chairs and reviewers (with a few exceptions (Ali and Shaban 2025; Ebadi et al. 2025)). Our work bridges this gap, centering their perspectives and contextualizing them within social media discourse. Our social media study captures public and semi-professional framings of GenAI in peer review, while interviews provide in-depth accounts of how reviewers and organizers understand and navigate these issues within their roles. Our interview protocol was developed independently and was not informed by the social media analysis, and both components were analyzed individually before putting the findings in conversation, using a convergent parallel mixed methods design (Demir and Pismek 2018; Kerrigan 2014).

Methodology

Social Media Study

We collected social media posts discussing the use of GenAI in peer review from LinkedIn, Reddit, and Twitter/X to capture diverse audiences and communication styles: LinkedIn for professional and academic discourse; Reddit for discussion-oriented and experiential exchanges; and Twitter/X for rapid, public-facing commentary. Using platform-specific search and filtering tools, we collected posts and comments containing keywords including "AI + peer review", "problems with AI peer review", "ChatGPT + peer review", and "LLM + peer review". Data collection focused on English language content between September 2024 - August 2025 and excluded private, locked, or deleted posts. Retweets without additional commentary and duplicate content were removed. Our final dataset consisted of 448 comments, approximately evenly distributed across these three platforms: 149 posts each from LinkedIn and Twitter/X; and, 150 from Reddit. We did not collect or attempt to identify individual identities. Our corpus is not intended to be representative of all platform discourse as it is a search-retrieved, English-language convenience sample shaped by our keyword queries and platform ranking algorithms. To improve coverage, we used multiple synonymous search strings across platforms and deduplicated results. Our findings are intended to be interpreted as prevalent framings within the retrieved corpus, not population prevalence estimates.

We conducted qualitative discourse analysis (Zajda 2020; Mogashoa 2014) on social media data to identify discussion patterns around AI in peer review, through an iterative coding process combining theory-informed deductive cate-

gories prevalent in scholarship on AI governance and academic evaluation, and inductive refinement based on recurring patterns.

Interview study

We conducted 14 semi-structured interviews (Varanasi et al. 2022; Thakkar et al. 2022) with top AI and HCI conference ACs and PC chairs to understand their perceptions, experiences of the current peer review process, as well as how and where AI can be useful and where it should not be used in peer review.

Recruitment. We sent emails and social media messages on LinkedIn and Twitter/X to reach diverse academic researchers from AI/HCI conferences, including PhD students, post-doctoral researchers, assistant professors, and industry researchers. We recruited 14 participants (P1-P14) (see Table 1 for demographic information). Recruitment messages contained information about the study, expected interview length, compensation, inclusion criteria, and a link to the screener survey. We also employed snowball sampling to directly recruit ACs and PC chairs.

ID	Gender	Position	AC/PC	Conference
P1	Female	Postdoc	AC	HCI
P2	Female	PhD	AC	HCI
P3	Male	Asst Prof	AC	HCI
P4	Male	Asso Prof	AC, PC	AI
P5	Female	Postdoc	AC	HCI
P6	Female	Asso Prof	AC, PC	AI
P7	Female	Industry	AC	HCI
P8	Male	Asst Prof	AC	HCI
P9	Male	Asso Prof	AC, PC	HCI
P10	Male	Asso Prof	AC, PC	AI
P11	Male	PhD	AC	AI
P12	Female	Asst Prof	AC	HCI
P13	Male	PhD	AC	HCI
P14	Male	PhD	AC	AI

Table 1: Gender, position, area chair (AC)/ program chair(PC), and conference type of interview participants.

Interview protocol. 30-60 minute semi-structured interviews (Adams 2015) were conducted virtually in September-November 2025. All interviews were recorded, and transcripts for analysis were generated via Zoom. The interview protocol was designed to elicit participants' experiences with peer review and their perspectives on the role of GenAI. We began by asking participants about their academic background, experience conducting peer reviews, and the tools or processes they typically use when assessing manuscripts. We invited them to describe whether and how they had used GenAI systems to support reviewing tasks, and which aspects of evaluation they believed these tools might enhance. We then explored perceptions of the benefits, risks, and appropriate boundaries of GenAI in peer review, asking about potential advantages, drawbacks, and areas where AI involvement might feel inappropriate or unsafe. Finally, we connected these experiences to broader eth-

ical and governance concerns, prompting participants to reflect on guidelines, policies, or institutional norms they believe should shape the responsible use of GenAI in scholarly review.

Interviews were thematically analyzed (Blandford, Furniss, and Makri 2016; Ding et al. 2022) as authors read transcripts multiple times before individually open-coding them and then grouping codes across authors into preliminary themes that reflected emergent conceptual groupings relevant to the research questions. The first two authors developed the codebook through iterative open coding, individually followed by discussion. Themes were refined through regular discussions among authors until they were finalized, as authors went back to the transcripts to verify that each theme was supported by rich data excerpts and sufficiently represented variations across participants. We present findings from each dataset independently before synthesizing them through triangulation. We begin with the social media discourse analysis. Our sampling of AI and HCI venues was purposive, as these two communities work very differently from each other. Our goal is to generate theoretical insight into emerging governance practices rather than estimate prevalence. To ensure analytic consistency, approximately 30% of the dataset was independently double-coded by both annotators.

Findings from Social Media: Six Themes

1. Trust Shifting Toward Procedural Mechanisms: Across the dataset, discourse was dominated by concerns about governance, enforceability, and verification, rather than purely evaluative judgments about whether AI is “good” or “bad.” The most frequent theme was trust shifting toward procedural mechanisms, present in 27.5% of posts. Posts coded in this theme emphasized the need for formalized processes to manage uncertainty introduced by AI use, including calls for disclosure rules, institutional policies, enforcement mechanisms, and editorial or program committee oversight. These discussions often framed procedural safeguards as necessary to preserve the credibility of peer review and to restore confidence in the provenance and reliability of reviews.

2. Ethical Risk and Unequal Harm: The second most prevalent theme was ethical risk and unequal harm, appearing in 23.2% of posts. Ethical concerns clustered around confidentiality and privacy, the risk of hallucinated or incorrect statements, and broader worries about unfairness and bias. A subset of posts also raised concerns about uneven consequences and vulnerability, including the possibility that ambiguous rules or enforcement practices could disproportionately burden less powerful actors. Overall, ethical risk discourse was not limited to abstract principles; it frequently focused on concrete practices.

3. Boundary Setting around AI: A third major theme involved boundary setting, present in 21.7% of posts. These posts drew explicit lines between acceptable and unacceptable uses of AI within peer review. Common acceptable uses included support tasks such as summarizing papers, improving grammar or clarity, structuring feedback, or helping reviewers organize their thoughts. In contrast, many posts re-

jected AI use for evaluative judgments, especially decisions about novelty, contribution, correctness, and accept/reject recommendations. This boundary was often framed as a defense of expert judgment and accountability, emphasizing that peer review requires contextual understanding and disciplinary responsibility that AI systems cannot reliably provide.

4. Responsibility and Care Work: The fourth most frequent theme was responsibility and care work, appearing in 17.2% of posts. Posts coded in this theme portrayed reviewing as a professional obligation rather than a mechanical task. They highlighted accountability, diligence, and the value of constructive feedback. This discourse also included discomfort with the idea that AI could normalize minimal-effort reviewing or reduce peer review to a procedural formality. While responsibility language sometimes overlapped with policy debates, it was distinct in its focus on norms of scholarly conduct and the ethical meaning of reviewing labor.

5. Detection Culture and Policing: Detection culture and policing of AI use appeared in 12.3% of posts. These posts centered on claims that AI-written or AI-assisted reviews are recognizable and debated whether detection is feasible, fair, or desirable. Some argued for strict sanctions, bans, or the use of detection tools, while others warned about false positives and reputational harms. Across these discussions, a shared concern was that uncertainty about whether reviews were authored by humans undermines confidence in the process. Detection discourse, therefore, functioned less as a purely technical conversation and more as an indicator of shifting norms: provenance becomes a matter of suspicion and verification rather than assumed integrity.

6. Adversarial Peer Review and Prompt Injection: Discourse related to adversarial peer review and prompt injection were present in 10.3% of posts. Posts in this category discussed tactics in which hidden or embedded instructions are inserted into manuscripts to manipulate LLM-based readers or reviewers (e.g., “prompt injection” using invisible text). This discourse framed peer review as entering a more adversarial ecosystem, where actors anticipate and exploit vulnerabilities introduced by AI-mediated reading and summarization. Discussions often extended from describing attacks to proposing countermeasures, such as stripping hidden text, filtering inputs, instituting procedural checks, or discouraging or prohibiting certain AI use cases.

Findings from Interviews: Six Dominant Themes

Pre- and Post-AI Structural Strain in Peer Review

All participants consistently mentioned peer review as a system that was already under significant strain, long before the emergence of GenAI. Across roles, they pointed to reviewer fatigue and overload, noting that a small pool of reviewers is repeatedly asked to evaluate a growing number of submissions. Reviewing was described as largely unpaid and undervalued labor, often added on top of heavy workloads that include teaching, research, service, and administrative

responsibilities. Thus, reviewers feel stretched thin and have limited time and energy for each manuscript.

It's challenging to find reviewers. Reviewers are really overloaded. All committee members have, like, 8 to 15 reviews that they're assigned to for one conference. This is a real problem of overload, which, I think, tempts people towards trying to use GenAI to reduce and manage their overloaded work. --P1

P1 also added that these problems increase when you work as an AC or PC because, at that time, you need to check that your subcommittee members are not using AI for review.

GenAI use becomes especially risky at higher levels of conference leadership because when area chairs or subcommittee members use GenAI to draft meta-reviews, it can undermine the integrity of decisions that are supposed to reflect expert judgment --P1

Participants also emphasized strong inconsistencies in review quality. While some reviews were described as careful, detailed, and constructive, others were brief and vague. This variability was linked not only to individual differences among reviewers but also to the lack of clear or shared reviewing standards.

Always, human reviews are not of high quality. Sometimes the reviews are short and low in quality. --P6

Within this already fragile system, participants did not view GenAI as the root cause of these problems. Instead, they described GenAI as amplifying existing weaknesses. When reviewers are overloaded, pressed for time, and uncertain about standards, AI tools may appear attractive as a way to cope with demands. At the same time, participants worried that this could further normalize minimal effort reviewing and deepen existing quality gaps. P5 describes the huge number of submissions as a major problem these days, as people use GenAI to paraphrase and do multiple submissions.

There has been a real uptake in the number of submitted papers. For example, with AAAI, I heard this year, 20,000 papers were submitted, and CHI were 6000. It's getting harder to tell what ethical submissions are. People were using AI to paraphrase their papers and submitting multiple copies of it. --P5

P4 also added that as the number of submissions is getting higher day by day, it is difficult to find experienced reviewers. Getting reviewed by early-career PhD or master's students can produce more low-quality reviews. He also added details about how the use of generative AI is introducing challenges for senior reviewers.

At a senior role, I am on the opposite side when the other reviewers are saying that this is an AI-generated review or something. And also, we got directives that, if we are looking at other reviews as a senior reviewer. We need to make sure that this is not AI-generated. So, it now adds one more thing, we had to make sure that the reviews are high quality, the reviews are polite, the reviews are collegial, now we also have to make

sure that the reviews are not AI-generated, because AI-generated reviews are also very polite. Like, most of the cases, that's what we have seen. So that makes it very hard to, hard to counter. --P4

Boundary Setting around AI

Across interviews, 13 out of 14 participants, except P7, consistently engaged in what can be described as boundary setting around AI, actively defining which aspects of peer review could appropriately involve GenAI and which must remain the responsibility of human reviewers. Rather than expressing rejection of AI, most participants articulated clear limits on its use. AI was generally seen as acceptable for supportive or preparatory tasks, such as writing manuscripts, checking false references, improving grammar or clarity, organizing thoughts, or helping structure a review. These uses were framed as ways to reduce cognitive load without replacing the reviewer's own reasoning.

A lot of my students who are English-as-a-second-language speakers, their writing and their ability to express their ideas in English have gotten significantly better, and I know they're co-writing with AI because they tell me. They used to pay professional copyeditors to rewrite papers, now they don't need it. --P10

P10 also mentions that GenAI can be super useful for checking grammar and typos, which can eventually save time for the reviewers, and they can use that time to check the technicality and ideas, as did P8.

I also have been using AI to proofread my language, and to correct grammar mistakes, and sometimes my review is too rambling, so I asked AI to shorten my writing without cutting any original meanings and writing that I had there, so that was very helpful for me to make sure my language is polished. --P8

In contrast, participants drew firm boundaries around evaluative tasks. Uses of GenAI to assess a paper's novelty, scientific contribution, methodological soundness, or overall merit, and especially to accept or reject recommendations, were widely described as inappropriate. Participants emphasized that these tasks require deep disciplinary expertise, familiarity with ongoing debates in the field, and sensitivity to context that cannot be reduced to patterns in training data. Many expressed concern that GenAI cannot understand why a contribution matters within scholarly conversations.

LLMs or generative AI have been used extensively nowadays, but they should not replace human judgment. The final decision should be made by humans. LLMs can not accept or reject, or should not make any decisions in the future as well. When it goes to technical details they can not do it properly. --P12

Importantly, these boundaries were not justified solely in terms of GenAI's technical limitations. Instead, participants framed them in normative and professional terms. Judgment in peer review was described as inseparable from accountability; reviewers are expected to stand behind their evaluations, explain their reasoning, and take responsibility for the

consequences of their recommendations. Delegating judgment to AI was seen as undermining this responsibility.

Ethical Risk and Unequal Harm

Participants consistently raised ethical concerns about the use of GenAI in peer review, most commonly pointing to risks such as hallucinated or incorrect claims or false references, amplification of existing biases, misuse of AI, and potential breaches of confidentiality. All participants except P11 mentioned their concerns about the growing use of AI in Peer Review. Many worried that AI tools could introduce errors or misleading statements into reviews, particularly when reviewers rely on them without carefully verifying outputs.

These are unpublished works that are being fed into systems that can potentially ingest, maintain, and redistribute that information to other users or train themselves on it, so you end up kind of feeding in potentially invalid information. There are copyright issues involved in that as well. A review that I got on a paper that I wrote about, it was about family studies, one of the reviewers mentioned that the paper was about children with ADHD. But it wasn't. There was no mention of ADHD or neurodivergence throughout the entire study, and this was a review that was recommending rejection. That made me wonder if maybe this was a detail that was hallucinated by GenAI? --P1

P4 and P5 mentioned a critical ethical problem of duplicate submissions. They both came across duplicate submissions, as using GenAI, it is very easy to paraphrase.

One thing that AI can be useful for is selecting and flagging duplicate submissions. I have played senior roles in multiple conferences. I have noticed that there are duplicate submissions, and the duplicates are done in a very clever way using AI, so that it is not identical; the title will be slightly different, the abstract will be slightly different, and the introduction will be slightly different. It's almost deliberately meant to game the system, so everything will be slightly different, but they are basically the same paper. --P4

Some participants expressed concern that AI systems may reproduce or intensify biases present in training data, shaping evaluations in ways that disadvantage certain topics, methods, or groups of scholars. But many participants are not very concerned about biases, as human review systems also have their own biases. Misuse of AI and its hallucination is a far more important concern mentioned by the participants.

An important concern is whether these AI-generated reviews are subsequently evaluated or verified by human reviewers. This is particularly critical due to the well-known issue of hallucinations in large language models. Without human oversight, it is unclear whether all the comments produced by the AI are valid, meaningful, or factually correct. Based on our

own investigation of the AI-generated review for our paper, we found that some comments were not meaningful, and others were simply incorrect. This raises serious concerns about the reliability of AI-based reviews and highlights the need for human validation to ensure the accuracy and usefulness of the feedback. --P6

P7 mentioned that this is gradually taking the critical thinking capabilities from humans, as peer review needs a lot of critical thinking. Beyond these general ethical risks, participants emphasized that the consequences of AI use would not be evenly shared across the academic community. Instead, they described a system in which some groups, particularly junior scholars, would bear a disproportionate share of the potential harm. Early-career authors were seen as especially vulnerable to low-quality or AI-assisted reviews, as they often have less power to contest unfair evaluations or request reconsideration. Participants noted that when feedback is vague, inaccurate, or inconsistent, junior authors may be more likely to internalize negative judgments or face career consequences without recourse.

Responsibility and Care Work

This is also another emerging theme that came across the interviews, as 11 out of 14 participants mentioned this, except P5, P10, and P14. Participants frequently described peer review as more than a technical process for deciding which papers should be accepted or rejected. Instead, many framed it as a form of scholarly responsibility that involves mentorship, care for authors, and the reproduction of good work and professional norms within a field. Providing thoughtful feedback, explaining reasoning, and helping authors improve their work were described as central to what it means to be a good reviewer, particularly for early-career scholars.

People who take it seriously and in good faith aren't going to use [GenAI]... It's really about finding people who are qualified to take it seriously. But I am afraid that people are gradually losing their responsibility and care for peer review. --P9

Within this understanding, participants expressed concern that GenAI could gradually undermine the sense of responsibility associated with reviewing. Several worried that reliance on AI tools might normalize perfunctory or surface-level reviews, especially in an already overloaded system. When reviews are produced quickly or partially delegated to AI, participants feared that reviewers may feel less accountable for the content of their feedback, even if they technically approve or edit the final text.

Previously, long reviews meant that the reviewer had thoroughly reviewed and done a great job, but now producing a longer review has become very easy... a long review doesn't necessarily mean the reviewer put in a lot of work anymore. --P4

Strategic Ambiguity as Institutional Governance

Across interviews, 9 out of 14 participants, P2, P3, P5, P8, P9, P10, P11, P12, and P14, pointed to the lack of clear,

enforceable policies governing the use of GenAI in peer review. Participants explained that this ambiguity was motivated by several concerns. First, they noted that GenAI technologies are evolving rapidly, making it difficult to define stable rules that would remain relevant over time. Clear or rigid policies were seen as likely to become outdated quickly or to unintentionally prohibit legitimate uses. Second, participants described fears of institutional liability or controversy. Committing to firm positions was viewed as risky, especially in the absence of consensus across the scholarly community. By keeping policies vague, journals and conferences could avoid backlash from reviewers, authors, or professional societies. At the same time, participants recognized that this strategic ambiguity came with costs. When expectations are not clearly defined, responsibility for interpretation shifts downward to individual scholars.

There are copyright issues, there are ethical issues, and there's a real concern about liability. If an institution makes a strong call and it turns out to be the wrong one, then someone has to take responsibility for that. And I think that's why nobody really wants to be the one making that definitive decision. --P1

Several participants noted that ambiguity also complicates enforcement. Without clear standards, it becomes difficult to determine when AI use crosses a line or warrants intervention.

If there isn't a clear standard, then it becomes very hard to say when something actually crosses the line. You might feel that a review is AI-generated or inappropriate, but without a clear policy, it's difficult to justify intervening or taking action, because there's no agreed-upon boundary. --P14

Cross-Community Differences

HCI vs AI Reviewers Across interviews (P1–P14), we did not observe a sharp disciplinary split in participants' overall stance toward GenAI in peer review. Participants across both HCI and AI communities broadly converged on core positions; peer review is already under strain, GenAI may be acceptable for limited supportive tasks, and evaluative judgment, e.g., assessing novelty, contribution, or accepting or rejecting recommendations, should remain a human responsibility. However, we did observe consistent differences in emphasis; that is, participants from HCI contexts tended to frame the problem in more relational and normative terms, whereas participants from AI contexts more often framed it as an infrastructural and governance challenge shaped by scale.

HCI participants more frequently described peer review as a social and moral practice, emphasizing mentorship, constructive feedback, and the relational responsibilities of reviewing. In contrast, AI participants more often situated the issue within the realities of high-volume conference reviewing and severe reviewer scarcity. Their accounts emphasized the scale and throughput pressures that make AI assistance important, alongside pragmatic questions about feasibility; what policies can realistically be enforced, how to manage

disclosure, whether detection is workable, and how procedural systems might be designed to protect integrity without collapsing under administrative burden.

There's a sense that reviewing is part of how we take care of each other in the community... it's not just about saying yes or no. --P5, HCI researcher

This year they received almost 40,000 papers... and it's hard to even get three human reviews for each paper. If you have two human reviewers, then the third one can be an AI reviewer... not for ratings, but at least to provide comments. --P6, AI researcher

Senior vs. Junior Reviewers Across the interviews, we observe a clear difference in how senior reviewers and more junior or early-career reviewers describe what is at risk in AI-mediated peer review. Senior voices tend to frame the problem as one of system governance and meta-review integrity; their attention is on whether the review pipeline can still reliably filter, weight, and synthesize evaluations under scale and AI assistance. P10 illustrates the “system reliability” framing through a concrete failure mode in AI-generated meta-review; when LLMs summarize discussions, they struggle to downweight the low-quality or unserious reviewer input. P10 describes frustration that there was no way to tell the system that this review is terrible, and it should not actually consider this review in the summary, because LLMs treat content in the context window as having roughly equal importance.

One of the things the AAAI conference did do was they used the AIs to help template or draft meta-reviews. That was a summary of the discussions. I think it really failed there. I think it was really bad there, because... there was no way for me to go in and say, This review is terrible. Like, you should not actually consider this review in the summary. --P10, associate prof

In contrast, early-career participants focused strongly on the labor demands of peer review and reviewing no longer being a space for learning. P5, a postdoctoral researcher, described a sharp rise in her reviewing workload, being asked to review an unusually large number of papers within a very short period of time. She emphasized that this increase could not be explained simply by her growing seniority, but reflected broader structural pressures on the review system. P5 also highlighted how increased submission volume has weakened existing mentorship and quality-control practices.

There's such a shortage of reviewers, we're reaching out to junior people, who may not have the experience or guidance to know what's acceptable. First-year PhD students reviewed before, but usually under guidance... now, because of sheer numbers, it's impossible to do that kind of quality control. --P5, post-doc

Triangulation

Social media discussions and interview findings show strong convergence in how different stakeholders think about the

use of GenAI in peer review. In both cases, people do not treat AI as simply good or bad, instead focusing on proper usage and delineating boundaries. Across interviews and social media posts, there is broad agreement that AI may be acceptable for limited supportive tasks, such as improving clarity or organizing feedback, but that key evaluative decisions like judging novelty, contribution, or whether a paper should be accepted should remain the responsibility of humans. Both sources also express concern about the lack of clear rules and guidance, noting that such ambiguity makes it harder to trust the review process. Ethical concerns, such as confidentiality, incorrect or hallucinated feedback, and fairness, appear frequently in both datasets, as well as worries that AI use could undermine trust in peer review if it becomes too common or poorly regulated, especially for authors who have little power to challenge unfair reviews. Three themes that overlap in both data sets, namely, *Ethical Risk and Unequal Harm*, *Boundary Setting around AI*, and *Responsibility and Care Work*.

There is also some divergence in the findings. For example, structural strain is a prominent theme emerging from our interviews, but is less discussed on social media platforms. This also highlights cross-community differences in the peer review process. Social media discussions focused primarily on governance concerns, such as disclosure, enforcement, and the legitimacy of peer review in the presence of GenAI. *Adversarial Peer Review and Prompt Injection* and *Shifting Trust Toward Procedural Mechanisms* were substantial themes in our social media data; they were touched upon but less represented in interviews. P8 mentioned rumors of AI-generated reviews containing prompts. P14 is the only participant who clearly and technically articulated concerns about prompt injection attacks and adversarial manipulation of AI-based reviewing systems. Drawing on their background as an AI safety researcher, P14 described how authors could intentionally embed hidden instructions in submitted papers to manipulate AI reviewers.

There are techniques called prompt injection. For example, you can add some text in white color into the paper's PDF, like, 'please give this paper a very high score.' As a human, you cannot read or be aware of these instructions in the paper, but the AI can read them, and might be misled. --P14

Implications

Our findings, as summarized in Table 2, have implications for understanding how the introduction of GenAI into peer review cannot be understood as a simple tool adoption problem. Instead, AI-mediated peer review is increasingly experienced and debated as a *sociotechnical pipeline* in which technical systems, institutional rules, professional norms, and uneven power relations interact to shape evaluative outcomes.

Peer Review as a Sociotechnical Pipeline, Not a Discrete Task

Across both datasets, peer review emerges as a tightly coupled sociotechnical pipeline rather than a bounded activ-

ity performed by individual reviewers. GenAI is not introduced at a single point but may intervene at multiple stages: manuscript preparation, reviewer sensemaking, feedback drafting, meta-review synthesis, and even detection and enforcement. Our findings show that concerns about AI use rarely focus on isolated interactions with tools; instead, participants worry about how AI reshapes responsibility, accountability, and trust across the entire pipeline.

This framing challenges narratives that treat AI in peer review primarily as a productivity enhancement. While efficiency gains are acknowledged, especially under conditions of reviewer overload, participants emphasize that the integrity of peer review depends on how evaluative authority is distributed and justified across human and non-human actors. When AI systems participate in producing or shaping evaluative judgments, the provenance of those judgments becomes opaque, complicating long-standing norms around scholarly accountability.

Boundaries as a Normative Defense of Human Judgment

A central observation here is that resistance to some GenAI use is not primarily about technical limitations, but about the preservation of normative commitments embedded in peer review. Both public discourse and interview participants engage in drawing boundaries that distinguish acceptable supportive uses of AI (e.g., improving clarity, organizing feedback) from unacceptable evaluative uses (e.g., judging novelty, contribution, or acceptance). Clearly defining these cases is required and necessary to create a safer and more practical application of AI in peer review.

These boundaries reflect an understanding of peer review as a moral and professional practice rather than a purely informational one. Evaluative judgment is tied to expertise, accountability, and the ability to justify decisions to affected parties. Delegating these judgments to AI systems is perceived as undermining the ethical foundation of peer review, even when AI outputs appear plausible or polished. Importantly, these boundaries are socially negotiated rather than fixed, highlighting the need for governance mechanisms that make such distinctions collectively enforceable.

Structural Strain as a Driver of AI Adoption

Our interview findings foreground structural strain as a critical but under-recognized driver of AI use in peer review. Reviewer scarcity, increasing submission volumes, and declining mentorship capacity create conditions in which AI tools become attractive coping mechanisms. However, social media discourse rarely centers these labor conditions, instead focusing on governance, enforcement, and legitimacy.

This divergence has important implications. Without addressing structural strain, governance efforts that rely solely on restriction or detection risk misdiagnosing the problem. Participants do not describe AI as causing the erosion of review quality; rather, AI amplifies existing weaknesses in an already fragile system. This suggests that sustainable governance must attend to workload distribution, reviewer development, and institutional incentives.

Uneven Risk and the Redistribution of Harm

We highlight how AI in peer review creates uneven risks: junior authors are more vulnerable to vague or incorrect AI-shaped feedback due to limited power to contest decisions; junior reviewers face ambiguity about acceptable AI use while being disproportionately pressured to accept review requests; and senior reviewers and chairs inherit new forms of invisible labor, including monitoring AI usage in reviewing. These findings complicate universal claims about fairness or efficiency. AI does not simply introduce new risks; it redistributes existing ones along lines of seniority, institutional authority, and visibility. Any governance framework that fails to account for these asymmetries risks reinforcing, rather than mitigating, inequities in academic evaluation.

Recommendations: Governing AI-Mediated Peer Review as a Sociotechnical System

Our findings indicate that GenAI is being introduced into peer review under structural strain (e.g., reviewer overload, rising submissions) while institutions rely on strategic ambiguity to avoid committing to enforceable rules. This produces predictable failure modes: (1) reviewers use AI to cope with time pressure; (2) chairs inherit new monitoring labor without reliable standards (and face difficulty because AI-generated reviews can appear ‘polite’ and plausible); (3) harms concentrate on those with least recourse (especially junior authors and junior reviewers); and (4) procedural trust becomes the dominant legitimacy mechanism, shifting peer review from an assumption of integrity to an ecology of verification.

For Reviewers: As primary producers of evaluative text, reviewers should *treat GenAI as assistive scaffolding without delegating evaluation*, using GenAI for supportive tasks only (e.g., improving clarity, grammar, tone; outlining review structure; summarizing their own notes) and not for passing judgments of quality or novelty of the work. If at all using it for any evaluative work, reviewers should consider *adopting verification norms for AI-assisted claims* about such content, e.g., if AI produces claims such as “the method fails to control X” or “the authors cite Y incorrectly,” reviewers should manually verify the accuracy of such claims, without taking them at face value. Reviewers should also *preserve confidentiality of manuscripts* and treat unpublished work as high-risk. They should avoid pasting or uploading unpublished manuscripts into third-party tools or popular chatbots, unless the venue explicitly approves that workflow (and the tool provides adequate contractual/privacy protections), and only look to institutionally-approved systems.

Generally, reviewers should use GenAI to *improve the communication of critique*, in light of our findings around participants valuing AI for politeness and brevity. This can be especially useful for tone calibration, ensuring reviews (especially negative ones) are measured. Above all, our findings reflect a concerted desire that *reviewers make labor visible* in the age of AI-generated content, demonstrating to authors that they have spent the appropriate time and mental effort in preparing structured, evidence-linked feedback (e.g.,

major/minor concerns, and actionable revisions).

For Area/Program Chairs: Area/Program Chairs should *replace “AI detection” with “review quality adjudication” procedures* based on review quality rather than provenance speculation, triggering escalation (e.g., request revision, solicit an additional review, downweight) of reviews that are generic, ungrounded, or inconsistent with manuscript content, irrespective of human or AI-production. This “*downweighting*” mechanism could also be institutionalized across venues, with ACs/PCs having mechanisms to downweight unreliable/low-quality reviews in final decision-making.

ACs/PCs should also *provide reviewer-facing micro-guidance at the moment of review submission*, such as (i) a concise permitted/prohibited list of AI usage, (ii) a reminder about confidentiality, and (iii) a requirement that major critiques cite manuscript sections. This directly targets the shallow/hallucinated review risk while reducing interpretive burden on chairs. Further, this could be part of a larger effort to support junior reviewers, while also *protecting them through “safe harbor” clarification channels* to address uncertainty about acceptable AI use among early-career reviewers. Venues could provide confidential guidance channels (e.g., “ask the chair” policy interpretations) for learning-oriented compliance, and emphasize that disclosure or clarification requests will not be penalized.

For Conferences and Journals: Our results suggest publication institutions should *move from strategic ambiguity to “adaptive specificity”* by defining a two-layered policy structure: a set of rigid principles (e.g., accountability, confidentiality, transparency, and proportional enforcement) and some operational guidance (e.g., permitted/prohibited uses, tool categories, disclosure requirements, and enforcement processes) that is adapted to each review cycle and emergent AI capabilities. Instead of banning AI usage completely in reviewing, policies should *codify a boundary between support and judgment* by explicitly distinguishing between permitted support and restricted usage to generate decisions.

Conference/journals should *require minimal disclosures of AI usage to avoid unequal punishment*, given concerns about unevenly distributed harms of AI disclosure, focusing on improving governance instead of creating stigma by designing policies that support transparency and auditability such as bounded disclosure, where reviewers select from a short checklist of AI-assisted activities (e.g., “grammar/tone,” “outline,” “none”) without requiring tool names or full prompts. Where provenance concerns arise, institutions should *introduce due-process protections and proportional remedies* that prioritize corrective actions over punishment. This can be done by requesting review revisions, adding extra reviewers, or downweighting unreliable reviews. Severe sanctions should require clear evidence and procedural safeguards to reduce reputational harm from false accusations, consistent with detection-culture concerns in social media discourse. At the same time, *adversarial risks should be treated as a part of review integrity, not edge cases*, and venues that allow AI tooling should incorporate basic integrity controls (e.g., PDF sanitization, detection of hidden text layers, and warnings about copying full manuscripts into LLMs). Even venues that prohibit AI use should ac-

Opportunities for GenAI in Peer Review	Risks of GenAI in Peer Review	Recommendations for GenAI Use
1. GenAI is viewed as useful for supportive and preparatory tasks such as improving clarity, grammar, tone, and structuring feedback, reducing reviewer cognitive load. 2. GenAI is seen as a tool that can lower participation barriers, particularly for non-native English speakers and reviewers with limited time. 3. AI assistance is framed as potentially helping sustain peer review in a context of reviewer overload and submission growth. 4. Procedural mechanisms (e.g., disclosure rules, oversight) are seen as necessary to maintain trust as AI use becomes more common. 5. Public discourse frames AI governance as a matter of procedural legitimacy rather than technical capability alone.	1.Epistemic Risk: AI-generated reviews may be shallow, generic, or hallucinated, undermining meaningful evaluation and scholarly rigor. 2.Ethical and Accountability Risk: Use of AI blurs responsibility for review content, making it unclear who is accountable for errors or harms. 3.Care and Quality Risk: Reliance on AI may normalize minimal-effort reviewing and erode peer review as a form of scholarly care and mentorship. 4.Unequal Harm Risk: Ambiguous rules and low-quality AI-assisted reviews disproportionately harm junior authors and early-career reviewers. 5.Adversarial Risk: Prompt injection and manipulation of AI-based reviewers introduce new attack surfaces in peer review.	1. Clearly distinguish supportive uses of AI from evaluative judgment, explicitly reserving decisions about novelty, contribution, and acceptance for humans. 2. Require meaningful human oversight and responsibility, where reviewers remain accountable for all feedback submitted under their name. 3. Reframe peer review as care work, emphasizing diligence, explanation, and constructive engagement rather than review length or polish. 4. Develop governance frameworks that explicitly consider uneven impacts on junior scholars and provide recourse for contesting flawed reviews. 5. Treat AI-mediated peer review as a potentially adversarial system, incorporating safeguards such as input sanitization, disclosure norms, and incident-response procedures.

Table 2: Synthesis of core findings across research questions, showing how opportunities and risks of GenAI in peer review motivate calls for clearer boundaries, accountability, and governance in AI-mediated evaluation.

knowledge adversarial possibilities, because reviewers may still use tools informally. Finally, keeping in mind that overload is a driver of AI adoption and AI restrictions might simply relocate strain and incentivize covert use, policies should *address structural strain as governance* through labor-facing reforms, including reviewer load caps, incentives/recognition for high-quality reviews, reviewer training programs, and expanded reviewer pools.

For AI Tool Designers: For designers of AI tools to be used in reviewing contexts, they should *focus on accountable assistance that makes choices visible*, in a way that clearly delineates AI contributions through tracked suggestions, provenance markers, and editable drafts that require reviewer confirmation. Such tools must contain *guardrails that operationalize the support/judgment boundary* through interaction patterns that discourage evaluative delegation (e.g., block generating accept/reject recommendations, require evidence citations for claims, and prompt users to add manuscript-grounded justification). They must also be robust enough to *anticipate adversarial behavior and prompt injection*, detecting hidden layers, unusual formatting, or suspicious instruction-like text in PDFs, and surface appropriate warnings. Finally, given widespread confidentiality concerns, designers should *build confidentiality-preserving workflows* and prioritize privacy-by-design: local/on-device processing when possible, contractual guarantees for non-retention/non-training, and clear notices about data handling.

Limitations

This study has limitations. First, all data are English language and largely reflect Western, primarily USA based academic contexts, which may limit applicability to other scholarly systems. Second, the social media dataset is not statistically representative; it is a keyword based, search retrieved convenience sample shaped by platform algorithms, and theme frequencies should not be interpreted as prevalence estimates. Third, findings are grounded mainly in AI and HCI conference contexts.

Conclusion

This study examined how GenAI is being used in academic peer review and how reviewers and chairs understand its role. Across interviews and social media discussions, we find broad agreement that GenAI may be useful for limited support tasks, but that key evaluative decisions such as judging novelty, contribution, and acceptance should remain human responsibilities. Participants also raised concerns about inaccurate feedback, unclear accountability, and new risks introduced by AI use. Our findings show that these concerns are shaped by existing pressures in peer review, including reviewer overload and unclear policies, which place greater burdens on individuals, especially early-career scholars. Together, these perspectives highlight that AI in peer review is not just a technical issue, but a governance challenge. We argue that responsible use of GenAI requires clear boundaries, meaningful human oversight, and attention to the labor and care involved in scholarly evaluation.

References

- Adams, W. C. 2015. Conducting semi-structured interviews. *Handbook of practical program evaluation*, 492–505.
- Al Sawi, I.; and Alaa, A. 2024. Navigating the impact: a study of editors' and proofreaders' perceptions of AI tools in editing and proofreading. *Discover Artificial Intelligence*, 4(1): 23.
- Alalag, A. S. 2024. AI tools assisting in the proofreading and scientific review of research papers. *ScienceOpen Preprints*.
- Ali, S. I.; and Shaban, M. 2025. Nursing Academic Reviewers' Perspectives on AI-Assisted Peer Review: Ethical Challenges and Acceptance. *International nursing review*, 72(3).
- Austin, T.; Rawal, B. S.; Diehl, A.; and Cosme, J. 2023. AI for equity: Unpacking potential human bias in decision making in higher education. *AI, Computer Science and Robotics Technology*, (13).
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and 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*, 610–623.
- Berdejo-Espinola, V.; and Amano, T. 2023. AI tools can improve equity in science. *Science*, 379(6636): 991–991.
- Berkenkotter, C. 1995. The power and the perils of peer review. *Rhetoric review*, 13(2): 245–248.
- Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 173–184.
- Biswas, S.; Dobaria, D.; and Cohen, H. L. 2023. ChatGPT and the future of journal reviews: a feasibility study. *The Yale Journal of Biology and Medicine*, 96(3): 415.
- Blandford, A.; Furniss, D.; and Makri, S. 2016. *Qualitative HCI research: Going behind the scenes*. Morgan & Claypool Publishers.
- Bornmann, L.; Weymuth, C.; and Daniel, H.-D. 2010. A content analysis of referees' comments: how do comments on manuscripts rejected by a high-impact journal and later published in either a low-or high-impact journal differ? *Scientometrics*, 83(2): 493–506.
- Borsuk, R. M.; Aarssen, L. W.; Budden, A. E.; Koricheva, J.; Leimu, R.; Tregenza, T.; and Lortie, C. J. 2009. To name or not to name: The effect of changing author gender on peer review. *BioScience*, 59(11): 985–989.
- Bougie, N.; and Watanabe, N. 2024. Generative adversarial reviews: When LLMs become the critic. *arXiv preprint arXiv:2412.10415*.
- Burnham, J. C. 1990. The evolution of editorial peer review. *Journal of the American Medical Association*, 263(10): 1323–1329.
- Chawla, D. S. 2024. Is ChatGPT corrupting peer review? Telltale words hint at AI use. *Nature*, 628(8008): 483–484.
- Checco, A.; Bracciale, L.; Loreti, P.; Pinfield, S.; and Bianchi, G. 2021. AI-assisted peer review. *Humanities and social sciences communications*, 8(1): 1–11.
- Chen, S.; Brumby, D.; and Cox, A. 2025. Envisioning the Future of Peer Review: Investigating LLM-Assisted Reviewing Using ChatGPT as a Case Study. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, 1–18.
- Chidambaram, V. 2025. The role of LLMs in academic reviewing. *The ACM Special Interest Group in Operating Systems*.
- Chitale, M. P.; Shetye, K. M.; Gupta, H.; Chaudhary, M.; and Varma, V. 2025. AutoRev: Automatic Peer Review System for Academic Research Papers. *arXiv preprint arXiv:2505.14376*.
- Chung, K. C.; Shauver, M. J.; Malay, S.; Zhong, L.; Weinstein, A.; and Rohrich, R. J. 2015. Is double-blinded peer review necessary? The effect of blinding on review quality. *Plastic and reconstructive surgery*, 136(6): 1369–1377.
- Conroy, G. 2023. How ChatGPT and other AI tools could disrupt scientific publishing. *Nature*, 622(7982): 234–236.
- Demir, S. B.; and Pismek, N. 2018. A Convergent Parallel Mixed-Methods Study of Controversial Issues in Social Studies Classes: A Clash of Ideologies. *Educational Sciences: Theory and Practice*, 18(1): 119–149.
- Ding, X.; Kou, Y.; Xu, Y.; and Zhang, P. 2022. "As Uploaders, We Have the Responsibility": Individualized Professionalization of Bilibili Uploaders. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Donker, T. 2023. The dangers of using large language models for peer review. *The Lancet Infectious Diseases*, 23(7): 781.
- Doskaliuk, B.; Zimba, O.; Yessirkepov, M.; Klishch, I.; and Yatsyshyn, R. 2025. Artificial intelligence in peer review: enhancing efficiency while preserving integrity. *Journal of Korean medical science*, 40(7).
- Drori, I.; and Te'eni, D. 2024. Human-in-the-loop AI reviewing: feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1): 98–109.
- Drozd, J. A.; and Ladomery, M. R. 2024. The peer review process: past, present, and future. *British Journal of Biomedical Science*, 81: 12054.
- Du, J.; Wang, Y.; Zhao, W.; Deng, Z.; Liu, S.; Lou, R.; Zou, H. P.; Venkit, P. N.; Zhang, N.; Srinath, M.; et al. 2024. LLMs assist NLP researchers: Critique paper (meta-) reviewing. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, 5081–5099.
- Ebadi, S.; Nejadghanbar, H.; Salman, A. R.; and Khosravi, H. 2025. Exploring the impact of generative AI on peer review: Insights from journal reviewers. *Journal of Academic Ethics*, 1–15.
- Feetham-Walker, L.; Brent-Jones, E.; Chorlton, E.; Dixon, M.; Coombs, A.; Powell, S.; Holst, F.; and Giles, K. 2025. AI and Peer Review 2025. Technical report, IOP Publishing.
- Gao, X.; Ruan, J.; Zhang, Z.; Gao, J.; Liu, T.; and Fu, Y. 2025. ReviewAgents: Bridging the gap between human and AI-generated paper reviews. *arXiv preprint arXiv:2503.08506*.

- Ghosh, S. 2024. ChatGPT as a tool for equitable education in engineering classes. In *2024 ASEE Annual Conference & Exposition*.
- Goldberg, A.; Ullah, I.; Khuong, T. G. H.; Rachmat, B. K.; Xu, Z.; Guyon, I.; and Shah, N. B. 2024. Usefulness of LLMs as an Author Checklist Assistant for Scientific Papers: NeurIPS'24 Experiment. *arXiv preprint arXiv:2411.03417*.
- Gruda, D. 2025. Three AI-powered steps to faster, smarter peer review. *Nature*, 10.
- Haffar, S.; Bazerbachi, F.; and Murad, M. H. 2019. Peer review bias: a critical review. In *Mayo Clinic Proceedings*, volume 94, 670–676. Elsevier.
- Helmer, M.; Schottdorf, M.; Neef, A.; and Battaglia, D. 2017. Gender bias in scholarly peer review. *elife*, 6: e21718.
- Idahl, M.; and Ahmadi, Z. 2025. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, 550–562.
- Kerrigan, M. R. 2014. A framework for understanding community colleges' organizational capacity for data use: A convergent parallel mixed methods study. *Journal of mixed methods research*, 8(4): 341–362.
- Kohnke, S.; and Zaugg, T. 2025. Artificial intelligence: an untapped opportunity for equity and access in STEM education. *Education Sciences*, 15(1): 68.
- Kronick, D. A. 1990. Peer review in 18th-century scientific journalism. *Journal of the American Medical Association*, 263(10): 1321–1322.
- Laufer, B.; Jain, S.; Cooper, A. F.; Kleinberg, J.; and Heidari, H. 2022. Four years of FAccT: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 401–426.
- Lee, C. J.; Sugimoto, C. R.; Zhang, G.; and Cronin, B. 2013. Bias in peer review. *Journal of the American Society for information Science and Technology*, 64(1): 2–17.
- Liang, W.; Izzo, Z.; Zhang, Y.; Lepp, H.; Cao, H.; Zhao, X.; Chen, L.; Ye, H.; Liu, S.; Huang, Z.; et al. 2024. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *arXiv preprint arXiv:2403.07183*.
- Lipworth, W.; and Kerridge, I. 2011. Shifting power relations and the ethics of journal peer review. *Social epistemology*, 25(1): 97–121.
- Liu, R.; and Shah, N. B. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.
- Marsh, H. W.; Jayasinghe, U. W.; and Bond, N. W. 2008. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *American psychologist*, 63(3): 160.
- Mogashoa, T. 2014. Understanding critical discourse analysis in qualitative research. *International Journal of Humanities Social Sciences and Education*, 1(7): 104–113.
- Neshaei, S. P.; Rietsche, R.; Su, X.; and Wambsganss, T. 2024. Enhancing peer review with AI-powered suggestion generation assistance: Investigating the design dynamics. In *Proceedings of the 29th international conference on intelligent user interfaces*, 88–102.
- Nixon, N.; Lin, Y.; and Snow, L. 2024. Catalyzing equity in STEM teams: Harnessing generative AI for inclusion and diversity. *Policy Insights from the Behavioral and Brain Sciences*, 11(1): 85–92.
- Palmer, G. 1908. Editorial Individuality. In *American Medical Editors' Association*, 57–63.
- Pontille, D.; and Torny, D. 2014. The blind shall see! The question of anonymity in journal peer review. *Ada: A Journal of Gender, New Media, and Technology*, (4).
- Rastogi, C.; Song, X.; Jin, Z.; Stelmakh, I.; Daumé III, H.; Zhang, K.; and Shah, N. B. 2024. A randomized controlled trial on anonymizing reviewers to each other in peer review discussions. *PloS one*, 19(12): e0315674.
- Robertson, Z. 2023. GPT4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492*.
- Roshanaei, M.; Olivares, H.; and Lopez, R. R. 2023. Harnessing AI to foster equity in education: Opportunities, challenges, and emerging strategies. *Journal of Intelligent Learning Systems and Applications*, 15(4): 123–143.
- Russo, G.; Horta Ribeiro, M.; Davidson, T. R.; Veselovsky, V.; and West, R. 2025. The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–28.
- Saad, A.; Jenko, N.; Ariyaratne, S.; Birch, N.; Iyengar, K. P.; Davies, A. M.; Vaishya, R.; and Botchu, R. 2024. Exploring the potential of ChatGPT in the peer review process: an observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 18(2): 102946.
- Sebastian, F.; and Baron, R. 2024. Artificial intelligence: What the future holds for multilingual authors and editing professionals. *Science Editor*, 47(2): 38–42.
- Shiflett, L. 1988. A difficult balance: Editorial peer review in medicine.
- Stelmakh, I.; Shah, N. B.; Singh, A.; and Daumé III, H. 2021. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–17.
- Su, X.; Wambsganss, T.; Rietsche, R.; Neshaei, S. P.; and Käser, T. 2023. Reviewer: AI-generated instructions for peer review writing. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, 57–71.
- Sukpanichnant, P.; Rapberger, A.; and Toni, F. 2024. Peer-Arg: Argumentative peer review with LLMs. *arXiv preprint arXiv:2409.16813*.

- Suleiman, A.; von Wedel, D.; Munoz-Acuna, R.; Redaelli, S.; Santarisi, A.; Seibold, E.-L.; Ratajczak, N.; Kato, S.; Said, N.; Sundar, E.; et al. 2024. Assessing ChatGPT's ability to emulate human reviewers in scientific research: A descriptive and qualitative approach. *Computer Methods and Programs in Biomedicine*, 254: 108313.
- Sun, Z. 2025. Large language models in peer review: challenges and opportunities. *Scientometrics*, 1–44.
- Tachoyotin, P.; and Acuna, D. 2025. REMOR: Automated Peer Review Generation with LLM Reasoning and Multi-Objective Reinforcement Learning. *arXiv preprint arXiv:2505.11718*.
- Tan, C.; Lyu, D.; Li, S.; Gao, Z.; Wei, J.; Ma, S.; Liu, Z.; and Li, S. Z. 2024. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*.
- Thakkar, D.; Ismail, A.; Kumar, P.; Hanna, A.; Sambasivan, N.; and Kumar, N. 2022. When is machine learning data good?: Valuing in public health datafication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Thakkar, N.; Yuksekgonul, M.; Silberg, J.; Garg, A.; Peng, N.; Sha, F.; Yu, R.; Vondrick, C.; and Zou, J. 2025. Can LLM feedback enhance review quality? A randomized study of 20k reviews at ICLR 2025. *arXiv preprint arXiv:2504.09737*.
- Thelwall, M.; and Yaghi, A. 2024. In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. *arXiv preprint arXiv:2409.16695*.
- Tregenza, T. 2002. Gender bias in the refereeing process? *Trends in Ecology & Evolution*, 17(8): 349–350.
- Varanasi, R. A.; Siddarth, D.; Seshadri, V.; Bali, K.; and Vashistha, A. 2022. Feeling Proud, Feeling Embarrassed: Experiences of Low-income Women with Crowd Work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–18.
- Wang, Q.; Zeng, Q.; Huang, L.; Knight, K.; Ji, H.; and Rajani, N. F. 2020. ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, 384–397. Dublin, Ireland: Association for Computational Linguistics.
- Weng, Y.; Zhu, M.; Bao, G.; Zhang, H.; Wang, J.; Zhang, Y.; and Yang, L. 2025. CycleResearcher: Improving Automated Research via Automated Review. In *The Thirteenth International Conference on Learning Representations*.
- Whittaker, R. J. 2008. Journal review and gender equality: a critical comment on Budden et al. *Trends in ecology & evolution*, 23(9): 478–479.
- Young, M.; Katell, M.; and Krafft, P. 2022. Confronting power and corporate capture at the FAccT Conference. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1375–1386.
- Yu, S.; Luo, M.; Madasu, A.; Lal, V.; and Howard, P. 2024. Is your paper being reviewed by an LLM? Investigating AI text detectability in peer review. *arXiv preprint arXiv:2410.03019*.
- Zajda, J. 2020. Discourse analysis as a qualitative methodology. *Educational Practice and Theory*, 42(2): 5–21.
- Zeng, S.; Tian, K.; Zhang, K.; Wang, Y.; Gao, J.; Liu, R.; Yang, S.; Li, J.; Long, X.; Ma, J.; et al. 2025. ReviewRL: Towards Automated Scientific Review with RL. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 16942–16954.
- Zhou, H.; Huang, X.; Pu, H.; and Qi, Z. 2024. May generative AI be a reviewer on an academic paper. In *EEKE-AII*.
- Zhou, R.; Chen, L.; and Yu, K. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, 9340–9351.
- Zhuang, Z.; Chen, J.; Xu, H.; Jiang, Y.; and Lin, J. 2025. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124(C).