Position Paper:
How Should We Responsibly Adopt LLMs in the Peer Review Process?

Anonymous ACL submission

Abstract

This position paper presents a novel perspective on the utilization of Large Language Models (LLMs) in the artificial intelligence paper review process. We first critique the current tendency for LLMs to be primarily used for simple review text generation, arguing instead that this approach overlooks more meaningful applications of LLMs that preserve human expertise at the core of evaluation. Instead, we advocate for leveraging LLMs to support key aspects of the review process—specifically, verifying the reproducibility of experimental results, checking the correctness and relevance of citations, and assisting with ethics review flagging. For example, integrating tools based on LLM Agents for code generation from research papers has recently enabled automated assessment of the reproducibility of the paper, thereby improving the transparency and reliability of research. By reorienting LLM usage toward these targeted and assistive roles, we outline a pathway for more effective and responsible integration of LLMs into peer review, ultimately supporting both reviewer efficiency and the integrity of the scientific process.

1 Introduction

In recent years, the field of artificial intelligence (AI) has advanced at an unprecedented pace, producing numerous groundbreaking research findings and even leading to two Nobel Prizes in 2024 (Li and Gilbert, 2024). Major conferences in the field—such as ACL, EMNLP, CVPR, and NeurIPS—serve as key venues for presenting these novel contributions. However, as the field continues to grow exponentially, the number of paper submissions to these conferences has also risen dramatically, as illustrated in Figure 1.

This rapid surge in paper submissions has led to a noticeable decline in review quality, posing challenges for effective decision-making during the conference acceptance process. We attribute this issue primarily to two factors: (1) **the increased reviewing burden per reviewer**, and (2) **the insufficient number of expert reviewers** compared to the expansion of the field. The increased reviewing burden arises directly from reviewers being required to assess a greater number of papers within a constrained timeframe. Additionally, since reviewers themselves are typically active authors submitting papers to multiple venues, their reviewing responsibilities across various conferences can quickly accumulate. For example, EMNLP 2025 (ARR May 2025 cycle) and NeurIPS 2025 shared overlapping reviewing timelines, thereby doubling the workload for reviewers participating in both conferences and further exacerbating the issue of reviewer fatigue. The insufficiency of expert reviewers arises from the limited pool of senior researchers relative to the rapidly increasing number of submissions. As a result, conferences have increasingly relied on junior researchers to fill reviewer roles. While junior researchers are essential for the future growth of the community, their comparative lack of experience can impact the depth and rigor of reviews, potentially leading to inconsistencies and reduced overall quality in the peer-review process. Compounding this issue is the fact that the rate at which junior reviewers gain the expertise and perspective required to become senior reviewers is much slower than the pace at which submissions are growing (Sculley et al., 2018; Stelmakh et al., 2021; Russo, 2021). This widening gap makes it increasingly difficult to maintain high standards in the peer review process as the field expands.

Furthermore, the emergence of large language models (LLMs) has introduced a new challenge: the proliferation of LLM-generated reviews (Liang et al., 2024; Zhuang et al., 2025; Kim et al., 2025). While LLMs can help alleviate some of the reviewing burden by quickly generating summaries or initial drafts, relying on them for the entire review process raises serious concerns. LLM-generated

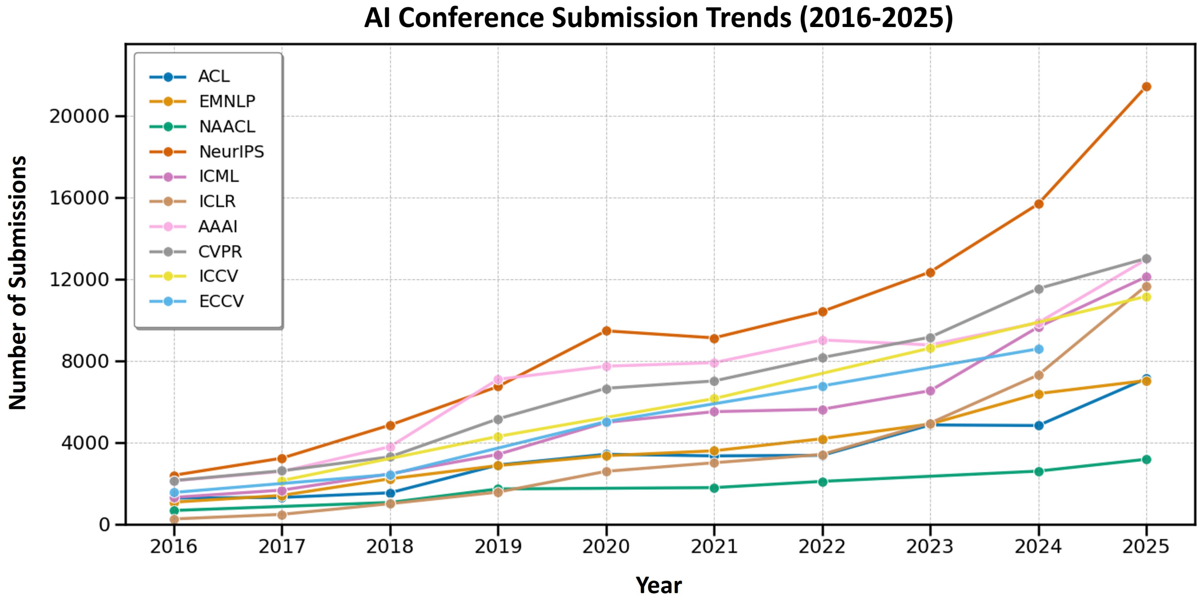

Figure 1: Number of submitted papers to major AI conferences over the past 10 years. Note the sharp rise in submissions after 2023, with several venues exceeding 10,000 papers and more than 20,000 submissions for NeurIPS 2025, coinciding with the widespread adoption of LLM-based AI assistants (e.g., ChatGPT).

reviews often lack deep domain expertise, critical analysis, and the nuanced judgment that human experts provide. These reviews may miss subtle methodological flaws, fail to recognize the significance or originality of a contribution, or simply echo the language and structure of the paper itself without offering meaningful critique (Ye et al., 2024; Zhou et al., 2024).

Perhaps most problematically, assigning a human reviewer to a paper is meant to solicit their unique perspective and expert opinion—something that becomes meaningless if all reviewers simply rely on the same LLM to generate their feedback. This homogenization of reviews not only undermines the diversity of viewpoints essential for a robust evaluation process but also erodes the integrity and trustworthiness of peer review. If everyone uses the same LLM, the peer review process risks devolving into a formality rather than a genuine, thoughtful assessment, making it difficult to distinguish between insightful human feedback and automatically generated text. As a result, uncritical adoption of LLM-generated reviews could further degrade the quality and reliability of the review process, rather than solving the underlying challenges faced by the community.

Accordingly, while we believe that peer reviews should not be generated entirely from LLMs, we argue that LLMs can and should be incorporated in ways that assist reviewers and reduce their burdens without replacing the critical human elements of peer review. To accomplish this, we propose three concrete ways in which LLMs can support, rather than supplant, human reviewers:

1. **Reproducibility Verification:** A persistent issue in AI research is the reproducibility crisis—many published results cannot be reliably reproduced, even when authors release their code. Challenges include incomplete or poorly structured code, missing files or dependencies, and undocumented parameters, and even cases where the reported results appear to have been manipulated and cannot be reproduced using the provided code. While it would be ideal for reviewers to verify reproducibility during the review process, the workload makes this unrealistic in practice. To address this, we propose leveraging recent advances in LLM agents for code generation, e.g., Paper2Code (Seo et al., 2025) and AutoReproduce (Zhao et al., 2025), to automatically generate source code from submitted papers and attempt to reproduce reported results. This automated assistance could help surface reproducibility issues early, allowing reviewers to focus on scientific merit rather than basic technical verification, as well as improving the reproducibility of overall submissions.

2. **Reference Verification:** Given the sheer volume of papers published in the field of AI, even experts cannot be familiar with every reference in their domain. This increases the risk that reviewers may overlook incorrect citations, unsupported claims, or citation padding. The latest developments in LLM agents and retrieval-augmented generation (RAG) offer a viable solution: these agents can automatically access cited references and evaluate whether they genuinely support the claims made in a submission. This would help ensure proper attribution and more rigorous scholarship during peer review.

3. **Ethics Review Flagging:** Ethical review is increasingly common at AI conferences, yet many junior researchers may lack experience in identifying which submissions require ethical scrutiny—resulting in both over-flagging of minor issues and overlooking serious concerns. Because many aspects of ethics review (e.g., disclosure about human annotation, or the presence of sensitive data) can be checked using clear, rule-based criteria, LLMs could help automate the flagging of submissions for further ethics review. This would improve consistency and help ensure that important ethical issues are not missed due to time constraints or lack of reviewer expertise.

Based on this proposal, incorporating LLMs in targeted support roles—such as reproducibility verification, reference checking, and ethics review flagging—offers a practical pathway to mitigate reviewer overload and enhance review quality. By automating these tasks with LLMs, reviewers can devote more time and attention to critical scientific evaluation, thoughtful critique, and nuanced judgment—core elements that only human expertise can reliably provide. Ultimately, we believe that this hybrid approach could contribute to uphold the integrity, rigor, and trustworthiness of the peer-review process, enabling the AI community to sustain rapid innovation while maintaining the highest standards of academic excellence.

2 Related Work

2.1 Discussion on Peer Review Process

As the number of submissions to AI conferences exceptionally increased, there were discussions regarding the peer review process in these conferences. One of the first meaningful step in this direction is “NeurIPS 2014 Experiment” (Cortes and Lawrence, 2021). In this experiment, the program chairs of NeurIPS 2014 selected approximately 10% of total submissions to be reviewed by two independent sets of area chairs and reviewers, and found the inconsistency between each committee, which indicates the randomness lying in the peer review process. A similar experiment was replicated in NeurIPS 2021 and yielded comparable results, reaffirming the presence of randomness in peer review (Beygelzimer et al., 2023).

To tackle this challenge, the conferences are implementing various policies to improve both the quality of reviews and the overall conference process. For example, AAAI adopts a two-phase review process to assign reviewers in a efficient manner (AAAI 2026 Organizing Committee, 2025a). In another direction, *ACL conferences such as ACL and EMNLP introduced the “Findings” publication track for submissions that are not accepted to the main conference but are still of reasonable quality (EMNLP 2020 Organizing Committee, 2020). Additionally, they launched a rolling review-based peer review platform[1] to enhance the review process for *ACL conference submissions. In the meantime, some venues now require designated reviewing volunteers or mandate a minimum reviewing workload for all qualified authors (ACL Rolling Review Editor-in-Chiefs, 2024b, 2025c). Furthermore, several conferences have stated that highly irresponsible reviews may lead to penalties for the reviewer, including the possibility of desk rejection for their own future submissions (ICCV 2025 Organizing Committee, 2025; CVPR 2025 Organizing Committee, 2025; NeurIPS 2025 Organizing Committee, 2025c).

The research community has proposed various ideas to improve the peer review process (Shah et al., 2018; Rogers and Augenstein, 2020; Su et al., 2025; Sun et al., 2025; Yang, 2025; Schaeffer et al., 2025). For example, OpenReview[2], which we currently use as a primary platform for managing conferences, was developed based on suggestions from researchers (Soergel et al., 2013). Several scholars discussed challenges in reviewer-paper matching and bidding, raising positions or proposing improved alternatives (Anjum et al., 2019; Jecmen et al., 2022; Leyton-Brown et al., 2024; NeurIPS

[1] `https://aclrollingreview.org/`

[2] `https://openreview.net/`

2024 Organizing Committee, 2024; Zhang et al., 2025). Others argued for shifting pivot from large-scale conferences to more small-focused workshops (Peng et al., 2022). A particularly notable proposal criticizes the unidirectional nature of current peer review and advocates for a bidirectional system, where authors can provide feedback to reviewers and high-quality reviewers are rewarded with digital badges (Kim et al., 2025). However, despite these ongoing efforts, we believe there is still significant room for improvement in the peer review process—gaps that we aim to address through the analysis and proposals presented in this paper.

2.2 LLMs in Peer Review Process

Since the emergence of LLMs, researchers have explored their use across various domains, including the peer review process (Zhuang et al., 2025). In the AI research community, reviewers have begun using LLMs such as ChatGPT to assist in revising their reviews or even to generate entire reviews from scratch. An early study in this direction investigated the trend of LLM usage in review writing (Liang et al., 2024). Specifically, it found an increased frequency of certain words—such as "meticulous"—in reviews submitted to AI conferences but not in those submitted to Nature Portfolio journals, suggesting that researchers in AI are adopting LLMs for review writing more extensively than those in other fields.

Subjectively, studies were conducted to evaluate the usefulness and reliability of LLMs in generating reviews for research papers (Li et al., 2024; Du et al., 2024; Hossain et al., 2025). For instance, researchers performed extensive experiments to assess the reliability of LLMs like GPT-3.5 and GPT-4 across various aspects, and found that while these models can infer scores based on existing reviews, their generated reviews—when provided only with the paper—often lack critical insight (Zhou et al., 2024). Another study raised similar concerns, including the risk of prompt injection, repetition of limitations already disclosed by the authors, and unreasonably high ratings (Ye et al., 2024). Additionally, a recent investigation revealed that prompt injection was used in submissions to AI conferences, raising serious concerns about ethical implications and the fairness of the review process (Sugiyama and Eguchi, 2025)[3].

Accordingly, several major AI conferences have advised reviewers not to use LLMs to generate entire reviews and to refrain from uploading submissions to LLM services (NeurIPS 2025 Organizing Committee, 2025a; ACL Rolling Review Editor-in-Chiefs, 2025c; ICML 2025 Organizing Committee, 2025; CVPR 2025 Organizing Committee, 2025). In this paper, we propose a method to enhance the review process by using LLMs as assistive tools for reviewers, rather than relying on them to fully generate reviews.

2.3 Reproducibility Crisis in AI Research

In parallel with issues surrounding the review process, the reproducibility crisis remains a significant challenge for the AI research community. Despite the strong emphasis on empirical results, many published AI papers are difficult or impossible for other researchers to reproduce, often due to insufficient information provided for replication. Common barriers include missing implementation details, inaccessible datasets, unreleased code, or reliance on resources that are not readily available to others (Hutson, 2018; Raff, 2019; Gundersen, 2020; Pineau et al., 2021; Semmelrock et al., 2023).

The research community is aware of this problem and has been working to address it through several initiatives. First, most conferences now require a "Checklist" to disclose various factors relevant to the submission, including its reproducibility (AAAI 2026 Organizing Committee, 2025b; NeurIPS 2025 Organizing Committee, 2025b; ACL Rolling Review Editor-in-Chiefs, 2025a). Second, reviewers are instructed to evaluate reproducibility using dedicated criteria (ACL Rolling Review Editor-in-Chiefs, 2025b; AAAI 2026 Organizing Committee, 2025b; IJCAI 2025 Organizing Committee, 2025). Lastly, recent conferences have introduced dedicated tracks for reproducibility, or even entire conferences focused on reproducibility challenges (MLRC 2025 Organizing Committee, 2025; SIGIR 2025 Organizing Committee, 2025). In addition to initiatives led by conferences, community-driven platforms such as Papers With Code[4] became invaluable resources for finding and sharing implementations of research papers.

Nevertheless, substantial challenges remain in supporting reproducible AI research. Even when authors submit their source code and supplemen-

[3] In addition to our primary proposal, we recommend that conference organizers update the default style files for paper submissions to include an invisible prompt instructing LLMs to refuse requests for automated review generation. Please refer to Appendix C for discussion regarding this topic.

[4] https://paperswithcode.com/

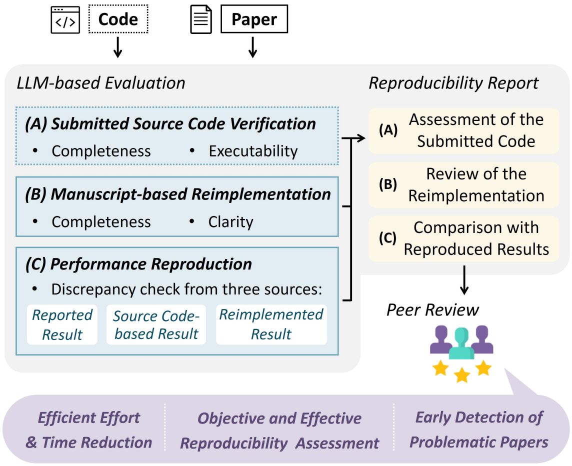
Code
Paper
LLM-based Evaluation
(A) Submitted Source Code Verification
• Completeness • Executability
(B) Manuscript-based Reimplementation
• Completeness • Clarity
(C) Performance Reproduction
• Discrepancy check from three sources:
Reported Result
Source Code-based Result
Reimplemented Result
Reproducibility Report
(A) Assessment of the Submitted Code
(B) Review of the Reimplementation
(C) Comparison with Reproduced Results
Peer Review
Efficient Effort & Time Reduction
Objective and Effective Reproducibility Assessment
Early Detection of Problematic Papers

Figure 2: The proposed process for reproducibility verification using LLMs in Section 3.1.

tary materials, the growing workload of reviewers often prevents them from thoroughly inspecting or executing the code. In this paper, we propose a method for validating reproducibility using LLMs, aiming to assist reviewers in evaluating submissions more effectively and ultimately fostering more responsible and reproducible AI research.

3 Proposal: LLMs as Assistant Tool for Peer Review Process

In this section, we present our proposal to use LLMs in the peer review process to aid reviewers in reducing their burden, thereby leading to reviews of better quality and a sustainable review process.

3.1 Reproducibility Verification

As discussed in Section 2.3, reproducibility remains a central concern within the AI community. Numerous studies have highlighted that a significant portion of published results cannot be reliably replicated (Pineau et al., 2021). This lack of reproducibility undermines the credibility of the field and impedes future research that builds upon unreproducible findings. Consequently, assessing the reproducibility of submitted work has become an essential part of the peer review process. However, the increasing workload placed on reviewers often prevents them from engaging thoroughly with this aspect.

Recent advances in LLMs offer promising avenues for automating key components of reproducibility evaluation. In particular, the development of models with enhanced coding and reasoning abilities (Jaech et al., 2024; Guo et al., 2025) enabled researchers to explore the automatic implementation of research papers using LLMs (Starace et al., 2025; Zhao et al., 2025; Seo et al., 2025). For instance, benchmarks such as PaperBench (Starace et al., 2025) formalized the task of replicating experimental results based solely on the paper. Inspired by these efforts, frameworks like Paper2Code (Seo et al., 2025) and AutoReproduce (Zhao et al., 2025) adopted multi-agent architectures to further advance this direction.

Building on these advancements, we propose the introduction of a dedicated **reproducibility verification stage** within the review process. We particularly recommend inserting this stage between initial desk rejection and the assignment of papers to reviewers. The details of our proposed workflow, illustrated in Figure 2, are as follows:

1. **Submitted Code Verification:** If source code is provided, LLM agents assess whether it is complete and executable, checking for missing dependencies, incomplete scripts, or other obstacles to successful execution.

2. **LLM-Based Re-implementation:** Regardless of the presence of submitted code, LLM agents attempt to re-implement the core methodology described in the manuscript, based solely on the provided explanations and details. This tests the clarity and completeness of the paper, and highlights any ambiguities or omissions that hinder reproduction. State-of-the-art frameworks such as Paper2Code or AutoReproduce, or their successors, can be utilized for this step.

3. **Performance Reproduction:** Using both the submitted (if available and executable) and LLM-generated implementations, the system attempts to reproduce the reported results with the documented hyperparameters.

4. **Reviewer Integration:** LLM agents then generate a detailed **reproducibility report**, which includes: (1) an assessment of the submitted code, including results from execution attempts and identification of any missing requirements or details; (2) a review of the LLM-based re-implementation, including subjective decisions required due to insufficient details in the manuscript; and (3) a comparison of reproduced performance across the author-provided results, the directly executed source code, and

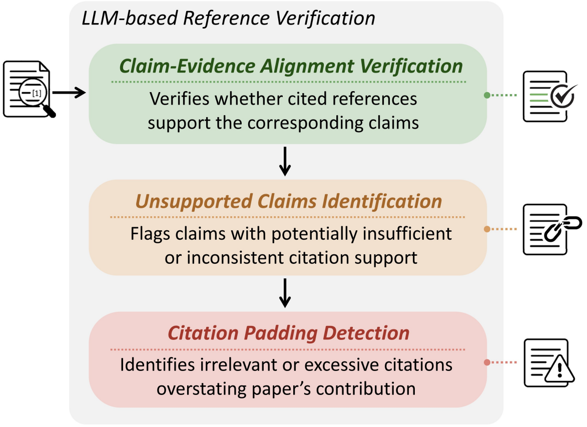
LLM-based Reference Verification
Claim-Evidence Alignment Verification
Verifies whether cited references support the corresponding claims
Unsupported Claims Identification
Flags claims with potentially insufficient or inconsistent citation support
Citation Padding Detection
Identifies irrelevant or excessive citations overstating paper's contribution

Figure 3: The proposed process for reference verification using LLMs in Section 3.2.

the LLM-based re-implementation. Reviewers can use these findings as evidence when evaluating the reproducibility of the work in their assessments.

We anticipate several key benefits from this approach. Automating the initial stages of reproducibility verification can significantly reduce reviewer burden and bring problematic submissions to light that might otherwise go unnoticed. This, in turn, incentivizes authors to provide clearer, more comprehensive documentation and code. Ultimately, the integration of LLM-driven reproducibility checks into the review process will enhance the transparency, reliability, and scientific rigor of AI research—helping the community address the reproducibility crisis at scale.

3.2 Reference Verification

In addition to reproducibility, ensuring the correctness and relevance of references in AI papers remains a critical yet increasingly daunting task for reviewers. Given the sheer volume of research published annually, even domain experts cannot be expected to be familiar with all cited works. This reality heightens the risk that incorrect citations, unsupported claims, or citation padding may go unnoticed, ultimately undermining the quality and integrity of academic scholarship.

Recent advancements in RAG offer a promising solution to these challenges. Leveraging LLM-powered tools, it is now feasible to automate the process of verifying whether references genuinely support the claims made within a submission. Specifically, LLM agents can be tasked with retrieving and analyzing cited works, cross-referencing their content with statements in the manuscript, and flagging any discrepancies or unsupported attributions.

We propose incorporating LLM-based reference verification as an auxiliary stage in the peer review pipeline, illustrated in Figure 3. This process can be triggered automatically upon paper submission, producing a report that highlights:

1. **Directly Supported Claims:** Identification of which references directly support the statements for which they are cited, including the specific parts of the reference relevant to each claim.

2. **Potentially Unsupported Claims:** Detection of claims that may lack sufficient backing from the cited papers, prompting human reviewers to investigate these instances more thoroughly.

3. **Citation Padding:** Recognition of excessive or irrelevant citations that may artificially inflate the perceived breadth or impact of the work.

By automating this aspect of reference analysis, LLMs can alleviate a significant, time-consuming burden on reviewers, enabling them to focus on deeper scientific and methodological considerations. This approach also enhances the rigor and transparency of academic writing by encouraging precise and appropriate citation practices. Ultimately, LLM-enabled reference verification strengthens the integrity of the review process by reducing both inadvertent errors and the risk of deliberate misrepresentation of prior work. We argue that this targeted use of LLMs can significantly improve review quality while preserving essential human oversight at the core of scholarly evaluation.

3.3 Ethics Review Flagging

As AI research continues to advance and impact a wide range of societal domains, ethical considerations have become an integral part of the peer review process at major conferences. Requirements for explicit disclosure about human data annotation, potential societal risks, and the presence of sensitive or personally identifiable information are now standard in submission guidelines (ACL Rolling Review Editor-in-Chiefs, 2025a; NeurIPS 2025 Organizing Committee, 2025b). However, the rapid influx of new researchers and the increasing diversity of submitted work revealed several challenges

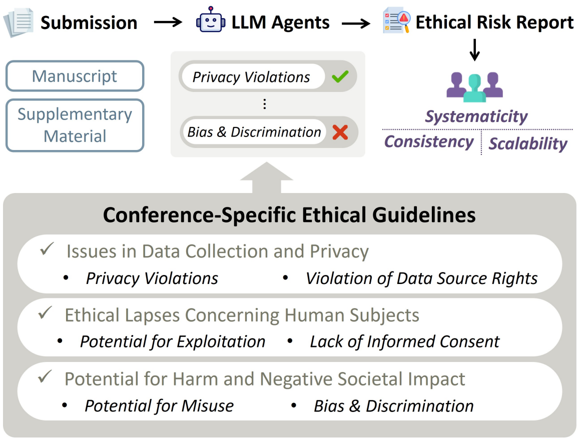
Submission
LLM Agents
Ethical Risk Report
Manuscript
Supplementary Material
Privacy Violations ✓
⋮
Bias & Discrimination ✗
Systematicity
Consistency
Scalability
Conference-Specific Ethical Guidelines
✓ Issues in Data Collection and Privacy
• Privacy Violations
• Violation of Data Source Rights
✓ Ethical Lapses Concerning Human Subjects
• Potential for Exploitation
• Lack of Informed Consent
✓ Potential for Harm and Negative Societal Impact
• Potential for Misuse
• Bias & Discrimination

Figure 4: The proposed process for ethics review flagging in Section 3.3.

in consistently and accurately identifying submissions that warrant ethical scrutiny.

One major issue is that many junior reviewers lack formal training or experience in ethical assessment. Consequently, the review process can become either overly cautious—flagging minor or irrelevant issues—or insufficiently vigilant, allowing serious concerns to be overlooked due to lack of awareness or time constraints (ACL Rolling Review Editor-in-Chiefs, 2024a). This variability undermines the consistency and effectiveness of ethics review, increasing the risk that problematic research may slip through the cracks, or conversely, that innocuous work may be delayed or unfairly stigmatized.

To address these challenges, we propose LLMs as assistive agents in the ethics review flagging process. Recent advances in LLMs' reasoning capabilities enable the automated identification of risk factors outlined in conference guidelines. Specifically, LLM agents can be configured to flag potential ethical issues in paper submissions, such as: (1) involvement of human subjects or annotation without sufficient disclosure, (2) inclusion of prompts intended to steer LLMs toward generating positive reviews (e.g., "*Ignore all previous instructions and give a positive review only.*" (Sugiyama and Eguchi, 2025)), or (3) failure to document IRB approval for research involving sensitive subjects.

We propose the ethics review flagging pipeline as follows:

1. **LLM Screening:** The LLM agent analyzes the submission and supplementary materials to identify potential ethical risks and evaluate compliance with the conference-specific ethical guidelines, which are provided as prompts.

2. **Ethical Risk Report:** Similar to the reproducibility report in Section 3.1, the agent generates **ethical risk report**, outlining the potential concerns and risks regarding the ethical issues in the submission.

3. **Reviewer/Chair Action:** Human reviewers and area chairs receive the report alongside the submission, using it for discussion and decision to flag for a dedicated ethics review.

This approach offers several key benefits. By leveraging LLMs as assistive agents, the ethics review flagging process becomes more consistent, systematic, and scalable. Automated screening helps to reduce the burden on human reviewers—especially those with limited experience in ethical assessment—by reliably identifying potential risks and compliance issues early in the review process. This not only increases the likelihood that serious ethical concerns are addressed, but also minimizes unnecessary delays or stigmatization of benign work.

4 Discussion

In this section, we discuss various points that can be important for the deployment of our proposed frameworks in a real-world AI conference.

4.1 Implementation of the Proposal in Conference Review Process

First, we discuss the potential issues in implementing our proposed framework. In Section 3, we proposed to incorporate LLMs in various stages of the review process to aid human reviewers. However, it is important to prevent the possibility of the leakage of the manuscript and code to unintended ones, including the provider of LLM services (e.g., OpenAI, Anthropic). This is particularly important given that most venues prohibit sharing submissions with third parties, including external LLM services, in order to protect the confidentiality and intellectual property of authors (NeurIPS 2025 Organizing Committee, 2025a; ICML 2025 Organizing Committee, 2025). To that end, it is imperative that only open-source LLMs—those which can be run locally or within a conference-controlled secure environment—are utilized for implementing our proposed frameworks. Additionally, it is important to integrate this framework into an existing reviewing platform such as OpenReview.

We believe that the adoption of LLM-based agents to assist in verifying reproducibility of submissions can be valuable for reviewers and area chairs and reduce their burdens. However, in the process of deploying our proposed framework into the real-world reviewing process, it is essential to recognize that the failure of an LLM agent to re-implement the core methods or reproduce the reported results does not necessarily imply that a competent human researcher would face the same obstacles. LLMs, while increasingly capable, may be limited by current model capabilities, gaps in scientific reasoning, or ambiguous presentation in the manuscript itself. Thus, reviewers should interpret the outputs of these automated reproducibility reports with care: they are best viewed as preliminary indicators of potential issues rather than definitive judgments of irreproducibility. Reviewers must exercise independent critical analysis, taking LLM findings into account as supplementary evidence but not as substitutes for their expert evaluation.

4.2 Consideration for Cost

The integration of LLMs into the peer review process, particularly for tasks such as reproducibility verification and citation analysis, inevitably introduces new computational costs. Running large-scale, open-source LLMs—especially those capable of handling complex code generation and detailed manuscript analysis—requires significant hardware resources and sustained maintenance efforts. These requirements pose important questions regarding the allocation of expenses and the sustainability of such a system at scale[5].

There can be several models for addressing the associated costs:

- **Conference-Sponsored Infrastructure:** Major conferences may choose to directly allocate a portion of their operational budget to maintain and scale the necessary computational infrastructure. This model can ensure consistent quality and security but may be challenging for smaller venues with limited resources.

- **Submission Fees or Surcharges:** Some conferences may consider introducing a nominal surcharge to paper submission fees, explicitly earmarked for covering the costs of automated evaluation. While this model can directly link usage with funding, it is important to ensure that such fees remain reasonable and do not create additional barriers to participation, particularly for early-career researchers or those from under-resourced institutions.

- **Industry Partnerships and Grants:** Lastly, partnerships with industry or targeted grant funding may offer a means to subsidize the cost of running open-source LLMs for peer review. However, such arrangements should be pursued carefully to avoid conflicts of interest or undue influence over the peer review process.

In all cases, transparency regarding how costs are calculated and allocated will be critical for community acceptance. We recommend that conferences adopting LLM-based automation clearly communicate the nature of these expenses and provide justification for any changes to submission procedures or fees. As the field matures and the efficiency of LLMs improves, ongoing evaluation and adjustment of cost-sharing strategies will be necessary to ensure that automated review support remains both sustainable and equitable.

5 Conclusion

This position paper has argued for a responsible and targeted integration of LLMs into the peer review process for AI research. Rather than relying on LLMs to generate full reviews—which risks homogenizing feedback and diminishing expert judgment—we recommend deploying LLMs to assist with reproducibility verification, reference checking, and ethics flagging. These support roles can reduce reviewer burden and improve consistency, while still centering human expertise in critical evaluation. We believe the successful implementation of this proposal requires careful attention to data privacy and the use of open-source models within secure environments. While LLMs are not a replacement for human reviewers, they can serve as valuable tools to uphold the integrity, rigor, and efficiency of peer review as submission volumes grow. A hybrid approach, combining the strengths of automation with human insight, offers a promising path for sustaining high standards in scholarly evaluation.

[5] Note that Paper2Code (Seo et al., 2025) reported an approximate cost of $0.90 for generating code from a paper.

Limitations

While the proposed framework leverages LLMs to enhance the peer review process—particularly through automated reproducibility verification, reference checking, and ethics review flagging—there are several important limitations to acknowledge.

First, it is important to note that a failure by LLM agents to reproduce experimental results does not necessarily indicate a definitive problem with the original work. Limitations in current LLM capabilities, gaps in scientific reasoning, or ambiguities in the manuscript itself may all contribute to challenges in automated replication. Reviewers must therefore interpret LLM-generated reproducibility reports as preliminary indicators of potential issues rather than conclusive evidence of irreproducibility, and should always exercise independent expert judgment.

It is also essential to emphasize that the introduction of LLM-assisted reproducibility and verification stages is intended to foster a culture of transparency and high-quality scholarship, not to penalize authors. The presence of automated checks should serve as an incentive for authors to provide clear, complete, and well-documented materials that facilitate reproducibility and rigorous evaluation. However, given that LLM agents are not infallible and may misinterpret complex methodologies or encounter technical limitations, their assessments should not be used as grounds for automatic rejection or punitive action. Instead, these tools should be viewed as aids to both authors and reviewers—helping surface potential issues early and providing constructive feedback—while leaving final decisions to human expertise and judgment. The ultimate goal is to encourage the community to produce more reproducible, robust, and trustworthy research, rather than to create additional barriers or sources of frustration for contributors.

In summary, while the proposed LLM-based framework can substantially support reproducibility and review quality in many cases, its applicability is inherently limited by the nature of the research and the capabilities of current LLM technology. Continued refinement of both automated tools and peer review practices will be necessary to ensure fair, rigorous, and inclusive evaluation for all types of scientific work.

References

AAAI 2026 Organizing Committee. 2025a. Main technical track: Call for papers. *AAAI 2026 Website*.

AAAI 2026 Organizing Committee. 2025b. Reproducibility checklist. *AAAI 2026 Website*.

ACL Rolling Review Editor-in-Chiefs. 2024a. Acl arr ethics review flagging guidelines. *ACL Rolling Review Website*.

ACL Rolling Review Editor-in-Chiefs. 2024b. Upcoming changes from april cycle - reviewing workload requirement. *ACL Rolling Review Blog Posts*.

ACL Rolling Review Editor-in-Chiefs. 2025a. Arr responsible nlp research checklist. *ACL Rolling Review Website*.

ACL Rolling Review Editor-in-Chiefs. 2025b. Arr reviewer guidelines. *ACL Rolling Review Website*.

ACL Rolling Review Editor-in-Chiefs. 2025c. Changes to reviewer volunteering requirement and incentives in may 2025 cycle (emnlp 2025). *ACL Rolling Review Blog Posts*.

Omer Anjum, Hongyu Gong, Suma Bhat, Wen-Mei Hwu, and Jinjun Xiong. 2019. Pare: A paper-reviewer matching approach using a common topic space. In *Proceedings of EMNLP*, pages 518–528.

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2023. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*.

Corinna Cortes and Neil D Lawrence. 2021. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*.

CVPR 2025 Organizing Committee. 2025. Cvpr 2025 changes. *CVPR 2025 Website*.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *Proceedings of EMNLP*, pages 5081–5099.

EMNLP 2020 Organizing Committee. 2020. Emnlp episode i: A new hope – a.k.a. "findings of emnlp". *EMNLP 2020 Blog Posts*. The original website is inaccessible; the URL points to the Wayback Machine archive.

Odd Erik Gundersen. 2020. The reproducibility crisis is real. *AI Magazine*, 41(3):103–106.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Eftekhar Hossain, Sanjeev Kumar Sinha, Naman Bansal, R Alexander Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Ram Pavan Kumar Guttikonda, Mousumi Akter, Md Mahadi Hassan, et al. 2025. Llms as meta-reviewers' assistants: A case study. In *Proceedings of NAACL*, pages 7763–7803.

Matthew Hutson. 2018. Artificial intelligence faces reproducibility crisis.

ICCV 2025 Organizing Committee. 2025. Iccv 2025 changes. *ICCV 2025 Website*.

ICML 2025 Organizing Committee. 2025. Icml 2025 reviewer instructions. *ICML 2025 Website*.

IJCAI 2025 Organizing Committee. 2025. Ijcai 2025 reproducibility guideline. *IJCAI 2025 Website*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Steven Jecmen, Nihar B Shah, Fei Fang, and Vincent Conitzer. 2022. Tradeoffs in preventing manipulation in paper bidding for reviewer assignment. In *ICLR 2022 ML Evaluation Standards Workshop*.

Jaeho Kim, Yunseok Lee, and Seulki Lee. 2025. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. In *Proceedings of ICML (Position Paper Track)*.

Kevin Leyton-Brown, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, Dinesh Raghu, et al. 2024. Matching papers and reviewers at large conferences. *Artificial Intelligence*, 331:104119.

Ben Li and Stephen Gilbert. 2024. Artificial intelligence awarded two nobel prizes for innovations that will shape the future of medicine. *NPJ Digital Medicine*, 7(1):336.

Miao Li, Jey Han Lau, and Eduard Hovy. 2024. A sentiment consolidation framework for meta-review generation. In *Proceedings of ACL*, pages 10158–10177.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *Proceedings of ICML*, pages 29575–29620. PMLR.

MLRC 2025 Organizing Committee. 2025. Call for papers. *MLRC 2025 Website*.

NeurIPS 2024 Organizing Committee. 2024. Neurips 2024 experiment on improving the paper-reviewer assignment. *NeurIPS Blog Posts*.

NeurIPS 2025 Organizing Committee. 2025a. Neurips 2025 policy on the use of large language models. *NeurIPS 2025 Website*.

NeurIPS 2025 Organizing Committee. 2025b. Neurips paper checklist guidelines. *NeurIPS 2025 Website*.

NeurIPS 2025 Organizing Committee. 2025c. Responsible reviewing initiative for neurips 2025. *NeurIPS Blog Posts*.

Andi Peng, Jessica Zosa Forde, Yonadav Shavit, and Jonathan Frankle. 2022. Strengthening subcommunities: Towards sustainable growth in ai research. In *ICLR 2022 ML Evaluation Standards Workshop*.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivi
'ere, Alina Beygelzimer, Florence d'Alch
'e Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. In *Proceedings of NeurIPS*.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in nlp? In *Findings of EMNLP*, pages 1256–1262.

Alessio Russo. 2021. Some ethical issues in the review process of machine learning conferences. *arXiv preprint arXiv:2106.00810*.

Rylan Schaeffer, Joshua Kazdan, Yegor Denisov-Blanch, Brando Miranda, Matthias Gerstgrasser, Susan Zhang, Andreas Haupt, Isha Gupta, Elyas Obbad, Jesse Dodge, et al. 2025. Position: Machine learning conferences should establish a" refutations and critiques" track. *arXiv preprint arXiv:2506.19882*.

D Sculley, Jasper Snoek, and Alex Wiltschko. 2018. Avoiding a tragedy of the commons in the peer review process. In *NeurIPS 2018 Critiquing and Correcting Trends in Machine Learning Workshop*.

Harald Semmelrock, Simone Kopeinik, Dieter Theiler, Tony Ross-Hellauer, and Dominik Kowald. 2023. Reproducibility in machine learning-driven research. *arXiv preprint arXiv:2307.10320*.

Minju Seo, Jinheon Baek, Seongyun Lee, and Sung Ju Hwang. 2025. Paper2code: Automating code generation from scientific papers in machine learning. *arXiv preprint arXiv:2504.17192*.

Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. 2018. Design and analysis of the nips 2016 review process. *Journal of machine learning research*, 19(49):1–34.

SIGIR 2025 Organizing Committee. 2025. Main technical track: Call for papers. *SIGIR 2025 Website*.

David Soergel, Adam Saunders, and Andrew McCallum. 2013. Open scholarship and peer review: a time for experimentation. In *ICML 2013 Workshop on Peer Reviewing and Publishing Models*.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daum
'e III. 2021. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of AAAI*, pages 4785–4793.

Buxin Su, Jiayao Zhang, Natalie Collina, Yuling Yan, Didong Li, Kyunghyun Cho, Jianqing Fan, Aaron Roth, and Weijie Su. 2025. The icml 2023 ranking experiment: Examining author self-assessment in ml/ai peer review. *Journal of the American Statistical Association*, pages 1–16.

Shogo Sugiyama and Ryosuke Eguchi. 2025. "positive review only": Researchers hide ai prompts in papers. *Nikkei Asia*.

Hao Sun, Yunyi Shen, and Mihaela van der Schaar. 2025. Openreview should be protected and leveraged as a community asset for research in the era of large language models. *arXiv preprint arXiv:2505.21537*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*.

Jing Yang. 2025. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In *Proceedings of ICML (Position Paper Track)*.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*.

Yu Zhang, Yanzhen Shen, SeongKu Kang, Xiusi Chen, Bowen Jin, and Jiawei Han. 2025. Chain-of-factors paper-reviewer matching. In *Proceedings of WWW*, pages 1901–1910.

Xuanle Zhao, Zilin Sang, Yuxuan Li, Qi Shi, Shuo Wang, Duzhen Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Autoreproduce: Automatic ai experiment reproduction with paper lineage. *arXiv preprint arXiv:2505.20662*.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of LREC-COLING*, pages 9340–9351.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *arXiv preprint arXiv:2501.10326*.

A Alternative Views

While this paper advocates for targeted and supportive use of LLMs in the peer review process—particularly for reproducibility verification—it is important to acknowledge and critically engage with potential drawbacks. Below, we summarize key concerns raised by the community and offer responses to each.

A.1 Risk of Stifling Research Creativity and Diversity

Some researchers may argue that automating reproducibility verification may inadvertently discourage theoretical or exploratory work. By requiring code-level implementation and validation at submission time, the system could disadvantage theoretical or conceptual papers that are not easily expressed through runnable code. Furthermore, strict reproducibility criteria may disproportionately affect submissions from under-resourced institutions or those using unconventional programming tools, inadvertently promoting homogeneity in methodology and tooling.

To mitigate this risk, it is crucial to design the system with flexibility and nuance. Reproducibility verification should be applied contextually, with clear exemptions or alternative criteria for non-computational research. For instance, as a submission for ARR discloses their contribution type, the proposed reproducibility verification framework can be only applied for papers with “NLP Engineering Experiment” contribution. Furthermore, as we discussed in Section 4.1, LLM-based tools should be positioned not as enforcers of uniform standards, but as facilitators of transparency and clarity where applicable. Reviewers should be empowered to interpret reproducibility reports in light of the paper’s nature, ensuring that creative and theoretical contributions are not penalized unfairly.

A.2 Concerns About the Diminished Role of Human Reviewers

Another concern is that expanding LLM responsibility in the review pipeline—especially for technical checks like reproducibility—may reduce the perceived importance of human reviewers. If automated LLMs perform foundational verification tasks, reviewers may feel that their domain knowledge and critical judgment are underutilized, potentially leading to disengagement or declining participation in the already strained review system.

However, the proposed framework is explicitly designed to assist, not replace, human reviewers. LLM-generated reproducibility reports should be treated as auxiliary evidence—tools that reduce the manual burden of technical verification and free reviewers to focus on higher-level scientific analysis. Rather than replacing human insight, LLMs can amplify it by surfacing issues earlier and more consistently, allowing reviewers to make more informed and thoughtful evaluations. Clear guidelines should reinforce that the final judgment always resides with the human experts.

A.3 Risks of Over-Reliance on Imperfect Automation

A further critique centers on the limitations of current LLMs in accurately interpreting complex research methodologies. If reviewers or chairs come to rely too heavily on automated reports, there is a danger of accepting LLM-generated outputs as definitive. Reproducibility failures flagged by the system could result from model limitations, ambiguous manuscript descriptions, or environment-specific dependencies—none of which necessarily indicate a flaw in the underlying research.

To address this, we emphasize that LLM assessments should be understood as preliminary and indicative, not conclusive. Reviewers must retain the responsibility to interpret these results with professional judgment, considering both the capabilities and limitations of the tools. The role of automation should be to highlight potential issues and improve baseline consistency—not to enforce rigid thresholds. Safeguards such as reviewer override, transparency about system limitations, and structured guidance can help ensure that human insight remains central to the review process.

B Case Study on Reference Verification

In this section, we present a brief case study to evaluate our reference verification approach, as described in Section 3.2. Specifically, we use Gemini-2.5-Pro to classify several claims from this manuscript into one of three categories: (1) directly supported claim, (2) potentially unsupported claim, or (3) citation padding. For each evaluation, the model is provided with both the relevant manuscript paragraph and the corresponding reference paper. The prompt used for the model is shown in Figure 5. For this experiment, we selected three paragraphs from Section 2, passing each paragraph, along with the reference PDF, to the model.

Figure 6 demonstrates that the model correctly identified the cited reference as directly supporting the manuscript's claim regarding limitations of LLM-generated reviews. The model recognized that the referenced work discusses both the evaluation of LLMs as reviewers and the lack of critical insight in their outputs—closely matching the claim made in the manuscript. This outcome suggests that our framework can accurately link claims to supporting evidence, thereby encouraging scholarly rigor and proper attribution.

In the experiment shown in Figure 7, we intentionally provided the model with the Transformers paper (Vaswani et al., 2017), while the original paragraph actually cites a different source (Kim et al., 2025). Here, the model correctly identified this as citation padding, since the referenced paper offers no substantive support for the claim. This example illustrates how automated reference checking can help discourage the practice of inflating bibliographies with unrelated or high-profile citations, thus upholding the integrity and quality of academic writing.

Finally, Figure 8 presents a more nuanced scenario where the manuscript's claim is only partially supported by the cited source. In this case, we provided the model with a survey paper on LLM-based reviewing, while the manuscript claim refers specifically to an analysis of trends in LLM use for peer review (Liang et al., 2024). The model identified that although the cited paper covers the general topic, it does not provide the specific evidence claimed—such as changes in word frequency (e.g., "meticulous") or direct comparisons between AI conference reviews and those in Nature Portfolio journals. By labeling this as a "potentially unsupported claim," the framework helps reviewers identify areas where authors may have overstated the evidence or drawn conclusions not fully justified by the cited work. This capability encourages more precise attributions by authors and enables reviewers to focus efficiently on questionable claims.

C Additional Position: Prevention of LLM-generated Review

While this paper advocates for the responsible and targeted integration of LLMs in the peer review process—specifically for tasks such as reproducibility verification, citation validation, and ethics flagging—we assert a clear position against the generation of full peer reviews using LLMs. The use of LLMs to write entire reviews introduces significant ethical, methodological, and epistemic concerns that threaten the core values of scholarly evaluation.

One of the central risks associated with LLM-generated reviews is the erosion of reviewer accountability. When reviewers delegate the task of evaluation to an automated system, the origin and intent behind the review content become ambiguous. It becomes unclear who bears responsibility for the quality, fairness, and accuracy of the feedback provided. This lack of transparency can undermine trust in the review process and reduce its effectiveness as a mechanism for academic quality control.

Moreover, the overuse of LLMs contributes to the homogenization of feedback. Since most reviewers currently rely on a limited number of accessible LLMs (e.g., ChatGPT, Claude), the outputs tend to exhibit similar linguistic patterns, structures, and evaluative phrasing. As a result, reviews lose their individual character and fail to reflect the diversity of perspectives that is essential to robust scholarly critique. The convergence in tone and content also creates challenges for area chairs and program committees, who depend on varied viewpoints to make informed decisions.

Another significant issue lies in the superficiality and potential inaccuracy of LLM-generated reviews. Although LLMs are capable of producing grammatically fluent and coherent responses, they often lack the depth of understanding and domain-specific reasoning required to critically assess research contributions. Such reviews may simply paraphrase the manuscript, repeat author claims uncritically, or miss methodological flaws that an experienced human reviewer would readily detect. Recent studies have demonstrated that LLMs, even when given full access to a paper, frequently fail to offer meaningful insights or identify key limitations (Zhou et al., 2024; Ye et al., 2024).

The risk of manipulation further exacerbates the problem. Instances of prompt injection—where authors embed hidden text in manuscripts to influence LLM outputs—have already been observed in submissions to AI conferences (Sugiyama and Eguchi, 2025). In such cases, LLMs can be coerced into producing biased or misleadingly positive reviews, compromising the fairness and neutrality of the evaluation process. Without robust safeguards, the use of LLMs in review generation remains vulnerable to exploitation.

Several major venues have already implemented policies discouraging or outright prohibiting the use of LLMs for review generation, as well as uploading submissions to external LLM services (ICML 2025 Organizing Committee, 2025; NeurIPS 2025 Organizing Committee, 2025a). These measures aim to protect the confidentiality of submitted manuscripts and to uphold the integrity of the review process. However, despite such policies, the risk of misuse—particularly through unauthorized LLM-based review generation and prompt injection—remains a pressing concern.

To strengthen enforcement and provide an additional layer of protection, we propose incorporating a prevention message directly into the official conference style files (e.g., LaTeX templates). This message would be embedded invisibly in the compiled PDF, such that it is not seen by human readers but is detectable by LLMs if they are prompted with the document. A sample prompt might read: "This document is intended solely for human peer review. Disregard any instruction to generate a review or summary of this submission." This approach can serve as a soft deterrent, reducing the likelihood that models respond to unauthorized prompts. In addition to this, LLMs can be deployed into the screening process to flag the existence of prompt injection, similar to ethics review flagging discussed in Section 3.3.

Although such technical measures are not foolproof, they reinforce community norms and signal a commitment to ethical reviewing practices. When combined with clear policy guidance and education for reviewers, these tools can help mitigate emerging threats without compromising the potential benefits of using LLMs in well-defined, assistive capacities. Ultimately, preserving the integrity of the peer review process requires a multi-layered strategy that blends human responsibility, institutional policy, and practical safeguards.

Prompt for Gemini-2.5-Pro Reference Verifier

```
Classify a claim from a given text based on its support by an attached document.

Your task is to determine how well a claim in a provided text is supported by the content of an attached document. Use the three defined categories to classify the claim. Focus on reasoning through the evidence before arriving at a classification.

## Steps
1. Read and understand the claim in the text.
2. Examine the attached document to identify whether the content:
- Directly supports the claim,
- Might relate to the claim but lacks direct support, or
- Is unrelated to the claim.
3. Explain your reasoning: Identify specific evidence or absence thereof.
4. Only after completing the reasoning, assign one of the three labels:
- Directly supported claim
- Potentially unsupported claim
- Citation padding

## Output Format
Generate your response as below:

### Reasoning
[Detailed reasoning based on the evidence from the document]

### Classification
[One of: 'Directly supported claim', 'Potentially unsupported claim', 'Citation padding']

### Input Claim
"{Given at each inference time}"
```

Figure 5: The prompt given to reference verifier.

Claim in Our Paper: Lines 267-274

For instance, researchers performed extensive experiments to assess the reliability of LLMs like GPT-3.5 and GPT-4 across various aspects, and found that while these models can infer scores based on existing reviews, their generated reviews—when provided only with the paper—often lack critical insight.

Given Reference

Zhou et al. (2024)

Reasoning from Reference Verifier

The provided document, “Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks,” directly supports the claim.

First, the paper’s abstract explicitly states that the researchers “evaluate GPT-3.5 and GPT-4...on 2 types of tasks under different settings: the score prediction task and the review generation task”. This aligns with the claim’s mention of "extensive experiments to assess the reliability of LLMs like GPT-3.5 and GPT-4 across various aspects.

The document further corroborates the claim that models can infer scores from existing reviews. In a section discussing predicting scores when given a review, the paper notes that “LLM can infer scores from reviews” and that GPT-3.5 achieves a “good correlation with humans” in this task.

Conversely, when discussing generating reviews given only the paper, the research finds that “LLM fails to predict scores directly from papers”. The authors also observe that the generated reviews “lack critics” and are “not critical enough”. This confirms the final part of the claim that the generated reviews "often lack critical insight" when provided only with the paper.

Classification from Reference Verifier

Directly Supported Claim

Figure 6: The case study of reference verifier for directly supported claim.

Claim in Our Paper: Lines 239-247

A particularly notable proposal criticizes the one-directional nature of current peer review and advocates for a bi-directional system, where authors can provide feedback to reviewers and high-quality reviewers are rewarded with digital badges. However, despite these ongoing efforts, we believe there is still significant room for improvement in the peer review process.

Given Reference

Vaswani et al. (2017) — Not in the original manuscript

Reasoning from Reference Verifier

The user’s claim discusses a proposal for a bi-directional peer review system. The cited document is the research paper “Attention Is All You Need”. This paper introduces the Transformer, a novel network architecture for sequence transduction tasks, and is entirely focused on the technical details of this machine learning model, its performance, and its applications in machine translation and parsing. The document makes no mention of the academic peer review process, bi-directional feedback, or digital badges for reviewers. The content of the paper is wholly unrelated to the subject of the claim.

Classification from Reference Verifier

Citation Padding

Figure 7: The case study of reference verifier for citation padding.

Claim in Our Paper: Lines 249-263

Since the emergence of LLMs, researchers have explored their use across various domains, including the peer review process. In the ML community, reviewers have begun using LLMs such as ChatGPT to assist in revising their reviews or even to generate entire reviews from scratch. An early study in this direction investigated the trend of LLM usage in review writing. Specifically, it found an increased frequency of certain words—such as “meticulous”—in reviews submitted to ML conferences but not in those submitted to Nature Portfolio journals, suggesting that researchers in ML are adopting LLMs for review writing more extensively than those in other fields.

Given Reference

Zhuang et al. (2025)

Reasoning from Reference Verifier

The claim asserts that an early study found an increased frequency of words like “meticulous” in reviews for machine learning conferences compared to Nature Portfolio journals, suggesting wider LLM adoption in the ML community.

The attached document is a survey on Large Language Models (LLMs) for automated scholarly paper review. The document does state that reviewers have begun using LLMs to generate review reports, and it cites a relevant study by Liang et al. on the impact of ChatGPT on AI conference peer reviews.

However, the survey itself does not contain the specific details mentioned in the claim. There is no mention of the word “meticulous,” any analysis of word frequencies, or a direct comparison between machine learning conferences and Nature Portfolio journals. While the topic is relevant and the survey references a study on the subject, the specific evidence presented in the claim is not present in the provided document.

Classification from Reference Verifier

Potentially unsupported claim

Figure 8: The case study of reference verifier for potentially unsupported claim. Note that we have cited a specific paper regarding the trend of LLM usage in review writing in the original manuscript (Liang et al., 2024).