SAFEGUARDING PUBLICATION CALIBER AMIDST MASSIVE AI CONFERENCE SUBMISSIONS IN THE ERA OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The explosion of AI research, accelerated by Large Language Models (LLMs) and agent-powered tools, has led to an unprecedented surge in conference submissions. This deluge imposes a tremendous burden on the peer-review process, triggering a **vicious cycle** characterized by low-quality reviews, stochastic decision-making, and a reproducibility crisis. In this paper, we analyze the structural failures of current AI reviewing paradigms and contrast them with established practices in other computer science subfields. We find that existing mechanisms in AI conferences fail to effectively constrain irresponsible behavior from both authors and reviewers in the LLM era. To break this cycle, we propose the "carrot and stick" methods that introduce lightweight yet high-impact modifications for authors, reviewers, and conference organizers. Our method incorporates mandatory revision plans, a real-name shepherding system, and artifact evaluation (AE) for high-tier papers including oral and spotlight. By optimizing workflows and UI features, such as explicit score-increase justifications, our practical, cost-effective approach aims to safeguard publication caliber and restore the academic reputation of AI conferences.

1 INTRODUCTION

The number of researchers in the Artificial intelligence (AI) field has continued to increase, which has promoted the development of AI. Meanwhile, the ease of use of Python, various open-source codebases, convenient deep learning frameworks like PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2016), and various plugin tools (TensorFlow, 2026) have made baseline reproduction, paper code implementation, and experimental implementation much easier. The development of various LLM-based programming assistants, such as Codex (Chen et al., 2021), Claude Code (Guo et al., 2024), and DeepSeek-Coder (DeepSeek-AI et al., 2024), significantly accelerates the coding and experimental phases of research papers. With the development of large language models (LLMs), many researchers use LLMs to assist with paper writing, and even use various agent-powered paper writing tools to enhance efficiency, e.g., Prism (OpenAI, 2026) and PaperDebugger (Hou et al., 2026). These approaches can generate a paper within a few minutes (Hu et al., 2024; Wang et al., 2024). This self-reinforcing progress in the AI field makes completing a research paper in the AI field increasingly easy, which has also led all major AI conferences to receive a large number of submissions every year.

While a vast number of practitioners and papers have significantly accelerated the development of the AI field, the annually surging number of AI conference submissions has imposed a tremendous burden on the peer review process. Current reviewing methodologies do not fully meet the impact of the massive volume of submissions. This not only poses challenges for reviewers but also creates a poor experience for authors. Worse still, this situation potentially not only brings pressure to conference organizers and but also forms a **vicious cycle** as Figure 1, which ultimately leads to a decline in the quality of some papers within the publication. This in turn affects the reputation of the AI field and the credibility of papers (§2). At the same time, as the influence of the AI field becomes stronger and the number of participants increases, this also brings negative public opinion to conferences (Reddit, 2025).

We further review the current reviewing methodologies commonly used in top AI conferences. Although conference organizers have made various improvements to address these challenges in the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

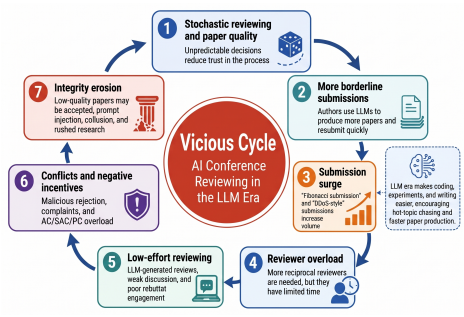


Figure 1: Current vicious circle.

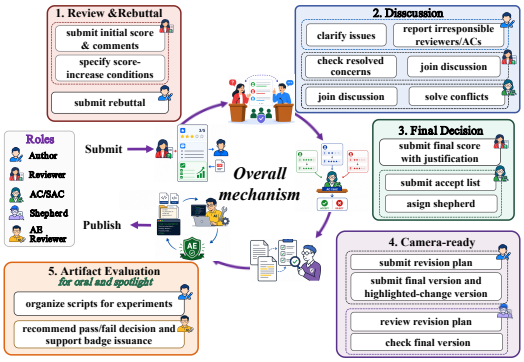


Figure 2: Our overall mechanism.

LLM era, current methods still fail to guarantee the caliber of papers (§3). Based on this analysis, we discuss existing works regarding suggested improvements for AI conference reviewing mechanisms, along with existing auxiliary projects aimed at enhancing the experience (§4), and further explore reviewing paradigms in other subfields within computer science (§5). However, naïvely applying these methods to current conferences cannot fully resolve the issues.

Based on the aforementioned analysis and the specific characteristics of current AI conference submissions, we propose a method that involves lightweight changes to existing reviewing mechanisms and remains relatively simple. We only introduce slight additional requirements for the three primary parties: reviewers, authors, and conference organizers. Reviewers and authors are asked to take on minor extra responsibilities and procedures, while conference organizers make slight modifications to the workflow and add several features to the reviewing platform. Specifically, our approach shown in Figure 2 mainly incorporates clearer reviewer accountability, concise rebuttal and discussion workflows, revision plans, shepherding, and artifact evaluation, along with improved reviewing platform support. This approach aims to achieve a high-cost-performance solution to guarantee and enhance the caliber of accepted papers in the LLM era (§6).

Then, we analyze alternative views (§7). Finally conclude our paper (§8). The potential impact of our method along with possible improvements is discussed (Appendix B).

2 The Vicious Cycle of Current AI Conference

Without a reasonable reviewing mechanism to guarantee caliber of papers, the massive volume of AI submissions in LLM era has brought negative impacts to authors, reviewers, conference organizers.

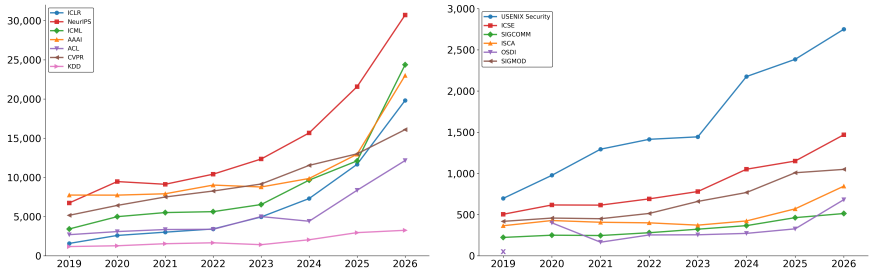
2.1 The Predicament of Authors

Due to the soaring submissions, there is an inevitable influx of inexperienced and irresponsible reviewers (cHHillee, 2019) and low-quality feedback. Consequently, authors currently face the following practical issues that severely compromise the submission experience: (1) Authors are increasingly confronted with nonsensical reviews, such as the notorious "Who is Adam" (2prime_PKU, 2026). (2) The final outcomes of submitted papers often depend on whims of stochastic reviewer decisions rather than substantive academic exchange. For example, ICLR 2025 Best Paper was previously rejected at NeurIPS 2024 despite having no significant changes (Mittal, 2024)[1]. (3) In LLM era, it is even more challenging for authors, as some reviews are now entirely generated by LLMs. exemplified by "40 weaknesses and 40 questions" review in ICLR 2026 (Anonymous, 2025a;c;b). Another critical concern with such synthetic feedback is that different LLM prompts can yield inconsistent conclusions for acceptance (Liang et al., 2024). (4) The rebuttal and discussion phases often prove to be futile (Cortes & Lawrence, 2021). Regardless of the quality of the rebuttal, it ultimately only impacts 1% to 4.4% of the final decisions (Yao, 2023). Score adjustments are frequently dictated by bellwether effect (Jin et al., 2024; Kargaran et al., 2025).

2.2 The Predicament of Reviewers

From reviewers, they inevitably face a massive influx of low-quality submissions. This phenomenon results in a heavy reviewing workload, wastes time of volunteer reviewers and dampens their enthu-

[1]The reviewer randomness experiments and statistical analyses are provided in Appendix A.



(a) AI Conference Submission Trends (b) Other CS Subfields Submission Trends

Figure 3: Submission trends in computer science top conference (2019–2026).

siasm for rigorous reviewing as well as participating in the rebuttal and discussion phases: (1) Some papers exhibit blatant plagiarism and "paper washing". For instance, the paper (Li & Chen, 2020) plagiarizes the work (Sun et al., 2019), Li et al. (Li & Gao, 2021) plagiarizes paper (Sander et al., 2021), and the TMI paper (Zhang et al., 2024) is "washed" by the paper (Guo et al., 2025). (2) Irresponsible authors could use LLMs and various agentic tools to assist in producing a large number of low-quality papers. For example, authors irresponsibly utilize LLMs to generate manuscripts, which leads to the inclusion of hallucinated citations (Anonymous, 2024). (3) Although many papers do not upload code or only provide key snippets during the submission stage to prevent being *scooped*, a significant number of authors promise to open-source their code after acceptance and even provide GitHub links. However, years after acceptance, these GitHub repositories remain empty or only contain a "coming soon" README file (Momo5913, 2024). In some cases, the open-sourced code is completely unable to reproduce the results in the paper, or the results are even obtained by including the test set in the training process. Such behavior is likely to cause serious concerns about the actual quality of the research. (4) Papers intentionally *cherry-pick* data or even misreport experimental results. The improvements promised during the rebuttal and discussion periods are often not incorporated into final camera-ready. Such situations occur even in high-profile papers under public scrutiny, such as the ICLR oral paper (Minh et al., 2024) (213 out of 3704, Top 1.82%) or the NeurIPS best paper runner-up (Kirsch, 2024) (4 out of 4035, Top 0.10%). These instances involving prestigious papers suggest it is even more prevalent among the wider pool of accepted manuscripts.

2.3 The Vicious Cycle

Due to the existence of these issues, both reviewers and authors face problems such as the stochastic nature of paper quality and reviewing decisions, which may lead them to become less serious about prestigious AI conferences.

Authors may use the LLM tools to produce more papers and submit a large number of borderline papers to counter the stochasticity of the reviewing process. When a paper is rejected, authors often resubmit to the next AI conference without serious revision, which is a potential driver of the "Fibonacci submission method" and causes the number of submissions to increase annually. This "DDoS-style" submission approach further discourages reviewers from being diligent.

The surge in submissions necessitates a larger number of reviewers, and authors are typically required to serve as reciprocal reviewers. When authors are overwhelmed with numerous submissions, they remain preoccupied with producing new papers and fail to dedicate sufficient effort to reviewing. Consequently, they are more likely to irresponsibly utilize LLMs to generate reviews and find themselves unable to attend to the requirements of the rebuttal and discussion phases when they are handling their own extensive rebuttals. This is especially true when their own papers are rejected, leading them to either disappear from the discussions entirely or engage in more negative reviewing behaviors. Furthermore, this situation can lead to a zero-sum game. Since acceptance rates generally remain stable, reviewers within the same field may maliciously reject a competitor's paper to increase the acceptance probability of their own work or to reduce the number of baselines they must compare against in the future. Such behaviors often trigger a flood of complaints from authors; however, Program Chairs (PCs), Senior Area Chairs (SACs), Area Chairs (ACs) are unable to handle the massive volume of conflicts. This exacerbates the stochasticity and poor experience of the review process, creating a **vicious cycle**, as shown in Figure 1.

Even worse, this stochasticity may allow low-quality papers or those with untrustworthy experimental results to be accepted (Minh et al., 2024; Kirsch, 2024), further discouraging authors from

Table 1: Comparison of review processes across different CS subfields.

Conference	Field	Subm. #	Early Rej.	Rebuttal	Discuss.	Revision	AE	Shepherd	Recip. Re	PC Meeting
USENIX Security	Security	2,750	✓	✓	✗	✓	✓	✓	✗	✓
ICSE	SE	1,150	✗	✓	✗	✓	✓	✓	✗	✓
SIGCOMM	Network	366	✓	✓	✗	✓	✓	✓	✗	✓
ISCA	Arch.	570	✓	✓	✗	✓	✓	○	✗	✓
OSDI	System	327	✓	✓	✗	✓	✓	✓	✗	✓
SIGMOD	Database	1,049	✗	✓	✗	✓	✓	○	✗	✓
NeurIPS	AI	30,709	✗	✓	✓	✗	✗	✗	✓	✗

AE is Artifact Evaluation; Recip. Re = Reciprocal Reviewer requirement (mandatory author-as-reviewer); ○ = partial applied.

maintaining integrity. On one hand, some authors may inject specific LLM prompts in hopes of obtaining positive ratings from LLM-assisted reviews (Retraction Watch, 2025). On the other hand, some authors may collude with reviewers by bidding to gain unfair advantages (SIGSAC, 2024). This behavior leads to a situation where "bad money drives out good", potentially allowing low-quality papers to get accepted and further intensifying the vicious cycle.

Apart from the reasons, the number of AI practitioners increases and tasks like writing code, conducting experiments, and drafting papers become much simpler in the LLM era, authors tend to rush into chasing various hot topics. As it is difficult to get a paper accepted once it is *scooped*, they work overtime to complete their papers. This trend increases the volume of submissions to some extent and makes the papers less rigorous, which further intensifies the aforementioned vicious cycle.

The confluence of the aforementioned factors fuels a vicious cycle that potentially leads to a decline in publication quality amidst the deluge of large-scale submissions in the LLM era.

2.4 The Predicament of Conference Organizers

Figure 3a illustrates trends of AI conference submissions since 2019. Our data focuses on long papers and valid submissions. Based on the results, we find that not only general AI conferences but also natural language processing (ACL), computer vision (CVPR), and data mining (KDD) exhibit a similar pattern. The number of papers has been growing steadily from 2019 to 2023, the maturation of LLMs and agents leads to a massive explosion in submission growth starting from 2024.

Beyond securing additional venues and hotels for the vast number of participants and managing larger-scale logistics, the aforementioned vicious cycle and the explosion in submission numbers impose immense pressure on AI conference organizers. Organizers must search for emergency reviewers when original reviewers go missing and manage numerous comments from authors. Given the massive scale of the reviewer pool, there are inevitably reviewers, ACs, and SACs who submit their results late, which leads to delays in paper notifications, such as ACL 2026 (ACL, 2026).

As authors and reviewers inflict mutual harm within this vicious cycle, conference organizers also face unexpected challenges. For instance, the leak of reviewer identities during the ICLR 2026 rebuttal phase (Reddit Community, 2026a) has turned the entire review process into a severe mess and imposed further burdens on the overall reviewing and organizational efforts, the similar deanonymization happens again in ICLR 2027 (Reddit Community, 2026d).

However, the most critical issue arising from this vicious cycle is that publication quality cannot be guaranteed. If we allow this vicious cycle to continue, academic conferences lose their authority and potentially face ranking downgrades, similar to the case of IJCAI in CCF ranking (feifei88087, 2026; Atom Community, 2026). Therefore, the focus of this paper centers on how to break the vicious cycle with these objectives in mind.

2.5 Our Goal

It's worth noting that safeguarding publication caliber, we can largely break the vicious cycle: **Safeguarding publication caliber** encompasses (1) paper reproducibility, (2) resolving collusion between reviewers and authors, (3) addressing plagiarism, (4) reducing stochasticity in the review process, ensuring reviewer accountability, and making the rebuttal process genuinely effective. **Paper reproducibility** ensures the open-sourcing of code and the inclusion of revisions promised by the authors in the camera-ready, which significantly improves the quality of the paper. **Reducing collusion** and **plagiarism** also leads to a decrease in low-quality papers. Meanwhile, **the reduction of low-quality reviews and the active participation of reviewers in the rebuttal and discussion phases** are essential to minimize subpar reviews, which further ensures that low-quality papers are difficult to accept. These factors allow authors to focus on improving the quality of their work rather than utilizing LLM tools to generate low-quality papers. Consequently, authors do not have to strug-

gle with an excessive number of submissions and can dedicate more efforts to the review process, which further enhances the quality of peer review. There is no need to employ "clever methods" by inserting prompts to combat low-quality reviews generated by LLM, thereby forming a **virtuous cycle**. All of these are fundamentally related to the **reviewing methodologies**.

We believe that this is not merely for the purpose of reducing the number of papers. Instead, it conserves the time of both authors and reviewers, leads to more high-quality papers, and allows conference organizers to reduce their workload. Therefore, the focus of this paper is on **safeguarding publication caliber for publication by improving reviewing mechanisms**.

3 Existing Reviewing Mechanisms

The organizers of current AI conferences have made improvements to the review mechanism:

Current conference organizers have punishments regarding **paper plagiarism and collusion** (NeurIPS, 2026b). They generally ban all authors for several years and report to the relevant institutions to pursue further action. There is even cross-conference linkage. Thus, as long as someone provides evidence for a report, these issues can be resolved to a large extent. However, some papers with plagiarism issues that are explicitly discussed on the internet, such as "Dynamic Hierarchical Mimicking ... Optimization Objectives" (Li & Chen, 2020) has not been retracted in CVPR 2020. Although the paper "m-revnet: Deep reversible ... with momentum" (Li & Gao, 2021) has been voluntarily retracted, all authors have not been punished. This lax enforcement fails to provide a sufficient deterrent for authors who engage in plagiarism and collusion.

Some conferences require authors to fill out a **reproducibility** checklist and even computing reporting form (CVPR, 2025; NeurIPS, 2026b). However, this practice still fails. For the complete reproducible code mentioned above, authors can still engage in fake open-sourcing or empty promises. Meanwhile, peer reviewers provide voluntary labor, and conference organizers cannot mandate reviewers to reproduce results of the papers. Further, it is impossible to mandate reviewers possess computational resources required to replicate results, which still leads to poor reproducibility.

Organizers impose **more requirements on reviewers and ACs**, mandating that they are more responsible during the review, rebuttal, and discussion phases. For instance, conferences like CVPR 2026 perform desk rejects on papers from irresponsible reviewers (CVPR, 2026). However, this action sometimes still fails to fully constrain reviewer behavior, such as when a reviewer's own paper has already been rejected and they do not complete the review process seriously. ICML 2026 desk rejects the submissions of reviewers who violate their pledge against using LLMs (ICML Program Chairs, 2026), yet these LLM-generated reviews still trouble the authors. Additionally, reviewers are required to indicate whether the rebuttal has resolved the issues; however, authors complain that even when all concerns are addressed, reviewers may still maliciously lower their scores by claiming the issues remain unresolved, especially since ACs cannot see the original ratings. NeurIPS 2026 requires reviewers and ACs to see their own feedback before they complete their tasks (NeurIPS, 2026b), yet this still allows low-quality reviews.

Although AI conferences increasingly transition to the **OpenReview system** (OpenReview, 2026) to conduct reviewing and completing paper submission, review, rebuttal, and final submission, even this system fails to adequately resolve the problems faced mentioned earlier. When the review quality is low or when malicious rejections occur, although one can view the history of review comments, it remains insufficiently intuitive. While it is possible to submit confidential comments to ACs, SACs, and PCs, the massive amount of information and emails, combined with the non-intuitive UI, still leaves conference organizers difficult to handle these issues.

Consequently, the methods of current AI conferences still cannot truly improve quality of accepted papers. It is worth noting that due to the complexity of paper submissions in the LLM era, guidelines for each submission and review become increasingly long and sometimes even inconsistent, which instead causes confusion. For example, the placement of NeurIPS checklist has triggered extensive discussions among authors (Reddit Community, 2026c). Therefore, it is highly valuable to propose simple and feasible methods that better improve publication quality and enhance the review process.

4 Related Works

Currently, relevant position papers propose adding a new "Refutations and Critiques" track (Schaeffer et al., 2025) or rejecting AI-generated survey papers (Lin et al., 2025). These papers only express suggestions for improving paper quality in certain AI topics. The paper (Goldberg et al., 2025) fo-

cuses on scalable strategies for improving review quality. However, it is still difficult to truly solve the vicious cycle in this paper, improve the experience of authors and reviewers, and alleviate the burden on conference organizers.

Although some papers express analysis of reviewer randomness (Cortes & Lawrence, 2021; Jin et al., 2024), the randomness of LLM reviews (Liang et al., 2024; Zhuang et al., 2025), and ineffectiveness of rebuttals and discussions (Yao, 2023; Jin et al., 2024; Kargaran et al., 2025), these papers still cannot solve the related problems. ICML 2026 introduces voluntary self-ranking in reviews (Su et al., 2025) to better help conference organizers analyze the reviewing mechanism (ICML, 2026).

The project (Momo5913, 2024) analyzes that all the 4035 accepted papers in NeurIPS 2024, yet in 2026, 98 papers still involve fake open-sourcing. A similar project (Tan, 2026) evaluates the reproducibility of ICML 2026 oral papers, finding that only eight papers reproduced more than 80% of their reported conclusions. Since these projects are not official, they cannot solve issues where open-sourced code is irreproducible or where only a small portion of core code is released.

5 Reviewing Paradigms Beyond AI Conferences

We analyze the reviewing paradigms of other subfields in computer science besides AI. We select several top-tier conferences across different subfields to analyze their reviewing paradigms: USENIX Security (security), ICSE (software engineering), SIGCOMM (network), ISCA (architecture), OSDI (system), and SIGMOD (database).

Figure 3b illustrates the submission trends for these from 2019 to 2026, where "x" denotes that the conference was not convened in that year. The overall trend from 2019 to 2023 exhibits fluctuating growth; submissions have an explosive surge since 2024, facilitated by LLM-assisted writing and coding. Although the volume of papers in these domains is also rising, the reviewing paradigms remain distinct. Table 1 shows the digest of reviewing workflows of these conferences and highlights the differences compared to AI (since the number of valid submissions to ICLR 2027 has not yet been finalized, we use the recently concluded NeurIPS 2026 as a representative example).

These conferences still have only two tiers of review, which are Program chairs and Program Committee members, and Program Committee members are basically senior researchers. AI conferences adopt a reciprocal reviewer mechanism and generally have four tiers: PCs, SACs, ACs and reviewers. These conferences may use early rejection for papers to reduce the workload of reviewers, keep the rebuttal, and require authors to submit a revision plan after acceptance or provide paper notifications, such as major or one-shot revision. Moreover, most accepted papers are assigned a shepherd with high expertise to first assess the quality of the revision plan, then ensure the quality of the post-revision camera-ready and monitor adherence to the plan. Regarding the reproducibility of papers, there is generally a voluntary artifact evaluation (AE) phase, and corresponding badges are given after passing (Association for Computing Machinery, 2020). However, due to the limited number of reviewers, there is generally no discussion phase with authors, but there are offline or online PC meetings to discuss the decision of every paper.

Although these methods can guarantee quality of publication to a certain extent, it is worth noting that the number of submissions for any single conference still does not exceed 2800 in 2026, which is only one-eleventh of 30,709 submissions for NeurIPS 2026 and even relatively small compared to 6743 submissions for NeurIPS 2019. The review pressure on senior researchers remains controllable. For example, the number of papers for a SIGCOMM Program Committee member to review is 8-10 in 2024 and only increases to 16-18 in 2026. For AI conferences, if reciprocal reviewers are not adopted, the number of reviews for reviewers may exceed an order of magnitude, which is fundamentally impractical. Such a large scale of reviewers and a massive number of papers make PC meetings impractical as well. Even if AE in these conferences is voluntary, AI conferences still cannot ensure the completion of voluntary AE for all accepted papers due to the inconsistent quality of reviewers as mentioned previously and the shortage in computing resources. Even for these conferences, there are still instances where reviewers and authors collude (SIGSAC, 2024), or reviewers are irresponsible and do not conduct reviews, leading to the need for external emergency reviewers.

Therefore, these reviewing paradigms are difficult to apply directly to AI conferences.

6 Methodology

Drawing lessons from reviewing methods from other subfields, we propose to make modifications based on the characteristics of AI conferences. We find that conference organizers can use a carrot

and stick approach to regulate reviewers and authors. Meanwhile, making some necessary modifications to the review website or process can eventually help reviewers, authors, and conference organizers break the vicious cycle at a relatively small cost to guarantee quality of publication.

6.1 Optimizing Workflows from Reviewer

For reviewers and ACs, the **sticks** are as follows: (RSt1) Reviewers must be more responsible. In addition to current requirements such as mandatory participation without using LLMs for reviews, without the requirement for authors to compare with contemporaneous papers, active participation in rebuttals and discussions, and completion of the final justification, additional rules are introduced. If these tasks are not completed, not only are their own submitted papers desk rejected, but their submissions to this conference are also rejected for many future years. These penalties may even be linked with other conferences. At the same time, review comments cannot be modified arbitrarily. The same strategy is applied to irresponsible ACs. (RSt2) Reviewers must explicitly specify the points for which they are willing to increase the score (which can be empty). When authors resolve the issues, the score must be increased. (RSt3) A reviewer with the highest expertise for the paper is selected as the shepherd. This person is responsible for checking the implementation status of the promised open sourcing, the revision plan of authors after acceptance and ensuring that the promised content is included in the camera-ready. (RSt4) The shepherd is also responsible for conducting searches. If plagiarism is discovered, the shepherd reports it to the conference to reject. (RSt5) New reviewers for artifact evaluation are added. Since there are too many accepted papers in AI conferences, oral ($0.5 - 1\%$) and spotlight ($2 - 3\%$) are generally high-score papers with a low proportion, AE is imposed on these papers. We certainly allow papers not to open source relevant code, such as company papers with trade secrets. For such papers, either a reasonable justification must be written or company-related signed documents must be provided to obtain an exemption. If there are too many experiments, sampling detection of multiple experiments can be conducted to ensure reproducibility of the paper.

The **carrots** are as follows: (RCa1) The names of the responsible shepherd and AE reviewers are included in the paper. (RCa2) They also receive corresponding certificates that prove their professional expertise in this field. (RCa3) For high-quality reviewers and emergency reviewers, certificates are given and registration fees are waived. They can even participate in a lottery for travel support or certain remuneration (Reddit Community, 2026b).

6.2 Optimizing Workflows from Authors

For authors, the **sticks** are as follows: (ASt1) If collusion and plagiarism occur, in addition to previous punishment methods (§3), conference officials publicly disclose the names on the website. (ASt2) If open sourcing is promised, it must be implemented, or the paper is rejected and all authors of the paper are banned from the conference for three years. (ASt3) Both rebuttal and discussion have only one response opportunity, which requires being concise to save the effort of reviewers, with approximately 5000-10000 characters, and no additional information can be added unless reviewers continue to ask questions. (ASt4) A revision plan must be written after the paper is accepted. (ASt5) Two versions must be submitted at the camera-ready stage, which include a final version and a version with the modified parts highlighted. (ASt6) Oral and spotlight papers must participate in AE unless there is a reasonable justification, and the authors must provide computing resources that can run all experiments or at least complete sampling detection experiments for three days. When participating in AE, the specific environment setup and corresponding scripts for reproducing the resulting tables and figures should be provided. (ASt7) The names of the shepherd and AE authors must be added to the acknowledgment in the camera-ready version.

The **carrots** are as follows: (ACa1) Authors can report irresponsible reviewers, which potentially results in multi-year bans for them, e.g., maliciously lowering scores after the rebuttal, when reviewers require comparisons with contemporaneous work (NeurIPS, 2026a) or even their own papers on arXiv, use LLMs for reviewing, provide malicious rejection opinions, or do not participate in the rebuttal and discussion. Authors can even report irresponsible ACs. (ACa2) Irresponsible review opinions are deleted. (ACa3) Additionally, obtaining badges provides better proof of paper quality.

6.3 Optimizing Workflows from Conference Organizations

To better guarantee the quality of publication and optimize the reviewing process, in addition to the current reviewing process, conference organizers also need to make certain improvements:

The UI Improvements: Unlike current systems, organizers can learn from ChairingTool (ChairingTool, 2026) to provide a dedicated UI for reporting LLM-generated review opinions. This also applies to reporting low-quality review opinions and the lack of responses for reviewers or ACs during the rebuttal and discussion phases (RSt1, ACa1). Moreover, a specific text box should be added for reviewers to specify which exact resolved questions would lead to a score increase (RSt2). The rating and review of a paper should not be modified in place, and the original should remain unchanged. A new text box for final review and rating should be provided, similar to the current final justification, to help ACs and SACs identify whether reviewers have changed the score as promised (RSt1, ACa1).

Modification of functionality: Responses to rebuttals are released simultaneously to all reviewers to mitigate the bellwether effect, while there is still a multi-stage rebuttal process for author-reviewer discussions that imposes strict deadlines for reviewer responses and score updates before the joint discussion phase. On both submission website and conference homepage, a leaderboard similar to "Retraction Watch" (Retraction Watch, 2026) is provided to publicly disclose the authors and reviewers involved in plagiarism and collusion in previous years (ASt1). For camera-ready, two submission links should be provided, including one for the final version and one for a version where modified parts are highlighted (ASt5). For high-quality reviewers, AE reviewers and shepherds, corresponding certificates are automatically issued (RCa2). Bonuses are provided to high-quality reviewers and ACs and a dedicated page on the website specifically acknowledges them (RCa3). Irresponsible reviews and their scores may be removed from consideration so that ACs and authors can disregard low-quality feedback, subject to joint approval by ACs, SACs, and PCs, while the removed reviews remain internally archived to ensure oversight and preserve reviewer accountability (ACa2). For papers that complete AE, corresponding badges are automatically added (ACa3).

In addition to these, conference organizers should highlight new changes in submission and review guidelines every year and provide a TL;DR for them. Irresponsible reviewers and ACs should be banned for three years. For individuals involved in plagiarism and collusion, in addition to previous penalties, attempts are made to report the information to university ranking organizations such as U.S.News and QS to help these institutions improve integrity education.

6.4 Final Paradigms

By applying the aforementioned requirements and improvements for reviewers, authors, and conference organizers, we find that no excessive requirements are imposed on the primary parties, yet they better achieve the goal (§2.5) and enhance the quality of publication.

By drawing lessons from the writing of revision plans and adding shepherds and AE in other subfields, since authors already have a deep understanding of their papers and should have prepared for open sourcing, writing a revision plan and organizing scripts for experiments do not add excessive workload. Instead, these actions significantly improve the reproducibility of papers and enhance their persuasiveness. Since there are many AI conference publications, reproducing all papers consumes significant computational resources and recruiting such a large number of AE reviewers is also difficult. Therefore, performing AE for a small number of oral and spotlight papers is reasonable, and providing certificates as recognition of their expertise likely does not add too much recruitment workload. Although shepherds and AE increase the burden, it remains within a reasonable range and imposes little additional recruitmentoverhead.

Meanwhile, this method, which decreases the randomness with new AE reviewers and requires shepherds to be responsible with their real names, makes it more difficult for papers to achieve plagiarism and bidding collusion, as AE and shepherds act as double insurance for papers such as oral and spotlight. It reduces the occurrence of cases like the ICLR oral (Minh et al., 2024) and NeurIPS best paper runner-up cases (Kirsch, 2024). The severe crackdown on plagiarism and collusion between reviewers and authors, along with the disclosure of names, provides a better deterrent. For reputation, shepherds also have more motivation to report plagiarism even years after acceptance when it is discovered.

Clarifying the score increase issues for reviewers and the constraints on reviewer behavior does not increase the workload. A strict three-year ban ensures that reviewers remain responsible, rather than ignoring their duties because they have no submissions this year. Disclosing all replies only after the rebuttal ends reduces stochasticity and prevents reviewers involved in collusion from using

the bellwether effect to influence honest reviewers. Immutable original scores and reviews help ACs/SACs assess whether authors have resolved the reviewers' queries and identify if reviewers are maliciously introducing new issues or lowering ratings. As low-quality reviews can be deleted, this also reduces the randomness and encourages authors to take the submission and review process more seriously and focus on high-quality work. They can also obtain badges to gain stronger influence, which reduces the pressure to generate a large number of papers to counter stochasticity.

For conference organizers, we believe this method does not bring excessive overhead. Once functions and the UI are improved and the paradigm is optimized, it can be utilized once and for all.

We believe that our proposed method can help break the previously mentioned vicious cycle, improve the experience of authors, reviewers, and organizers, better guarantee the quality of papers in AI conferences, and enhance the reputation of AI conferences.

Overall mechanism. After the reviews end, the rebuttal phase begins. After the rebuttal ends, the responses from reviewers are released and the discussion phase begins to further clarify issues. During this period, authors can provide feedback regarding irresponsible reviewers and ACs. Finally, reviewers input their final scores. ACs and SACs can better identify the complaints based on the authors' rebuttal, score increase issues, and the changes in reviewer scores before and after. After the paper is accepted, all authors must provide a revision plan and the paper is assigned a shepherd. Oral and spotlight papers are also provided with AE guidelines and one AE reviewer. The AE badge is issued after passing. The overall procedure is shown in Figure 2.

More discussions are shown in Appendix B.

7 Alternative Views

Restricting the number of submissions per author: Although this can significantly reduce the number of papers, as in ICLR 2027 (ICLR, 2027), this may lead to the fact that some advisees of a professor cannot submit papers or the professor cannot be credited on papers supported by their own funding, as a professor may provide funding to many advisees. Since the rapid development of AI is also related to the increase in the number of papers and practitioners, this is not reasonable. Perhaps it is feasible to restrict the number of first-author submissions per author. **Charging for submissions**: IJCAI 2026 charges 100 dollars for each submission beyond one per author (IJCAI, 2026), which could be unfair for authors. **Changing the reviewing system**: IJCAI 2026 uses ChairingTool (ChairingTool, 2026) and WWW 2026 uses EasyChair (EasyChair, 2026). However, reviewing systems essentially cannot constrain reviewers and authors. A better reviewing mechanism is necessary to better guarantee the quality of publication. **Using AI-assisted reviewing**: AAAI 2026 uses AI for reviewing (AAAI, 2025), but this still cannot truly improve the reviewing experiences (Zhuang et al., 2025). **Changing the reviewing method to a journal-like style**: UbiComp (ACM, 2026), VLDB (VLDB, 2026) and ACL (ACL Rolling Review, 2026) have adopted journal-like reviewing methods, but still face the problems mentioned previously. Some conferences allow reviewers to **see previous reviews for resubmitted papers** (ACL Rolling Review, 2026). However, this bias from initial impressions or malicious reviews makes revisions from authors ineffective in changing the outcome. **Revealing reviewer names** risks triggering the chaos as ICLR 2026 (Reddit Community, 2026a), which is distinct from disclosing shepherd names, given that shepherds are high-expertise reviewers who support acceptance of the paper. **Requiring ACs to participate in reviewing**: Currently, the volume of AI submissions is massive. ACs need to be responsible for 15-30 papers during the review process and must read the review opinions and rebuttal discussions for each paper. It is not reasonable to require ACs to participate in the review due to the heavy workload.

8 Conclusion

In this paper, we analyze the vicious cycle faced by AI conferences in the LLM era, where massive submissions, low-quality reviews, weak reproducibility guarantees, plagiarism, collusion, and ineffective rebuttals jointly undermine the publication caliber and the experience of authors, reviewers, and organizers. To address these challenges, we propose a lightweight reviewing paradigm that introduces limited but targeted responsibilities for reviewers, authors, and conference organizers. We believe these changes can help transform the current vicious cycle into a more sustainable and quality-oriented reviewing ecosystem for AI conferences.

AI USE STATEMENT

In this work, we used generative AI tools for [assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results]. We have not used generative AI tools for [generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods], and [] are not applicable to this work.

Additionally, we used generative AI tools for [create or modify scientific figures or images, create or edit software code, creation of artifacts, sourcing/searching for information, edit a research paper to improve readability, format references, suggest a structure for a research paper, propose a title or keywords for a research paper]. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper discusses ethical and community-level concerns arising from the increasing use of large language models in AI conference reviewing and scholarly communication. In particular, we highlight potential risks associated with rapidly growing submission volumes, LLM-assisted writing and reviewing, limited reviewer accountability, and increasing pressure on authors, reviewers, and conference organizers. These trends may affect publication quality, reproducibility, fairness, transparency, and trust in the peer-review process.

Our goal is not to discourage the responsible use of LLMs in research or reviewing, nor to prescribe a single reviewing model for all conferences. Instead, we aim to encourage discussion on how peer-review processes can evolve while preserving human accountability, transparency, fairness, and scholarly integrity. The workflow changes discussed in this paper, including clearer reviewer accountability, concise rebuttal and discussion procedures, revision plans, shepherding, selective artifact evaluation, and improved platform support, are intended as possible directions rather than mandatory policies.

We recognize that any reform of conference reviewing may introduce trade-offs among reviewer workload, author burden, accessibility, inclusiveness, and review quality. Therefore, such changes should be evaluated carefully, introduced transparently, and designed to avoid disproportionately disadvantaging particular groups of researchers or institutions.

REPRODUCIBILITY STATEMENT

We provide all source code in the supplementary material, which analyzes the data.

REFERENCES

2prime_PKU. Discussion on AI conference submission deluge and reviewing mechanisms, 2026. https://x.com/2prime_PKU/status/1948549824594485696, Accessed: 2026-04-28.

AAAI. AAAI Launches AI-Powered Peer Review Assessment System, 2025. https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/, Accessed: 2026-04-28.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265–283, 2016.

ACL. The paper acceptance notifications will be out by the 6th of april, aoe. the pcs are working hard throughout the holiday season to finalize the decisions. Twitter (X), April 2026. https://x.com/i/status/2040938767637508289, Accessed: 2026-04-28.

ACL Rolling Review. ACL Rolling Review (ARR), 2026. https://aclrollingreview.org/, Accessed: 2026-04-29.

ACM. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), 2026. https://dl.acm.org/journal/imwut, Accessed: 2026-04-29.

Anonymous. Review decision for "semantic calibration in media streams". OpenReview, 2024. https://openreview.net/forum?id=6YTEDVd80T&noteId=j2MXmfAI15, Accessed: 2026-04-28.

Anonymous. Review of 40 weaknesses and 40 questions for "mixture-of-visual-thoughts: Exploring context-adaptive reasoning mode selection for general visual reasoning". OpenReview, 2025a. https://openreview.net/forum?id=8qk6eUnvbH&noteId=r1Pd9RmCgF, Accessed: 2026-04-28.

Anonymous. Review of 40 weaknesses and 40 questions for " object hierarchy guides visual relationship reasoning". OpenReview, 2025b. https://openreview.net/forum?id=GlXyFjUbfN&noteId=5dbAhbyjEr, Accessed: 2026-04-28.

Anonymous. Review of 40 weaknesses and 40 questions for "vgr: Visual grounded reasoning". OpenReview, 2025c. https://openreview.net/forum?id=kDhAiaGzrn&noteId=XzScUnmDGs, Accessed: 2026-04-28.

Association for Computing Machinery. Artifact Review and Badging, 2020. https://www.acm.org/publications/policies/artifact-review-badging, Accessed: 2026-04-28.

Atom Community. CCF Recommended Conference and Journal List Deadlines, 2026. https://ccf.atom.im/, Accessed: 2026-04-28.

ChairingTool. ChairingTool: Conference Management System, 2026. https://chairingtool.com/, Accessed: 2026-04-29.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

cHHillee. Some metadata for those curious about their #iclr2020, 2019. https://twitter.com/cHHillee/status/1191823707100131329, Twitter post.

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

CVPR. Cvpr 2025 suggested practices for authors, 2025. https://cvpr.thecvf.com/Conferences/2025/AuthorSuggestedPractices, Accessed: 2026-04-28.

CVPR. CVPR 2026 Reviewer Guidelines, 2026. https://cvpr.thecvf.com/Conferences/2026/ReviewerGuidelines, Accessed: 2026-04-28.

DeepSeek-AI et al. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*, 2024. URL https://arxiv.org/abs/2406.11931.

EasyChair. EasyChair: Conference Management System, 2026. http://easychair.org/, Accessed: 2026-04-29.

feifei88087. On march 5, 2026, china computer federation (ccf) downgraded ijcai to a b-level conference, while iclr has been upgraded to ccf-a. there are only seven a-level conferences in the ai domain. ijcai is dead, so who is the next? Twitter (X), 2026. https://x.com/feifei88087/status/2029507407899209981, Accessed: 2026-04-28.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PloS one*, 20(4):e0320444, 2025.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, Ning Li, et al. Deepseek-coder: When the large language model meets programming – the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024. URL https://arxiv.org/abs/2401.14196.

Shuai Guo, Yang Gu, Yuan Ma, Yingwei Zhang, Weining Weng, Jun Liu, Weiwei Dai, and Yiqiang Chen. Semantic-oriented visual prompt learning for class incremental learning. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887945.

Junyi Hou, Andre Lin Huikai, Nuo Chen, Yiwei Gong, and Bingsheng He. Paperdebugger: A plugin-based multi-agent system for in-editor academic writing, review, and editing. In *Companion Proceedings of the ACM Web Conference 2026*, pp. 144–147, 2026.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. Taxonomy tree generation from citation graph. *arXiv preprint arXiv:2410.03761*, 2024.

ICLR. Iclr 2027 author guidelines, 2027. https://iclr.cc/Conferences/2027/AuthorGuidelines,Accessed: 2026-09-25.

ICML. Introducing ICML 2026 Policy for Self-Ranking in Reviews, January 2026. https://blog.icml.cc/2026/01/12/introducing-icml-2026-policy-for-self-ranking-in-reviews/, Accessed: 2026-04-29.

ICML Program Chairs. On Violations of LLM Review Policies. ICML Blog, March 2026. https://blog.icml.cc/2026/03/18/on-violations-of-llm-review-policies/, Accessed: 2026-04-28.

IJCAI. IJCAI-ECAI 2026 Call for Papers: Main Track, 2026. https://2026.ijcai.org/ijcai-ecai-2026-call-for-papers-main-track/, Accessed: 2026-04-29.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1208–1226, 2024.

Amir Hossein Kargaran, Nafiseh Nikeghbal, Jing Yang, and Nedjma Ousidhoum. Insights from the iclr peer review and rebuttal process. *arXiv preprint arXiv:2511.15462*, 2025.

Andreas Kirsch. Important prior work attribution: RhoLoss and Rho-1. OpenReview Public Comment, 2024. https://openreview.net/forum?id=0NMzBwqaAJ&noteId=dZcOZBIIVe, Accessed: 2026-05-01.

Duo Li and Qifeng Chen. Dynamic hierarchical mimicking towards consistent optimization objectives. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7639–7648, 2020. URL https://api.semanticscholar.org/CorpusID:214623002.

Duo Li and Shang-Hua Gao. m-revnet: Deep reversible neural networks with momentum. *arXiv preprint arXiv:2108.05862*, 2021.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024.

Jianghao Lin, Rong Shan, Jiachen Zhu, Yunjia Xi, Yong Yu, and Weinan Zhang. Stop ddos attacking the research community with ai-generated survey papers. *Neural Information Processing Systems*, 2025.

Nguyen Huu Nhat Minh, Andrew Baker, Clement Neo, Allen G. Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:270870613.

Prateek Mittal. ICLR 2025 best paper was rejected at NeurIPS 2024 with no significant changes. Twitter (X), December 2024. https://x.com/prateekmittal_/status/1918350357144420859, Accessed: 2026-04-28.

Momo5913. NeurIPS 2024 Coming Soon: A collection of papers without code after acceptance, 2024. https://momo5913.github.io/NeurIPS_24_Coming_Soon/, Accessed: 2026-04-28.

NeurIPS. NeurIPS 2026 Main Track Handbook: Contemporaneous Work, 2026a. https://neurips.cc/Conferences/2026/MainTrackHandbook#ContemporaneousWork, Accessed: 2026-04-28.

NeurIPS. NeurIPS 2026 Main Track Handbook, 2026b. https://neurips.cc/Conferences/2026/MainTrackHandbook, Accessed: 2026-04-28.

OpenAI. Prism: An AI-powered LaTeX workspace for scientific writing, 2026. https://prism.openai.com, Accessed: 2026-04-27.

OpenReview. OpenReview website, 2026. https://openreview.net/, Accessed: 2026-04-29.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Reddit. [D] NeurIPS 2025 rebuttals, 2025. https://www.reddit.com/r/MachineLearning/comments/1mea5g0/d_neurips_2025_rebuttals/, Accessed: 2026-04-28.

Reddit Community. [d] possible solutions after the iclr 2026 identity-leak incident, 2026a. https://www.reddit.com/r/MachineLearning/comments/1p95g7t/d_possible_solutions_after_the_iclr_2026/.

Reddit Community. [D] IJCAI-ECAI 2026 piloting "primary paper" and secondary artifacts. Reddit r/MachineLearning, 2026b. https://www.reddit.com/r/MachineLearning/comments/1pe8rix/d_ijcaiecai_2026_piloting_primary_paper_and, Accessed: 2026-04-29.

Reddit Community. [D] NeurIPS Camera-ready Checklist. Reddit r/MachineLearning, 2026c. https://www.reddit.com/r/MachineLearning/comments/1oc52j1/d_neurips_camerarady_checklist/, Accessed: 2026-04-28.

Reddit Community. ICLR 2027 de-anonymization [D], 2026d. https://www.reddit.com/r/MachineLearning/comments/1wptsvx/iclr_2027_de_anonymization_d/, Reddit r/MachineLearning discussion, accessed September 26, 2026.

Retraction Watch. The retraction watch leaderboard, 2025. https://www.reddit.com/r/singapore/comments/1lw0sxy/academic_paper_by_nus_researchers_withdrawn_from/, Accessed: 2026-04-28.

Retraction Watch. The Retraction Watch Leaderboard, 2026. https://retractionwatch.com/the-retraction-watch-leaderboard/, Accessed: 2026-04-29.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Momentum residual neural networks. In *International Conference on Machine Learning*, pp. 9276–9287. PMLR, 2021.

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

SIGSAC. Sigsac committee on preserving professional conduct and academic ethics (protect), 2024. https://www.sigsac.org/protect/, Accessed: 2026-04-28.

Buxin Su, Natalie Collina, Garrett G. Wen, Didong Li, Kyunghyun Cho, Jianqing Fan, Bingxin Zhao, and Weijie J Su. How to find fantastic ai papers: Self-rankings as a powerful predictor of scientific impact beyond peer review. 2025. URL https://api.semanticscholar.org/CorpusID:281725172.

Dawei Sun, Anbang Yao, Aojun Zhou, and Hao Zhao. Deeply-supervised knowledge synergy. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6990–6999, 2019. URL https://api.semanticscholar.org/CorpusID:173990799.

Chenhao Tan. How much science is verifiable? results from replicating ICML 2026 oral papers, 2026. https://x.com/ChenhaoTan/article/2079969545629118737, accessed September 25, 2026.

TensorFlow. Tensorboard, 2026. https://github.com/tensorflow/tensorboard, Accessed: 2026-04-27.

VLDB. VLDB 2026 Call for Research Track, 2026. https://www.vldb.org/2026/call-for-research-track.html, Accessed: 2026-04-29.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. Autosurvey: Large language models can automatically write surveys. *Neural Information Processing Systems*, 2024.

Danfeng Yao. Rebuttal how-to: Strategies, tactics, and the big picture in research. *Communications of the ACM*, 67(1):47–49, 2023.

Yuhan Zhang, Xiao Ma, Kun Huang, Mingchao Li, and Pheng-Ann Heng. Semantic-oriented visual prompt learning for diabetic retinopathy grading on fundus images. *IEEE Transactions on Medical Imaging*, 43:2960–2969, 2024. URL https://api.semanticscholar.org/CorpusID:268873308.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, 124:103332, 2025.

Table 2: Comparison of review scores for minimally revised papers rejected by ICLR 2026 and subsequently accepted by ICML 2026.

	Number	**ICML Score**	**ICLR Score**	**Pearson r**	**Spearman ρ**
All Papers	731	4.15	4.41	-0.008	-0.014
Spotlight	55	4.59	4.38	0.038	0.064
Poster	676	4.12	4.41	-0.007	-0.020

A Empirical Evidence of Review Outcome Variability

To further examine the variability of peer-review outcomes across major AI conferences, we collect 777 papers that were rejected by ICLR 2026 and subsequently accepted by ICML 2026. Among these papers, 718 were accepted as Posters and 59 as Spotlights at ICML. We further identify 731 papers that underwent only minimal revisions between the two submissions, allowing us to reduce the potential confounding effect of substantial manuscript changes and more directly compare their review outcomes across the two venues.

Table 2 summarizes the results. Across all 731 papers, the average review score is 4.41 at ICLR and 4.15 at ICML. Interestingly, papers that were eventually accepted as Spotlights at ICML received an average ICLR score of 4.38, slightly lower than the 4.41 average of papers accepted as ICML Posters. This discrepancy suggests that review scores at one venue may provide limited indication of how the same work will be evaluated at another venue.

More importantly, we observe almost no correlation between the review scores assigned by the two conferences. Across all papers, the Pearson correlation is $r = -0.008$ and the Spearman correlation is $\rho = -0.014$. The correlations remain close to zero when considering Spotlight and Poster papers separately, with all absolute correlation coefficients below 0.07. These results provide quantitative evidence of substantial variability and weak cross-venue consistency in the evaluation of essentially unchanged papers, highlighting the stochastic component of current conference reviewing processes.

B Discussions

Our method may reduce the number of submissions and reviewers in the short term, but they help guarantee the quality of papers in the LLM era in the long term. Moreover, our method is generalizable to almost all AI academic conferences, such as NeurIPS, ICML, ICLR, and AAAI, which helps better maintain the academic reputation of conferences. Our method can be combined with other existing methods. For example, papers from benchmark track can help verify reported results.

Of course, our method remains incomplete. For instance, it still fails to strictly avoid collusion and plagiarism. The research on conferences in other subfields in this paper is likely incomplete. Furthermore, this method may not be very effective for conferences in other subfields. We can investigate more conferences and reviewing methods in the future. Our method only mitigates the problem of the vicious cycle and cannot resolve it completely. The effect may be more significant for higher-level papers, such as oral and spotlight. However, due to the massive volume of papers, it is difficult to recruit enough AE reviewers to complete AE for accepted Poster papers as well.

Public naming will be strictly confined to cases with indisputable evidence of deliberate plagiarism or collusion. To eliminate legal ambiguity, we propose integrating a pre-submission consent agreement, where authors and reviewers explicitly agree to potential disclosure if grave misconduct is conclusively proven. We also acknowledge that blanket multi-year bans are overly punitive for first-time or accidental infractions. Based on the specific requirements from the conference organizer, we could revise our policy to: 1. First-time noncompliance (e.g., missing code release): Paper rejection alone without submission bans; 2. Repeat offenses: An escalating submission ban structure applied only to persistent offenders across multiple venues; 3. Proportionality: Clear distinction between accidental negligence (handled via formal warnings) vs. intentional malpractice. Furthermore, investigatory and punitive resources will be strictly focused on high-severity cases backed by indisputable evidence (e.g., clear plagiarism or collusion), rather than expending limited resources on ambiguous edge cases.

We also recognize that tasks such as verifying whether authors and reviewers have fulfilled revision requirements and determining whether a review is of sufficiently low quality to be disregarded may introduce additional workload for ACs, SACs, and PCs. LLM-based agents and supporting tools could help reduce this burden by assisting with evidence collection, consistency checks, and preliminary assessments, thereby providing more objective support for resolving disagreements.

Inspired by your insight, we plan to explore automated methods to combat plagiarism in future work, such as deploying specialized AI agents to analyze textual and figure similarity against existing literature, proactively flagging high-risk candidates for ACs and reviewers to inspect. Similarly, future conferences could leverage AI-based methods to automate plagiarism detection and enforcement. Dedicated AI agents could analyze textual and graphical similarities with existing literature and proactively flag high-risk cases for further inspection by program committees and reviewers, thereby substantially reducing their workload.

Conference submission and review platforms could also undergo multiple rounds of stress testing to reduce risks such as reviewer identity leakage and other system vulnerabilities.