Evaluating AI Reviewers' Ability to Assess Soundness for Deployment in the TMLR Journal

Vishisht Rao and Nihar B. Shah

June 30, 2026

Abstract

We evaluate candidate AI reviewers for a planned experimental deployment at the Transactions on Machine Learning Research (TMLR) journal, in which an AI-generated review will be included alongside standard human reviews. We focus on whether AI reviewers can assess the soundness of submitted papers, because checking soundness is well suited for AI reviewers and is also directly aligned with TMLR's primary acceptance criterion. We evaluate four candidate AI reviewers on recent TMLR submissions with Action Editor decisions and on papers containing inserted claim-breaking errors. We find that soundness-focused AI reviews can provide useful assistance in real peer-review workflows. Based on these evaluations and manual inspection of generated reviews, we identify a candidate system suitable for experimental deployment at TMLR.

1 Introduction

Peer review is central to maintaining the quality and reliability of scientific research. Over the past few years, the capabilities of AI for writing reviews for scientific papers have grown significantly. This raises the question of whether AI-generated reviews can be useful inside real reviewing workflows, and how they can be used best keeping in mind the necessity of human-written reviews. We study these questions from the point of view of an actual proposed deployment — in the Transactions on Machine Learning Research (TMLR) journal. TMLR is known for a high-quality review process [Cho et al., 2024, Survey Section] and is looking to retain this quality while preventing excessive burdens on reviewers and action editors (AEs). Our approach follows Shah [2025a, Part 1]:

(i) Emphasizing conducting evaluations of AI reviewers; and

(ii) Focusing on AI evaluating the soundness of submitted papers.

We discuss this approach below, followed by our actual evaluation methods and results.

There are several possible approaches to evaluate the quality of AI-generated reviews. One popular approach is to generate reviews of previously published papers and compare the AI-generated scores with past human review scores. This approach has been taken in many past works [Yuan et al., 2022, Checco et al., 2021, Idahl and Ahmadi, 2025, Shcherbiak et al., 2024, Thelwall and Yaghi, 2025, Chitale et al., 2025, Shin et al., 2025]. Importantly, these works perform such evaluations based on subjective human reviews, that is, where the reviews were a combination of a variety of criteria such as potential impact and perceived interestingness. These human review scores on such subjective criteria are themselves highly noisy and potentially biased, and may not reflect the true objectives of peer review.

Here is a concrete example via an experiment conducted at a top-tier ML/AI conference by the second author of this report. A key objective of peer review is to identify whether the submitted paper is technically sound. To measure the extent to which human-provided reviews meet this objective, the experimenter created three versions of a paper, and inserted one fatal flaw in each version. The papers comprised approximately 8 pages of main text, with no supplementary material. The program chairs of the conference, who were collaborating in this experiment, obtained a total of 79 reviews for these papers (approximately 26 for each

version). The primary outcome variable was the number of reviews that detected the flaw. Only one out of the 79 reviews pointed out the flaw, and very few reviews had comments on the methods. Furthermore, all reviews were annotated on various review quality indices. It was found that the review quality was uncorrelated with the confidence and expertise self-reported by the reviewers. See Shah [2025b, Section 10.1.2] for more details.

A second common approach for evaluating AI reviews is to ask humans to judge the AI-generated reviews subjectively [Liang et al., 2023, D'Arcy et al., 2024, Tyser et al., 2024]. Some works taking this approach ask authors of the reviewed papers for their evaluation of the reviews. However, a large-scale experiment Goldberg et al. [2025] at the NeurIPS 2022 conference found that authors are (perhaps unsurprisingly) significantly biased towards reviews that are positive towards their paper. Alternatively, one may ask other experts in the field to evaluate the reviews. A randomized controlled trial at the same conference Goldberg et al. [2025] found that evaluations of reviews are significantly affected by the style of the reviews. In particular, in this experiment, the evaluator was either shown the original review or a version of the review that was made two to three times as long but without any additional information. Evaluators rated the longer reviews much better, not just in terms of overall quality, but also rated the longer reviews as exhibiting better understanding of the paper, showcasing better coverage of the paper, substantiating criticisms better and being more constructive. All in all, these results suggest that both past review scores and subjective evaluations of reviews may fail to measure whether a review correctly assesses the technical soundness of a paper.

A third approach is to objectively evaluate AI reviewers against the goals of peer review. One key goal of peer review is to ensure soundness of published research. There is evidence that AI reviews can help support this goal. In the experiment described above (three versions of a paper with 79 reviewers, [Shah, 2025b, Section 10.1.2]), GPT-4 was the frontier model at the time and was evaluated on the inserted errors. This allowed for a comparison between human reviewers in a real review process and an AI reviewer on the same soundness-checking task [Shah, 2025b, Section 11.3.2]. When GPT-4 was asked specifically to evaluate the correctness of individual results, it identified one error consistently, identified another error only with additional prompting, and did not identify the third error. In another study [Liu and Shah, 2023], GPT-4 performed better on focused reviewing tasks than on generic review generation, and detected inserted errors in 7 out of 13 short computer science papers. These results motivate using error detection as a central component in evaluating AI reviewers. Recent benchmarks [Xi et al., 2025] also focus on the error detection ability of AI reviewers.

With this motivation, we evaluate the quality of AI-generated reviews by focusing directly on the core goal of peer review: checking soundness of submitted papers. Fortunately, this objective of peer review is perfectly aligned with the objectives of the Transactions on Machine Learning Research (TMLR). The primary criterion for acceptance in TMLR is soundness of the submission: *Are the claims made in the submission supported by accurate, convincing and clear evidence?*

In our evaluations, we consider four candidate AI reviewers (CSPaper's AI Reviewer [Cao et al., 2025], ReviewerToo's AI Reviewer [Sahu et al., 2025], Anthropic's Claude Opus 4.7, and OpenAI's GPT-5.5) and evaluate them across recent TMLR submissions and papers with inserted claim-breaking errors. Our goal is to select the best performing AI reviewer to deploy in TMLR, and subsequently get broader feedback from all participants (focused on whether claims in a submission are met with evidence, and not other criteria).

While our work is focused on deployment at TMLR, the same motivation applies more broadly to conferences and other journals. The review processes in these venues ask reviewers to assess several criteria at once. However, the experiment described above (three versions of a paper with 79 reviewers, [Shah, 2025b, Section 10.1.2]) found that reviews in such a top-tier conference hardly contain substantial comments on the methods, and tend to focus more on the problem and final results. A soundness-focused AI review can help fill this void and complement human reviewers.

2 Methods

In this section, we describe the construction of the evaluation dataset and the four AI reviewers we evaluate.

2.1 Evaluation Dataset

We construct two types of evaluation data for the claims and evidence criterion. The first consists of recent TMLR submissions for which we have access to the final reviews and decisions. These papers allow us to evaluate whether AI reviewers align with the claims and evidence judgment made in the actual TMLR review process. The second follows the FLAWS methodology [Xi et al., 2025], where we insert claim-breaking errors into published papers and evaluate whether AI reviewers can identify the subtle but important errors in the resulting manuscripts.

For the TMLR submissions, the Action Editor (AE) must provide a decision on each paper, based on the reviews and their own judgment. The primary contributor to this decision is a "yes"/"no" answer to the question of whether the claims made in the submission are supported by accurate, convincing, and clear evidence. We use this AE response as the ground truth label for the claims and evidence criterion. The use of this data as a label has three key differences from the generic use of conference review scores criticized earlier. First, conference reviews include subjective criteria such as novelty, interest, and reviewer taste in addition to soundness. It is also known that when asked to evaluate multiple criteria, evaluators perform poorly on individual criteria, compared to when asked for just one criterion at a time [Lane et al., 2024]. Our criterion is more objective, and the absence of other criteria helps focus the review on this criterion of soundness. Second, conferences require reviewers to review 3-6 papers in a short duration whereas TMLR only assigns reviewers one paper at a time, thereby allowing reviewers to spend more time on a paper. Third, TMLR is known for its consistently higher quality reviews as compared to top-tier conferences in the field. Due to these reasons, we use the AE's decisions on the claims and evidence criterion as labels.

For our evaluations, we consider three categories of TMLR papers, detailed below. We select the 50 most recent TMLR papers in each category, with decisions released as of April 2026. The first category comprises papers that are `accepted` for publication in TMLR. A necessary condition for publication is that the AE answered "yes" to the claims and evidence question. The second category is the `claims_only_no` category, which consists of papers rejected by TMLR, where the AE has answered "no" to the claims and evidence question (but yes to the second question of whether the paper has a relevant audience in the TMLR community). The third is the `both_no` category, which consists of papers rejected by TMLR, where the AE has answered "no" to both questions. In all three categories, we compare the AI reviewer's answer to the AE's answer to the claims and evidence question.

Before constructing this evaluation set, we randomly selected a separate calibration set from each of the above three categories, used to refine the AI reviewer prompts. We drew this calibration set randomly, without reference to AE labels, decisions, or model outputs, and is disjoint from the evaluation set described above. Because the selection was random and done before we looked at any labels, the excluded papers are not biased toward any particular outcome. Keeping the two sets disjoint also means there is no leakage between calibration and evaluation.

The other type of evaluation data we construct is a new version of the FLAWS dataset [Xi et al., 2025]. The original FLAWS dataset was released publicly in December 2025. Hence, to avoid data contamination, we construct a new version. Here, we take papers published in NeurIPS 2025 and insert a subtle but claim-breaking error into these manuscripts. We construct 67 such papers using Gemini 2.5 Pro and 50 such papers using OpenAI GPT-5. These papers with inserted errors constitute a new data set that has never been seen before by any AI reviewer. We use the answer to the claims and evidence question as a signal, where the ground truth answer to this is "no" since the error inserted specifically breaks a claim made in the paper. We also manually annotate some of the reviews to check whether the reviews that marked "no" to the claims and evidence question actually identified the inserted error or marked it "no" for an unrelated reason, and to check whether the reviews that marked "yes" genuinely missed the inserted error as opposed to finding it too trivial to mark as "no".

2.2 AI Reviewers

We consider four AI reviewers, consisting of two AI-review systems whose pipelines have been built specifically to review papers, and two general-purpose AI models.

The two specialized AI-review systems are CSPaper's AI reviewer [Cao et al., 2025] and ReviewerToo's AI reviewer [Sahu et al., 2025]. As the pipelines for these reviewers were built by CSPaper and ReviewerToo respectively, we requested API access to their models, which allowed us to generate reviews in the TMLR review format by uploading a PDF of the manuscript. They were informed of the nature of the experiments we were looking to conduct. Since they provided us with the models, we did not have to input additional prompts to generate the reviews.

The two general-purpose AI models are Anthropic's Claude Opus 4.7 and OpenAI's GPT-5.5. These two models have been shown to top scientific research benchmarks [Bragg et al., 2026] and other reasoning and intelligence benchmarks. In both these models, we set the reasoning effort to 'high' and prompted them to produce a review of the presented manuscript as per the TMLR review guidelines. The prompts used can be found in the GitHub repository[1].

The prompts were constructed through an iterative process using TMLR Action Editor (AE) reviews from a separate set of papers, disjoint from those used in the final evaluations, for calibration. The starting point for both prompts was the same. We described the two TMLR acceptance criteria, emphasized the distinction between a fundamental gap and an addressable weakness, and included calibration examples based on real AE decisions. These examples were chosen to cover both straightforward cases and more subtle cases where AI reviewers may make mistakes, such as being overly positive about papers with weak baselines or overly negative about papers with secondary technical issues. We refined the prompts by running each model on a labeled set of accepted and rejected papers and inspecting the resulting errors. In particular, we looked for false positives, where rejected papers were incorrectly judged positively, and false negatives, where accepted papers were incorrectly judged negatively. When consistent error patterns appeared, we revised the prompt to better specify the intended TMLR-style judgement. The final prompts for the two models differed in structure, reflecting the prompting guidance and observed behavior of each model.

The calibration set consisted of 9, 6, and 4 papers from the `accepted`, `claims_only_no`, and `both_no` categories respectively, and prompt refinement was stopped when no consistent error patterns remained on this set. We note that this iterative refinement means the prompts for GPT-5.5 and Opus 4.7 are, to some degree, calibrated to AE judgment. The strict disjointness between calibration and evaluation sets is the primary safeguard against this biasing the reported results. For CSPaper, we did not modify any prompts, though its pipeline may itself have been tuned to similar reviewing rubrics during its own development, which should be kept in mind when interpreting its stronger alignment with TMLR outcomes.

We run the four models on the categories of papers described in Section 2.1. For the main comparison between CSPaper, GPT-5.5, and Opus 4.7, we report results only on a common held-out evaluation subset. Papers whose reviews or decisions were used during prompt calibration were excluded from this subset. ReviewerToo is reported separately as a partial evaluation because reliability and runtime constraints prevented us from running it on the full set of categories.

3 Results and Discussion

Table 1 contains evaluation results on the categories where the ground truth answer to the claims and evidence criterion is known. The percentages correspond to the percentage of responses where the AI reviewer answered "yes" to the claims and evidence question.

We were unable to include the results from ReviewerToo in the final evaluation. In initial experiments, the ReviewerToo pipeline repeatedly failed even on small batches of papers. After communication with the ReviewerToo team, they made changes that allowed us to run the system on a subset of papers. However, these runs took approximately one day, compared to roughly ten minutes for each of the other AI reviewers on comparable batches. Because of this substantially higher runtime and the earlier reliability issues, we decided not to include ReviewerToo in the full evaluation.

In Table 1, we see that on `accepted` papers, CSPaper has the highest accuracy, while GPT-5.5 and Opus 4.7 are tied. On clear claims and evidence failures (`claims_only_no` and `both_no`), CSPaper is much more

[1] `https://github.com/Vishisht-rao/ai-reviewer-evals`

AI reviewer	`accepted` Yes ↑	`claims_only_no` No ↓	`both_no` No ↓	`flaws_gemini` No ↓	`flaws_openai` No ↓
CSPaper	**92.7%** ($n = 41$)	**9.1%** ($n = 44$)	**0.0%** ($n = 46$)	82.1% ($n = 67$)	**70.0%** ($n = 50$)
GPT-5.5	82.9% ($n = 41$)	65.9% ($n = 44$)	32.6% ($n = 46$)	**76.1%** ($n = 67$)	80.0% ($n = 50$)
Opus 4.7	82.9% ($n = 41$)	34.1% ($n = 44$)	23.9% ($n = 46$)	86.6% ($n = 67$)	94.0% ($n = 50$)
ReviewerToo*	59.4% ($n = 32$)	–	20.0% ($n = 45$)	–	–

Table 1: Evaluations on categories where ground truth labels on the claims and evidence question are available (given by "Yes"/"No" in column headers). Each entry reports the percentage of reviewed papers for which the AI reviewer answered "yes" to the claims and evidence criterion. The realized n in the `accepted`, `claims_only_no`, and `both_no` categories falls below 50 because of the calibration exclusion described in Section 2.1. *ReviewerToo was only run on a subset of categories due to reliability and runtime challenges, and is therefore included only as a partial evaluation rather than as part of the main comparison. Further analysis of the results from the FLAWS dataset is present in Table 2.

conservative and closer to the expected "no" label. GPT-5.5 and Opus 4.7 are more likely to incorrectly answer "yes" on papers which contain issues that resemble rejected TMLR papers. On FLAWS papers, all models have low accuracy. This is not surprising, since these papers contain claim-invalidating errors that are designed to be subtle and substantive. The FLAWS construction process explicitly filters out invalid, trivial, or superficial errors, leaving errors that are challenging to identify [Xi et al., 2025].

To examine the results from the FLAWS papers in more detail, we manually annotate reviews on the `flaws_gemini` category. We consider 20 reviews where the AI reviewer answered "yes" to the claims and evidence question and all reviews where the AI reviewer answered "no" to the claims and evidence question. For reviews where the model answered "yes", we distinguish between cases where the model identified the inserted error but dismissed it as trivial, and cases where the model missed the inserted error. For reviews where the model answered "no", we distinguish between cases where the model identified the inserted error with the correct explanation, identified the inserted error with the wrong explanation, or answered "no" for an unrelated reason. The results can be seen in Table 2.

	AI answered "Yes"			AI answered "No"			
AI Reviewer	n	Found error, called trivial ↑	Missed error ↓	n	Found error, correct reason ↑	Found error, wrong reason ↑	No for other reason ↓
CSPaper	20	70.0%	30.0%	12	25.0%	58.3%	16.7%
GPT-5.5	20	40.0%	60.0%	16	43.8%	25.0%	31.2%
Opus 4.7	20	25.0%	75.0%	9	11.1%	77.8%	11.1%

Table 2: Manual annotation of AI-generated reviews on the `flaws_gemini` category. We annotate 20 reviews where each AI reviewer answered "yes" to the claims and evidence question, and all reviews where the AI reviewer answered "no". Percentages under "AI answered Yes" and "AI answered No" are computed over n reviews.

Unlike the original FLAWS evaluation [Xi et al., 2025], which prompts models to output ranked candidate error excerpts and then compares those excerpts to ground truth error locations, our evaluation prompts each AI system to generate a full review in the TMLR format. Therefore, the model outputs are not directly comparable to the excerpt based FLAWS outputs, and we use manual annotation of the generated reviews rather than the automated FLAWS excerpt-matching metric.

These annotations show that the FLAWS results can be interpreted in two ways. Under a strict interpretation, where the AI reviewer is used as a standalone reviewer, the desired behavior is to answer "no" and correctly explain the inserted error. On the `flaws_gemini` set, this occurs for 3 papers for CSPaper, 7 papers for GPT-5.5, and 1 paper for Opus 4.7, corresponding to 4.5%, 10.4%, and 1.5% of the 67 papers respectively.

Under this interpretation, GPT-5.5 performs the best. However, the authors of the FLAWS paper motivate the design of their dataset as a way to evaluate AI reviewers that can assist and be used alongside human reviewers. Under this more lenient assistance interpretation, cases where the model identifies the inserted error but classifies it as trivial, or identifies the error but gives the wrong explanation, may still be useful because they bring the human reviewer's attention to the relevant part of the paper. From this perspective, CSPaper appears more useful than the strict metric alone suggests. It identifies the inserted error in 70% of the sampled reviews where the AI reviewer answered "yes", even though it often dismisses the error as trivial, and in the "no" reviews it points to the inserted error more often than it gives an unrelated reason.

To check whether these differences may be partly explained by verbosity, we also measured the length of the claims and evidence text used in the evaluation. On `flaws_gemini`, the median claims and evidence length was 400 words for CSPaper, 254 words for GPT-5.5, and 323 words for Opus 4.7. On `flaws_openai`, the corresponding medians were 419, 256, and 310 words. Thus, CSPaper's higher lenient error identification rate should be interpreted in light of its longer claims and evidence comments. However, review length alone does not measure precision. Assessing whether a model raises many false, trivial, or otherwise non-meaningful issues would require annotating all points raised in the claims and evidence section of the review.

Finally, the last author, who is also a co-Editor-in-Chief of TMLR, read through several reviews provided by CSPaper. Based on all this evidence, we believe that this system is suitable for an experimental deployment.

4 Conclusions

While no system emerges as an unambiguous winner across all of our evaluations, CSPaper is the most suitable candidate for the deployment we envision, which would be in parallel with human reviewers and with further feedback to be collected. ReviewerToo was ruled out early for reliability and runtime reasons, and Opus 4.7 was dominated by CSPaper across our metrics. In comparison with GPT-5.5, CSPaper agreed more consistently with prior TMLR outcomes, even though none of the systems is a strong detector of the inserted errors in the FLAWS data. Moreover, the manual annotation in Table 2 indicates that even when CSPaper errs on the FLAWS papers, it frequently brings up the part of the manuscript containing the inserted error. A read-through of several CSPaper reviews by the last author, a co-Editor-in-Chief of TMLR, found them to be of sufficient quality to move forward with this experimental deployment.

Recent work has begun to study AI assistance in real peer review workflows. At NeurIPS 2024, an AI checklist assistant was deployed to help authors identify issues in their checklist responses before submission [Goldberg et al., 2024]. At ICLR 2025, a randomized study tested whether AI-generated feedback could help reviewers improve the clarity and actionability of their reviews [Thakkar et al., 2025]. At AAAI-26, an AI-generated review was produced for every main-track submission that entered the full-review stage [Biswas et al., 2026]. These studies show that the community is increasingly moving from offline evaluation of AI reviewers to field experiments in real review processes. This makes it important to evaluate AI reviewers on criteria that directly reflect the objectives of peer review, as we do here.

Our goal of verifying soundness of submissions is also essential for conferences and other journals. Evaluations and subsequent deployments like ours that focus on soundness can help complement current conference reviews. It can also be a part of broader changes to the review process like the four-step process proposed in [Shah, 2026] that has soundness-verification as an initial step.

We do not consider AI reviewers for evaluating papers on subjective criteria, such as excitingness, for four reasons. First, such criteria are not well aligned with TMLR's acceptance criteria, which primarily emphasize soundness. Second, although TMLR also includes a secondary criterion on whether a paper is likely to find a relevant audience within the TMLR community, our pilot experiments found that AI reviewers aligned poorly with human reviewer opinions on this dimension. Third, using AI reviewers to assess subjective criteria raises broader questions that require further community discussion. For example, should excitingness be judged by humans, given that it is inherently about human perception Shah [2026]? How should we mitigate risks such as reduced diversity of opinions or "hivemind" effects Baumann et al. [2026]? And fourth, given our focus on soundness, we also think it is less susceptible to adversarial attacks such as [Lin et al., 2025, Hsieh

et al., 2025, Sugiyama and Eguchi, 2025, Rao et al., 2025], although we have not evaluated the adversarial attack aspect yet.

We look forward to an experimental deployment of the AI reviewer in TMLR. Our next steps are to integrate it with OpenReview to provide one AI-review per submission, and then survey AEs, reviewers, and authors about their perceptions of the AI review. Going ahead, it is also of interest to move beyond review of just the paper, to reviewing associated data/code and other intermediate artifacts. Given the increasing use of AI for automation in research, there are new challenges in verifying whether evidence supports the claims, since some issues may not be detectable from the paper alone and may require intermediate traces and code [Luo et al., 2025]. Our future work will aim to tackle these new challenges.

Acknowledgments

We thank Celeste Martinez Gomez and OpenReview for working with us on the LLM reviewer deployment. We thank Kevin Wu, James Zou, and former TMLR EiC Hugo Larochelle for helpful discussions in the initial parts of this project. This evaluation was funded by NSF 1942124 and a Gemini Academic Program Award.

References

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation. In *International Conference on Machine Learning*, 2026.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. AI-assisted peer review at scale: The AAAI-26 AI review pilot, 2026.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefan Candra, Jason Dunkelberger, Dan Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Cecile Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. AstaBench: Rigorous benchmarking of AI agents with a scientific research suite. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=M7TNf5J26u`.

Lele Cao, Lei You, and CSPaper Review R&D Team. CSPaper Review: Fast, rubric-faithful conference feedback. In *Proceedings of the 18th International Natural Language Generation Conference: System Demonstrations*, pages 3–7, 2025. URL `https://aclanthology.org/2025.inlg-demos.2/`. Website: `https://cspaper.org/`.

Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(25), 2021. doi: 10.1057/s41599-020-00703-8.

Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, and Vasudeva Varma. AutoRev: Automatic peer review system for academic research papers, 2025.

Kyunghyun Cho, Raia Hadsell, Gautam Kamath, Hugo Larochelle, and Paul Vicol. 2023 TMLR Annual Report. `https://docs.google.com/document/d/13mknNOEvwkibxxrES-RrvgbWIuXY6u0PpMjnMrvp6h0/edit`, February 2024. Released February 16, 2024. Accessed June 13, 2026.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers, 2024.

Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. Usefulness of LLMs as an author checklist assistant for scientific papers: NeurIPS'24 experiment, 2024.

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. doi: 10.1371/journal.pone.0320444. URL `https://doi.org/10.1371/journal.pone.0320444`.

Janet Hsieh, Aditi Raghunathan, Nihar B Shah, et al. Vulnerability of text-matching in ml/ai conference reviewer assignments to collusions. In *USENIX Security*, 2025.

Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, 2025.

Jacqueline N. Lane, Simon Friis, Tianxi Cai, Michael Menietti, Griffin Weber, and Eva C. Guinan. Greenlighting innovative projects: How evaluation format shapes the perceived feasibility of novel ideas. Harvard Business School Technology & Operations Management Unit Working Paper, 2024. URL `https://ssrn.com/abstract=4769282`.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? A large-scale empirical analysis, 2023.

Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, Philip S Yu, and Hong-Han Shuai. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. *arXiv preprint arXiv:2506.11113*, 2025.

Ryan Liu and Nihar B. Shah. ReviewerGPT? An exploratory study on using large language models for paper reviewing, 2023.

Ziming Luo, Atoosa Kasirzadeh, and Nihar B. Shah. The More You Automate, the Less You See: Hidden pitfalls of AI scientist systems, 2025. URL `https://arxiv.org/abs/2509.08713`.

Vishisht Srihari Rao, Aounon Kumar, Himabindu Lakkaraju, and Nihar B Shah. Detecting llm-generated peer reviews. *PLoS One*, 20(9):e0331871, 2025.

Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. ReviewerToo: Should AI join the program committee? A look at the future of peer review, 2025. URL `https://reviewertoo.org/`. Website: `https://reviewertoo.org/`; arXiv: `https://arxiv.org/abs/2510.08867`.

Nihar B. Shah. AI Meets Peer Review: The Good, The Bad, and The Ugly. Invited talk at The Role of AI in Scientific Peer Review, NeurIPS 2025 Social, December 2025a. URL `https://cs.cmu.edu/~nihars/tutorials/AI_meets_Peer_Review_2025.pdf`. Presented on December 3, 2025.

Nihar B. Shah. An overview of challenges, experiments, and computational solutions in peer review. `https://www.cs.cmu.edu/~nihars/preprints/SurveyPeerReview.pdf`, 2025b. Extended version; current version dated August 7, 2025.

Nihar B. Shah. Position: Peer review in ML/AI conferences should separate publication from presentation and offer non-anonymous review tracks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Anna Shcherbiak, Hooman Habibnia, Robert Böhm, and Susann Fiedler. Evaluating science: A comparison of human and AI reviewers. *Judgment and Decision Making*, 19:e21, 2024. doi: 10.1017/jdm.2024.24.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. Automatically evaluating the paper reviewing capability of large language models, 2025.

Shogo Sugiyama and Ryosuke Eguchi. Positive review only: Researchers hide ai prompts in papers. `https://asia.nikkei.com/business/technology/artificial-intelligence/positive-review-only-researchers-hide-ai-prompts-in-papers`, 2025.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025, 2025.

Mike Thelwall and Abdallah Yaghi. Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms. *Scientometrics*, 130:5285–5307, 2025. doi: 10.1007/s11192-025-05287-1.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. AI-driven review systems: Evaluating LLMs in scalable and bias-aware academic reviews, 2024.

Sarina Xi, Vishisht Rao, Justin Payan, and Nihar B. Shah. FLAWS: A benchmark for error identification and localization in scientific papers, 2025. URL `https://arxiv.org/abs/2511.21843`.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.