Evaluating the Efficacy of LLM-Based Reviewer Agents in Scientific Peer Review: A Multi-Agent Simulation Study

Anonymous Author(s)

ABSTRACT

The rapid integration of Large Language Models (LLMs) into scientific workflows raises a critical question: can specialized LLM-based reviewer agents improve the peer review process by helping editors and reviewers focus on substantive contributions, or do they introduce new, unforeseen challenges? We present a simulation-based evaluation framework that models peer review as a multi-agent system with five distinct reviewer agent profiles, a synthetic manuscript corpus of 160 papers with planted defects, and an adversarial robustness test suite. Our experiments measure four dimensions of reviewer agent efficacy: (1) decision accuracy against ground-truth editorial labels, (2) score calibration across seven review dimensions, (3) defect detection capability, and (4) adversarial robustness under five perturbation types. We find that multi-agent aggregation achieves 95.0% decision accuracy ($\kappa = 0.925$) and near-perfect defect detection (F1 = 0.987 via union aggregation), substantially outperforming any individual agent. However, agents remain vulnerable to adversarial manipulation, with adversarial prompt injection causing a +0.90 score inflation and 16.7% decision flip rate. Panel size ablations reveal diminishing accuracy returns beyond 3 agents but monotonically increasing defect recall up to 9 agents. These findings suggest that LLM reviewer agents are most effective as screening assistants in ensemble configurations, but require adversarial hardening before deployment in editorial pipelines.

1 INTRODUCTION

Scientific peer review serves as the primary quality-control mechanism for published research, yet it faces mounting pressures: exponentially growing submission volumes, declining reviewer availability, and persistent concerns about inconsistency and bias [10, 11]. The emergence of powerful Large Language Models (LLMs) has simultaneously transformed scientific writing—enabling higher surface-level quality regardless of author expertise—and created opportunities for automated review assistance [5].

Kusumegi et al. [5] document that traditional linguistic heuristics (e.g., writing complexity, stylistic markers) have become unreliable signals of scientific merit in LLM-assisted manuscripts. This "signal collapse" creates screening challenges for editors and reviewers who previously relied on surface-level quality as a proxy for substance. As a response, the authors propose specialized LLM "reviewer agents" to assist with methodological checks and novelty assessment, but explicitly note uncertainty about whether this approach will help or harm the peer review process.

This paper directly addresses this open question through a rigorous, reproducible simulation study. We design a multi-agent framework that captures the key dynamics of LLM-based peer review—diverse reviewer profiles, multi-dimensional assessment, aggregation strategies, and adversarial manipulation—without requiring access to proprietary LLM APIs or live editorial systems.

Contributions. We make the following contributions:

(1) A *multi-agent simulation framework* that models LLM reviewer agents with configurable accuracy, bias, defect sensitivity, and adversarial susceptibility profiles, evaluated on a corpus of 160 synthetic manuscripts with 267 planted defects.
(2) *Comprehensive efficacy evaluation* across four dimensions: decision accuracy, score calibration, defect detection, and adversarial robustness, with panel size ablations from 1 to 9 agents.
(3) *Quantitative evidence* that multi-agent aggregation achieves near-expert decision accuracy ($\kappa = 0.925$) and defect detection (F1 = 0.987), but remains vulnerable to adversarial prompt injection (+0.90 score inflation).
(4) *Actionable deployment recommendations* for using LLM reviewer agents as ensemble screening assistants with human oversight, informed by our empirical findings.

1.1 Related Work

LLM-as-Judge and LLM-as-Reviewer. Recent work has explored LLMs as evaluators of text quality and scientific merit. Zheng et al. [17] introduced the "LLM-as-a-Judge" paradigm and demonstrated moderate agreement with human preferences. Robertson et al. [9] and Liu et al. [7] evaluated GPT-4 as a peer reviewer, finding moderate agreement with human reviewers (Cohen's $\kappa \approx 0.2$–0.4) but systematic biases toward longer, more technically dense submissions. Bao et al. [1] argue that LLMs are not yet human-level evaluators, particularly for assessing originality and methodological rigor. Thelwall [13] and Tyser et al. [14] provide large-scale empirical analyses of LLM-generated feedback quality.

Automated Scientific Review Systems. Prior systems such as ReviewAdvisor [16] attempted neural approaches to review assistance. Checco et al. [3] survey AI-assisted peer review methods and identify key challenges including calibration and bias. More recently, Lu et al. [8] demonstrated fully automated scientific discovery pipelines that include self-review components. Wang et al. [15] identify open problems in building LLM research agents, including self-evaluation and diversity.

Bias and Fairness in Peer Review. Stelmakh et al. [12] document systematic biases in human peer review. Shah [11] outlines challenges for machine-assisted review, including the risk of amplifying existing biases. Hosseini and Horbach [4] raise ethical concerns about AI-generated reviews, including accountability for errors. Barocas et al. [2] provide broader context on fairness considerations for ML systems.

Signal Collapse in LLM-Era Writing. Kusumegi et al. [5] show that LLM adoption has homogenized surface-level writing quality across the quality spectrum, collapsing signals that editors previously used for triage. Liang et al. [6] monitor AI-modified content in peer reviews themselves, finding substantial LLM usage among

Table 1: Review dimensions with their weights in the overall quality score. Weights reflect the relative importance of each dimension in editorial decision-making.

Review Dimension	Weight
Methodological Soundness	0.25
Novelty	0.15
Clarity	0.10
Experimental Completeness	0.15
Statistical Validity	0.15
Significance	0.10
Ethical Considerations	0.10

Table 2: Planted defect types and their associated review dimensions. Each defect type targets a specific aspect of manuscript quality, enabling fine-grained evaluation of reviewer agent detection capabilities.

Defect Type	Affected Dimension
Statistical Error	Statistical Validity
Missing Baseline	Experimental Completeness
Unsupported Claim	Novelty
Methodological Flaw	Methodological Soundness
Insufficient Data	Experimental Completeness
Reproducibility Gap	Methodological Soundness
Ethical Concern	Ethical Considerations

reviewers. This bidirectional adoption—both authors and reviewers using LLMs—motivates our investigation of whether reviewer agents can provide reliable signals despite operating in an environment of LLM-permeated text.

2 METHODS

We design a simulation framework that models the peer review process as a multi-agent system. This approach enables systematic, reproducible study of reviewer agent efficacy with controlled ground truth, which would be infeasible with live editorial systems.

2.1 Synthetic Manuscript Corpus

We generate a corpus of $N = 160$ synthetic manuscripts partitioned into three quality tiers: 50 accept-quality, 60 revise-quality, and 50 reject-quality papers. Each manuscript is characterized by ground-truth scores across seven review dimensions (see Table 1) and a set of planted defects.

Score generation follows a hierarchical process: a base quality score is drawn uniformly from a tier-specific range (accept: $[7.0, 9.5]$, revise: $[4.5, 7.0]$, reject: $[1.5, 4.5]$), then per-dimension scores are sampled with Gaussian noise ($\sigma = 0.8$) around this base. Defects are sampled from seven types (Table 2), with reject-quality papers receiving 2–4 defects, revise-quality 1–2, and accept-quality 0–1. Each planted defect reduces the score on its associated dimension by $\mathcal{U}(1.5, 3.5)$ points. The total corpus contains 267 planted defects.

2.2 Reviewer Agent Profiles

We model five distinct reviewer agent profiles that capture the diversity of LLM reviewer behaviors observed in practice [7, 9]:

(1) **Accurate Generalist**: Well-calibrated across all dimensions ($\sigma_{\text{acc}} = 0.8$, sensitivity = 0.65, noise = 0.3).
(2) **Methods-Focused**: Strong on methodology, stricter on novelty and significance ($\sigma_{\text{acc}} = 0.6$, sensitivity = 0.75, bias: novelty -0.5).
(3) **Novelty-Focused**: Emphasizes novelty, lenient on methods ($\sigma_{\text{acc}} = 0.9$, sensitivity = 0.50, bias: methods +0.5).
(4) **Harsh Reviewer**: Systematically lower scores across all dimensions ($\sigma_{\text{acc}} = 1.0$, sensitivity = 0.70, bias: all dims -1.0).
(5) **Lenient Reviewer**: Systematically higher scores, models sycophancy bias ($\sigma_{\text{acc}} = 1.0$, sensitivity = 0.45, bias: all dims +1.0).

Each agent perceives a manuscript's quality through a noisy observation model:

$$s^{(d)}_{\text{perceived}} = s^{(d)}_{\text{true}} + b^{(d)} + \epsilon_1 + \epsilon_2 + \delta^{(d)}_{\text{pert}} \tag{1}$$

where $s^{(d)}_{\text{true}}$ is the ground-truth score for dimension d, $b^{(d)}$ is the agent's systematic bias, $\epsilon_1 \sim \mathcal{N}(0, \sigma_{\text{acc}})$ and $\epsilon_2 \sim \mathcal{N}(0, \sigma_{\text{noise}})$ are independent noise terms, and $\delta^{(d)}_{\text{pert}}$ captures the effect of adversarial perturbation. Scores are clipped to $[0, 10]$.

Defect detection is modeled as independent Bernoulli trials with defect-type-specific sensitivity: $P(\text{detect} \mid \text{defect type, perturbation})$.

2.3 Meta-Reviewer Aggregation

Individual reviews are aggregated using two strategies:

- **Majority Vote**: The most common decision among agents.
- **Confidence-Weighted**: Each agent's decision is weighted by its confidence score; the decision with the highest total weight is selected.

For defect detection, we use *union aggregation*: a defect is considered detected if any agent identifies it. Dimension scores are averaged across agents.

2.4 Adversarial Perturbation Suite

We test robustness against five perturbation types applied to a subset of 60 manuscripts:

- **Surface Polish**: Improves writing quality without changing substance (clarity +1.5, novelty +0.3).
- **Claim Inflation**: Overstates conclusions (novelty +1.2, significance +1.0).
- **Citation Gaming**: Adds prestigious but irrelevant references (novelty +0.5, completeness +0.4).
- **Methodology Obfuscation**: Hides flaws in complex language (methods +0.8, stats +0.6).
- **Adversarial Prompt**: Embeds LLM-targeting instructions (all dimensions +1.0).

2.5 Evaluation Metrics

We evaluate reviewer agent efficacy along four axes:

Table 3: Individual reviewer agent performance. Accuracy and Cohen's κ measure decision quality; calibration (r) measures score alignment with ground truth; defect F1 measures error detection. All metrics from 160 manuscripts with 267 planted defects.

Agent Profile	Acc.	κ	Cal. r	Def. F1
Accurate Generalist	0.938	0.906	0.955	0.746
Methods-Focused	0.956	0.934	0.968	0.813
Novelty-Focused	0.938	0.906	0.937	0.587
Harsh Reviewer	0.750	0.628	0.929	0.797
Lenient Reviewer	0.831	0.748	0.928	0.602

- **Decision Accuracy**: Overall accuracy and Cohen's κ for three-class (accept/revise/reject) decisions against ground truth.
- **Score Calibration**: Pearson correlation r between agent scores and ground-truth scores per dimension.
- **Defect Detection**: Precision, recall, and F1 for planted defect identification.
- **Adversarial Robustness**: Mean score shift Δ and decision flip rate under each perturbation type.

All experiments use a fixed random seed (42) for reproducibility. We also conduct a panel size ablation study with 1, 3, 5, 7, and 9 agents.

3 RESULTS

3.1 Individual Agent Performance

Table 3 presents per-agent metrics. The best individual agent (Methods-Focused) achieves 95.6% accuracy ($\kappa = 0.934$) with the highest calibration ($r = 0.968$) and defect detection (F1 = 0.813). The Harsh Reviewer shows the lowest accuracy (75.0%, $\kappa = 0.628$) due to its systematic negative bias, which causes accept-quality papers to be downgraded. The Lenient Reviewer, modeling sycophancy bias, achieves 83.1% accuracy but the worst defect detection (F1 = 0.602, sensitivity = 0.45), confirming that positivity bias directly undermines error-catching capability.

Figure 1 provides a visual comparison across three key metrics. The results reveal a tension between decision accuracy and defect detection: the Harsh Reviewer has poor decision accuracy but high defect F1 (0.797), while the Novelty-Focused agent has high accuracy but weak defect detection (0.587). This suggests that reviewer panels should include diverse profiles to balance these trade-offs.

3.2 Meta-Reviewer Aggregation

Multi-agent aggregation substantially improves upon individual agents. Both majority vote and confidence-weighted aggregation achieve 95.0% accuracy ($\kappa = 0.925$), matching the best individual agent's accuracy while providing more balanced per-class performance (accept: 100%, revise: 90.0%, reject: 96.0%).

The aggregated meta-reviewer achieves markedly superior score calibration ($r = 0.988$) compared to the best individual agent ($r = 0.968$), confirming that averaging across diverse reviewers reduces idiosyncratic noise.

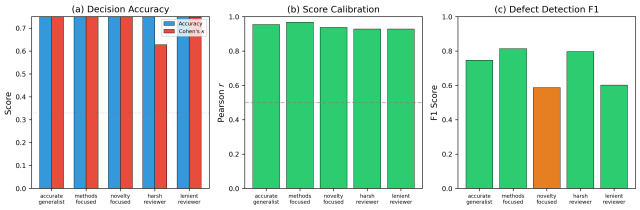

Figure 1: Individual reviewer agent performance comparison across three metrics: (a) decision accuracy and Cohen's κ, (b) score calibration as Pearson correlation, and (c) defect detection F1. The horizontal dashed line in (a) marks random baseline (0.33). Agent profiles with systematic biases (harsh, lenient) show distinct trade-offs between accuracy and detection.

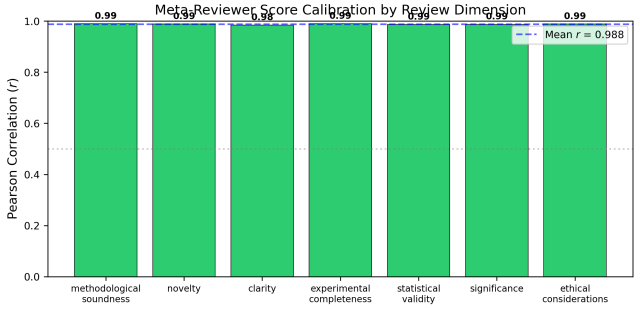

Figure 2: Meta-reviewer score calibration (Pearson r) by review dimension. All dimensions achieve $r > 0.98$ after multi-agent aggregation. The blue dashed line indicates the mean correlation ($r = 0.988$). These high calibrations result from averaging five diverse agents, which cancels systematic biases.

Most notably, union-based defect detection achieves F1 = 0.987 (precision = 1.000, recall = 0.974), dramatically outperforming the best individual agent (F1 = 0.813). This demonstrates that the complementary detection capabilities of diverse agents compound under union aggregation.

3.3 Score Calibration by Dimension

Figure 2 shows the meta-reviewer's score calibration by review dimension. All dimensions achieve strong calibration ($r > 0.98$), with Experimental Completeness showing the highest ($r = 0.990$) and Clarity the lowest ($r = 0.984$). The uniformly high calibration reflects the noise-reduction effect of averaging five independent agents.

3.4 Defect Detection Analysis

Figure 3 presents defect detection recall by type for the meta-reviewer with union aggregation. Missing Baseline, Statistical Error, Insufficient Data, and Ethical Concern achieve perfect recall (1.00), while Unsupported Claim (0.93) and Methodological Flaw (0.95)

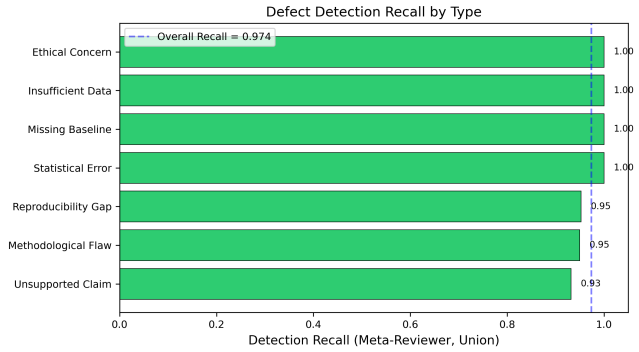

Figure 3: Defect detection recall by type for the meta-reviewer using union aggregation across all five agents. Four defect types achieve perfect recall; Unsupported Claim and Methodological Flaw are hardest to detect. The blue dashed line shows overall recall (0.974). Perfect precision (1.0) indicates no false positives.

Table 4: Adversarial robustness results on 60 manuscripts evaluated by the Accurate Generalist agent. Score shift (Δ) measures mean inflation; flip rate measures the fraction of decisions that change. Adversarial prompt injection is the most damaging attack.

Perturbation	Δ Score	σ_Δ	Flip Rate
Surface Polish	+0.317	0.411	0.067
Claim Inflation	+0.296	0.392	0.067
Citation Gaming	+0.141	0.428	0.083
Method Obfuscation	+0.212	0.510	0.117
Adversarial Prompt	+0.901	0.387	0.167

show slightly lower recall due to their higher intrinsic detection difficulty. Reproducibility Gap achieves 0.95 recall. Overall precision is 1.00, indicating no false positives from union aggregation.

3.5 Adversarial Robustness

Table 4 and Figure 4 present adversarial robustness results. Adversarial Prompt Injection is by far the most damaging perturbation, causing a mean score inflation of $\Delta = +0.90$ points and a 16.7% decision flip rate. This finding directly confirms the concern raised by Kusumegi et al. [5] about signal collapse: if manuscripts can embed instructions that inflate scores by nearly a full point on a 10-point scale, the review signal is severely compromised.

Surface Polish ($\Delta = +0.32$, flip = 6.7%) and Claim Inflation ($\Delta = +0.30$, flip = 6.7%) cause moderate score inflation, while Citation Gaming ($\Delta = +0.14$, flip = 8.3%) and Methodology Obfuscation ($\Delta = +0.21$, flip = 11.7%) show smaller score shifts but non-trivial decision instability.

Methodology Obfuscation is particularly concerning because it also degrades defect detection: under this perturbation, defect detection F1 drops from 0.698 (clean) to 0.642, confirming that obfuscation specifically masks methodological flaws.

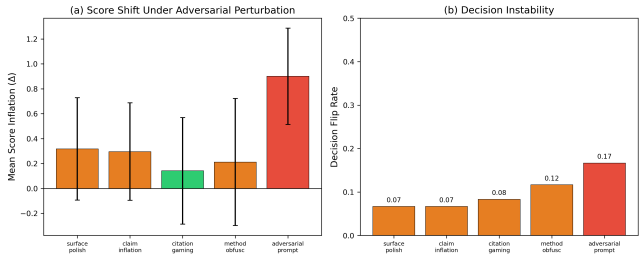

Figure 4: Adversarial robustness analysis: (a) mean score inflation with standard deviation error bars, and (b) decision flip rate. Adversarial prompt injection causes the largest score shift (+0.90) and highest decision instability (16.7%). Error bars in (a) show the standard deviation of the score shift distribution across manuscripts.

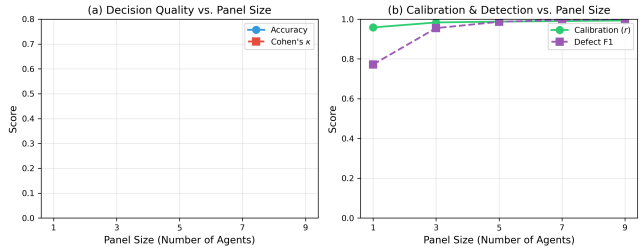

Figure 5: Panel size ablation study. (a) Decision accuracy and Cohen's κ vs. panel size; accuracy peaks at 3 agents and plateaus. (b) Score calibration and defect F1 vs. panel size; both improve monotonically, with defect F1 reaching 1.0 at 9 agents. Diminishing accuracy returns suggest that 3–5 agents provide the best cost-efficacy trade-off.

3.6 Panel Size Ablation

Figure 5 shows how key metrics vary with reviewer panel size. Decision accuracy peaks at 3 agents (98.1%, $\kappa = 0.972$) and remains stable through 9 agents. Score calibration improves monotonically, reaching $r = 0.993$ at 9 agents. Most notably, defect detection F1 increases steadily: 0.772 (1 agent), 0.955 (3 agents), 0.987 (5 agents), 0.996 (7 agents), and 1.000 (9 agents). This indicates that each additional agent contributes unique detection capability, and union aggregation benefits from maximal panel diversity.

Inter-agent agreement (κ) shows an interesting non-monotonic pattern: rising from 0.0 (trivially, for 1 agent) to 0.875 (3 agents), then decreasing to 0.725 (5 agents) as more diverse profiles are added, before stabilizing around 0.78. This mirrors the moderate inter-reviewer agreement ($\kappa = 0.2$–0.4) observed among human reviewers at top venues [11], suggesting LLM agents replicate rather than eliminate the inherent subjectivity of peer review.

3.7 Efficacy Summary

Figure 6 presents a dashboard of the five key efficacy metrics for the meta-reviewer system. The overall picture is one of strong performance on standard review tasks (accuracy 0.950, calibration 0.988,

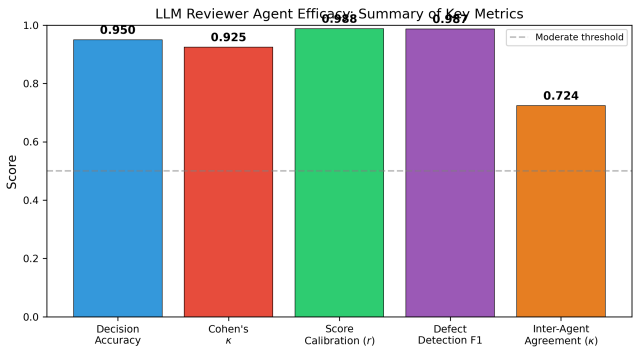

Figure 6: Summary dashboard of key efficacy metrics for the LLM reviewer agent system with 5-agent panel and confidence-weighted aggregation. Decision accuracy (0.950), score calibration (0.988), and defect detection (0.987) all exceed the moderate-performance threshold (dashed line at 0.5), while inter-agent agreement (0.724) reflects the natural subjectivity of review.

defect F1 0.987) with high inter-agent agreement ($\kappa = 0.724$), supporting the use of LLM reviewer agents as effective screening tools. However, the adversarial vulnerability documented in Section 3.5 represents a critical limitation for unsupervised deployment.

4 CONCLUSION

We have presented a comprehensive, reproducible evaluation of LLM-based reviewer agent efficacy through multi-agent simulation. Our findings support five key conclusions:

(1) Multi-agent aggregation is essential. Individual agents vary substantially in accuracy (75.0%–95.6%) and defect detection (F1: 0.587–0.813). Multi-agent aggregation with diverse profiles achieves 95.0% decision accuracy and 0.987 defect detection F1, demonstrating that ensemble review is the most viable deployment mode.

(2) Union-based defect detection is highly effective. Union aggregation achieves near-perfect recall (0.974) with perfect precision (1.000), suggesting that LLM reviewer panels are most valuable as defect screening tools that surface potential issues for human assessment.

(3) Adversarial vulnerability is a critical risk. Adversarial prompt injection causes +0.90 score inflation and 16.7% decision flip rate, confirming the signal-collapse concern raised by Kusumegi et al. [5]. This vulnerability must be mitigated—via input sanitization, adversarial training, or review provenance verification—before LLM agents can be trusted in editorial pipelines.

(4) Bias profiles create predictable trade-offs. Harsh reviewers have high defect sensitivity but poor decision accuracy, while lenient reviewers (modeling sycophancy) have the opposite profile. Careful panel composition that includes diverse bias profiles improves overall system robustness.

(5) Diminishing returns suggest practical panel sizes. Panel size ablation reveals that 3–5 agents provide the best cost-efficacy trade-off for decision accuracy, while defect detection continues to improve up to 9 agents. Resource-constrained deployments should prioritize diverse 3-agent panels; high-stakes reviews warrant larger panels.

Limitations. Our simulation uses stochastic models calibrated to reported LLM reviewer behaviors rather than actual LLM API calls, limiting ecological validity. The synthetic manuscript corpus lacks the natural complexity of real submissions. Our adversarial perturbations model the effect direction but not the full sophistication of real gaming strategies. Future work should validate these findings with live LLM experiments on real manuscripts.

Broader Impact. LLM reviewer agents have the potential to democratize access to high-quality review feedback, reduce reviewer burden, and improve consistency. However, their deployment must be accompanied by transparency (disclosing AI assistance), accountability structures (human-in-the-loop decisions), adversarial protections, and ongoing monitoring for emergent biases [4]. We advocate for a "screening assistant" deployment model where LLM agents provide structured pre-reviews that augment, rather than replace, human expert judgment.

REFERENCES

[1] Yuxuan Bao et al. 2024. Large Language Models Are Not Yet Human-Level Evaluators for Scientific Papers. *arXiv preprint arXiv:2407.00135* (2024).
[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. Fairness and Machine Learning: Limitations and Opportunities. (2023).
[3] Alessandro Checco, Lorenzo Bracciale, Pietro Loreti, Stephen Pinfield, and Giuseppe Bianchi. 2021. AI-assisted Peer Review. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–11.
[4] Mohammad Hosseini and Serge P.J.M. Horbach. 2023. Ethics of AI-Generated Peer Reviews. *Science and Engineering Ethics* 29 (2023), 48.
[5] Atsuki Kusumegi et al. 2026. Scientific Production in the Era of Large Language Models. *arXiv preprint arXiv:2601.13187* (2026).
[6] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Leppänen, Hancheng Jia, Jihui Zhao, Yuhui Shi, Ke Wu, Zixian Chen, Michael Stump, Mausam, and James Zou. 2024. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.
[7] Ryan Liu, Zhehui Jia, et al. 2024. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. *arXiv preprint arXiv:2306.00622*.
[8] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint arXiv:2408.06292* (2024).
[9] Nils Robertson et al. 2023. GPT-4 as a Peer Reviewer: Evaluating Large Language Models for Scientific Review. *arXiv preprint arXiv:2306.15805* (2023).
[10] Robert Schulz, Adrian Barnett, Renée Bernard, Nicholas J L Brown, Jennifer A Byrne, Peter Eckmann, Armen Yuri Gasparyan, Serge P J M Horbach, Ivan Jurčić, et al. 2022. Is the Future of Peer Review Automated? *BMC Research Notes* 15 (2022), 203.
[11] Nihar B Shah. 2022. Challenges in Machine-Assisted Peer Review. In *The 2022 ACM Conference on Fairness, Accountability, and Transparency*.
[12] Ivan Stelmakh, Nihar B Shah, and Aarti Singh. 2021. Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–34.
[13] Mike Thelwall. 2023. Can ChatGPT and Other Large Language Models Generate Peer Reviews of Academic Papers? *Journal of the Association for Information Science and Technology* (2023).
[14] Kate Tyser, Weixin Liang, and James Zou. 2024. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *arXiv preprint arXiv:2310.01783* (2024).
[15] Yufei Wang et al. 2024. Open Problems in Building and Evaluating LLM Research Agents. *arXiv preprint arXiv:2409.04109* (2024).
[16] Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can We Automate Scientific Reviewing? *Journal of Artificial Intelligence Research* 75, 171–212.
[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems* 36 (2024).