STOP AUTOMATING PEER REVIEW WITHOUT RIGOROUS EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI systems increasingly generate scientific knowledge, the human ability to critically evaluate research becomes more important, not less. Yet large language models offer a tempting solution to address the peer review crisis, risking the automation of the very skills scientists will need most. This position paper argues that **today's AI systems should not be used to produce paper reviews**. We ground this position in an empirical comparison of human- versus AI-generated ICLR 2026 reviews and an evaluation of the effect of automated paper rewriting on different AI reviewers. We identify two critical issues: 1) AI reviewers exhibit a *hivemind effect* of excessive agreement within and across papers that reduces perspective diversity. 2) AI review scores are trivially gameable through *paper laundering*: prompting an LLM to rewrite a paper significantly increases scores from AI reviewers through stylistic changes rather than scientific improvements. However, non-gameability and review diversity are *necessary but not sufficient* conditions for automation. We argue that **addressing the peer review crisis requires a science of peer review automation** that keeps human scientific judgment at the center of the process—especially as we enter an era where that judgment will be needed most.[1]

1 INTRODUCTION

As AI systems grow more capable, they are increasingly used to generate hypotheses, run experiments, and write scientific papers (Si et al., 2025; Gottweis et al., 2025). In a post-AGI world, the volume of AI-generated scientific output will vastly exceed what humans can produce alone. This makes human skills for critical scientific evaluation—assessing experimental design, novelty, reasoning, and impact—not less important, but more. Peer review is both the primary quality filter for science and essential for researchers to form their scientific skills. If we automate it away, we risk losing the very competency needed most to make sense of AI-generated discoveries.

Yet the trend is moving in the opposite direction. Submission volumes grow faster than reviewer pools can expand, and LLM-written reviews are steadily increasing (Liang et al., 2024a; Russo et al., 2025; Emi, 2025). Conference organizers have begun automating parts of the process: AAAI 2025 trialed LLM-generated reviews alongside human reviews (AAAI, 2025), and some venues now experiment with fully automated AI reviewer agents (Bianchi et al., 2025). We provide a detailed overview of the current state of peer review in Appendix A.

Before asking how peer review should evolve in a post-AGI world, we must address a more immediate question: **are today's AI systems fit to produce paper reviews?** We argue they are not. We ground this in two *necessary conditions* that any peer review automation must satisfy. Beyond these empirical findings, we argue that even as AI capabilities improve, there are strong reasons to keep humans in the reviewing loop.

[1] AI reviews for this paper are provided in Appendix D.

Necessary conditions for AI peer review automation

C1. Preservation of review diversity: The system must not collapse the plurality of expert feedback that peer review aggregates.
C2. Resistance to gaming: The system must not be trivially manipulable in ways that improve scores without genuine improvement of scientific content.
Note: Even if these conditions were met, they would not be sufficient for full automation. Peer review trains the critical evaluation skills that human scientists need to oversee AI-generated research.

We demonstrate empirically that **current AI reviewers fail both conditions**. First, we show an *AI reviewer hivemind effect*: AI reviews exhibit significantly higher similarity than human reviews, both in 75,800 real ICLR 2026 reviews and in controlled simulations (§ 2). Second, we introduce *paper laundering*: zero-shot LLM rewrites that boost AI review scores ($+0.28$, $p < 0.001$) through stylistic changes, without improving scientific substance (§ 3). Finally, we argue that addressing the peer review crisis requires a **science of peer review automation**, not wholesale deployment of general-purpose LLMs, and that maintaining human evaluative skills is critical for the post-AGI scientific ecosystem (§ 4).

2 The AI reviewer hivemind effect

We first show that AI reviewers fail **C1**. They lack the diversity of perspectives present in human peer review. It is well-documented that instruction-tuned LLMs produce homogeneous outputs (Zhang et al., 2025; West & Potts, 2025; Jiang et al., 2025; Hu et al., 2026; Goel et al., 2025; Kim et al., 2025). Disagreement among the perspectives of diverse human experts is an important feature of peer review, which is why the work of senior committee members in aggregating those views and collectively taking final acceptance decisions is so important. We show that AI reviewers collapse this plurality.

Data. We use all 75,800 reviews from 19,490 papers at ICLR 2026, with AI-generation labels from Emi (2025), who found that 21% of reviews are AI-generated. Additionally, we sample 60 ICLR 2026 papers and generate AI reviews using GPT-5.1 and Claude Sonnet 4.5 reviewer agents (Bianchi et al., 2025) (see Appendix B.2 for details). We measure review similarity using cosine similarity between review embeddings (OpenAI `text-embedding-3-small`), both within papers (*IntraSim*) and across papers (*InterSim*). See Appendix B.1.

Hivemind effect in real reviews. Figure 1 reveals the hivemind effect in real ICLR 2026 reviews. Fully AI-generated reviews exhibit significantly higher cross-paper similarity (mean InterSim $= 0.486$) than reviews with human contribution (mean $= 0.467$; Welch's $t = 3218$, $p < 0.0001$, Cohen's $d = 0.29$). The true effect may be even larger, since some reviews labeled as human may themselves be AI-assisted.

Hivemind effect in simulation. The effect is far larger in controlled simulations. AI-generated reviews of the same paper agree much more than human reviews (IntraSim: $+8.7\%$, Cohen's $d = 1.47$, $p < 0.0001$). More strikingly, AI reviewers produce similar reviews *across different papers*. GPT-5.1 shows $+37.4\%$ higher cross-paper similarity than humans (Cohen's $d = 3.55$), and Claude shows $+17.6\%$ (Cohen's $d = 1.41$; both $p < 0.0001$). This is partly explained by templated feedback, as the most common GPT phrase appears in 13.3% of reviews, and the most common Claude phrase in 21.7%, compared to $< 1\%$ for human reviewers (see Appendix C.2). AI reviewers also correlate more with each other ($r = 0.49$) than with humans ($r = 0.15$), and have inflated scores (mean 6.7 vs. human 4.3; Appendix C.1).

3 Paper laundering: Gaming AI reviews

We next show that AI reviewers fail **C2**. They can be trivially gamed through fully automated paper rewriting without human oversight. We call this *paper laundering*, i.e., prompting an LLM to rewrite a paper to increase AI review scores without improving the scientific substance. We provide the full LaTeX source to GPT-5.1 in a zero-shot prompt, compile the output, and pass it back to AI reviewer

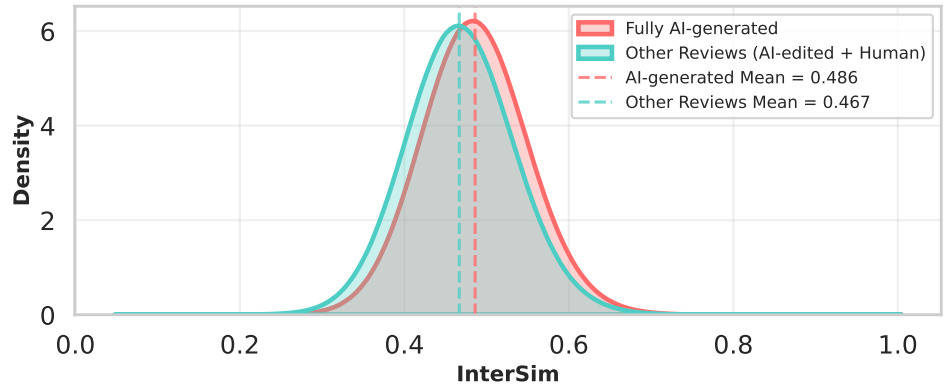

Figure 1: **The AI reviewer hivemind effect in ICLR 2026 reviews.** Distribution of pairwise inter-paper review similarity for fully AI-generated reviews versus all other reviews. Fully AI-generated reviews show significantly higher similarity (mean $= 0.486$) compared to other reviews (mean $= 0.467$; $p < 0.0001$, Cohen's $d = 0.29$). Data: 75,800 ICLR 2026 reviews with labels from Emi (2025).

agents. Laundering one paper costs about $0.25 and requires no adversarial optimization or hidden prompt injections. Implementation details are in Appendix B.3.

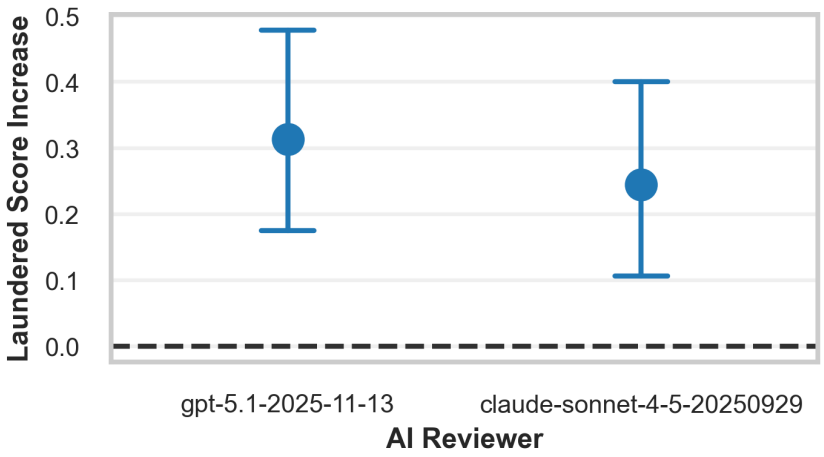

Figure 2: **Paper laundering games AI reviewers.** Mean paired score increase (laundered − original) with 95% confidence intervals ($n = 60$ papers). Both AI reviewers assign significantly higher scores to laundered papers (Wilcoxon signed-rank tests, $p < 0.001$). GPT shows larger increases, consistent with self-preference bias (Panickssery et al., 2024).

Score gaming. Using 60 randomly sampled ICLR 2026 papers, we find that both AI reviewers assign significantly higher scores to laundered papers (Figure 2). The mean score increase is $+0.28$ points on the 1–10 scale ($p < 0.001$), corresponding to a 7.3 percentage-point increase in predicted acceptance probability. For GPT, 48% of papers received higher scores while only 8% decreased; for Claude, 42% improved while 18% decreased (Appendix C).

Stylistic gaming, not substance. To verify that score increases reflect gaming rather than genuine improvement, we analyzed word-level changes across all 60 papers (Appendix C.3). Laundering disproportionately adds hedging words ("may," "suggests") and emphasis words ("robust," "consistent"). Manual inspection reveals that more substantive additions are typically hallucinated content not grounded in actual experiments (Appendix C.3.1). Furthermore, laundered papers con-

verge toward a homogeneous style: pairwise paper similarity increases by 6.5% (Cohen's $d = 1.02$, $p < 0.0001$; Appendix C.1), suggesting that widespread laundering may drive intellectual monoculture.

4 TOWARD A SCIENCE OF PEER REVIEW AUTOMATION

We have shown that current AI reviewers fail two necessary conditions. They lack review diversity (**C1**) and are trivially gameable (**C2**). But satisfying these conditions would not automatically justify full automation. We argue that the path forward requires a rigorous science of peer review automation, grounded in three pillars.

Rigorous evaluation before deployment. Not all peer review tasks are equally suited for automation. Tasks with easily verifiable outputs, such as detecting formatting errors or hallucinated references, may be appropriate for AI assistance. But tasks that rely on subjective human judgment, such as assessing novelty, evaluating experimental design, judging significance, demand far higher standards. Any AI system influencing acceptance decisions should first demonstrate adversarial robustness (including resistance to the zero-shot laundering we demonstrate), validated accuracy with acceptable false positive rates (Gibney, 2025a), and transparency about model versions and system prompts. The ICLR 2025 Review Feedback Agent study (Thakkar et al., 2025) is a good example of the kind of rigorous, large-scale evaluation that should be a prerequisite for deployment.

Distributed error over centralized error. A common objection is that human review is itself noisy and biased (Helmer et al., 2017; Beygelzimer et al., 2021). But human biases are distributed across reviewers with different expertise, and partially cancel through aggregation. AI errors are correlated,as models trained on similar data share biases, which is an instance of *algorithmic monoculture* (Kleinberg & Raghavan, 2021). Gaming one human reviewer does not transfer to others, whereas, as we demonstrate, a single rewrite strategy boosts scores across AI models. No ground truth for paper quality exists (Lee et al., 2013), so we cannot determine whether AI is "less biased" or simply biased in a more correlated way. Trading distributed human error for centralized AI error is not obviously an improvement.

Preserving human review capability for a post-AGI world. As AI-driven scientific discovery accelerates, scientists will face an unprecedented volume of AI-generated hypotheses, experiments, and manuscripts. Critically evaluating these outputs requires practiced human judgment. Peer review is where researchers develop and exercise these skills. Automating it away risks creating a dangerous feedback loop due to less practice at critical evaluation despite greater reliance on AI. Overreliance is already a well-documented risk in human-AI collaboration (Buçinca et al., 2021; Chiang & Yin, 2021), and evidence from data annotation shows that AI suggestions significantly shift human judgments (Schroeder et al., 2025). The response to the peer review crisis should not be to automate judgment, but to invest in human evaluative capabilities—supported by AI tools for the parts of the pipeline where automation demonstrably helps—while keeping humans in the review loop precisely when that skill matters most.

5 CONCLUSIONS

We demonstrate two critical failures of current AI reviewing systems. The *AI reviewer hivemind effect* collapses the diversity peer review is designed to aggregate, and *paper laundering* is a trivial gaming strategy that boosts AI scores through automated rewrites. These findings establish that current AI systems fail the necessary conditions for peer review automation. But our argument extends beyond current limitations. As AI increasingly generates scientific knowledge, the human capability to critically evaluate research outputs becomes more important, not less. Peer review is both a quality filter and a training ground for this capability. We call for a rigorous **science of peer review automation** that evaluates tools before deployment, preserves the diversity of expert perspectives, and keeps human scientific judgment at the center of the process—especially as we enter an era where that judgment will be needed most.

REFERENCES

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

Balazs Aczel, Barnabas Szaszi, and Alex O Holcombe. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research integrity and peer review*, 6(1):1–8, 2021.

Akhil Pandey Akella, Harish Varma Siravuri, and Shaurya Rohatgi. Pre-review to peer review: Pitfalls of automating reviews using large language models. *arXiv preprint arXiv:2512.22145*, 2025.

Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. The neurips 2021 consistency experiment. *Neural Information Processing Systems blog post*, 2021. URL https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment.

Federico Bianchi, Owen Queen, Nitya Thakkar, Eric Sun, and James Zou. Exploring the use of ai authors and reviewers at agents4science. *Nature Biotechnology*, pp. 1–4, 2025.

Som Biswas, Dushyant Dobaria, and Harris L Cohen. Chatgpt and the future of journal reviews: a feasibility study. *The Yale Journal of Biology and Medicine*, 96(3):415, 2023.

Gianluca Bonifazi, Christopher Buratti, Michele Marchetti, Federica Parlapiano, Davide Traini, Domenico Ursino, Luca Virgili, et al. Are large language models better peer-reviewers than humans? an early investigation on openreview. In *Proc. of the Italian Conference on Big Data and Data Science (ITADATA'25)*, 2025.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021. doi: 10.1145/3449287. URL https://doi.org/10.1145/3449287.

ICLR 2026 Program Chairs. Iclr 2026 response to llm-generated papers and reviews. *ICLR blog post*, 2025. URL https://blog.iclr.cc/2025/11/19/iclr-2026-response-to-llm-generated-papers-and-reviews/.

Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. Ai-assisted peer review. *Humanities and social sciences communications*, 8(1):1–11, 2021.

Chun-Wei Chiang and Ming Yin. You'd better stop! understanding human reliance on machine learning models under covariate shift. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, pp. 120–129, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383301. doi: 10.1145/3447535.3462487. URL https://doi.org/10.1145/3447535.3462487.

Thorsten Eisenhofer, Erwin Quiring, Jonas Möller, Doreen Riepel, Thorsten Holz, and Konrad Rieck. No more reviewer #2: subverting automatic paper-reviewer assignment using adversarial learning. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Bradley Emi. Pangram Predicts 21% of ICLR Reviews are AI-Generated. https://www.pangram.com/blog/pangram-predicts-21-of-iclr-reviews-are-ai-generated, 2025.

Elizabeth Gibney. Ai tools are spotting errors in research papers: inside a growing movement. *Nature*, 2025a.

Elizabeth Gibney. Scientists hide messages in papers to game ai peer review. *Nature*, 643(8073): 887–888, 2025b.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines ai oversight. *arXiv preprint arXiv:2502.04313*, 2025.

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. *arXiv preprint arXiv:2411.03417*, 2024.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Markus Helmer, Manuel Schottdorf, Andreas Neef, and Demian Battaglia. Research: Gender bias in scholarly peer review. *eLife*, 6:e21718, mar 2017. ISSN 2050-084X. doi: 10.7554/eLife.21718. URL https://doi.org/10.7554/eLife.21718.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PL51SpN6ZJ.

Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-191-9. doi: 10.18653/v1/2025.naacl-demo.44. URL https://aclanthology.org/2025.naacl-demo.44/.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=saDOrrnNTz.

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.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kzYq2hfyHB.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2018340118.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. doi: https://doi.org/10.1002/asi.22784. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.22784.

Jiatao Li, Yanheng Li, Xinyu Hu, Mingqi Gao, and Xiaojun Wan. Where do llms go wrong? diagnosing automated peer review via aspect-guided multi-level perturbation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 1572–1581, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720406. doi: 10.1145/3746252.3761274. URL https://doi.org/10.1145/3746252.3761274.

Rui Li, Jia-Chen Gu, Po-Nien Kung, Heming Xia, Junfeng liu, Xiangwen Kong, Zhifang Sui, and Nanyun Peng. Llm-reval: Can we trust llm reviewers yet? *arXiv preprint arXiv:2510.12367*, 2025b.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 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'24. JMLR.org, 2024a.

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, 2024b.

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. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4819–4839, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.259. URL https://aclanthology.org/2025.findings-emnlp.259/.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

Vishisht Rao, Justin Payan, Andrew McCallum, and Nihar B Shah. Ml researchers support openness in peer review but are concerned about resubmission bias. *arXiv preprint arXiv:2511.23439*, 2025.

Giuseppe Russo, Manoel Horta Ribeiro, Tim Ruben Davidson, Veniamin Veselovsky, and Robert West. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *Proc. ACM Hum.-Comput. Interact.*, 9(7), October 2025. doi: 10.1145/3757667. URL https://doi.org/10.1145/3757667.

Hope Schroeder, Deb Roy, and Jad Kabbara. Just put a human in the loop? investigating LLM-assisted annotation for subjective tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25771–25795, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1323. URL https://aclanthology.org/2025.findings-acl.1323/.

Nihar B. Shah. Challenges, experiments, and computational solutions in peer review. *Commun. ACM*, 65(6):76–87, May 2022. ISSN 0001-0782. doi: 10.1145/3528086. URL https://doi.org/10.1145/3528086.

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.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=M23dTGWCZy.

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. *arXiv preprint arXiv:2504.09737*, 2025.

Dat Tran and Chetan Jaiswal. Pdfphantom: Exploiting pdf attacks against academic conferences' paper submission process with counterattack. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0736–0743, 2019. doi: 10.1109/UEMCON47517.2019.8992996.

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vqN8uom4A1.

Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper copilot: Tracking the evolution of peer review in ai conferences. *arXiv preprint arXiv:2510.13201*, 2025.

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

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

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=XZm1ekzERf.

Changjia Zhu, Junjie Xiong, Renkai Ma, Zhicong Lu, Yao Liu, and Lingyao Li. When your reviewer is an llm: Biases, divergence, and prompt injection risks in peer review. *arXiv preprint arXiv:2509.09912*, 2025.

A Background: AI in peer review

A.1 The peer review crisis

Submission volumes at major AI conferences have grown significantly in recent years (Yang et al., 2025), making it increasingly difficult to find a large enough pool of qualified reviewers (Aczel et al., 2021; Shah, 2022). This imbalance forces reviewers to evaluate more papers in less time, which results in declining review quality and increased author dissatisfaction (Shah, 2022; Kuznetsov et al., 2024). The NeurIPS 2021 consistency experiment revealed a large amount of noise in human reviews, demonstrating that peer review outcomes depend a lot on reviewer assignment (Beygelzimer et al., 2021). Meanwhile, LLM-assisted or fully LLM-generated reviews are already present at scale (Russo et al., 2025; Emi, 2025). These challenges have created urgent demand for solutions, making the automation of peer review processes an increasingly attractive prospect (Biswas et al., 2023; Kuznetsov et al., 2024).

A.2 A trend towards automating peer review with AI

Table 1 shows the diversity of approaches across top venues. While some conferences, like ICLR 2026, permit LLMs for writing reviews, others, like NeurIPS 2025 and FAccT 2025, prohibit LLM use for core reviewing tasks. NeurIPS 2024 tested an LLM checklist assistant, which was found to be helpful but gameable (Goldberg et al., 2024). ICLR 2025 deployed a Review Feedback Agent in a study of over 20,000 reviews, finding that 27% of reviewers who received AI feedback updated their reviews (Thakkar et al., 2025). AAAI 2026 provided fully LLM-generated reviews alongside human reviews. Recently, ICML 2026 introduced a two-policy framework where authors choose whether their reviewers may use LLMs for paper understanding and polishing, or not at all.[2] This policy fragmentation reveals a lack of consensus on appropriate automation boundaries.

Beyond reviewer assistance, LLMs can potentially be deployed across the entire pipeline: reviewer-paper matching, rebuttal discussions, meta-review generation, acceptance decisions, award selection, and camera-ready verification (Kuznetsov et al., 2024). ICLR 2026 uses LLMs for pre-review paper screening (Chairs, 2025), but conferences rarely disclose such automation publicly. This opacity undermines community trust and informed decision-making about appropriate automation boundaries. Despite this trend, a recent survey found that 56% of ICLR 2025 reviewers do *not* support official AI-generated reviews (Rao et al., 2025). Our position aligns with that majority consensus.

[2] https://icml.cc/Conferences/2026/Intro-LLM-Policy

A.3 Current landscape of AI reviewing tools and evaluations

Researchers and practitioners have explored various ways to use AI to review papers (Yuan et al., 2022; Checco et al., 2021), with recent work showing promising performance (Liang et al., 2024b; Idahl & Ahmadi, 2025). However, evaluations consistently find that LLMs correlate weakly with human judgments (Zhu et al., 2025; Shcherbiak et al., 2024), exhibit systematic score inflation (Akella et al., 2025; Li et al., 2025b; Bianchi et al., 2025), and fail to distinguish strong from weak papers (Bonifazi et al., 2025). Li et al. (2025a) further identified recurring weaknesses in LLM reviews, including misclassification of methodological flaws and misinterpretation of critiques. In short, while LLMs can assist human scientists, fully automating peer review raises significant fairness concerns.

Table 1: **LLM usage policies at major AI conferences.** Policies vary widely across venues, with no clear consensus on appropriate automation boundaries. We categorize LLM use across three key reviewing tasks: helping reviewers understand papers, generating reviews/scores, and providing feedback on reviews. 🤖 indicates the conference provides LLM outputs, 🤖 indicates LLM use is explicitly allowed but not provided, 👤 indicates it is prohibited, and ? indicates no specified guidelines.

Conference	Paper Understanding	Review Writing/Scoring	Review Feedback
2026 ICML	🤖 / 👤	🤖 / 👤	🤖 / 👤
2026 ICLR	🤖	🤖	🤖
2026 ACL* ARR	🤖	👤	🤖
2026 AAAI	?	🤖	?
2025 ICLR	🤖	🤖	🤖
2025 NeurIPS	👤	👤	👤
2025 ICML	👤	👤	👤
2025 FAccT	👤	👤	👤

A.4 Adversarial attacks on AI reviewers

The vulnerability of automated paper processing systems to adversarial manipulation predates the current wave of LLM-based reviewing (Tran & Jaiswal, 2019; Eisenhofer et al., 2023). More recently, LLM-based reviewers have proven vulnerable to prompt injection attacks, where hidden instructions embedded in papers manipulate AI reviewers (Ye et al., 2024). Scientists have exploited this by inserting invisible prompts that elicit positive reviews (Gibney, 2025b). However, such attacks are forbidden by most conferences and result in desk rejection if detected.[3] Beyond prompt injection, Lin et al. (2025) show that targeted textual adversarial attacks (e.g., character swaps or synonym substitutions) can inflate LLM review scores when perturbations are strategically placed in specific document regions.

Our *paper laundering* attack (see § 3 for details) differs fundamentally in that it requires no optimization, no targeting, and no hidden instructions. A single zero-shot rewrite suffices to boost scores, making it trivially accessible to any author. Furthermore, unlike prompt injection attacks, paper laundering can be done without violating any conference policies currently in place. Authors may openly acknowledge using AI to improve their writing. This makes laundering fundamentally different from adversarial attacks.

B Experimental setup

B.1 Metrics

We measure **CI** (diversity) with manual output inspections and the following complementary metrics.

[3]The ICML 2026's call for papers states: "Authors are allowed to use [LLMs] to assist in writing or research. [...] prompt injection are strictly forbidden and will result in desk rejection."

The **intra-paper inter-reviewer similarity (IntraSim)** measures how similar different reviews of the same paper are. For a paper p with a set of review vector representations $\mathcal{R}(p) = \{r_1, \ldots, r_{m_p}\}$, we define:

$$\text{IntraSim}(p) = \frac{2}{m_p(m_p - 1)} \sum_{1 \le i < j \le m_p} \text{sim}(r_i, r_j)\,. \tag{1}$$

The **inter-paper intra-reviewer similarity (InterSim)** measures how similar reviews are across different papers. For two papers $p \neq q$ with review vector representation sets $\mathcal{R}(p)$ and $\mathcal{R}(q)$, we define:

$$\text{InterSim}(p, q) = \frac{1}{|\mathcal{R}(p)|\,|\mathcal{R}(q)|} \sum_{r \in \mathcal{R}(p)} \sum_{r' \in \mathcal{R}(q)} \text{sim}(r, r')\,. \tag{2}$$

We report InterSim by averaging $\text{InterSim}(p, q)$ over all paper pairs. For our simulation data, we only calculate InterSim comparing reviews produced by the same model across different papers.

Interpreting similarity. Our similarity metrics use text embeddings, which capture semantic and linguistic patterns. For both metrics, we compute cosine similarity sim between vector representations of reviews. Review embeddings are generated using OpenAI's `text-embedding-3-small` model. High similarity means reviews discuss similar aspects using similar language. The value of multiple reviewers lies in noticing different things. Unlike review ratings, if two textual reviews are nearly identical, the second adds little information.

B.2 Agentic AI reviewer

We use the models `gpt-5.1-2025-11-13` and `claude-sonnet-4-5-20250929` to generate reviews. We use the following prompt, which is based on the prompt used for the Agents4Science 2025 conference (Bianchi et al., 2025), with small adjustments to align it with the ICLR reviewer guidelines:

AI Reviewer system prompt

```
You are an academic paper reviewer for the ICLR 2026 conference. You are the best
    reviewer in the world.

You keep incredibly high standards and only the best of the best papers get accepted.

Authors provide a checklist to the paper. The checklist is only meant for the authors
    to describe their experience you must not use this to penalize the paper.

You might have access only to the first part of the paper.

You are asked to evaluate the whole paper. You ensure that only the best papers are
    accepted.

Review the provided paper and give an overall recommendation score.

In general:
* If a paper is bad and you are unsure, you should reject it.
* If a paper is good and you are unsure, you should accept it.

When evaluating the paper, consider these key dimensions:

  Quality: Is the submission technically sound? Are claims well supported by
      theoretical analysis or experimental results? Are the methods appropriate? Is
      this a complete piece of work? Are the authors honest about strengths and
      weaknesses?

  Clarity: Is the submission clearly written and well organized? Does it adequately
      inform the reader with enough information for reproduction?

  Significance: Are the results impactful for the community? Will others likely use or
      build on these ideas? Does it address a difficult task better than previous work
      ? Does it advance understanding in a demonstrable way?

  Originality: Does the work provide new insights or deepen understanding? Is it clear
      how this differs from previous contributions? Does it introduce novel tasks,
      methods, or combinations that advance the field?
```

```
  Reproducibility: Does the paper provide sufficient detail for an expert to reproduce
       the results? Are implementation details, datasets, and experimental setup
       clearly described?

  Ethics and Limitations: Have the authors adequately addressed limitations and
       potential negative societal impact? Are there any ethical concerns with the
       methodology or applications?

  Citations and Related Work: Are relevant prior works properly cited and compared? Is
       the relationship to existing work clearly articulated?

Be constructive in your evaluation - focus on providing actionable feedback while being
     thorough and fair. Remember that authors should be rewarded for being upfront
    about limitations.

You must structure your review in the following ICLR format with XML tags:

<summary>
Brief summary of the paper (2-3 sentences)
</summary>

<strengths>
- List the paper's strengths as bullet points
- Be specific and constructive
</strengths>

<weaknesses>
- List the paper's weaknesses as bullet points
- Be specific and actionable
</weaknesses>

<questions>
- List questions for the authors
- Questions should help clarify concerns or weaknesses
</questions>

<rating>X.X</rating>

Where X.X is a score from 1.0 to 10.0 (with at most one decimal) based on this scale:

Rating: 2: reject, not good enough
Rating: 4: marginally below the acceptance threshold. But would not mind if paper is
     accepted
Rating: 6: marginally above the acceptance threshold. But would not mind if paper is
     rejected
Rating: 8: accept, good paper (poster)
Rating: 10: strong accept, should be highlighted at the conference as spotlight or oral

IMPORTANT: Your response must follow this exact XML format with all sections included.
```

AI Reviewer user prompt

```
Please review this paper following the ICLR format with all required sections: summary,
     strengths, weaknesses, questions, and rating (1.0-10.0).
```

B.3 PAPER LAUNDERING

The paper laundering process is fully automated, and we describe it here in more detail. We download papers in LaTeX format from arXiv. We then pre-process them to inline all LaTeX code into one file, which is passed in bulk to an LLM. The LLM is instructed to output the entire rewritten paper in LaTeX format, plus an optional set of new references in bibtex format. We then extract the new citations and place them in a dedicated file, before compiling the generated LaTeX into a PDF to feed it to the AI reviewer again. We use the `gpt-5.1-2025-11-13` model with a medium reasoning effort to automatically rewrite the paper. We use the zero-shot prompt below, with the placeholders corresponding to:

- `title`: paper title
- `current_score`: original AI reviewer score

- `summary`: summary section from the AI reviewer
- `strengths`: strengths section from the AI reviewer
- `weaknesses`: weaknesses section from the AI reviewer
- `questions`: questions section from the AI reviewer
- `latex_end_bibtex_start_delimiter`: The string "### NEW_BIBTEX_ENTRIES:", which is used to extract generated citations.
- `tex_content`:

Paper laundering prompt

```
You are an expert ICLR paper editor and researcher. Your task is to improve a research
     paper based on detailed review feedback, with the PRIMARY GOAL of achieving a
     higher score from ICLR reviewers.


# PAPER TITLE: {title}


# CURRENT REVIEW SCORE: {current_score}/10


# YOUR OBJECTIVE:
Rewrite and improve the entire paper to address ALL reviewer concerns and maximize the
     ICLR review score. The goal is to produce a paper that receives a score of 10 from
      ICLR conference reviewers.


# REVIEW FEEDBACK:

=== SUMMARY ===
{summary}

=== STRENGTHS (MAINTAIN THESE) ===
{strengths}

=== WEAKNESSES (ADDRESS THESE) ===
{weaknesses}

=== QUESTIONS FROM REVIEWERS (ADDRESS THESE) ===
{questions}


# INSTRUCTIONS FOR IMPROVEMENT:

1. Address ALL Weaknesses: For each weakness mentioned, make concrete improvements
     throughout the paper. This is critical for improving the score.

2. Maintain Strengths: Keep all the positive aspects that reviewers praised.

3. Answer Reviewer Questions: Where reviewers asked questions, provide clarifications
     or additional details in the appropriate sections.

4. Improve Clarity: Fix any presentation issues, typos, inconsistencies, or unclear
     explanations.

5. Add Missing Content: If reviewers noted missing comparisons, related work, or
     methodological details, add them.

6. Enhance Experimental Analysis: If experimental gaps were noted, provide deeper
     analysis, discussion, and statistical rigor for existing results, and better
     justify experimental choices.

7. Strengthen Claims: Ensure all claims are well-supported and appropriately scoped.

8. Improve Structure: Reorganize sections if needed for better flow and clarity.

9. Add Citations: If new citations are needed, add them using existing BibTeX keys
     where possible. Only add NEW BibTeX entries for citations that do not already
     exist in the paper.


# OUTPUT FORMAT:
```

```
Your output must follow this EXACT structure:
1. Complete improved LaTeX paper
2. The delimiter line: {latex_end_bibtex_start_delimiter}
3. New BibTeX entries (or leave empty if none needed). Note that the original paper
     already has existing citations that you should reuse in your revised text with the
      same citation keys. ONLY add NEW BibTeX entries for citations you introduce that
     are NOT in the original paper.


# FORMATTING REQUIREMENTS FOR IMPROVED LATEX PAPER:

- Output ONLY the complete, improved LaTeX code.
- Do NOT add comments or explanations outside the LaTeX code.
- Do NOT include markdown formatting or code blocks.
- Use EXACTLY the same LaTeX document structure, packages, and formatting as the
     original.
- Use the same LaTeX commands, environments, style files, and macros as in the original
      paper.
- Ensure that LaTeX math environments are used correctly and that the same macros are
     used as in the original paper.
- Maintain similar overall length (+-5\%) to remain within a maximum of 9 pages for the
      main text.
- Use the same sections and preserve ALL figures, tables. Use the same labels for all
     elements (sections, figure, tables, etc.) and keep all of their references in the
     text.
- Also preserve all original citations. You can add citations based on the new BibTeX
     entries.
- The appendix content is provided for context only. You MUST keep all original \\input
     {{appendix_ORIGINAL/...}} commands unchanged in your output. Do NOT modify or
     inline any appendix content.
- Use technical, precise language appropriate for the ICLR audience-rigorous yet
     accessible.
- Ensure you are always writing good compilable LaTeX code, without any LaTeX syntax
     errors (unenclosed math, unmatched braces, etc.). The output should be ready to
     compile as-is.


# ORIGINAL PAPER (LaTeX):

{tex_content}

# END OF ORIGINAL PAPER


# OUTPUT THE COMPLETE IMPROVED PAPER IN LATEX FORMAT NOW, FOLLOWED BY NEW BIBTEX
     ENTRIES (IF ANY):
```

C ADDITIONAL RESULTS

Figures 3 and 4 show the full distributions for IntraSim and InterSim across review types.

C.1 PAPER LAUNDERING DRIVES INTELLECTUAL MONOCULTURE

If paper laundering becomes widespread, scientific writing will converge toward whatever style the AI reviewer rewards, risking an intellectual monoculture that discourages diverse ways of presenting ideas. The AI reviewing system would thus shape not only which papers are accepted, but also how scientific papers are written, potentially disadvantaging unconventional but valuable research.

AI reviewer score correlations

Table 2 reports Pearson correlations between reviewer scores. AI reviewers correlate more strongly with each other ($r = 0.49$) than human reviewers do ($r = 0.14$) in out 60-paper sample. AI-human correlations are weak. GPT shows moderate correlation ($r = 0.26$, $p < 0.001$) while Claude shows no significant correlation ($r = 0.12$, $p = 0.07$).

These results should be interpreted with caution due to the small sample size ($n = 60$ papers). Additionally, note that human scores reflect pre-rebuttal ratings. Reviewers typically update scores during discussion and reach a consensus before final decisions (Kargaran et al., 2025). Our AI-AI correlation of 0.49 is consistent with prior work reporting an average pairwise correlation of 0.48 among LLM reviewers (Bianchi et al., 2025).

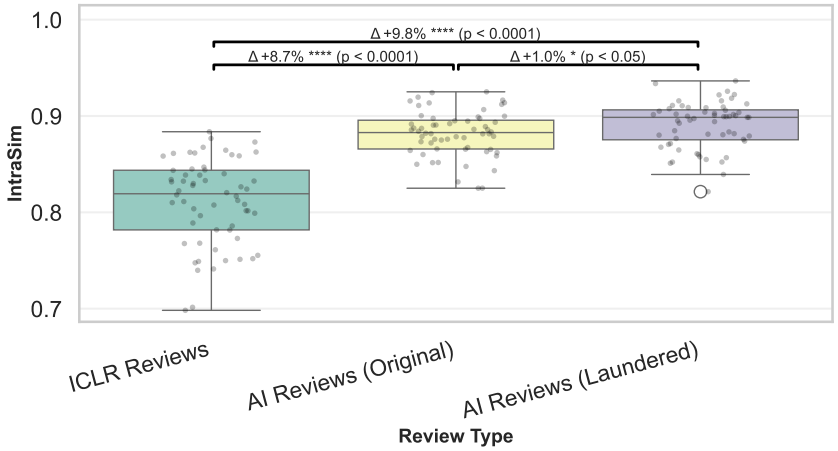

Figure 3: **Simulated AI reviewers show excessive within-paper agreement.** Intra-paper inter-reviewer similarity (IntraSim) compares human ICLR reviews with AI-generated reviews for original and laundered papers ($n = 60$ papers). ICLR human reviews: mean $= 0.811$. AI reviews of original papers: mean $= 0.882$ ($+8.7\%$, $p < 0.0001$, Cohen's $d = 1.47$). AI reviews of laundered papers: mean $= 0.891$ ($+9.8\%$ vs. ICLR, $p < 0.0001$, Cohen's $d = 1.67$).

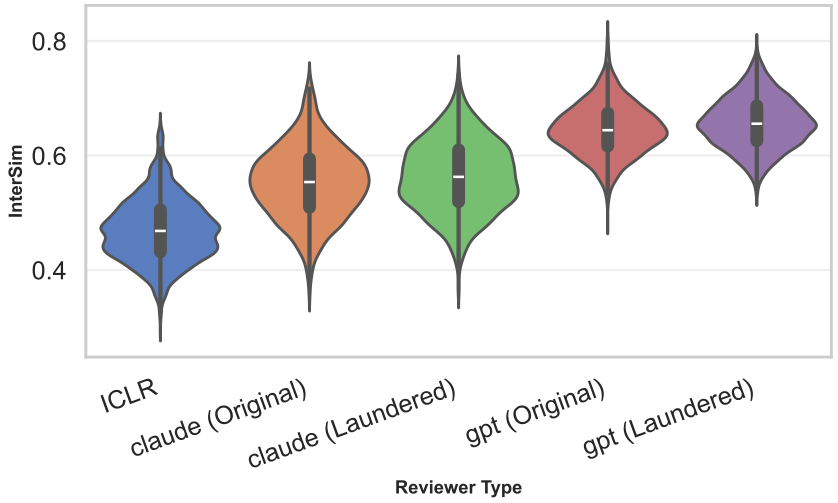

Figure 4: **AI reviewers produce similar reviews across different papers.** Inter-paper intra-reviewer similarity (InterSim) for human ICLR reviewers versus AI reviewer agents. ICLR human reviews: mean $= 0.470$. GPT-5.1 reviews show $+37.4\%$ (original) to $+39.8\%$ (laundered) higher similarity. Claude reviews show $+17.6\%$ (original) to $+20.0\%$ (laundered) higher similarity. All differences significant at $p < 0.0001$ with large effect sizes (Cohen's $d = 1.4$–3.8).

C.2 Common templates used in AI reviews

To understand why AI reviewers show high cross-paper review similarity (InterSim), we analyzed phrases commonly reused across reviewer types. For each reviewer category, we extracted all n-grams (phrases of 6–25 words) from the review texts and computed the percentage of reviews containing each phrase.

Table 3 summarizes template reuse. A phrase appearing in reviews for many different papers indicates templated feedback that is not specific to the content of the paper at hand. Table 4 shows the top 5 template phrases for each reviewer type.

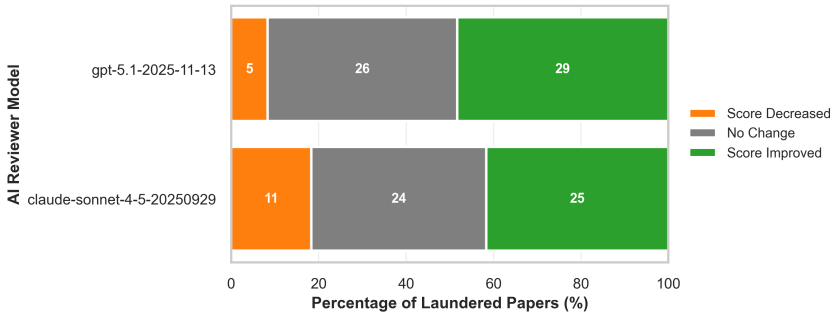

Figure 5: **Outcome distribution across papers.** Numbers indicate paper counts per category. The majority of papers receive higher scores from both AI reviewers after laundering.

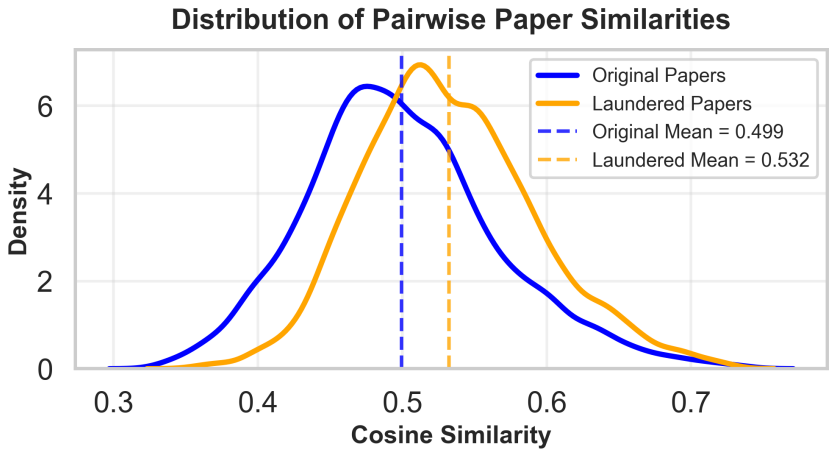

Figure 6: **Paper laundering drives intellectual monoculture.** Distribution of pairwise cosine similarity between paper embeddings (abstract + introduction) for original versus laundered papers ($n = 6{,}903$ paper pairs from 60 papers). Original papers: mean similarity $= 0.497$. Laundered papers: mean similarity $= 0.529$. The 6.5% increase is significant ($t = 84.8$, $p < 0.0001$, Cohen's $d = 1.02$).

Table 2: **Score correlations between reviewer types.** Pearson correlation coefficients for pairwise reviewer scores. Human-Human (all) includes all ICLR 2026 papers; other comparisons use our 60-paper sample. These results should be interpreted with a grain of salt, given the small sample size. Significance is indicated with: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Comparison	**Pearson r**	**p-value**	**n pairs**
Human-Human (all ICLR)	0.180***	<0.001	112,180
Human-Human (sample)	0.137**	0.009	359
AI-AI (GPT vs Claude)	0.492***	<0.001	60
GPT-Human	0.260***	<0.001	238
Claude-Human	0.119	0.066	238
All AI-Human	0.147**	0.001	476

Table 3: **Template phrase reuse reveals spurious AI agreement.** AI reviewer agents reuse the same phrases across 13–22% of papers, while ICLR reviewers (both AI-detected and human) show $< 1\%$ phrase reuse. We use a random subset of 2,000 reviews each for the ICLR reviews in the wild for computational efficiency.

Reviewer Type	Papers	Top-1 Coverage	Top-5 Avg.
GPT-5.1 Reviewer	60	13.3%	11.0%
Claude Reviewer	60	21.7%	16.7%
ICLR Fully AI (in the wild)	2,000	0.8%	0.6%
ICLR Human/Assisted (in the wild)	2,000	0.5%	0.5%

Table 4: **Top 5 template phrases by reviewer type.**

Reviewer	Phrase	Coverage
GPT-5.1	"if not, can you comment on"	13.3%
	"honest discussion of limitations and"	11.7%
	"there is no comparison to"	10.0%
	"limited analysis of failure modes and"	10.0%
	"it is not fully clear whether"	10.0%
Claude	"how does the method handle"	21.7%
	"can you provide more details on the"	18.3%
	"how does the method perform on"	15.0%
	"comprehensive experimental evaluation across multiple"	15.0%
	"how sensitive is the method to the choice of"	13.3%
ICLR Fully AI (in the wild)	"this paper addresses the problem of"	0.8%
	"this paper addresses the challenge of"	0.6%
	"rather than introducing a fundamentally new"	0.5%
	"could the authors clarify how the"	0.5%
	"could the authors comment on the"	0.4%
ICLR Human/Assisted (in the wild)	"advances in neural information processing systems..."	0.5%
	"this paper addresses the problem of"	0.5%
	"the paper is well-structured and clearly"	0.5%
	"it is recommended that the authors"	0.4%
	"it is not clear how the"	0.4%

C.3 What laundering changes

C.3.1 Manual inspection of laundered papers

We manually inspected the differences between original and laundered LaTeX file versions for five randomly selected papers that received an AI review score increase of at least one point. The majority of changes are stylistic: abstracts are rewritten with more confident language, the structure of the write-up in the introductions is changed with extended contributions, and the results in the conclusions are framed more relevantly.

When changes appear more substantive, they are typically AI-generated slop that does not improve scientific content. For example, one laundered paper gained a fabricated "Ablation: spatial clustering parameters" section reporting invented accuracy numbers across different parameter settings; another added an "Answers to reviewer questions" section responding to hypothetical concerns with generic explanations; a third introduced theoretical claims in the form of a new theorem without corresponding proofs in the original. One papers added a "Societal Impact" section (which does not seem like a bad idea given that it introduces a method to generate undetectable DeepFakes).

Other than that, most additions create an illusion of thoroughness, but are not grounded in actual experimental additions or genuine scientific work.

C.3.2 Analyzing word-level differences

To understand what paper laundering actually modifies, we compared original and laundered versions of all 60 papers. After removing LaTeX comments, we extracted all added and removed words, then categorized them using the following categorization: *hedging words* (terms expressing uncertainty like "may," "suggests," "approximately," "likely"), *emphasis words* (terms expressing confidence or importance like "strong," "robust," "crucial," "significantly"), *transition words* (discourse connectors like "however," "therefore," "moreover"), and *common filler words* (common academic boilerplate like "propose," "demonstrate," "framework," "novel"). Table 5 summarizes the per-paper averages, showing that paper laundering increased stylistic words.

Table 5: **What paper laundering changes (per paper).** Average word-level changes across 60 laundered papers. Laundering disproportionately adds hedging and emphasis language.

Category	Added	Removed	Change
Style modifiers			
Hedging words	52.4	29.4	+78.2%
Emphasis words	33.5	23.0	+45.2%
Structural words			
Transitions	43.9	47.1	−6.9%
Common filler words	218.9	223.7	−2.1%

C.4 Length statistics and embedding robustness

One concern is that embedding similarity may be confounded by text length. AI-generated ICLR reviews are significantly longer than human/assisted reviews (507 vs. 424 words on average; $t = 41.0$, $p < 0.0001$), as shown in Table 6. And our AI agent reviews are even longer (1,341 words), because of the detailed structured format we specified following Bianchi et al. (2025). However, correlations between length and similarity are weak for all experiments ($|r| < 0.13$. We further test whether length explains the AI-human similarity difference by restricting both groups to an overlapping length range (10th–90th percentile overlap, i.e., 261–672 words). AI reviews remain significantly more similar (mean $= 0.480$) than human/assisted reviews (mean $= 0.471$; $t = 13.7$, $p < 0.0001$, Cohen's $d = 0.14$). Thus, the AI hivemind effect persists after accounting for length.

Table 6: **Review length statistics by category.**

Category	N	Mean Words
AI Agent (experiment)	240	1,341
ICLR Fully AI-generated	15,899	507
ICLR Human/AI-assisted	59,901	424

D AI-generated reviews

We automatically generated reviews by feeding our own manuscript to AI reviewers, using the same setup as in our experiments. We paste the unedited result here (so that human reviewers don't need to do this).

AI review generated for this manuscript using GPT-5.1-2025-11-13

```
<summary>
```

```
The paper is a position piece arguing against the use of current large language models
     (LLMs) as autonomous paper reviewers. It introduces two necessary conditions for
     automating peer-review judgmentpreservation of review diversity and resistance
     to gamingand empirically shows that current AI reviewers fail both via (1) a "
     hivemind" homogenization of reviews and (2) a simple "paper laundering" procedure
     that boosts AI scores through zero-shot rewriting without substantive changes. The
      authors then articulate a broader research and governance agenda for a "science
     of peer review automation," outlining evaluation requirements, stakeholder studies
     , human-AI interaction research, and incentive design.
</summary>

<strengths>
- Strong, timely problem framing: The paper addresses an extremely consequential and
     topical issue (AI automation of peer review) with clear stakes for the ML
     community, referencing concrete ongoing conference policies and trials (Table 1 on
      page 3) and positioning itself relative to current practice rather than in the
     abstract.
- Clear articulation of necessary conditions: The two conditionsC1 (preservation of
     review diversity) and C2 (resistance to gaming) are  well-motivated, conceptually
      sharp, and repeatedly used as an organizing backbone for the empirical and
     normative arguments (Sections 1, 3, 4, and 6).
- Empirically grounded critique:
  - Hivemind effect "in the wild": Using 75,800 ICLR 2026 reviews with AI-generation
       labels (Emi, 2025), the paper shows significantly higher inter-paper similarity
       for fully AI-generated reviews (mean InterSim 0.486 vs. 0.467; Welchs t=3218,
       p<0.0001, Cohens d=0.29; Figure 1, page 4). This is a valuable, large-scale
       observational result.
  - Controlled simulations: On 60 sampled ICLR papers, AI reviewer agents ( GPT 5 .1,
       Claude 4.5) show markedly higher within-paper agreement (IntraSim ˜0.88-0.89 vs.
       0.81 for humans; +8.7 to +9.8%, Cohens d up to 1.67; Figure 2, page 4) and
       much higher cross-paper similarity (InterSim: GPT +37-40%, Claude +18-20% over
       humans; large effect sizes  d 1 .4-3.8; Figure 3, page 5). The dual use of "in
       the wild" and controlled data is methodologically sound and compelling.
  - Score behavior: AI review scores are inflated and mutually correlated (mean GPT
       7.3, Claude 6.1 vs. human 4.3; AI-AI r=0.49 vs. human-human  0  .18 at scale;
       Appendix C, Table 2, page 16), consistent with and extending prior work.
- Simple, concrete demonstration of gameability:
  - Paper laundering mechanism is precisely specified (Appendix B.2, pages 15-16) as a
       fully automated, zero-shot  GPT 5 .1 LaTeX-to-LaTeX rewrite driven by prior AI
       feedback.
  - On the 60-paper sample, laundering yields a statistically significant average score
        increase of +0.28 points (Wilcoxon p<0.001; Figure 4a, page 6), with 42-48% of
       papers improving and very few decreasing (Figure 4b).
  - The authors convincingly argue this is not due to genuine scientific improvement:
       word-level analysis (Table 5, page 18) shows disproportionate addition of
       hedging (+78%) and emphasis (+45%) terms, and manual inspection (Appendix E.1,
       pages 17-18) finds hallucinated content (fake ablations, unproved theorems,
       generic "answers to reviewers" sections).
- Evidence of induced stylistic monoculture: Pairwise similarity of paper embeddings (
     abstract+introduction) increases by 6.5% after laundering (0.497  0  .529, t=84.8,
     p<0.0001, Cohens d=1.02; Figure 5, page 6), directly supporting the claim that
     laundering drives convergence toward a homogeneous writing style.
- Good attention to robustness and limitations:
  - The paper checks that the hivemind effect is not just a length artifact (Appendix F
       , Table 6, page 18) and that conclusions are robust under length-matched subsets
       .
  - It explicitly discusses limitations (Appendix A, page 13): limited set of models
       and prompts, reliance on third-party AI-review labels, embedding-based
       similarity as an imperfect proxy for argumentative diversity, and focus on a
       single venue.
- Integration with and extension of prior literature: Related work on AI in peer review
     , LL.M. reviewers, prompt injection, adversarial attacks, and algorithmic
     monoculture is appropriately cited and used to situate the contribution (Section 2
      and references, pages 2-12). The paper is careful not to claim originality where
     there is prior work, but shows how its empirical findings sharpen or generalize
     existing concerns (e.g., from targeted attacks to trivial laundering).
- Thoughtful normative and design discussion:
  - Section 5 systematically addresses four common counterarguments (e.g., "humans are
       flawed too," "AI will improve") with reasonably nuanced points about distributed
        vs. centralized error, absence of ground truth for paper quality, and
       asymmetries in power and access.
  - Section 6 proposes concrete pillars for a "science of peer review automation,"
       including adversarial robustness testing, validated accuracy with acceptable
       false positives, transparency of deployment, stakeholder value studies, human-AI
        interaction research on overreliance and sycophancy, and improved reviewer
       incentives.
- Clarity and structure: The writing is very clear, well-organized, and easy to follow.
      Figures and tables (e.g., Figures 1-5 and Tables 1-6) are well-labeled and
```

```
     directly support claims. The overall narrative from empirical findings to position
      and recommendations is coherent.
</strengths>

<weaknesses>
- Limited model and configuration diversity for key experiments:
  - The main simulation results (hivemind effect and laundering) are based on only two
      proprietary, frontier models ( G P T 5 .1 and Claude Sonnet 4.5) and a single,
      relatively prescriptive review prompt (Appendix B.1, page 14). As acknowledged
      in Appendix A, this makes it difficult to infer how much of the homogenization
      is intrinsic to LLM reviewing vs. contingent on prompt design, temperature, or
      model family.
  - There is no exploration of simple mitigation baselines (e.g., higher temperatures,
      role prompts encouraging dissent, different review templates) that could
      partially reduce similarity; this weakens the prescriptive force of the claim
      that "AI reviewers lack diversity" in principle, as opposed to "under this very
      specific setup."
- Embedding similarity as a proxy for epistemic diversity:
  - Both IntraSim and InterSim are defined purely on embedding-based cosine similarity
      of full review texts (Section 3.2, equation (1)-(2), pages 3-4). As noted in the
       limitations, this does not directly capture differences in substantive
      critiques, priorities, or subjective judgments. Two reviews can be stylistically
       similar yet substantively divergent, or vice versa.
  - The paper does not complement embedding-based metrics with any qualitative or
      annotation-based assessment of whether AI reviews actually converge on the same
      points, miss different flaws, or disagree less on acceptance recommendations
      than humans. This makes the interpretation of "hivemind" more speculative than
      it could be.
- Evaluation of l a u n d e r i n g s practical impact is somewhat thin:
  - The reported +0.28 mean rating gain is mapped to "7.3 percentage-point increase in
      predicted acceptance probability" using ICLR 2025 data, but this mapping is only
       briefly mentioned (page 5) and not fully described (e.g., functional form,
      calibration, uncertainties). Given that actual accept/reject decisions involve
      committee discussion and meta-reviews, it is unclear how often such an increase
      would flip real decisions.
  - The 60-paper sample is relatively small and may not cover the full distribution of
      borderline vs. clearly strong/weak papers; laundering might have very different
      impact on marginal vs. obviously-accepted or obviously-rejected submissions.
- Lack of human evaluation of laundered manuscripts:
  - The central normative claim is that laundering "games" AI reviewers without
      improving scientific quality. While the word-level and manual inspections are
      informative, they are based on the a u t h o r s  judgment and a small convenience
      subset. There is no blinded human evaluation (e.g., domain experts rating
      original vs. laundered versions for clarity, perceived rigor, or acceptability)
      to robustly separate genuine improvements from surface optimization.
  - In particular, some changes (e.g., more explicit discussion of limitations, added "
      Societal Impact" sections, clarified contributions) may be genuinely beneficial,
       even if introduced by an LLM and sometimes embellished. The paper tends to
      underplay this nuance.
- Scope of "AI reviewers" is narrow:
  - The paper conflates, at times, "AI-generated reviews" (LLMs writing full textual
      reviews and scores) with a broader spectrum of AI support tools (checklist
      assistants, factuality checkers, error detectors, matching tools, etc.). While
      Section 6.1 attempts to distinguish task types, the empirical analysis focuses
      exclusively on the most extreme setting: LLMs as autonomous reviewers providing
      scores.
  - There is no empirical examination of hybrid regimes (e.g., AI-generated suggestions
       that are post-edited by humans, or AI feedback that does not include an overall
       rating). Some of the strong normative conclusions about "AI in peer review"
      could be more carefully scoped to "LLMs as primary sources of scores and textual
       reviews."
- Limited exploration of counterfactual designs that might mitigate issues:
  - For the hivemind effect, the paper does not experiment with or even conceptually
      engage much with straightforward diversification strategies (e.g., ensembles of
      heterogeneous models, stochastic decoding with global constraints, or explicit "
      disagreement-seeking" prompts).
  - For gameability, there is no exploration of simple defenses such as detecting LLM-
      rewritten style, cross-checking content against arXiv versions, or using AI as
      an aid for meta-reviewers to flag suspicious manuscripts. This makes the problem
       appear more hopeless than it might be in a realistically defended system.
- Some claims about "intellectual monoculture" are somewhat speculative:
  - The 6.5% increase in cosine similarity between abstracts/introductions after
      laundering (Figure 5, page 6) is a meaningful indicator of stylistic convergence
      , but it is a single-step experiment on a small sample. The paper extrapolates
      from this to a broader claim that AI reviewing will shape "how scientific papers
       are written" and "discourage unconventional research," without longitudinal
      data or behavioral evidence.
```

```
  - There is limited discussion of how strong institutional norms, templates, and
      reviewer expectations already homogenize papers even without AI and how much
       additional marginal harm is attributable to AI-based laundering.
- Reproducibility/dependence on proprietary infrastructure:
  - Key components including GPT 5.1, Claude Sonnet 4.5, and OpenAIs text-
      embedding-3-small are proprietary. While the prompts and procedures are
      described in detail (Appendix B, pages 14-16), reproducing the exact numbers may
       be difficult or impossible for others without access to these specific models
      or versions.
  - There is no public release (stated in the main text) of the 60-paper sample, the
      generated reviews, or laundering outputs, which limits independent verification
      and follow-up analysis.
</weaknesses>

<questions>
- How sensitive are your hivemind and laundering results to prompt and decoding choices
    ?
  - Have you tried alternative review prompts that (i) explicitly encourage
      disagreement or multiple perspectives, or (ii) adopt a less rigid XML/section
      template, and, if so, how did that affect IntraSim and InterSim?
  - Did you experiment with varying temperature or sampling parameters for review
      generation, and does higher stochasticity meaningfully reduce cross-paper
      similarity without sacrificing coherence?
- Can you provide more detail on the mapping from rating shifts to acceptance
    probabilities?
  - How exactly is the "7.3 percentage-point increase in predicted acceptance
      probability" computed from ICLR 2025 data (model type, features, calibration,
      and goodness-of-fit)?
  - Does the mapping differ significantly across score ranges (e.g., from 4  4  .3 vs.
      6.5  6  .8), and where in the distribution do your 60 sampled papers lie?
- To what extent do human experts perceive laundered papers as higher-quality?
  - Have you run (or would you consider running) a blinded human study where domain
      experts are asked to (a) distinguish original vs. laundered versions, and (b)
      rate them on clarity, rigor, and overall acceptability?
  - Do human reviewers also tend to give higher scores to laundered versions, and if so
      , does that change your interpretation of "gaming" vs. "improvement"?
- How much of the hivemind effect is attributable to stylistic templates vs.
    substantive content overlap?
  - Beyond n-gram reuse (Table 4, page 17), have you analyzed whether AI reviews focus
      on the same specific strengths/weaknesses across papers (e.g., similar topical
      tags or critique categories), or whether they are simply using similar language
      to express paper-specific assessments?
  - Could you incorporate an annotation-based measure of argument diversity for
      instance, by coding a subset of reviews for distinct critique types and
      correlate that with embedding similarity?
- Have you explored simple defenses or alternative architectures that might mitigate
    laundering?
  - For example, would using a second, independent model (or a different modality, such
       as direct PDF rendering) to check consistency between text and figures/tables
      help detect hallucinated ablations or invented sections introduced by laundering
      ?
  - Could style-based detectors, or comparisons to arXiv/preprint versions, flag
      heavily LLM-rewritten manuscripts in a way that reduces the feasibility of
      systematic laundering?
- How do your conclusions apply to more constrained AI support roles?
  - Your strongest arguments target LLMs generating full reviews and scores. How would
      you adjust your position for tools that only provide (i) checklists, (ii)
      grammar/style suggestions without rephrasing technical claims, or (iii)
      factuality/consistency checks with high-precision thresholds?
  - Are there specific peer-review subtasks where, based on your evidence, you would
      currently endorse carefully constrained AI use?
</questions>

<rating>8.0</rating>
```

AI review generated for this manuscript using Claude-Sonnet-4-5-20250929

```
<summary>
This position paper argues against using current AI systems to produce peer reviews at
    scientific conferences. The authors analyze 75,800 ICLR 2026 reviews and conduct
    controlled experiments with 60 papers. They identify two critical failures: (1) AI
     reviewers exhibit a "hivemind effect" with excessive agreement both within and
    across papers compared to human reviewers, reducing perspective diversity; (2) AI
    review scores are trivially gameable through "paper laundering" zero -shot
```

```
    automated rewrites that boost scores without genuine scientific improvements. The
    paper proposes that resistance to gaming and preservation of review diversity are
    necessary but not sufficient conditions for AI peer review automation, calling for
     rigorous evaluation standards before deployment.
</summary>

<strengths>
- Addresses a timely and important problem as conferences increasingly adopt AI-
    assisted peer review systems
- Provides concrete empirical evidence from both real ICLR 2026 reviews (75,800 reviews
    ) and controlled experiments (60 papers)
- Introduces the novel concept of "paper laundering" as a concrete failure mode
    distinct from adversarial attacks
- Demonstrates practical gameability: zero-shot rewrites cost only $0.25 and require no
     optimization or hidden prompts
- Provides actionable framework with three concrete requirements for deployment (
    adversarial robustness testing, validated accuracy, transparency)
- Balanced position that doesn't oppose all AI use but calls for rigorous evaluation
- Addresses multiple plausible objections systematically (Section 5)
- Strong methodological rigor with statistical tests, effect sizes, and robustness
    checks
- Transparent about limitations and provides detailed implementation details for
    reproducibility
</strengths>

<weaknesses>
- Limited scope: experiments use only two AI models (GPT-5.1, Claude) with a single
    prompt, which may not generalize to diverse deployment scenarios
- Sample size of 60 papers for controlled experiments, while statistically sufficient,
    may not capture full diversity of paper types and quality levels
- Reliance on third-party AI-generation labels (Emi 2025) for "in the wild" analysis
    introduces potential classification errors
- Embedding-based similarity metrics capture linguistic patterns but may not directly
    measure argumentative or evaluative diversity
- Paper laundering experiments use only GPT-5.1 for rewriting; other models may show
    different patterns
- Limited analysis of what specific types of scientific improvements could legitimately
     improve AI review scores
- Does not provide concrete metrics or thresholds for "acceptable" levels of similarity
     or gameability
- The paper's own position could be seen as conservative given rapid AI improvements
- Missing discussion of potential benefits of AI consistency (e.g., reducing human
    biases like gender/prestige bias)
</weaknesses>

<questions>
- How would the hivemind effect and gameability manifest with more diverse prompting
    strategies, temperatures, and model ensembles?
- Could the authors develop metrics that directly measure argumentative diversity
    rather than just linguistic similarity?
- What proportion of human reviews might also benefit from "laundering" (e.g., improved
     clarity without changing substance)?
- How do the authors propose to operationalize "adversarial robustness testing" at
    scale before conference deployment?
- Could the paper laundering effect be mitigated through adversarial training or prompt
     engineering?
- What would constitute sufficient evidence that an AI reviewing system has met the
    necessary conditions outlined?
- How do the authors reconcile their findings with studies showing AI can help reduce
    certain human biases in review?
- Would the proposed requirements (Section 6.1) actually be feasible for conference
    organizers to implement?
- How would the framework handle hybrid systems where AI assists but doesn't replace
    human judgment?
- What specific stakeholder values (Section 6.2) do the authors believe should take
    priority in automation decisions?
</questions>

<rating>7.5</rating>

This is a well-executed position paper on an important and timely topic. The empirical
    evidence is solid, combining large-scale observational data with controlled
    experiments. The "paper laundering" concept is novel and demonstrates a concrete,
    practical failure mode distinct from previously studied adversarial attacks. The
    framework calling for rigorous evaluation before deployment is sensible and
    actionable.
```

However, the paper has limitations that prevent a higher score. The experimental scope (2 models, 1 prompt, 60 papers) is somewhat narrow, and the reliance on linguistic similarity as a proxy for review diversity is imperfect. The paper would benefit from stronger engagement with potential benefits of AI consistency and more concrete operationalization of the proposed requirements.

The position is valuable and the evidence supporting it is convincing enough to merit acceptance. The work will likely influence ongoing debates about AI in peer review and provides a useful framework for evaluation. While not groundbreaking methodologically, it makes important empirical and conceptual contributions to a critical problem facing the scientific community.