From Doxa to Logos in Scientific Peer Review

Negar Arabzadeh
Reviewerly, UC Berkeley
Berkeley, California, United States

Sajad Ebrahimi
Reviewerly
Toronto, Ontario, Canada

Alireza Daghighfarsoodeh
Reviewerly
Toronto, Ontario, Canada

Soroush Sadeghian
Reviewerly
Toronto, Ontario, Canada

Seyed Mohammad Hosseini
Reviewerly
Toronto, Ontario, Canada

Hai Son Le
Reviewerly
Toronto, Ontario, Canada

Mahdi Bashari
Reviewerly
Toronto, Ontario, Canada

Ebrahim Bagheri
University of Toronto, Reviewerly
Toronto, Ontario, Canada

Abstract

Peer review is central to scientific decision-making, yet it is rarely evaluated or audited at scale. Growing submission volumes and the increasing use of large language models (LLMs) in drafting reviews have introduced new challenges for transparency, accountability, and quality control. Today, peer reviews are often produced through hybrid human–AI workflows, where a reviewer may develop the core evaluative ideas while using an LLM to refine wording, restructure arguments, or improve fluency. This shift raises new questions beyond authorship detection alone: Are reviews constructive? Are reviewer claims grounded in the submitted paper? How can we quantify collaboration between human reasoning and AI-assisted writing, and distinguish whether the intellectual contribution or the surface text originates from humans or models?

In this industry talk, we present Reviewerly's retrieval-centered infrastructure for auditing peer review at scale. We describe three deployed systems: *Peeriscope*, which evaluates review quality across multiple interpretable dimensions; *Peerispect*, which uses retrieval-augmented generation (RAG) to verify whether reviewer claims are supported by evidence in the manuscript; and *PeerPrism*, which analyzes hybrid human–AI authorship by disentangling the origin of ideas from the origin of text in peer reviews, enabling measurement of how human reasoning and AI-generated writing interact within a review. We will share architectural design decisions, lessons learned from real-world deployment, and practical trade-offs among model complexity, interpretability, computational efficiency, and predictive accuracy. The talk will include short live demonstrations[1] [2] of our systems to illustrate how the combination of retrieval systems and LLM-based pipelines can improve peer review quality and strengthen transparency and research integrity across high-volume decision-making environments.

Keywords

Peer Review, Human-AI Collaboration, Research Integrity

ACM Reference Format:
Negar Arabzadeh, Sajad Ebrahimi, Alireza Daghighfarsoodeh, Soroush Sadeghian, Seyed Mohammad Hosseini, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. 2026. From Doxa to Logos in Scientific Peer Review. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] https://app.reviewer.ly/app/peerispect
[2] https://app.reviewer.ly/app/peeriscope

1 Introduction

Peer review has long been the main mechanism for evaluating scientific manuscripts and ensuring research quality [8, 12]. It shapes publication decisions, influences funding outcomes, and determines what knowledge becomes part of the scientific record. Yet despite its central role, peer review itself is rarely evaluated or audited at scale. In recent years, submission numbers at major conferences and journals have grown rapidly [11, 14]. Editors and program chairs now manage thousands of papers per cycle, making careful manual oversight increasingly difficult. This scaling bottleneck creates a gap in accountability since vague, biased, or unsupported critiques can strongly influence decisions, often without systematic verification [9, 21]. At the same time, what counts as a constructive and high-quality review varies widely across fields, and clear quality standards are rarely articulated. In practice, reviews are seldom audited, and there are limited mechanisms to address consistently low-quality feedback. Importantly, this is not due to a lack of effort from organizers, editors, area chairs, or program committee members. Rather, the scale of modern review processes makes comprehensive oversight extremely difficult.

At the same time, the rapid adoption of large language models (LLMs) has introduced new complexity into peer review [1, 22]. Reviewers increasingly use generative AI tools to draft, refine, or expand their feedback [13]. While these tools can improve fluency and organization, they also introduce risks. LLM-generated content may sound confident and well-structured while lacking conceptual depth, misrepresenting the paper, or including unsupported claims [5, 10, 19, 26]. Such issues are difficult to detect at scale and can further weaken transparency and reliability in high-volume editorial pipelines [17, 27]. We argue that the inability to systematically evaluate peer reviews is not merely an academic limitation—it is a critical infrastructure gap. Without scalable and transparent tools to assess review quality, verify claims, and surface ungrounded feedback, research communities lack effective mechanisms to detect when the peer review process fails.

To address these concerns, many communities have turned to AI text detection systems [15, 20, 23]. However, these approaches are limited by how they frame the problem. Most existing benchmarks treat authorship as a simple binary choice where text is either fully human-written or fully machine-generated. This assumption does not reflect current practice. In reality, many reviews are produced through collaboration, where a human reviewer develops the core

evaluative ideas and an LLM helps rewrite, expand, or polish the final text [18]. This creates a gray area between human and AI authorship. A reviewer may carefully analyze the paper, formulate thoughtful critiques, and then use an LLM only to improve clarity or writing style. The final review might appear "LLM-like" in tone, but the intellectual contribution remains human. Should such reviews be classified as AI-generated? Current detection systems offer little guidance for these hybrid scenarios.

More importantly, authorship alone is not the central issue. Even if we know that a review is fully human-written, we still need to assess whether it is constructive, high-quality, and grounded in the submitted paper. A human review can still be vague, biased, or factually unsupported. Conversely, an AI-assisted review might be well-structured but lack depth or misrepresent the manuscript. This highlights two distinct but related challenges: (1) evaluating the quality and groundedness of peer reviews, and (2) understanding the role of LLMs in review authorship. Focusing only on detection does not solve the broader problem of ensuring reliable and accountable evaluation in peer review. When detection systems do not distinguish between the origin of the ideas and the origin of the surface text, they risk misclassifying legitimate human contributions. As a result, human expertise may be penalized simply because the writing was refined with AI assistance.

At Reviewerly[3], we believe that supporting research integrity at scale requires more than binary AI detection [2, 3]. Over the past two years, we have designed a set of systems that combine large language models with Information Retrieval (IR) methods to systematically audit different aspects of the peer review process. We present three components of this infrastructure: `Peeriscope`, which evaluates review quality across multiple dimensions; `Peerispect`, which uses Retrieval Augmented Generation (RAG) pipelines to verify whether reviewer claims are grounded in the submitted paper; and `PeerPrism`, which analyzes hybrid human–AI authorship by disentangling idea origin from text origin.

Together, these systems reflect a key insight: reviews can no longer be meaningfully categorized as simply "human-written" or "AI-generated." In practice, reviewers often develop the core evaluative reasoning themselves while using LLMs to refine clarity or presentation. Authorship alone is therefore insufficient—regardless of who produced the text, reviews must be constructive, evidence-based, and grounded in the manuscript.

By jointly examining review quality, factual grounding, and human–AI collaboration, we move beyond detection toward structured, retrieval-centered evaluation. This talk will describe the architectural design of these systems, their evaluation methodologies, and the practical challenges encountered in building scalable auditing tools, along with lessons learned and trade-offs among interpretability, model complexity, and evaluation reliability in hybrid human–AI workflows.

Although our motivating case is academic peer review, it is worth mentioning that the need for scalable and automated assessment of expert judgment extends far beyond academic publishing. Similar challenges arise wherever expert feedback influences high-stakes decisions, including funding agencies, industry R&D evaluation, grant review panels, regulatory audits, and marketplace moderation systems. As such, a scalable automated retrieval-grounded approach to auditing expert judgment enables more transparent, scalable, and accountable decision-making across domains beyond academia.

[3] https://reviewer.ly/

2 Motivation and Design Principles

Designing infrastructure for review auditing requires rethinking how peer evaluation itself is measured. The goal is not simply to detect AI usage or assign a single quality score, but to construct a framework that reflects the structural complexity of modern review workflows. Two observations guide this reframing.

First, review quality is inherently multidimensional. A review may be fluent yet vague, critical yet unsupported, or detailed yet misaligned with the manuscript. Reliable evaluation therefore demands structured assessment across interpretable dimensions, such as specificity, justification, conceptual depth, argumentative coherence, tone, and factual grounding, rather than relying on surface-level signals or stylistic heuristics.

Second, authorship in contemporary peer review is increasingly hybrid. Reviews often emerge from collaboration between human expertise and AI assistance. A reviewer may develop the substantive evaluative reasoning, while an LLM refines clarity or reorganizes the presentation. In such settings, idea provenance and textual realization may originate from different sources. Binary detection frameworks fail to capture this nuance and risk penalizing legitimate human contributions simply because the final text appears stylistically "LLM-like."

Crucially, attribution alone does not determine trustworthiness. A fully human-written review may still lack grounding, while an AI-assisted review may vary widely in rigor and alignment with the manuscript. Ensuring accountability therefore requires evaluation beyond authorship detection. These observations motivate a multidimensional, retrieval-centered auditing framework with the following design principles:

- **Multidimensional Quality Assessment:** Evaluate reviews across interpretable dimensions such as specificity, justification, tone, conceptual depth, and argumentative coherence.
- **Grounded Claim Verification:** Verify whether reviewer claims are supported by evidence in the submitted manuscript, using retrieval-based methods to ensure factual alignment.
- **Human–AI Collaboration Identification:** Distinguish between idea provenance and text realization to better understand hybrid authorship regimes.
- **Interpretability for Stakeholders:** Provide transparent signals and explanations that editors, program chairs, funding agencies, and industrial evaluators can trust and act upon.

At Reviewerly, these requirements define the foundation for scalable infrastructure that strengthens transparency and accountability in high-volume peer review systems.

3 Multidimensional Peer Review Evaluation

To operationalize the design principles outlined in Section 2, we developed an interpretable auditing infrastructure that addresses three dimensions of review reliability of quality assessment, factual grounding, and human–AI collaboration. Across all components, we prioritize transparency, stability, and deployment-readiness over

opaque black-box scoring. In this talk, we will present the architectural design of each system, discuss key technical and deployment challenges, report empirical evaluation results, and provide short live demonstrations illustrating how these tools function in real editorial workflows.

3.1 Peeriscope: Multidimensional Peer Review Quality Assessment

Review quality is inherently multidimensional and subjective, making automated evaluation particularly challenging. To ground automated assessment in expert judgment, we build upon the RottenReviews framework [5], which serves as the foundation for `Peeriscope`[4]. This framework leverages a large-scale dataset comprising over 15,000 paper submissions from four academic venues, enriched with more than 9,000 reviewer scholarly profiles and comprehensive paper metadata [5]. To better understand which measurable signals align with expert perceptions of review quality, we curated a gold-standard subset of over 700 paper–review pairs, meticulously annotated by domain experts across 13 interpretable dimensions, including specificity, justification, tone, and argumentative depth. These annotations enable systematic analysis of the linguistic, structural, and reviewer-related factors that correlate with human judgments of review quality.

Building on this dataset, `Peeriscope` operates as a structured scoring engine based on transparent and quantifiable review-dependent and reviewer-dependent metrics. Lightweight regression models leverage interpretable signals such as reviewer–paper topical alignment, hedging frequency, lexical diversity, and section-level coverage. Extensive empirical evaluation revealed a notable insight: simple, interpretable models aligned more consistently with expert human judgments than advanced generative evaluators. Both zero-shot and fine-tuned LLMs (e.g., GPT-4o, Qwen-3, Phi-4) demonstrated limited agreement with expert assessments of peer review quality. Although LLM-based scoring improved modestly after fine-tuning, it remained sensitive to edge cases and required substantially higher computational resources. These findings highlight the importance of stability, transparency, and interpretability in high-stakes editorial environments, suggesting that carefully designed feature-based learning approaches can provide more reliable signals for quality auditing than opaque LLM-based evaluators.

3.2 Peerispect: Retrieval-Based Grounding Verification

Well-written reviews may still contain critiques that are subjective, rhetorical, or factually unsupported by the manuscript. Assessing whether review statements are verifiable is crucial for fairness, yet manually inspecting the grounding of such claims at the scale of modern conferences is infeasible. To address this, we developed `Peerispect`[5], an interactive, retrieval-centered verification system designed to operationalize claim-level verification in peer reviews.

`Peerispect` is architected as a modular information retrieval pipeline that supports the integration of alternative retrievers, rerankers, and verifiers. The system follows a structured retrieval-augmented generation workflow: first, an LLM parses the review to extract discrete, check-worthy claims; next, a neural retriever identifies and isolates relevant evidence passages from the submitted manuscript; finally, a Natural Language Inference (NLI) entailment model determines whether the retrieved textual evidence supports or contradicts each extracted claim.

To ensure rigorous evaluation, we construct a dedicated benchmark for this task, providing a standardized dataset for measuring how accurately models map complex reviewer feedback to specific manuscript evidence. Beyond backend verification, `Peerispect` presents its results through a visual interface that highlights the retrieved evidence directly within the paper. This enables editors and area chairs to perform rapid visual inspection and interpretation. By explicitly linking review statements to textual evidence, `Peerispect` allows editorial teams to flag unsupported critiques efficiently. This reduces the risk of penalizing authors for issues already addressed in the paper and strengthens factual accountability without requiring exhaustive manual review.

3.3 PeerPrism: Human–AI Collaboration Analysis

Authorship in modern review workflows increasingly reflects collaboration between human reasoning and AI-assisted text generation. Rather than treating authorship as a rigid binary classification problem, `PeerPrism` models it as a multidimensional spectrum, explicitly disentangling the provenance of evaluative ideas from the provenance of surface text.

To map this spectrum, we constructed a comprehensive benchmark of 20,690 peer reviews derived from 160 seed papers across high-tier venues (ICLR and NeurIPS). Our data collection ensures temporal diversity by capturing both pre-LLM (2021–2022) and post-LLM (2023–2024) reviewing behaviors. Using this foundation, we generated controlled transformations to simulate realistic collaboration regimes, including fully synthetic reviews as well as hybrid states in which human ideas were rewritten, expanded, or passed through an extract-and-regenerate pipeline by an LLM.

We evaluated a diverse set of state-of-the-art detection baselines against this corpus, spanning likelihood-based models (GLTR) [6], likelihood-ratio metrics (Binoculars), perturbation-based approaches (DetectGPT, Fast-DetectGPT, Lastde++) [4, 16, 24], supervised classifiers (RADAR), and context-aware embedding frameworks (Anchor) [7, 25]. Our large-scale evaluation revealed consistent and significant vulnerabilities in current AI detection methods. While baseline detectors successfully distinguished entirely human texts from entirely synthetic ones, their consensus collapsed under hybrid authorship regimes. This exposed a fundamental task fragmentation: likelihood- and statistical-based models primarily captured surface generation signals (identifying that an LLM generated the specific tokens), whereas embedding-based approaches captured semantic inheritance (identifying that the underlying argumentative logic originated from a human).

Our findings from both detection evaluations and stylistic and linguistic analyses comparing human-written and LLM-generated reviews demonstrate that standalone binary AI detection is conceptually flawed and insufficient for editorial deployment. Attribution scores proved highly sensitive to generator choice and model updates, introducing volatility that undermines reliability. In practice,

[4] https://github.com/Reviewerly-Inc/Peeriscope
[5] https://github.com/Reviewerly-Inc/Peerispect

editors require interpretable collaboration indicators that are contextualized alongside quality and grounding assessments, rather than isolated probabilistic labels that risk penalizing human reasoning simply because its presentation was refined by an AI.

4 Real-World Deployment Challenges

Deploying review auditing systems in real editorial environments revealed that the primary challenges lie not in incremental improvements to model accuracy, but in system-level constraints related to scale, efficiency, and operational reliability. Editorial platforms must process thousands of submissions under strict decision timelines, heterogeneous reviewing norms, and continuously evolving language models.

A central challenge is scalability, which fundamentally depends on efficiency and low-latency system design. Grounding verification requires high-recall retrieval over long manuscripts, where reviewer claims are short, implicit, and often only loosely aligned with document structure. Reliable evidence matching therefore demands structure-aware indexing and staged retrieval pipelines that balance recall with computational cost. At conference scale, naïvely applying expensive reranking or LLM reasoning to every claim becomes prohibitively slow and costly, making selective inference, caching strategies, and careful allocation of computation essential. Review auditing systems must operate within tight review cycles while supporting concurrent access by editors and program chairs. Precomputing representations, batching retrieval operations, and maintaining lightweight interpretable models proved critical for reducing latency and keeping infrastructure costs manageable. In practice, scalability depended less on increasing model complexity and more on designing efficient retrieval and inference pipelines that minimize computational overhead while preserving evaluation reliability.

Another key requirement is explainability. Editorial stakeholders require actionable evidence rather than opaque scores. Automated assessments must not only produce judgments, but also justify them by exposing interpretable artifacts such as extracted claims, retrieved passages, and dimension-level quality signals. For example, when a system identifies a review as low quality or flags a critique as unsupported, it must indicate which aspects lack specificity, justification, or grounding in the manuscript. Such transparent reasoning traces enable stakeholders to understand why a particular decision was made, supporting informed oversight and maintaining reviewer trust in high-stakes evaluation settings.

Another emerging challenge is the inherently multimodal nature of scientific documents. Manuscripts extend beyond plain text and often include figures, tables, mathematical expressions, and visual results that contain essential evidence supporting reviewer claims. Verifying groundedness therefore cannot rely solely on textual retrieval. Effective auditing systems must incorporate document structure understanding and multimodal parsing to interpret visual elements alongside text. This introduces additional complexity in indexing, retrieval, and reasoning pipelines, as evidence relevant to a reviewer's statement may reside in captions, diagrams, or tabular results rather than narrative sections alone.

Taken together, these experiences demonstrate that production-grade review auditing is fundamentally a joint LLM and IR systems problem where success depends not only on intelligent models, but on the effective integration of LLM reasoning with scalable, efficient, and interpretable retrieval infrastructure capable of operating reliably within real-world evaluation ecosystems.

5 Company Portrait

Reviewerly is a Toronto-based AI company and university spin-off dedicated to strengthening research integrity through scalable, retrieval-driven auditing tools for peer review. Founded by researchers with experience in both academia and industry, Reviewerly develops modular systems that evaluate review quality, verify reviewer claims, and support transparent editorial workflows. Its platform integrates large language models with information retrieval pipelines to analyze review reports, assess groundedness, and generate actionable insights for editors and stakeholders. Reviewerly interoperates seamlessly with widely used peer review systems such as Open Journal Systems (OJS) through robust APIs, enabling automated review analysis, reviewer vetting, and workflow support at scale.

The technology is used by academic conferences, journals, funding agencies, research institutions, and enterprise R&D teams seeking greater accountability and reliability in expert evaluation. By combining automation with interpretable analytics, Reviewerly helps stakeholders reduce administrative burden while improving transparency, fairness, and decision quality across high-stakes review processes.

6 Speaker's Bio

Dr. Negar Arabzadeh is Head of Data Science at Reviewerly and a Postdoctoral Researcher at the Sky Lab, UC Berkeley, supervised by Prof. Matei Zaharia. Her research focuses on information retrieval and retrieval-augmented generation systems, with a particular emphasis on evaluating and deploying LLM-powered systems in production. She received her Ph.D. from the University of Waterloo. She has authored over 60 publications in top venues including SIGIR, EMNLP, EACL, CIKM, WSDM, and ECIR. Her work has received multiple Best Paper Awards at WSDM, ECIR, and SIGIR-AP. She has delivered tutorials at SIGIR, WSDM, and ECIR, and has organized workshops and competitions at NeurIPS and SIGIR. She has also held research positions at Microsoft Research, Spotify Research, and Google Brain.

References

[1] Sangzin Ahn. 2024. The transformative impact of large language models on medical writing and publishing: current applications, challenges and future directions. *The Korean journal of physiology & pharmacology: official journal of the Korean Physiological Society and the Korean Society of Pharmacology* 28, 5 (2024), 393–401.

[2] Negar Arabzadeh, Sajad Ebrahimi, Ali Ghorbanpour, Soroush Sadeghian, Sara Salamat, Muhan Li, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. 2025. Building Trustworthy Peer Review Quality Assessment Systems. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 6863–6864.

[3] Negar Arabzadeh, Sajad Ebrahimi, Sara Salamat, Mahdi Bashari, and Ebrahim Bagheri. 2024. Reviewerly: modeling the reviewer assignment task as an information retrieval problem. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5554–5555.

[4] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *International Conference on Learning Representations*.

[5] Sajad Ebrahimi, Soroush Sadeghian, Ali Ghorbanpour, Negar Arabzadeh, Sara Salamat, Muhan Li, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. 2025. RottenReviews: Benchmarking Review Quality with Human and LLM-Based Judgments. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM 2025)*. Seoul, Korea. doi:10.1145/3746252.3761436

[6] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of ACL*.

[7] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR.

[8] Tim Hillard and Rod Baber. 2021. Peer review: the cornerstone of scientific publishing integrity. 107–108 pages.

[9] Hugo Horta and Jisun Jung. 2024. The crisis of peer review: Part of the evolution of science. *Higher Education Quarterly* 78, 4 (2024), e12511.

[10] Mohammad Hosseini and Serge PJM Horbach. 2023. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research integrity and peer review* 8, 1 (2023), 4.

[11] Odest Chadwicke Jenkins and Matthew E. Taylor. 2025. AAAI-26 Review Process Update: Scale, Integrity Measures, and Experimental Use of AI-Assisted Reviewing. https://aaai.org/conference/aaai/aaai-26/review-process-update/. AAAI-26 Program Chairs; with contributions from Bo An, Joydeep Biswas, David J. Crandall, Matthew Lease, Kiri L. Wagstaff, Sven Koenig, Eric Eaton, Kevin Leyton-Brown, and Stephen Smith. Accessed: 2025-09-15.

[12] Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. 2014. Peer review in scientific publications: benefits, critiques, & a survival guide. *Ejifcc* 25, 3 (2014), 227.

[13] Jisoo Lee, Jieun Lee, and Jeong-Ju Yoo. 2025. The role of large language models in the peer-review process: opportunities and challenges for medical journal reviewers and editors. *Journal of Educational Evaluation for Health Professions* 22 (2025).

[14] Seth S Leopold. 2015. Increased manuscript submissions prompt journals to make hard choices. *Clinical Orthopaedics and Related Research®* 473, 3 (2015), 753–755.

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

[16] Eric Mitchell et al. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Learning Representations*.

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

[18] Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences* 122, 8 (2025), e2422455122.

[19] Zachary Robertson. 2023. Gpt4 is slightly helpful for peer-review assistance: A pilot study. *arXiv preprint arXiv:2307.05492* (2023).

[20] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156* (2023).

[21] Carolina Tropini, B Brett Finlay, Mark Nichter, Melissa K Melby, Jessica L Metcalf, Maria Gloria Dominguez-Bello, Liping Zhao, Margaret J McFall-Ngai, Naama Geva-Zatorsky, Katherine R Amato, et al. 2023. Time to rethink academic publishing: the peer reviewer crisis. e01091–23 pages.

[22] Eva AM Van Dis, Johan Bollen, Willem Zuidema, Robert Van Rooij, and Claudi L Bockting. 2023. ChatGPT: five priorities for research. *Nature* 614, 7947 (2023), 224–226.

[23] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity* 19, 1 (2023), 1–39.

[24] Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. 2025. Training-free LLM-generated Text Detection by Mining Token Probability Sequences. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=vo4AHjowKi

[25] Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. 2025. Is Your Paper Being Reviewed by an LLM? Benchmarking AI Text Detection in Peer Review. https://api.semanticscholar.org/CorpusID:276647742

[26] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*. 9340–9351.

[27] Lingxuan Zhu, Yancheng Lai, Jiarui Xie, Weiming Mou, Lihaoyun Huang, Chang Qi, Tao Yang, Aimin Jiang, Wenyi Gan, Dongqiang Zeng, et al. 2025. Evaluating the potential risks of employing large language models in peer review. *Clinical and Translational Discovery* 5, 4 (2025), e70067.