Can LLMs Uphold Research Integrity? Evaluating the Role of LLMs in Peer Review Quality

Negar Arabzadeh
UC Berkeley and Reviewer.ly
Toronto, Canada

Sajad Ebrahimi
Reviewer.ly
Toronto, Canada

Soroush Sadeghian
Reviewer.ly
Toronto, Canada

Mohammad Hosseini
Reviewer.ly
Toronto, Canada

Alireza Daqiq
Reviewer.ly
Toronto, Canada

Hai Son Le
Reviewer.ly
Toronto, Canada

Mahdi Bashari
Reviewer.ly
Toronto, Canada

Ebrahim Bagheri
University of Toronto and Reviewer.ly
Toronto, Canada

Abstract

While large language models (LLMs) have been widely studied in scholarly workflows, e.g., for citation recommendation and literature summarization and more, their role in supporting research integrity remains underexplored. In this talk, we share our experience building and deploying two real-world systems that audit peer reviews and verify their factual grounding at scale.

We evaluate the capabilities and limitations of LLMs in two key tasks: (1) assessing review quality along dimensions like specificity and tone, and (2) verifying whether reviewer claims are supported by the submitted paper. Using expert-annotated benchmarks, we compare static metrics, ML baselines, zero-shot LLMs, and fine-tuned models to assess alignment with human judgment.

The talk will highlight methodological choices, deployment lessons, and empirical insights into where LLMs succeed and where hybrid approaches with interpretable ML and retrieval perform more reliably. We conclude with reflections on what infrastructure is needed to make use LLMs as a robust foundation for research integrity at scale.

1 Introduction

Peer review remains the cornerstone of scientific communication and quality control. It informs publication decisions, determines the allocation of funding, and filters what knowledge enters the scientific record. Yet despite its central role, peer review is rarely evaluated or audited at scale. The reasons are both practical and structural. With rapidly growing submission volumes across journals and conferences [5, 6], editors and program chairs face an overwhelming number of reviews, making manual oversight infeasible. This scaling bottleneck has created a vacuum of accountability, where vague, biased, or unsubstantiated feedback can shape outcomes without recourse or verification [3, 8]. At the same time, the rise of large language models (LLMs) has introduced new complexity. Reviewers increasingly use generative AI tools to help write, summarize, or refine reviews, which might improve fluency and coherence, but also carry their own risks. LLM-generated content can sound authoritative while lacking conceptual depth, or include hallucinated claims and irrelevant critiques [4, 7]. These issues are difficult to detect at scale and compound existing problems of transparency, factual grounding, and reviewer reliability in high-volume editorial pipelines [2, 9].

This is further complicated by the subjective and inconsistent nature of review evaluation itself. What counts as a "good" review varies across fields and venues. Most conferences and journals provide little guidance on quality criteria, and feedback mechanisms remain sparse. Reviews are rarely audited unless authors raise complaints, typically after decisions have been made.

We argue that the inability to evaluate reviews systematically is not just an academic oversight; rather, it is a critical infrastructure gap in the scientific ecosystem. Without scalable, transparent tools for evaluating reviewer behavior, verifying claims, or surfacing poor-quality feedback, research communities have no meaningful way to detect when the peer review process fails.

At Reviewer.ly, we view this as both a technical challenge and an opportunity for meaningful impact [1]. Over the past two years, we have developed and deployed two systems that address this gap in operational editorial environments, namely `Peeriscope`[1], a multi-dimensional framework for auditing review quality based on interpretable features and human-aligned metrics; and `Peerispect`[2], a claim verification tool that uses retrieval-augmented generation to assess whether reviewer statements are supported by evidence in the paper. These systems blend classic machine learning, neural-based retrievers, and LLM for structured, grounded evaluation. They are actively used by editorial teams, integrated via dashboards and APIs, and validated against expert-labeled benchmarks.

In this talk, we will explore the motivations, methodologies, and deployment lessons behind these systems. We reflect on the real-world capabilities and limitations of LLMs in the context of research integrity, highlighting where they succeed, where they fail, and what complementary strategies are needed to support trustworthy large-scale auditing. Our empirical findings suggest that while LLMs are helpful, simpler, interpretable models often show stronger alignment with expert judgment and offer clearer value in editorial workflows.

Ultimately, this talk aims to answer a pressing question: *to what extent can LLMs help uphold research integrity at scale? And what*

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

infrastructure is still missing to make AI-assisted auditing reliable, explainable, and deployable? Our goal is to offer a blueprint for building systems that move toward trustworthy peer-review to enable a more rigorous, transparent, and accountable future for science integrity.

2 System Overview and Methodological Insights

To meet the need for scalable review auditing, we developed two complementary systems: `Peeriscope`, which assesses review quality, and `Peerispect`, which verifies factual grounding. Both are deployed in real editorial settings and designed for transparency and extensibility.

In building these systems, we pursued two goals: (1) aligning review quality evaluation with expert judgment, and (2) making the decision process interpretable to editors, reviewers, and stakeholders. To do so, we adopted a layered evaluation framework that integrates different scoring functions, including the following:

Human-Annotated Quality Benchmarks. As a foundation, we constructed a high-quality dataset of over 700 paper–review pairs across computer science venues, annotated by domain experts along multiple quality dimensions such as specificity, justification, tone, and informativeness. This dataset served two purposes: training transparent models on interpretable signals and evaluating how well alternative approaches align with expert judgment.

Metric-Based Scoring and Baselines. We implemented a static scoring system using interpretable features inspired by the PeeriScope and Rotten Reviews frameworks [2]. These included lexical diversity, section-level coverage, hedging, alignment, and reviewer-paper topical proximity. Although simplistic, these metrics offered a surprisingly strong baseline, especially for highlighting low-effort or formulaic reviews. Building on these features, we trained a simple regression model to predict expert-assigned quality scores. This baseline outperformed both zero-shot and fine-tuned LLMs in terms of agreement with human judgments, especially on abstract dimensions like conceptual depth and argumentative structure.

LLM-Based Evaluation Strategies. We next tested general-purpose LLMs such as GPT-4o, Qwen, and Phi in zero- and few-shot settings, prompting them to evaluate review quality across dimensions. While these models occasionally surfaced insightful observations, they often overestimated quality and failed to penalize vague or unsupported claims. To push their capabilities further, we fine-tuned select LLMs on the expert-labeled dataset. Fine-tuning improved calibration but required heavy resources and still underperformed the interpretable baseline.

Claim Verification with Retrieval-Augmented Generation. We further extended the pipeline to address review grounding issues. Some reviews appear constructive on the surface but reference issues not actually present in the paper. This is problematic for editors and authors alike. To tackle this, `PeeriSpect` uses a claim extraction module (LLM-based) followed by retrieval of candidate evidence from the paper, and finally, a RAG entailment pipeline to evaluate whether the retrieved content supports the reviewer's claim. This enables editors to flag unsupported feedback and authors to understand where reviewers might have misinterpreted the work.

2.1 Challenges and Limitations

Designing and deploying these systems revealed a range of challenges that we believe are critical to acknowledge when considering the role of LLMs in research integrity infrastructure. First, extracting well-formed reviewer claims is inherently difficult due to vague phrasing, embedded references, or domain-specific shorthand that can obscure the intent or meaning of a statement. Even when claims are clearly identified, verifying their factual grounding is non-trivial; determining whether a claim is "supported" by the paper often requires reasoning over implicit knowledge or assumptions not explicitly stated in the text. LLMs, especially in zero-shot or even fine-tuned configurations, frequently produce evaluations that are fluent and confident but misaligned with expert judgment, which raise concerns about hallucination and consistency. Moreover, our deployments revealed that editors and stakeholders often preferred interpretable, feature-based outputs over LLM-generated scores, particularly in ambiguous or high-stakes cases.

By building `Peeriscope` and `Peerispect`, we provide concrete infrastructure to make review auditing actionable. These tools allow editorial teams to flag vague or ungrounded reviews, weigh reviewer input more selectively, and surface inconsistencies during decision-making. For example, reviews suggesting major revisions without sufficient grounding can be deprioritized in scoring or routed for follow-up by the editors. Beyond academic publishing, these capabilities extend to funding agencies, industrial R&D evaluations, and regulatory review panels, essentially any environment where expert judgment plays a critical role. Our systems enable these stakeholders to scale oversight, improve review quality, and increase confidence in evaluative processes.

3 Company Portrait

Reviewer.ly[3] is a Toronto-based AI company focused on strengthening research integrity through scalable, data-driven auditing tools for peer review. It offers modular systems for evaluating review quality, verifying reviewer claims, and supporting transparent editorial workflows. Reviewer.ly's technology is used by academic conferences, journals, research institutions, and industrial stakeholders, including funders and enterprise R&D teams seeking greater accountability and reliability in expert evaluation. By integrating with platforms like OJS, Reviewer.ly enables automated review analysis, reviewer vetting, and feedback generation at scale, helping stakeholders identify bias, detect low-quality feedback, and support evidence-based decision-making across the scientific ecosystem.

4 Speaker's Bio

Dr. Negar Arabzadeh is Head of Data Science at Reviewer.ly and a Postdoctoral Fellow at UC Berkeley, specializing in information retrieval and AI-driven evaluation systems with a specific focus on improving peer review quality and research integrity at scale using both LLMs and transparent machine learning. She has over 60 publications in venues such as SIGIR, CIKM, and EMNLP, and her background spans both academia and industry, including roles at Google Brain, Microsoft Research, and Spotify. She also has delivered tutorials at major conferences including SIGIR, WSDM, and ECIR.

[3] https://reviewer.ly/

References

[1] 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.

[2] 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. https://doi.org/10.1145/3746252.3761436

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

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.