Reliable Evaluation of AI Assisted Peer Review

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

Sajad Ebrahimi
Reviewerly
Toronto, Ontario, Canada

Shakiba Amirshahi
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

Large language models (LLMs) are rapidly moving from experimental prototypes into operational peer review workflows, where they are used to polish, rewrite, expand, synthesize, screen, and in some cases generate reviews. This shift creates an urgent challenge for the organizations that depend on expert assessment, including publishers, conferences, funding agencies, research institutions, and enterprise research teams. Rather than asking only whether AI can improve peer review, this talk argues that we must first confront a more fundamental evaluation question. Can we reliably measure what makes a review useful, fair, accurate, and decision-relevant? The talk examines this question through the lens of *AI-assisted peer review* and shows that commonly used signals, including recommendation agreement, similarity to reference reviews, writing fluency, and single score *LLM-as-a-judge* evaluation, often fail to capture substantive review quality. Drawing on real world deployment experience at Reviewerly, we present case studies in which evaluation metrics reward surface-level improvements such as polish, organization, and LLM like wording while overlooking the evidentiary content, specificity, correctness, and editorial usefulness of the review. We then discuss practical lessons for building reliable evaluation systems in real peer review workflows, including stress testing review metrics, separating surface sensitivity from practical robustness, comparing judge models under realistic perturbations, and designing layered evaluation pipelines that are interpretable, evidence-grounded, and aligned with editorial needs. Although peer review is the central case study, the broader message applies to many industry and organizational settings where AI is being introduced into expert judgment workflows. Responsible automation requires more than stronger generation models. It requires evaluation methods that measure the qualities we actually want AI systems to improve.

1 Motivation and Talk Objective

Peer review plays a central role in how research makes progress [12, 13]. It influences which papers are published, which projects receive funding, which ideas gain visibility, and how institutions and organizations make decisions about research quality [9, 11, 13]. As these evaluation workflows grow in scale and complexity, reviewers, editors, program chairs, and funders increasingly need ways to assess the quality and reliability of review feedback itself.

At the same time, LLMs are becoming part of the peer review process [10, 15, 21]. Reviewers may use LLMs to polish their writing, reorganize feedback, expand short comments, synthesize additional critiques, or even generate full reviews [14–16]. This shift naturally raises an important question: *does AI make peer review better?* However, before we can answer this question, we need to ask a more fundamental one: *how do we know what a better review is?*

This talk starts from a saying often attributed to Lord Kelvin, *"what we cannot measure, we cannot improve."* In the context of AI-assisted peer review, this premise becomes especially important. A review can be fluent, structured, and polished while still being shallow, unsupported, or unhelpful [3, 8]. Conversely, a review may be less polished but contain careful technical judgment, important criticism, and useful guidance. If evaluation metrics reward surface form rather than substantive assessment, we may conclude that AI improves peer review simply because it makes reviews sound better. This creates a measurement risk, where we may optimize for the appearance of quality rather than quality itself.

Measuring peer review quality is difficult because review quality is not a single property [5, 18]. A useful review should be technically accurate, specific, constructive, actionable, fair, grounded in the manuscript, and useful to authors, editors, program chairs, funders, and other decision-makers [3, 8]. Common evaluation shortcuts are therefore insufficient. The recommendation agreement does not tell us whether the feedback is useful [1]. Similarity to another review does not tell us whether the critique is correct [6]. Fluency does not imply depth [8]. Even LLM-as-a-judge metrics, which are increasingly used to score review quality, may be affected by style, structure, explicitness, or polish rather than the underlying quality of the evaluation [19, 20].

The objective of this talk is to examine the measurement foundations needed for trustworthy AI-assisted peer review. Rather than presenting another AI tool for peer review, we focus on the challenges of building reliable evaluation metrics, i.e., what they measure, where they fail, and how they can be stress-tested before being used in real workflows. We will showcase scenarios where review evaluation metrics fail, including cases where a review receives a higher score because it is better written or more organized, even when the underlying evaluative content has not changed.

A central technical question is whether peer-review evaluation metrics are robust to meaning-preserving changes in wording and presentation. Consider a human review that is somewhat informal

but identifies a valid methodological flaw. If an LLM rewrites the same review to make it smoother, clearer, and more structured while preserving the same critique, should the review receive a substantially higher quality score? If the answer is yes, we need to ask what the metric is actually measuring, whether it is the quality of the evaluation or the quality of the writing.

Drawing on our experience at Reviewerly, the talk will discuss practical approaches we tested for making LLM-as-a-judge evaluation more reliable for AI-assisted peer review. We will discuss how to separate surface-level presentation from substantive review quality, how to evaluate robustness under meaning-preserving rewrites, why broad holistic metrics are often fragile, and why more concrete, content-anchored metrics may provide more stable signals. We will also discuss the limitations of relying on a single LLM judge and the need for layered, interpretable, and evidence-grounded evaluation.

While peer review is the central case study, the broader lesson is not limited to academia. Many stakeholders, including publishers, conference organizers, funding agencies, research institutions, enterprise R&D teams, and organizations that rely on expert assessments, face the same challenge as they want to use AI to improve evaluation workflows, but first need reliable ways to measure whether those workflows are actually improving. In all of these settings, reliable measurement is the first step toward responsible automation.

2 Lessons for Reliable AI-Assisted Evaluation

The main lesson we want to share with the CIKM community from our experience automatically evaluating peer reviews at scale is that reliable AI-assisted review evaluation requires more than calling an LLM judge. LLM-as-a-judge metrics are attractive because they are flexible and scalable, but they must be validated before being used to compare human-written, AI-assisted, and AI-generated reviews. Without such validation, organizations risk rewarding reviews that sound better rather than reviews that reason better.

We will discuss several practical lessons for building more reliable evaluation systems. First, metrics should be stress-tested under meaning-preserving transformations to determine whether they capture substantive quality or surface presentation. Second, we will distinguish between two complementary reliability criteria. *Surface sensitivity* asks whether a metric systematically changes under meaning-preserving rewriting. *Robustness* asks whether any change is small enough to be practically negligible. This distinction is important because a metric may show a statistically detectable shift that is too small to matter, or it may fail to show a shift without providing positive evidence of stability.

Third, broad holistic metrics should be decomposed into more interpretable dimensions whenever possible. Many broad LLM-as-a-judge metrics, such as overall quality, fairness, constructiveness, and analytical depth, can be sensitive to changes in wording and organization. More concrete and content-anchored metrics tend to be more stable, but they are not sufficient on their own because high-quality reviewing also requires synthesis, prioritization, and contextual judgment.

Fourth, metric outputs should be tested for robustness across judge models. Relying on a single model can hide model-specific biases and instabilities, especially as LLMs differ in calibration, style preferences, and sensitivity to wording. Comparing results across multiple judge models helps identify which evaluation signals are stable and which are artifacts of a particular evaluator.

The broader lesson is that *measurement should come before automation*. Before deploying AI systems to improve review workflows, stakeholders need to understand what their metrics reward, where they fail, and whether they remain stable under realistic variations in writing style, structure, and AI assistance. Reliable evaluation therefore requires layered metrics, robustness testing, interpretable outputs, and evidence-grounded assessment.

3 Relevance to CIKM Industry Track

This talk addresses the practical question of how to build reliable measurement infrastructure for workflows where expert judgment is increasingly mediated by LLMs. While peer review is the main case study, the technical challenges are broadly relevant to knowledge management, information retrieval, evaluation, trustworthy AI, and human-AI collaboration. A key focus is on metrics and measurement techniques for production AI systems, including robustness testing for LLM-as-a-judge evaluation. Drawing on Reviewerly's deployment experience, we discuss how to validate metrics, identify failure modes, separate substance from presentation, and build evaluation systems that provide stakeholders with actionable evidence rather than opaque scores.

These questions are directly relevant to the CIKM audience, including information retrieval researchers in both academia and industry who are building, deploying, or evaluating LLM-powered systems. The lessons also extend beyond academic publishing to grant review, internal R&D evaluation, hiring and promotion workflows, regulatory assessment, and other high-stakes settings where organizations need to evaluate expert feedback at scale.

4 Company Portrait

Reviewerly is a Toronto-based AI company and university spin-off focused on strengthening research integrity through scalable auditing tools for peer review and expert evaluation. Founded by researchers with experience across academia and industry, Reviewerly combines active research with product development, regularly publishing work on peer review evaluation and AI-assisted reviewing while developing modular systems for review quality assessment, metric benchmarking, claim verification, reviewer recommendation, and transparent editorial workflows [2–4, 7, 17]. Reviewerly integrates with widely used peer review systems such as Open Journal Systems (OJS) through APIs, enabling automated review analysis, reviewer vetting, and workflow support at scale.

Reviewerly's technology and tools[1,2,3,4] are 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 in high-stakes review processes.

[1] https://app.reviewer.ly/app/peerfect
[2] https://app.reviewer.ly/app/peeriscope
[3] https://app.reviewer.ly/app/peerispect
[4] https://app.reviewer.ly/app/peerclear

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, ICML, EMNLP, EACL, CIKM, WSDM, WWW and ECIR. 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.

Reviewerly team has previously delivered industry talks on closely related topics at CIKM 2025 and WSDM 2026, and received the Best Industry Talk Award at WSDM 2026.

References

[1] Corinna Cortes and Neil D Lawrence. 2021. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774* (2021).

[2] Sajad Ebrahimi, Soroush Sadeghian, Ali Ghorbanpour, Negar Arabzadeh, Sara Salamat, Seyed Mohammad Hosseini, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. 2026. PeeriScope: A Multi-Faceted Framework for Evaluating Peer Review Quality. In *Companion Proceedings of the ACM Web Conference 2026*. 168–171.

[3] 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 '25)*. 5642–5649. doi:10.1145/3746252.3761506

[4] Sajad Ebrahimi, Sara Salamat, Negar Arabzadeh, Mahdi Bashari, and Ebrahim Bagheri. 2025. exHarmony: Authorship and Citations for Benchmarking the Reviewer Assignment Problem. In *European Conference on Information Retrieval*. Springer, 1–16. doi:10.1007/978-3-031-88714-7_1

[5] Daniel Garcia-Costa, Flaminio Squazzoni, Bahar Mehmani, and Francisco Grimaldo. 2022. Measuring the developmental function of peer review: a multi-dimensional, cross-disciplinary analysis of peer review reports from 740 academic journals. *PeerJ* 10 (2022).

[6] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research* 77 (2023), 103–166.

[7] Ali Ghorbanpour, Soroush Sadeghian, Alireza Daghighfarsoodeh, Sajad Ebrahimi, Negar Arabzadeh, Seyed Mohammad Hosseini, and Ebrahim Bagheri. 2026. Peerispect: Claim Verification in Scientific Peer Reviews. *arXiv preprint arXiv:2604.17667* (2026).

[8] Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one* 17, 1 (2022), e0259238.

[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] Sven E Hug and Mirjam Aeschbach. 2020. Criteria for assessing grant applications: A systematic review. *Palgrave Communications* 6, 1 (2020), 37.

[12] Terne Sasha Thorn Jakobsen and Anna Rogers. 2022. What factors should paper-reviewer assignments rely on? community perspectives on issues and ideals in conference peer-review. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4810–4823.

[13] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. *arXiv preprint arXiv:1804.09635* (2018).

[14] Sandeep Kumar, Samarth Garg, Sagnik Sengupta, Tirthankar Ghosal, and Asif Ekbal. 2025. MixRevDetect: Towards Detecting AI-Generated Content in Hybrid Peer Reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

[15] 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. 2024. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. In *Proceedings of the 41st International Conference on Machine Learning (ICML) (PMLR)*. arXiv:2403.07183.

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

[17] Soroush Sadeghian, Alireza Daqiq, Radin Cheraghi, Sajad Ebrahimi, Negar Arabzadeh, and Ebrahim Bagheri. 2026. PeerPrism: Peer Evaluation Expertise vs Review-writing AI. *arXiv preprint arXiv:2604.14513* (2026).

[18] Susan van Rooyen, Nick Black, and Fiona Godlee. 1999. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *Journal of clinical epidemiology* 52 7 (1999), 625–9.

[19] Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 297–312.

[20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

[21] 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.