Building Trustworthy Peer Review Quality Assessment Systems

Negar Arabzadeh
Reviewer.ly

Sajad Ebrahimi
Reviewer.ly

Ali Ghorbanpour
Reviewer.ly

Soroush Sadeghian
Reviewer.ly

Sara Salamat
Reviewer.ly

Muhan Li
Reviewer.ly

Hai Son Le
Reviewer.ly

Mahdi Bashari
Reviewer.ly

Ebrahim Bagheri
Reviewer.ly
University of Toronto

Abstract

Peer review is foundational to academic publishing, yet the quality of reviews remains difficult to assess at scale due to subjectivity, inconsistency, and the lack of standardized evaluation mechanisms. This talk presents our experience developing and deploying a scalable framework for assessing review quality in operational settings. We combine two complementary approaches: interpretable machine learning models built on quantifiable review- and reviewer-level features, and the application of large language models (LLMs), including Qwen, Phi, and GPT-4o, in zero- and few-shot configurations for textual quality evaluation. We also explore the fine-tuning of LLMs on expert-annotated datasets to examine their upper-bound capabilities. To benchmark these methods, we constructed a dataset of over 700 paper–review pairs labeled by domain experts across multiple quality dimensions. Our findings demonstrate that transparent, feature-based models consistently outperform LLMs in reliability and generalization, particularly when evaluating conceptual depth and argumentative structure. The talk will highlight key engineering choices, deployment challenges, and broader implications for integrating automated review evaluation into scholarly workflows.

1 Introduction

The peer review process is central to scholarly communication, yet the quality of individual reviews remains largely unmeasured and inconsistently assessed. Most academic venues rely on informal impressions to identify "helpful" reviewers, with little in the way of standardized criteria or scalable infrastructure. As conferences and journals grow to accommodate thousands of submissions, the variability and inconsistency of peer reviews have emerged as critical challenges for fairness, transparency, and the credibility of scientific evaluation [3, 5]. High-quality reviews can clarify contributions and improve decision-making, while vague, biased, or even AI-generated reviews risk undermining trust and delaying progress [2, 4, 6]. Despite increasing attention to the integrity of peer review, tools for evaluating review quality remain underdeveloped. In this work, we report our experience on how to build a scalable framework for assessing peer review quality in operational settings. We explain how to construct a high-quality, expert-annotated dataset of over 700 paper–review pairs and how to use it for benchmarking multiple evaluation strategies. Our approach integrates two complementary methods: interpretable machine learning models based on quantifiable review- and reviewer-level signals, and large language models (LLMs) such as Qwen, Phi, and GPT-4o applied in zero- and few-shot configurations. We will also present our approach for fine-tuning LLMs on the expert-labeled data to assess their upper-bound capabilities and also present evaluation methodologies in this context for assessing alignment with expert human judgments across multiple dimensions of review quality.

2 Significance of the Problem

Automating peer review quality assessment is crucial yet inherently complex, primarily due to the multifaceted and often subjective nature of defining and measuring *quality* [1]. Several substantial obstacles complicate the development of robust automated systems:
Subjectivity and Lack of Standards: In the absence of universally accepted criteria, review quality assessments are shaped by individual judgment rather than standardized reproducible guidelines, making it difficult to define ground truth.
Opaque Review Processes: While anonymity and confidentiality are central to peer review, they limit transparency and accountability, making it challenging to systematically audit reviews.
Data Scarcity for Annotation: High-quality, human-annotated datasets that explicitly assess review quality are extremely limited, posing a major barrier to develop learning-based methods.
Scalability of Manual Assessment: As submission volumes grow, manual oversight of review quality becomes impractical, necessitating automated solutions that can operate at scale.
Emergence of LLM-generated Reviews: The growing presence of AI-generated reviews introduces additional complexity, calling for methods that can not only evaluate quality but also detect and assess machine-written content.

In this talk, we share our experience developing a scalable framework for peer review quality assessment, grounded in real-world deployment and expert validation. By detailing our methodology, evaluation strategies, and key findings, we aim to inform and support stakeholders across academia, funding agencies, and research institutions in strengthening the consistency, transparency, and reliability of the peer review process. Our objective is also to demonstrate how such systems can be built systematically and effectively, providing a roadmap for future efforts to integrate automated assessment into scholarly workflows.

3 Exploring Review Quality Assessment

At Reviewer.ly, we built a scalable, modular framework for assessing peer review quality, drawing on both interpretable machine learning and large language models. This talk focuses on the theoretical foundations, design decisions and empirical approaches and findings we explored throughout the development process.

1) Interpretable Models with Quantifiable Metrics: We implemented two complementary sets of transparent features: *review-dependent* and *reviewer-dependent* metrics. Review-dependent metrics are extracted from the review and its associated paper and include indicators such as specificity, section-level coverage, semantic alignment, politeness, sentiment, hedging, lexical diversity, readability, and the use of clarifying questions. Reviewer-dependent metrics reflect characteristics of the reviewer, including citation count, academic seniority, and the semantic similarity between their publication history and the submission under review. Together, these interpretable signals supported models that consistently aligned with human judgments and outperformed more complex alternatives on core quality dimensions.
2) Zero and Few-Shot LLM Evaluation: We evaluated the effectiveness of using LLMs such as GPT-4o, Qwen-3, and Phi-4 in zero and few-shot settings to generate structured assessments of review quality. These models were prompted to evaluate dimensions such as informativeness, tone, and relevance without access to training data. While LLMs occasionally surfaced useful observations, their assessments were inconsistent when it came to deeper conceptual or argumentative structure. Correlation with expert labels was weak, underscoring the limitations of general-purpose LLMs in nuanced evaluation tasks.
3) Fine-Tuned LLMs on Expert-Labeled Data: To probe the upper limits of LLM-based assessment, we fine-tuned selected models on a curated dataset of over 700 expert-annotated paper–review pairs. Fine-tuning improved performance on mid-level criteria but required substantial annotation effort and computational resources. Despite these improvements, the fine-tuned models still underperformed relative to the simpler, interpretable baselines, reinforcing the practical and methodological advantages of transparent feature-driven approaches in high-stakes review workflows.

4 Assessment and Validation Strategy

A core focus of this talk will be the evaluation strategy we developed to assess peer review quality at scale. Unlike typical classification or scoring tasks, review quality lacks a fixed gold standard. It is inherently subjective, shaped by domain norms, reviewer expectations, and editorial context. As a result, validating automated assessment methods requires careful attention not only to model performance but also to how quality itself is defined, measured, and operationalized. We will present the full process we followed in designing and executing a robust evaluation pipeline. This included a large-scale annotation effort in which domain experts labeled over 700 paper and review pairs across multiple dimensions of quality. These annotations served as a foundation for assessing model alignment with expert judgment using rank-based measures such as Kendall's tau. Throughout the design of this evaluation strategy, we engaged with several key challenges. These included deciding how to define review quality in a way that is consistent and interpretable, determining how to distinguish superficial fluency from conceptual depth, and ensuring that the resulting evaluation framework would generalize across different venues and reviewing cultures. We will present three central findings from this process, namely:

- Zero and few-shot evaluations using large language models often failed to capture deeper aspects of review quality and showed weak correlation with human annotations
- Simple machine learning models based on transparent review and reviewer features aligned more closely with expert assessments and offered consistent, interpretable outputs
- Fine-tuned large language models provided limited performance gains and required substantial data and compute resources, underscoring tradeoffs between accuracy and scalability

By sharing this experience, our aim is to provide a clear and reproducible roadmap for developing, validating, and deploying peer review assessment systems in real-world editorial workflows.

5 Company Portrait

Reviewer.ly[1] is an AI-driven platform headquartered in Toronto, Canada, dedicated to improving the integrity, efficiency, and transparency of the peer review process. Leveraging advanced natural language processing, interpretable machine learning, and large language models (LLMs), Reviewer.ly delivers end-to-end solutions for reviewer assignment, review quality assessment, and editorial decision support. Its core technology enables the automatic analysis of review texts to identify conceptual strengths, detect vague or low-effort reviews, and generate targeted feedback for editors and reviewers. Reviewer.ly integrates seamlessly with editorial management systems, including Open Journal Systems (OJS), through robust APIs that support real-time data exchange and workflow automation. It is actively deployed by academic publishers, research institutions, and funding agencies to reduce administrative overhead, increase review quality, and support fair and diverse reviewer selection. By embedding explainable AI into critical decision points, the platform helps stakeholders move toward more consistent, evidence-based evaluation practices.

6 Speaker's Bio

Negar Arabzadeh is the Head of Data Science at Reviewer.ly. She is also a Postdoctoral Fellow at the University of California, Berkeley, specializing in Information Retrieval with the focus on evaluation. Negar has authored over 60 publications in top-tier venues, including SIGIR, CIKM, EACL, EMNLP, ECIR, and IP&M. Her industry experience includes research internships at Google Brain, Microsoft Research, and Spotify Research. She has delivered tutorials at SIGIR 2022, WSDM 2023, ECIR 2023 and 2024, and SIGIR AP 2024.

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 (CIKM '24)*.
[2] Sajad Ebrahimi, Sara Salamat, Negar Arabzadeh, Mahdi Bashari, and Ebrahim Bagheri. 2025. exHarmony: Authorship and Citations for Benchmarking the Reviewer Assignment Problem. In *Advances in Information Retrieval*. Cham.
[3] 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).
[4] Cheng Long, Raymond Chi-Wing Wong, Yu Peng, and Liangliang Ye. 2013. On good and fair paper-reviewer assignment. In *ICDM*. IEEE.
[5] 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.
[6] Jelte M Wicherts. 2016. Peer review quality and transparency of the peer-review process in open access and subscription journals. *PloS one* 11, 1 (2016), e0147913.

[1] https://reviewer.ly/