PeeriScope: A Multi-Faceted Framework for Evaluating Peer Review Quality

Sajad Ebrahimi
Reviewerly, University of Guelph

Soroush Sadeghian
Reviewerly

Ali Ghorbanpour
Reviewerly

Negar Arabzadeh
Reviewerly, UC Berkeley

Sara Salamat
Reviewerly

Seyed Mohammad Hosseini
Reviewerly

Hai Son Le
Reviewerly

Mahdi Bashari
Reviewerly

Ebrahim Bagheri
Reviewerly, University of Toronto

Abstract

The increasing scale of peer review in scholarly venues has created an urgent need for systematic, interpretable, and extensible tools to assess review quality. We present PeeriScope, a modular platform that integrates structured features, rubric-guided large language model assessments, and supervised prediction to evaluate peer review quality along multiple dimensions. Designed for openness and integration, PeeriScope provides both a public interface and an opensource web service, supporting practical deployment and research extensibility. The demonstration illustrates its use for reviewer self-assessment, editorial triage, and large-scale auditing, and it enables the continued development of quality evaluation methods within scientific peer review. PeeriScope is available as a live demo[1] and via API services[2] and is accompanied by a video tutorial[3].

1 Introduction

Peer review is a cornerstone of scholarly publishing, ensuring the quality and credibility of scientific communication. However, the quality of reviews varies widely, and most venues lack standardized or scalable mechanisms to assess them. As conferences and journals continue to grow, this inconsistency raises concerns about fairness, transparency, and trust in the evaluation process. Recent advances in large language models (LLMs) have further transformed the peer-review landscape [17]. Given their ability to produce fluent and well-structured prose, LLMs are increasingly used to draft review reports (as witnessed in the ICLR 2026 drama) [18]. Recent studies show, however, that although LLMs can mimic the surface form of expert feedback, their critiques often lack depth, domain-specific reasoning, and reliable factual grounding; they also struggle with providing actionable recommendations tailored to a paper's actual weaknesses [11, 22]. Importantly, evaluating these emerging LLM-assisted practices remains difficult because peer-review datasets are inherently sparse, fragmented across venues, and largely inaccessible due to confidentiality constraints. This lack of annotated, high-quality review data makes it challenging to benchmark LLM-generated reviews or to compare them meaningfully with human judgments. As such, these challenges underscore the need for scalable and interpretable frameworks capable of evaluating (human- and LLM-generated) peer-reviews.

[1] https://app.reviewer.ly/app/peeriscope
[2] https://github.com/Reviewerly-Inc/Peeriscope
[3] https://bit.ly/3LMobm8

A growing body of research has started exploring automated approaches to review analysis and generation. Recent studies have examined the politeness and engagement of reviews [1], the prevalence of superficial or "lazy" reviewing [13], and biases such as institutional or gender disparities [6, 19]. Other work has investigated the use of LLMs as reviewers or meta-reviewers [3], and proposed systems for summarizing or generating reviews [9, 10]. Together, these efforts highlight promising progress but remain fragmented, lacking a unified framework for comprehensive review quality assessment. In parallel, several tools explicitly deploy LLMs to support review quality at scale. For example, the ICLR 2025 Review Feedback Agent provides structured, optional feedback on clarity, specificity, and professionalism to thousands of reviewers in a randomized study [18]. Other systems explore automated peer-review generation and iterative review loops for academic writing [17]. We view PeeriScope as part of this emerging ecosystem rather than a stand-alone or definitive solution. PeeriScope offers an additional, complementary tool focused on post-hoc, multidimensional assessment of review helpfulness that can plug into existing reviewer training, monitoring, and decision-support workflows. PeeriScope integrates structured linguistic metrics, LLM-based scoring, and supervised modeling to capture diverse aspects of review helpfulness. Trained on expert-annotated reviews, it provides interpretable diagnostics and quantitative assessments through an accessible web interface and API. By combining interpretability with the power of foundation models, PeeriScope advances the development of trustworthy, automated review evaluation.

2 PeeriScope Overview

Design Requirements. Automated review quality assessment must satisfy both computational and workflow constraints. First, it must be *scalable*. Conferences and journals handle thousands of submissions, requiring efficient, high-throughput evaluation. Second, it must be *transparent* and *interpretable* to editors, reviewers, and authors who rely on these signals for quality control and decisions. Finally, it must be *compatible* with existing ecosystems and integrate smoothly with current infrastructures. Figure 1 summarizes the architecture of PeeriScope under these considerations.

Inputs and Ingestion. PeeriScope supports two input modes: In individual review mode, users provide the paper title, abstract, and full review text via a web form, intended for targeted editorial checks or reviewer self-evaluation. An optional OpenAlex identifier lets the system retrieve the reviewer's publication profile and derive

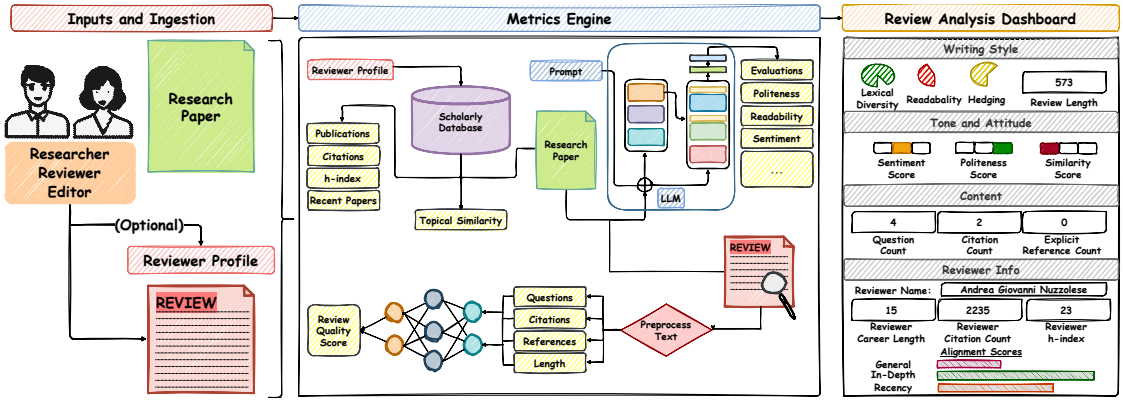

Figure 1: Overview workflow of PeeriScope.

topical-expertise and citation features. In OpenReview mode, users supply the URL of a public submission, and the system uses the OpenReview API to fetch manuscript metadata and all associated reviews. Both modes map inputs to a unified internal schema (Figure 1) with structured fields for paper metadata, review texts, and optional reviewer profiles.

Metrics Engine. PeeriScope evaluates peer review quality using three complementary groups of metrics: (1) structured metrics derived from the review text and optionally reviewer-profile metrics obtained from scholarly metadata, (2) rubric-guided LLM-based scores for abstract qualities such as constructiveness, and (3) a supervised overall quality estimator that integrates these signals into a single score. Further details on these metric categories are provided in Section 3.1. The output of metric engine should approximate human judgments on review quality from different perspectives while preserving interpretability and computational efficiency.

Review Analysis Dashboard. PeeriScope exposes all signals from the metrics engine through an interactive dashboard with an overall quality bar, metric cards, and drill-down tabs that group outputs into interpretable categories, favoring transparency over a single opaque score and move beyond a single recommendation label (e.g., weak accept) toward a richer, multidimensional view of review quality. For settings where a web interface is not desirable, we additionally provide a programmatic API that returns structured JSON outputs for single reviews or batches, enabling automated quality auditing at scale.

3 PeeriScope Metrics

PeeriScope evaluates peer review quality using three complementary sources of evidence, introduced in the following subsections.

3.1 Structured Metrics

Our work is inspired by a strong line of prior work on automated assessment of scientific peer reviews and review helpfulness [12], which has typically defined metrics around three classes: (i) writing style and readability (e.g., surface features, fluency, coherence) [14, 20], (ii) tone and reviewer attitude (e.g., sentiment, politeness, harshness) [1, 16], and (iii) the substantive content and coverage of the critique (e.g., section/aspect coverage, informativeness) [4, 15].

Writing Style includes metrics related to reviewer effort and communicative clarity. *Review length* is reported as a proxy for thoroughness. In addition, we report *hedging* , which marks uncertainty and plays an important role in balancing authority with caution [7]. *Hedging* is measured using a cue-based neural detector and captures the epistemic stance of the reviewer. *Lexical diversity*, measured as type–token ratio, reflects linguistic variation and effort, and *readability*, captured using the Flesch Reading Ease score, reflects how easily the review can be understood.

Tone and Attitude are modeled through *politeness*, *sentiment*, and *similarity* between the review and the paper. *Politeness* has been linked to perceptions of fairness and author receptivity, while *sentiment* polarity offers a coarse but informative indicator of evaluative direction. The general *similarity* of the review to the manuscript reflects the reviewer's overall domain proximity.

Review Content is also being assessed using metrics such as *mentions of manuscript structure*; e.g., references to figures or specific sections, which suggest close reading and submission-specific critique. Additionally, we consider *mentions of citations*, which provide support for reviewers' claims and help ground them. Engagement is further captured through the *presence of questions*. Constructive reviews often contain forward-looking or clarifying questions that invite reflection or revision. We identify these using a fine-tuned classifier trained on interrogative forms that indicate substantive reviewer intent.

Textual features alone cannot capture the credibility or relevance of the reviewer. We therefore incorporate **reviewer-based metrics** using metadata from OpenAlex. We measure *topical alignment* between the submission and the reviewer's publication history using SPECTER [2]. Reviewer standing is further characterized through *citation count* and *scholarly tenure*, a proxy for influence within a field. These features provide complementary views of seniority, continuity, and visibility in the research community.

3.2 LLM-based Metrics

Structured metrics provide interpretable signals of review quality from observable textual and contextual attributes. However, many important properties of peer reviews—such as fairness, constructiveness, factuality, and overall utility—are abstract and hard to reduce to surface-level proxies. To capture these dimensions, PeeriScope incorporates a second class of LLM-based evaluative signals. Prior work has shown that, when carefully prompted, LLMs can produce judgments that approximate expert annotations [5].

We adopt Qwen-3-8B, a multilingual instruction-tuned LLM deployed locally, as our primary LLM-based evaluator to ensure data privacy and fast inference, though PeeriScope can readily

Table 1: Review quality dimensions and Kendall's τ correlation between human and LLM judgments across three models.

Aspect	Description	GPT-4o	Phi-4	Qwen-3
Overall Quality	Holistic evaluation of the review's usefulness and professionalism.	0.359	0.241	0.252
Comprehensiveness	Covering all key aspects of the paper.	0.476	0.374	0.338
Actionability	Helpfulness of the review in suggesting clear next steps.	0.411	0.279	0.314
Sentiment Polarity	Overall sentiment conveyed by the reviewer.	0.407	0.397	0.428
Constructiveness	Whether the review suggests improvements rather than only criticism.	0.343	0.259	0.211
Use of Technical Terms	Using domain-specific vocabulary.	0.327	0.254	0.176
Objectivity	Presence of unbiased, evidence-based commentary.	0.298	0.215	0.186
Alignment	Relevance of the review to the scope of the paper.	0.295	0.204	0.105
Vagueness	Degree of ambiguity or lack of specificity in the review.	0.189	0.175	0.078
Fairness	Perceived impartiality and balance in judgments.	0.163	0.186	0.139
Politeness	Tone and manner of the review language.	0.128	0.053	0.106
Clarity and Readability	Ease of understanding the review, including grammar and structure.	0.124	0.038	0.117
Factuality	Accuracy of the statements made in the review.	0.115	0.006	0.089

swap in any comparable model. For each review, the LLM receives the full review text plus the title and abstract of the associated paper, and is prompted to score the review along several qualitative dimensions. Each dimension is paired with a rubric that specifies the scale and anchors scores in concrete criteria, adapted from editorial guidelines and prior review-quality annotation schemes [21]. The full rubric set is given in Table 1. These scores are not meant to replace human judgment, but to provide a complementary layer that captures latent qualities beyond traditional features.

3.3 Overall Quality Estimator

While structured and model-based metrics provide useful signals about specific dimensions of peer review quality, they do not by themselves yield a unified assessment that reflects how expert evaluators synthesize these attributes into an overall judgment. To produce such an aggregate estimate, PeeriScope incorporates a supervised scoring component trained on expert-annotated data. This module learns to map a heterogeneous feature set to a continuous quality score that approximates human assessments.

We use two classes of models to estimate overall quality. The first class consists of lightweight regressors trained on the full set of structured and LLM-derived features. These include a linear regression model, a random forest, and a two-layer multilayer perceptron. Their simplicity offers three advantages. They require no pretrained language model, which reduces computational overhead. They preserve feature-level interpretability, allowing users to trace model outputs back to specific input signals. They also support efficient inference, making them suitable for large-scale auditing tasks in conference or journal workflows.

The second class of models incorporates a fine-tuned large language model that receives the title, abstract, and review text as input and is trained to regress directly to the human-annotated overall quality score. We use LLaMA3 8B with parameter-efficient fine-tuning based on low-rank adaptation and eight-bit weight quantization. This architecture allows the model to integrate local linguistic context and high-level semantics without incurring prohibitive memory or deployment costs. We report agreement with human scores of both classes of models in §4.1.

4 Empirical Assessment of PeeriScope

We evaluate to what extend the quality evaluations produced by PeeriScope align with human expert judgments. We approach this by comparing the system's structured and LLM-based metrics, as well as its supervised quality estimators, against a set of reference annotations produced by trained human raters.

Dataset. We constructed a dataset of 753 peer reviews drawn from 200 papers published across OpenReview, F1000 Research, and the Semantic Web Journal. This data is available publicly on our GitHub for research purposes. Reviews were sampled to ensure coverage across venues and subject areas, with an emphasis on including both high- and low-quality examples. Each review was paired with the title and abstract of the corresponding paper and annotated independently by graduate students who were particularly trained for this task. Annotators scored each review along thirteen dimensions of quality, including comprehensiveness, fairness, clarity and more. An additional overall quality score was assigned using a continuous rubric designed to reflect editorial standards.

4.1 LLM-Human Alignment

To evaluate the degree to which PeeriScope's rubric-aligned LLM outputs capture meaningful aspects of review quality, we compare model-generated scores with human annotations for each of the abstract dimensions included in the evaluation rubric. In each case, the LLM receives the review text along with the paper title and abstract and is prompted to assign quality scores on a five-point ordinal scale. We compute Kendall's Tau correlation between the model's ranked outputs and the corresponding human judgments.

Table 1 presents the correlation results for three models. GPT-4o serves as a high-capacity commercial baseline, while Phi-4 and Qwen-3 represent open-source alternatives. Among the evaluated models, GPT-4o achieves the highest alignment across most dimensions. However, absolute scores remain modest, with the highest observed correlation for overall quality reaching only 0.359. Several dimensions, such as factuality, politeness, and clarity, show weak correlation across all models. These findings are consistent with recent work in LLM-based evaluation [8], which has shown that zero-shot judgments, while often fluent and plausible, can fail to track ground-truth assessments in complex subjective tasks. In peer review, this limitation is pronounced, as criteria are subtle and context-dependent. LLMs may capture broad patterns but still struggle with the nuanced reasoning typical of expert reviewers.

4.2 Comparison with Supervised Estimators

We evaluate the overall quality estimators introduced in Section 3.3 in terms of their alignments with human judgements. These include lightweight regressors trained on structured and LLM-based features, as well as a fine-tuned LLaMA3 8B model trained end-to-end on annotated review–paper pairs.

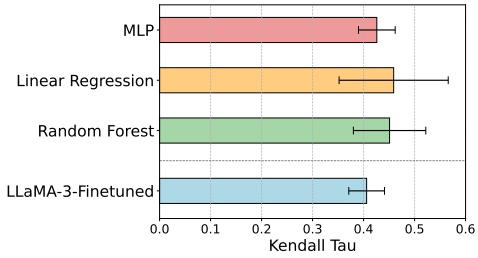

Figure 2: Kendall's τ correlation between human-evaluated and supervised overall quality estimators.

Each model produces a scalar quality score for every review in the dataset. We compute Kendall's Tau between the model predictions and human-assigned overall quality scores. Results on 10 fold cross validation are summarized in Figure 2, show that the structured-feature regressors outperform all zero-shot LLMs and also surpass the fine-tuned LLaMA model. Among these, a two-layer multilayer perceptron achieves the highest agreement, suggesting that relatively simple models can yield strong empirical performance when supported by carefully selected features. These results offer several insights. First, structured features grounded in theory and linguistic analysis retain high predictive utility despite their simplicity. Second, fine-tuning large models on small-scale review datasets may not yield robust improvements, especially when evaluative reasoning depends on multiple latent dimensions not easily captured in training signals. Finally, the consistent outperformance of supervised predictors relative to zero-shot LLMs supports the use of hybrid systems like PeeriScope, which combine interpretable metrics, model-based evaluation, and supervised calibration.

5 Implementation & Deployment Details

PeeriScope is built using a modular architecture that supports local and cloud deployment, with components implemented in FastAPI and ReactJS served by Docker containers. It exposes three REST endpoints for different review analysis tasks, integrates a locally hosted LLM (Qwen-3-8b in live demo via VLLM), and uses lightweight classifiers for complementary metrics. Reviewer metadata is sourced from OpenAlex[4] and stored in MongoDB, with all computations performed in-memory to ensure privacy. Incoming requests containing review or paper data are processed in-memory and discarded after metric computation. No content is stored or persisted beyond the duration of computation, thereby minimizing privacy exposure and satisfying lightweight compliance requirements. The frontend client is implemented in React 18 with TypeScript and the Vite toolchain. It interacts with the backend through Axios, renders visual outputs using Recharts, and maintains session state in Redux Toolkit. This architecture supports both standalone usage and embedded deployment in broader editorial systems.

6 Concluding Remarks

This demonstration introduces PeeriScope, a system for evaluating peer-review quality using structured metrics, LLM-based assessments, and supervised prediction. Grounded in discourse theory and validated against expert annotations, PeeriScope supports reviewer self-assessment, editorial triage, and large-scale audit. It offers a practical tool for authors, reviewers, and organizers seeking greater transparency and accountability in scientific evaluation through interpretable signals and interactive exploration. The system is openly accessible via a web interface and API, with a modular design that supports extensions to new metrics and domains. We hope it contributes to shared infrastructure for improving peer review and advancing research on scholarly evaluation and feedback.

[4]https://openalex.org/

References

[1] Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. 2024. PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation* 58, 4 (2024), 1291–1313.
[2] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180* (2020).
[3] Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. Llms assist nlp researchers: Critique paper (meta-) reviewing. (2024).
[4] 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.
[5] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).
[6] Markus Helmer, Manuel Schottdorf, Andreas Neef, and Demian Battaglia. 2017. Gender bias in scholarly peer review. *elife* 6 (2017), e21718.
[7] Ken Hyland. 1998. Hedging in scientific research articles. (1998).
[8] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594 [cs.AI]
[9] Miao Li, Eduard Hovy, and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. (2023).
[10] Ethan Lin, Zhiyuan Peng, and Yi Fang. 2024. Evaluating and enhancing large language models for novelty assessment in scholarly publications. (2024).
[11] 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. 2025. Breaking the Reviewer: Assessing the Vulnerability of Large Language Models in Automated Peer Review Under Textual Adversarial Attacks. (2025).
[12] Chengyuan Liu, Divyang Doshi, Muskaan Bhargava, Ruixuan Shang, Jialin Cui, Dongkuan Xu, and Edward Gehringer. 2023. Labels are not necessary: Assessing peer-review helpfulness using domain adaptation based on self-training. In *Proceedings of BEA 2023*.
[13] Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. 2025. LazyReview A Dataset for Uncovering Lazy Thinking in NLP Peer Reviews. *arXiv preprint arXiv:2504.11042* (2025).
[14] Shah Jafor Sadeek Quaderi and Kasturi Dewi Varathan. 2024. Identification of significant features and machine learning technique in predicting helpful reviews. *PeerJ Computer Science* 10 (2024), e1745.
[15] Lakshmi Ramachandran, Edward F Gehringer, and Ravi K Yadav. 2017. Automated assessment of the quality of peer reviews using natural language processing techniques. *International Journal of Artificial Intelligence in Education* (2017).
[16] Maria Sahakyan and Bedoor AlShebli. 2025. Disparities in Peer Review Tone and the Role of Reviewer Anonymity.
[17] Pawin Taechoyotin and Daniel Acuna. 2025. REMOR: Automated Peer Review Generation with LLM Reasoning and Multi-Objective Reinforcement Learning.
[18] Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025.
[19] Andrew Tomkins, Min Zhang, and William D Heavlin. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences* 114, 48 (2017), 12708–12713.
[20] Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 502–507.
[21] Andrii Zahorodnii, Jasper JF van den Bosch, Ian Charest, Christopher Summerfield, and Ila R Fiete. 2025. Paper Quality Assessment based on Individual Wisdom Metrics from Open Peer Review. *arXiv preprint arXiv:2501.13014* (2025).
[22] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *LREC-COLING 2024*. ELRA and ICCL.