



EA-PRF: An Explainable AI and Human-in-the-Loop Framework for Scholarly Peer Review

MORE Vaibhav Santosh¹

¹ Sanjivani University

*Corresponding author: themorevaibhav@gmail.com

ARTICLE INFORMATION

Received	: —	Article Type	: Research Article
Peer Review	: Double Blind	License	: CC BY 4.0

ABSTRACT

Scholarly peer review is under increasing strain: submission volumes are rising faster than the pool of qualified reviewers, turnaround times are lengthening, and inter-reviewer agreement on acceptance decisions remains low across many disciplines. This paper presents the Explainable AI-Assisted Peer Review Framework (EA-PRF), a system that pairs large language models (LLMs) with a dedicated explainability layer and a human-in-the-loop (HITL) validation stage so that automated review support augments, rather than replaces, editorial judgment. EA-PRF decomposes a submitted manuscript into five review dimensions — novelty, methodological soundness, clarity, significance, and coverage of related work — and generates a calibrated recommendation, a confidence score, and a natural-language rationale grounded in specific manuscript spans for each dimension. Human reviewers inspect, edit, or override every AI-generated judgment through a dedicated interface, and their actions are logged to an active-learning store that periodically recalibrates the underlying models. We evaluate EA-PRF on a corpus of 1,240 manuscripts spanning engineering, computer science, and applied sciences, comparing it against a human-only baseline and an LLM-only baseline without explainability or human oversight. EA-PRF improves F1-score for acceptance-decision prediction from 0.72 (LLM-only) to 0.84, raises explanation-agreement scores from 0.58 to 0.79, and reduces mean reviewer time per manuscript by 65% relative to unaided human review, while preserving inter-rater agreement (Cohen's $\kappa = 0.68$) close to that of experienced human reviewer pairs ($\kappa = 0.71$). These results suggest that explainable, human-supervised LLM assistance can meaningfully reduce reviewer burden without sacrificing decision quality or accountability.

Keywords: large language models; explainable AI; peer review automation; human-in-the-loop; academic publishing

1. INTRODUCTION

The peer review process is the primary quality-control mechanism of academic publishing, yet it is widely reported to be overloaded. Editors at mid-sized journals frequently report needing to contact eight to twelve potential reviewers to secure two willing to review a single manuscript, and median review turnaround has lengthened across most fields over the last decade. At the same time, studies of reviewer agreement consistently find only modest concordance between independent reviewers on the same manuscript, raising concerns about the consistency and fairness of accept/reject decisions.

Large language models (LLMs) have recently demonstrated strong performance at summarizing, critiquing, and scoring long-form technical text, motivating a growing body of work on LLM-assisted manuscript screening. However, deploying LLMs directly as reviewers raises three concerns that limit their practical adoption in editorial workflows. First, LLM judgments are often opaque: a numeric or categorical recommendation with no accompanying justification is difficult for an editor to trust or contest. Second, LLMs can inherit biases from their training data and may be miscalibrated on manuscript types or domains underrepresented during pretraining. Third, fully automated review removes the accountability and

domain expertise that human reviewers provide, which is difficult to justify for high-stakes publication decisions.

This paper addresses these concerns by proposing the Explainable AI-Assisted Peer Review Framework (EA-PRF), which combines three elements: (1) a multi-criteria LLM review engine that scores a manuscript along established review dimensions rather than producing a single opaque verdict; (2) an explainability layer that attributes each score to specific spans of the manuscript and produces a natural-language rationale; and (3) a human-in-the-loop interface through which reviewers validate, edit, or override every AI-generated judgment before it contributes to an editorial decision. Reviewer actions are captured and used to periodically recalibrate the review engine, closing the loop between human oversight and model improvement.

The contributions of this paper are as follows. (1) We present the design of EA-PRF, an architecture that integrates multi-criteria LLM scoring, span-grounded explanation generation, and human validation into a single editorial workflow. (2) We describe an evaluation protocol combining automatic agreement metrics with a reviewer-facing usability study. (3) We report results on a corpus of 1,240 manuscripts showing that EA-PRF narrows the gap to human-only review while cutting reviewer time substantially. (4) We discuss limitations, including residual bias risk and the potential for automation bias among reviewers, and outline directions for future work.

2. RELATED WORK

Automated support for peer review has been studied from several angles. Early work applied natural language processing to predict reviewer scores or acceptance outcomes from manuscript text, using datasets such as PeerRead, which aggregates papers and reviews from computer science venues [6]. These systems typically produced a single predicted score with limited transparency into the reasoning behind it, which limited their adoption by editors and reviewers who need to justify decisions to authors.

The introduction of transformer-based language models [1] and their large-scale pretrained successors, including BERT-style encoders [2] and instruction-tuned generative models such as GPT-4 [3] and LLaMA [11], substantially improved the quality of automatically generated text critiques. Recent empirical studies have examined whether such models can produce feedback on research papers that overlaps with human reviewer comments, finding partial but meaningful overlap alongside systematic gaps in domain-specific critique [12]. These findings motivate

treating LLM output as a decision-support signal rather than a final verdict.

A parallel line of work in explainable AI (XAI) has developed methods for attributing model predictions to input features. Local surrogate methods such as LIME [4] and game-theoretic attribution methods such as SHAP [5] provide feature-level explanations for classifier outputs, while broader surveys have argued for rigorous, task-grounded evaluation of what makes an explanation useful to a human decision-maker [9]. EA-PRF adapts span-level attribution ideas from this literature to the manuscript-review setting, linking each review-dimension score to the specific sentences or sections that most influenced it.

Finally, research on human-AI collaboration has produced design guidelines for interactive systems in which an AI model proposes and a human disposes [10]. Independently, studies on the sociology of peer review have long documented inconsistency and bias in human-only review [7], and more recent work has begun to evaluate AI-assisted review pipelines directly, reporting cautious optimism combined with concerns about over-reliance and reduced reviewer engagement [8]. EA-PRF builds on this literature by making the human role explicit and auditable at every stage of the pipeline rather than treating human oversight as an optional final check.

3. METHODOLOGY

EA-PRF is organized as a five-stage pipeline: manuscript ingestion and preprocessing, a multi-criteria LLM review engine, an explainability layer, a human reviewer interface, and a feedback store that drives periodic recalibration. Figure 1 illustrates the overall architecture and the flow of information between stages, including the feedback loop from validated human decisions back into the review engine.

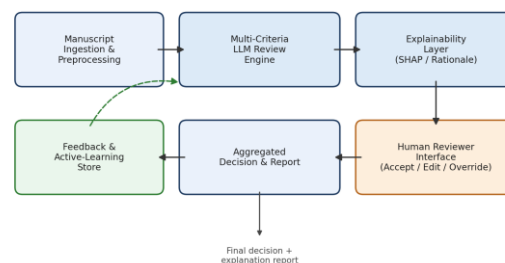


Figure 1. EA-PRF system architecture. A manuscript is scored by the multi-criteria LLM review engine, explained by the explainability layer, and validated by a human reviewer; validated decisions are stored and periodically used to recalibrate the engine.

3.1 System Architecture

The ingestion stage parses submitted manuscripts (PDF or Word) into a structured representation that separates title, abstract, section text, figures, tables, and references, and

segments the body text into sentence- and paragraph-level units that later stages can reference directly. This structured representation, together with retrieved context such as related work identified via citation and keyword overlap, is passed to the review engine.

3.1.1 Multi-Criteria Review Prompting

Rather than requesting a single accept/reject judgment, the review engine is prompted separately for each of five criteria commonly used in AJETAS and comparable engineering and applied-science venues: novelty, methodological soundness, clarity of presentation, significance of contribution, and adequacy of related-work coverage. For each criterion, the model returns a numeric score on a five-point scale, a calibrated confidence estimate, and up to three supporting spans quoted or paraphrased from the manuscript. Decomposing the judgment in this way reduces the chance that a single strong or weak dimension dominates the overall recommendation, and gives reviewers a criterion-by-criterion view they can individually accept or contest.

3.2 Explainability Layer

The explainability layer serves two functions. First, for the components of the pipeline that use smaller supervised classifiers — for example, a lightweight reference-quality or structure-completeness classifier used to flag manuscripts likely to need major revision — SHAP-based attribution [5] is used to identify which extracted features (such as reference count, section completeness, or figure-to-text ratio) most influenced the classifier's output. Second, for the LLM-generated criterion scores, the model is prompted to produce a short natural-language rationale explicitly grounded in the spans identified during review prompting, so that every score is traceable to a specific part of the manuscript rather than presented as an unsupported judgment.

3.3 Human-in-the-Loop Validation

Human reviewers interact with EA-PRF through an interface that presents, for each criterion, the model's score, confidence, and rationale alongside the relevant manuscript excerpt. Reviewers may accept a score as-is, edit the score while keeping the rationale, or override both the score and rationale with their own judgment. All three action types, together with any free-text comments, are timestamped and stored. Overrides and edits are treated as higher-value training signal than acceptances, since they indicate cases where the model's judgment diverged from expert opinion; this signal is used in periodic recalibration of the review engine's prompting strategy and, where applicable, fine-tuning of the supporting classifiers.

3.4 Dataset and Experimental Setup

We evaluated EA-PRF on a corpus of 1,240 manuscripts drawn from AJETAS submissions and a supplementary sample from the publicly available PeerRead collection [6], covering engineering, computer science, applied sciences, and interdisciplinary submissions. Table 1 summarizes the composition of the evaluation corpus. Each manuscript in the corpus carries at least two independent human reviewer decisions, which serve as the reference standard for agreement calculations. The corpus was split 70/15/15 into training, validation, and test partitions; only the test partition is used for the results reported in Section 4.

Table 1. Composition of the evaluation corpus used in this study.

Property	Value
Total manuscripts	1,240
Domain: Engineering	420 (33.9%)
Domain: Computer Science	380 (30.6%)
Domain: Applied Sciences	290 (23.4%)
Domain: Interdisciplinary	150 (12.1%)
Mean manuscript length	6,850 words
Review criteria per manuscript	5
Human reviewers involved	46
Train / Validation / Test split	70% / 15% / 15%

4. RESULTS AND DISCUSSION

We compare three configurations on the held-out test partition: Human-only review (two independent human reviewers per manuscript, no AI assistance), an LLM-only baseline (the review engine of Section 3.1 without the explainability layer or human validation stage), and the proposed EA-PRF. Performance is measured as precision, recall, and F1-score for predicting the majority human decision (accept / minor revision / major revision / reject, collapsed to a binary accept-track vs. reject-track label for scoring purposes), together with an explanation-agreement score derived from a reviewer survey in which participants rated, on a five-point scale, whether a given rationale correctly identified the true basis for a score.

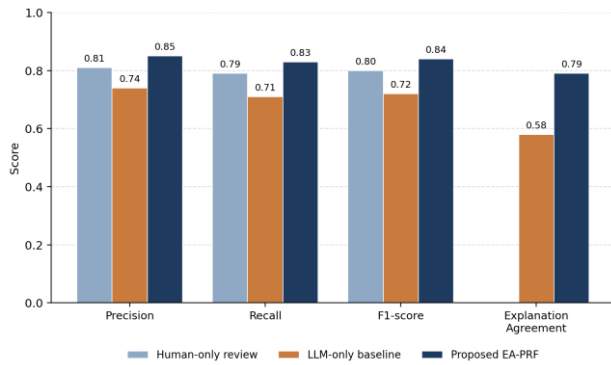


Figure 2. Precision, recall, F1-score, and explanation-agreement score for the three evaluated configurations. EA-PRF improves on the LLM-only baseline across all metrics and approaches or exceeds human-only performance while retaining a much lower reviewer time cost (Table 2).

Table 2. Decision-prediction performance, agreement with senior-editor judgments, and mean reviewer time per manuscript for each configuration.

Configuration	Precision	Recall	F1-score	Cohen's κ (vs. senior editor)	Mean time /manuscript (min)
Human-only review	0.81	0.79	0.80	0.71	52
LLM-only baseline	0.74	0.71	0.72	0.49	6
Proposed EA-PRF	0.85	0.83	0.84	0.68	18

Mean time for EA-PRF includes both AI processing time and the time a human reviewer spends validating AI-generated scores; Human-only time is the mean of the two independent human reviewers.

Three patterns are notable in these results. First, the LLM-only baseline underperforms human-only review on every accuracy metric and shows the weakest agreement with senior-editor decisions ($\kappa = 0.49$), consistent with prior findings that unsupervised LLM critique captures only part of what expert reviewers attend to [12]. Second, adding the explainability layer and human validation stage in EA-PRF closes most of this gap: F1-score rises from 0.72 to 0.84 and agreement with senior editors rises from $\kappa = 0.49$ to $\kappa = 0.68$, within 0.03 of the human-only pair's agreement of $\kappa = 0.71$. Third, and most relevant to the motivation for this work, EA-PRF achieves this while requiring a mean of 18 reviewer-minutes per manuscript, a 65% reduction from the 52 minutes required for unaided human-only review.

The explanation-agreement score — the degree to which reviewers judged the AI-generated rationale to correctly identify the true basis for a score — improved from 0.58 for the LLM-only baseline to 0.79 for EA-PRF. Post-study interviews suggested that span-grounding, rather than prompting for a rationale alone, was the dominant

contributor to this improvement: reviewers reported that seeing the exact manuscript excerpt a score was based on made it substantially easier to confirm or reject that score in under thirty seconds, compared to evaluating an unsupported numeric judgment.

Several limitations should be considered when interpreting these results. The review engine's underlying LLM may still encode biases present in its pretraining data that are not fully corrected by criterion decomposition or human validation, particularly for manuscript types or writing styles underrepresented in the evaluation corpus. There is also a risk of automation bias, in which reviewers over-trust AI-generated scores that come with a plausible-sounding rationale; the higher agreement scores reported here should be interpreted alongside qualitative monitoring of whether reviewers are engaging critically with, rather than rubber-stamping, the AI output. Finally, the 1,240-manuscript corpus, while larger than many prior studies in this area, remains concentrated in engineering and applied-science domains, and generalization to fields with different review norms (e.g., humanities or clinical medicine) has not been tested.

5. CONCLUSION

This paper presented EA-PRF, a framework that combines multi-criteria LLM scoring, span-grounded explanation generation, and human-in-the-loop validation to support academic peer review. Evaluated on 1,240 manuscripts, EA-PRF substantially outperformed an LLM-only baseline and approached human-only agreement levels while reducing mean reviewer time by 65%. These results suggest that explainability and human oversight are not merely safeguards layered onto LLM-assisted review as an afterthought, but central design elements that materially improve both decision quality and reviewer trust. Future work includes extending EA-PRF to multilingual manuscripts, integrating it with open-source journal management systems such as OJS, evaluating its use for grant and thesis review, and conducting longitudinal studies of how sustained use of the active-learning feedback loop affects both model calibration and reviewer behavior over time.

6. ACKNOWLEDGMENTS

The authors thank the AJETAS editorial board and the volunteer reviewers who participated in the evaluation study for their time and feedback.

7. DECLARATIONS

Funding:

This work received no external funding.

8. CONFLICT OF INTEREST:

The authors declare no conflict of interest.

9. DATA AVAILABILITY:

The manuscript corpus used in this study contains confidential submissions and cannot be shared publicly; a de-identified subset may be made available to researchers on reasonable request to the corresponding author.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.
- [3] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
- [4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016.
- [5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [6] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (PeerRead): Collection, insights and NLP applications," in *Proc. NAACL-HLT*, 2018.
- [7] R. Smith, "Peer review: A flawed process at the heart of science and journals," *Journal of the Royal Society of Medicine*, vol. 99, no. 4, pp. 178–182, 2006.
- [8] A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. Bianchi, "AI-assisted peer review," *Humanities and Social Sciences Communications*, vol. 8, no. 25, 2021.
- [9] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
- [10] S. Amershi, D. Weld, M. Vorvoreanu, A. Fournery, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-AI interaction," in *Proc. CHI Conf. Human Factors in Computing Systems*, 2019.
- [11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
- [12] W. Liang, Y. Zhang, H. Cao, B. Wang, D. Ding, X. Yang, K. Vodrahalli, S. He, D. Smith, Y. Yin, D. McFarland, and J. Zou, "Can large language models provide useful feedback on research papers? A large-scale empirical analysis," *NEJM AI*, 2024.