Fair and Unbiased AI-Assisted Peer Review

Vansh Ramani

October 2, 2025

1 Problem Formulation

1.1 Problem Definition

We formalize fair AI-assisted peer review as a constrained optimization problem balancing efficiency and equity.

1.1.1 Review Process Model

Let $P = \{p_1, \ldots, p_n\}$ be submitted papers and $R = \{r_1, \ldots, r_m\}$ be reviewers. Traditional assignment $\mathcal{A} : P \times R \to \{0, 1\}$ where $\mathcal{A}(p_i, r_j) = 1$ indicates reviewer r_j reviews paper p_i.

Each review produces $E_{i,j} = \langle S_{i,j}, C_{i,j} \rangle$ with score $S_{i,j} \in [1, 10]$ and comments $C_{i,j}$.

1.1.2 Bias Sources

Demographic Bias occurs when:

$$\mathbb{E}[S_{i,j}|D_a, Q_i] \neq \mathbb{E}[S_{i,j}|Q_i]$$

where D_a represents author demographics and Q_i is intrinsic paper quality.

Institutional Bias manifests as:

$$\mathbb{E}[S_{i,j}|I_a = \text{high}] > \mathbb{E}[S_{i,j}|I_a = \text{low}]$$

for equal quality papers from high vs. low prestige institutions.

1.1.3 AI-Assisted Framework

AI assistance function $\mathcal{AI} : P \to \mathcal{F}$ generates features:

- Factual consistency checks F_c
- Citation verification F_{cit}
- Bias indicators F_b

Human reviewers then produce: $E^{AI}_{i,j} = f(p_i, \mathcal{AI}(p_i), r_j)$

1.2 Fairness Requirements

1.2.1 Individual Fairness

Similar papers receive similar treatment: For $Q_i \approx Q_j$:

$$||\mathbb{E}[S_i^{AI}] - \mathbb{E}[S_j^{AI}]|| \leq \epsilon$$

1.2.2 Group Fairness (Shah's Core)

No community has incentive to withdraw: For any subset $N' \subseteq N$, no alternative assignment $\hat{R}$ exists where all members of N' are strictly better off.

1.3 Optimization Objective

$$\max_{\mathcal{AI},\mathcal{A}} \text{Efficiency}(\mathcal{AI}, \mathcal{A})$$

subject to:

$$\text{IndividualFairness}(\mathcal{AI}, \mathcal{A}) \leq \epsilon_1 \tag{1}$$

$$\text{GroupFairness}(\mathcal{AI}, \mathcal{A}) = \text{True} \tag{2}$$

$$\text{HumanOversight}(\mathcal{AI}) = \text{True} \tag{3}$$

This ensures AI assistance improves review speed and consistency while maintaining fairness guarantees.

2 Literature Review

2.1 Nihar Shah's Foundational Work on Peer Review Fairness

Shah's research program has systematically documented and addressed bias in peer review through both empirical analysis and algorithmic solutions, establishing the theoretical foundation for fair AI integration.

2.1.1 Group Fairness Theory

Shah et al. [7] introduced the core insight that peer review can be modeled as a matching problem with fairness constraints. Their key contribution is the core stability concept: no community of researchers should have incentives to withdraw from large conferences due to unfair treatment. This prevents the fragmentation of scientific communities while ensuring equitable review allocation. The core requires that for any subset of researchers forming their own conference, at least one member would not be strictly better off than in the original system, providing both fairness guarantees and system stability that maintains the collaborative nature of scientific discourse.

2.1.2 Empirical Evidence of Identity Bias

Shah's analysis of the ITCS 2023 "middle-ground" approach—initially anonymizing authors but revealing identities after initial reviews—revealed that 7.1% of reviews changed their overall merit scores after identity revelation [7]. This demonstrates persistent identity-based bias even in supposedly blind systems. Key findings include that reviewers often couldn't identify authors initially,

score changes correlated weakly with institutional prestige, and conflict of interest detection remains challenging under anonymization protocols. This work provides concrete evidence for the bias patterns that AI-assisted systems must address.

2.2 LLMs in Peer Review: Opportunities and Risks

Recent work by Kuznetsov et al. [2] surveys NLP applications in peer review, identifying both promising automation opportunities and critical limitations that inform our framework design. Proven capabilities include factual consistency checking across paper sections, citation accuracy and formatting verification, basic statistical validation, and structural presentation analysis. These technical strengths align with our identification of AI-suitable tasks in the review process workflow.

However, critical limitations constrain current AI applications. Novelty assessment requires deep domain understanding beyond current AI capabilities, while significance evaluation depends on research trends awareness that LLMs lack. Bias amplification from training data presents ongoing risks, and AI systems show limited reasoning about experimental design appropriateness. These limitations support our emphasis on preserving human authority over conceptual evaluation tasks.

Recent empirical studies reveal concerning vulnerabilities in AI-assisted review systems. Tyser et al. [8] demonstrate that LLMs can be manipulated through hidden prompt injection techniques, where authors embed invisible instructions in papers to bias AI reviewers toward positive evaluations. Li et al. [3] show that aspect-guided perturbations can significantly influence LLM review outcomes, while Lin et al. [4] reveal systematic vulnerabilities to textual adversarial attacks. These findings underscore the critical importance of bias monitoring and human oversight in our proposed framework.

2.3 Emerging Evidence of Systemic Bias

Large-scale empirical studies provide compelling evidence for the bias patterns our framework addresses. Sahakyan and AlShebli [6] conducted one of the most comprehensive linguistic analyses of peer review, examining over 80,000 reviews to uncover systematic disparities in review tone, sentiment, and supportive language across author demographics including gender, race, and institutional affiliation. Their analysis reveals that bias manifests not only in scores but in the language of evaluation itself, suggesting that AI bias detection systems must examine textual patterns beyond numerical ratings.

Müller et al. [5] analyze 39,280 review reports from the Swiss National Science Foundation, finding that gender and discipline significantly shape review length, content focus, and tone. Female reviewers write longer, more criterion-aligned reviews with more positive sentiment, while reviews in Social Sciences and Humanities are more critical compared to STEM fields. These findings highlight the complex interactions between reviewer characteristics, disciplinary norms, and evaluation outcomes that AI systems must navigate.

Giannakakos et al. [1] provide systematic review evidence that single-blind peer review systematically favors male authors, White researchers, US-based scholars, well-published scientists, and those from prestigious institutions. Crucially, their analysis reveals that double-blind review eliminates these advantages but does not necessarily improve outcomes for disadvantaged groups, suggesting that bias mitigation requires active intervention rather than simple anonymization.

2.4 AI Integration Challenges

Zhou et al. [9] document increasing LLM penetration in scholarly writing and peer review through their ScholarLens dataset and LLMetrica evaluation tool. Their findings reveal growing AI influence in scholarly processes while highlighting transparency and accountability challenges. This research emphasizes the need for explicit AI use disclosure and bias monitoring systems like those proposed in our framework.

Current literature reveals a critical gap between bias detection research and AI assistance implementation. Shah's group fairness framework provides theoretical foundations, while empirical studies document pervasive bias patterns, but few works integrate modern LLM capabilities with explicit fairness constraints. Our framework bridges this gap by embedding Shah's core stability principles directly into AI-assisted review systems, creating the first comprehensive approach to bias-aware LLM integration in peer review.

3 Review Process Analysis and AI Integration

3.1 Actual Peer Review Workflow

Based on empirical research on reviewer behavior, we analyze the real peer review process to identify optimal AI integration points:

3.1.1 Reviewer Cognitive Process

Figure 1 illustrates the actual cognitive workflow reviewers follow:

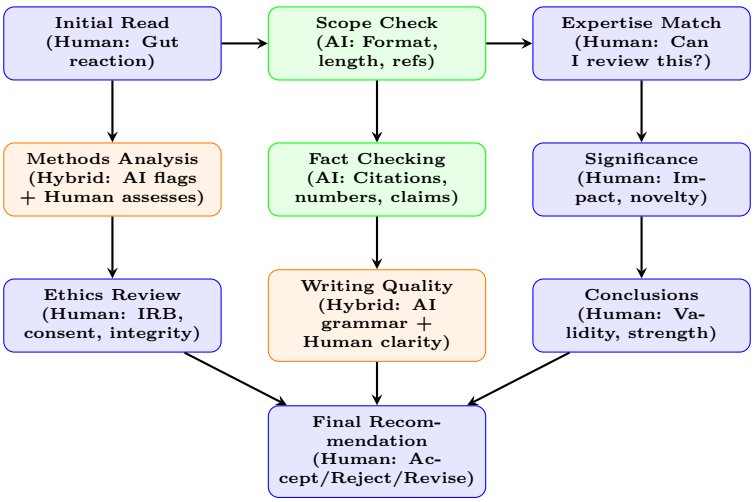

Figure 1: Realistic Peer Review Workflow: Blue=Human Essential, Green=AI Suitable, Orange=Hybrid

3.1.2 AI-Suitable Tasks (Green)

Administrative verification represents the most straightforward application of AI assistance in peer review. These mechanical tasks include reference formatting and completeness checking, word count and structural compliance verification, statistical notation consistency analysis, and citation

accuracy verification. AI systems excel at these pattern-matching activities that require minimal subjective judgment and can be performed consistently at scale.

Factual consistency checking presents another strong opportunity for AI assistance. Current LLMs demonstrate capability in cross-section consistency analysis (ensuring methods align with results), mathematical accuracy verification in calculations, data extraction from figures and tables, and basic statistical validation. These tasks leverage AI's strength in systematic verification while avoiding subjective interpretation.

3.1.3 Human-Essential Tasks (Blue)

Conceptual evaluation requires deep domain expertise and contextual understanding that remains beyond current AI capabilities. Research significance and impact assessment demands awareness of field trends, competitive landscape understanding, and intuition about breakthrough potential. Similarly, novelty determination within field context requires knowledge of subtle distinctions between related work, awareness of concurrent developments, and judgment about meaningful versus incremental advances.

Contextual judgment encompasses the most complex reviewer responsibilities. Ethical considerations beyond basic IRB compliance require nuanced understanding of participant protection, data use appropriateness, and potential societal implications. Expertise-specific methodology evaluation demands field-specific knowledge about appropriate research designs, valid measurement approaches, and acceptable trade-offs in experimental choices. Strategic research direction assessment and policy implications evaluation require broad perspective that current AI systems cannot provide.

3.1.4 Hybrid Tasks (Orange)

Methods assessment represents an optimal area for human-AI collaboration. AI can effectively flag potential methodological issues such as missing power analyses, unusual statistical tests, or incomplete experimental descriptions, while humans evaluate the appropriateness of methods for specific research questions and assess domain-specific validity concerns. This division leverages AI's systematic analysis capabilities while preserving human expertise for complex judgments.

Writing evaluation similarly benefits from hybrid approaches. AI excels at grammar checking, formatting consistency, and basic clarity analysis, while humans assess argument structure, logical flow, and field-appropriate communication standards. This collaboration can significantly improve review efficiency while maintaining quality standards for scientific discourse.

3.2 Integration Framework

Our framework implements three collaboration tiers that acknowledge the realistic boundaries of current AI capabilities. **Tier 1: AI-Primary** tasks include those where AI provides primary analysis with human validation checkpoints, primarily covering administrative and factual verification activities. **Tier 2: Human-Primary** encompasses tasks where human analysis is supported by AI-generated insights, particularly in methods assessment and writing evaluation. **Tier 3: Human-Only** reserves critical judgments requiring domain expertise and contextual understanding exclusively for human reviewers.

This realistic decomposition ensures AI enhances efficiency without compromising review quality or introducing bias through inappropriate automation. The framework acknowledges that effective

peer review requires human judgment for significance assessment, novelty evaluation, and contextual interpretation while leveraging AI strengths in systematic verification and pattern detection.

References

[1] Vasiliki P Giannakakos, Troy S Karanfilian, Antonios D Dimopoulos, and Anne Barmettler. Impact of author characteristics on outcomes of single- versus double-blind peer review: a systematic review of comparative studies in scientific abstracts and publications. *Scientometrics*, 130:399–421, 2025.

[2] Ivan Kuznetsov et al. Natural language processing for peer review automation: Opportunities and challenges. *arXiv preprint arXiv:2411.03417*, 2024. Available at: https://arxiv.org/abs/2411.03417.

[3] Jiatao Li, Yanheng Li, Xinyu Hu, Mingqi Gao, and Xiaojun Wan. Aspect-guided multi-level perturbation analysis of large language models in automated peer review. *arXiv preprint arXiv:2502.12510*, 2025. Available at: https://arxiv.org/abs/2502.12510.

[4] Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, et al. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks. *arXiv preprint arXiv:2506.11113*, 2025. Available at: https://arxiv.org/abs/2506.11113.

[5] Stefan Müller, Gabriel Okasa, Michaela Strinzel, et al. Gender and discipline shape length, content and tone of grant peer review reports. *arXiv preprint arXiv:2507.00103*, 2025. Available at: https://arxiv.org/abs/2507.00103.

[6] Maria Sahakyan and Bedoor AlShebli. Disparities in peer review tone and the role of reviewer anonymity. *arXiv preprint arXiv:2507.14741*, 2025. Available at: https://arxiv.org/abs/2507.14741.

[7] Nihar B Shah et al. Group fairness in peer review. *arXiv preprint arXiv:2503.15772*, 2025. Available at: https://arxiv.org/abs/2503.15772.

[8] Keith Tyser, Ben Segev, Gaston Longhitano, et al. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*, 2024. Available at: https://arxiv.org/abs/2408.10365.

[9] Li Zhou, Ruijie Zhang, Xunlian Dai, Daniel Hershcovich, and Haizhou Li. Large language models penetration in scholarly writing and peer review. *arXiv preprint arXiv:2502.11193*, 2025. Available at: https://arxiv.org/abs/2502.11193.