BiasChain: A Multi-Agent LLM Framework for Justified Peer Review Bias Detection

Anonymous ACL submission

Abstract

Peer review forms the cornerstone of academic quality control, yet it remains vulnerable to latent biases, topic preferences, and methodological disagreements, which can unfairly influence acceptance decisions. Manual bias audits are often resource-intensive and lack consistency. To address this, we propose a modular AI framework that leverages specialized large language model (LLM) agents to analyze sentiment and justification coherence, assess internal consistency, and evaluate inter-review alignment. These insights are then integrated into a schema-based module that identifies bias types, estimates confidence levels, and generates actionable recommendations. By automating these steps, our approach offers an added layer of confidence to editors and area chairs by providing a transparent and scalable tool for continuous bias monitoring in the peer review process. We have made the source code and supplementary materials publicly available[1] to support reproducibility and encourage future research.

1 Introduction

Peer review underpins the integrity of scholarly communication, yet it is far from immune to human biases. Seminal investigations revealed that demographic factors such as gender and institutional prestige can unduly sway reviewers' judgments: female grant applicants needed significantly stronger records to secure equivalent scores (Wennerås and Wold, 1997), and authors from top-tier universities enjoy a measurable advantage under single-blind review (Tomkins et al., 2017). Such distortions not only jeopardize fairness but also skew the trajectory of research fields by privileging certain topics or methodologies over others.

To mitigate bias, manual audits, where editors comb through text for unfair language or inconsistent criteria, are sometimes employed. However, these audits are labor-intensive and themselves subject to inter-auditor variability. Automated Natural Language Processing (NLP) techniques offer a promising alternative. Early sentiment-analysis tools can quantify tone and flag overly harsh or unduly lenient comments (Pang and Lee, 2008), while recent benchmarks such as ReviewEval provide frameworks to evaluate AI-generated reviews for coherence and relevance (Kirtani et al., 2025). Surveys of automated peer-review systems further catalog linguistic and structural checks (Zhuang et al., 2025), yet none integrates bias detection directly into an end-to-end pipeline.

Several prior works have targeted individual bias-detection subtasks. Manzoor and Shah (2020) propose statistical techniques to surface latent biases in review corpora, and AgentReview (Jin et al., 2024) uses LLM agents to simulate review dynamics and highlight decision variability. While informative, these approaches typically address only topic or rating bias in isolation, lacking mechanisms to cross-validate findings or generate actionable recommendations for reviewers and organizers.

Emerging large-language models and agentic AI architectures promise richer, multi-facet analysis. Chain-of-thought prompting has been shown to elicit more reliable multi-step reasoning, enabling coherence checks within free-form text (Wei et al., 2022). Bias-Aware Agent frameworks extend this by orchestrating specialized agents to identify and correct for skewed evidence (Singh and Ngu, 2025), and recent surveys of agentic AI for scientific discovery demonstrate how multi-agent workflows can tackle complex evaluation tasks (Gridach et al., 2025).

In this work, we introduce a modular, LLM-driven framework that (1) analyzes sentiment and justification coherence in individual reviews, (2) verifies internal logical consistency, (3) assesses inter-review alignment and contradiction, and (4) consolidates these signals into a schema-based

[1] BiasChain GitHub Repository

module that flags bias types, assigns calibrated confidence scores, provides supporting evidence, and suggests actionable improvements. We implement this pipeline on a sampled subset of the PeerSum dataset (Miao Li and Lau, 2023).

The remainder of this paper is structured as follows. Section 2 details our Methodology, covering dataset curation, agent designs, prompt schemas, and implementation specifics. Section 3 describes the results and does both qualitative and quantitative analysis. Section 4 discusses the limitations in detail while section 5 describes the future scope. Finally, Section 6 concludes and outlines directions for future research.

2 Methodology

Our research methodology employs a multi-stage pipeline designed to detect and analyze bias in academic peer reviews. The process involves dataset preparation, sequential analysis through specialized LLM-powered agents, and a final bias detection phase using purpose-built modules. We leveraged the *gemini-2.0-flash-thinking-exp* language model (Team et al., 2023) throughout the pipeline due to its robust reasoning capabilities and effectiveness in handling complex textual inference tasks. Each sub-task was addressed using a unique prompt.[2]

2.1 Dataset Acquisition and Preprocessing

We used the publicly available oaimli/PeerSum dataset (Li et al., 2023), specifically the train split containing over 14,000 academic papers with associated peer reviews. Due to computational constraints, we created a random subset of 200 papers, each with at least three reviews, yielding approximately 1,200+ reviews for analysis. This sampling approach provided sufficient data for meaningful analysis while keeping computational requirements manageable. From the original dataset schema, we extracted only fields directly relevant to our research questions: paper_id, paper_title, paper_abstract, review_id (filtering for 'official_reviewer' role only), and review_contents. This preprocessing step eliminated extraneous information and focused our analysis on the core review content.

[2]The prompts for all sub-tasks are available at an anonymous code repository along with the supplementary material.

2.2 Sentiment and Tone Analysis

The first analytical stage in our pipeline examined the affective dimensions of peer reviews. The model analyzed each review for sentiment and tone characteristics. The agent was instructed to function as an expert review analyzer.

The output schema captured sentiment classification (*Positive, Negative, or Neutral*) along with accompanying justifications, as well as tone classification (*Formal, Informal, Neutral, Supportive, Critical, or Balanced*) with corresponding rationales. This analysis provided the first layer of insight into reviewer disposition and potential affective biases.

2.3 Internal Consistency Analysis

The second analytical stage evaluated each review for internal logical consistency. Using the same model, we analyzed whether individual reviews maintained coherent arguments without self-contradiction.

The resulting structured output included a binary consistency assessment (*Yes/No*) accompanied by explanatory reasoning. This stage helped identify reviews containing inconsistent evaluations that might indicate bias in the reviewer's approach.

2.4 Inter-Review Comparison

The third stage involved a comparative analysis of reviews for the same paper to identify alignment patterns and outlier perspectives. The model was configured to simulate the role of an editor or area chair, comparing each review against all other reviews for the same paper:

This comparative analysis provided crucial context for determining whether individual reviews deviated significantly from peer consensus, indicating reviewer bias by giving a structured output containing the consistency label, alignment score, contradictions, bias flags, and a summary of differences.

2.5 Specialized Bias Detection

The final stage of our pipeline implements a complex, agent-based framework specifically designed to detect and categorize different types of bias within academic peer reviews. Our specialized bias detection system employs a hierarchical agent structure consisting of a parent agent that coordinates the entire bias detection process. It receives the comprehensive peer review data along with the metadata generated from previous pipeline stages.

The parent agent determines which specialized detectors to invoke based on initial content analysis. Then five purpose-built bias detection tools are implemented:

Novelty Bias Detector: Identifies cases where reviewers overemphasize or undervalue the novelty of approaches at the expense of practical or theoretical contribution (Wang et al., 2017).
Methodology Bias Detector: Detects bias toward particular methodologies, research paradigms, or trendy approaches independent of their objective merit.
Confirmation Bias Detector: Recognizes when reviewers favor papers that align with their preexisting beliefs, hypotheses, or prior work (Mahoney, 1977).
Positive Results Bias Detector: Identifies preference for papers with positive, significant, or state-of-the-art results over equally valid null or negative findings (Emerson et al., 2010).
Linguistic Bias Detector: Detects when reviewers penalize authors for linguistic or writing quality deficiencies, particularly affecting non-native English speakers (Politzer-Ahles et al., 2020).

3 Results and Analysis

3.1 Quantitative Analysis

3.1.1 Sentiment and Tone Analysis

Our first analysis stage revealed significant patterns in reviewer sentiment and tone distribution across the sampled peer reviews. Critical tone dominated the dataset (42%), followed by balanced (31%) and supportive tones (18%). Notably, informal tone was completely absent from our sample, suggesting a strong adherence to academic discourse conventions in peer review communications.

When examining the reasoning length generated by the AI agent for each tone category (see Figure 1), we observed that 'Critical' and 'Balanced' tones are associated with longer explanations (with an inter-quartile range of approximately 35-52 words). In contrast, 'Neutral' and 'Supportive' tones feature shorter justifications (typically in the 20-40 word range). This pattern suggests that the agent's model has learned that expressing negative or nuanced feedback requires more extensive justification to be comprehensive.

3.1.2 Internal Consistency Analysis

The second stage of our framework evaluated logical consistency within individual reviews, revealing that the vast majority of reviews (98.7%, n=1043) maintained internal consistency, with only a small fraction (1.3%, n=13) exhibiting contradictory statements. This finding aligns with expectations for high-quality academic peer review, where reviewers typically maintain logical coherence in their assessments.

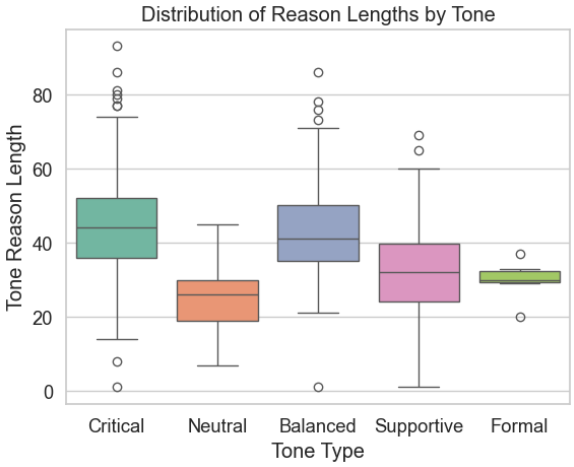

Figure 1: Distribution of Reason Lengths by Tone

3.1.3 Inter-Review Comparison

The third stage revealed compelling patterns in how individual reviews aligned with the consensus view for each paper. While the majority of reviews (75%) were judged consistent with other reviews of the same paper, a substantial minority (25%) diverged significantly from peer consensus. This divergence rate is particularly interesting as it quantifies the degree of reviewer disagreement in academic evaluation, which could stem from either bias or specialized expertise.

The kernel density estimation (KDE) plot (refer Fig. 2) further illustrated this relationship, with consistent reviews clustering toward higher alignment scores (7-10) and inconsistent reviews showing a broader distribution centered at lower values (3-6). This pattern suggests that while consistency with other reviewers typically correlates with higher alignment scores, inconsistent reviews exhibit greater variance in their alignment characteristics, potentially indicating multiple different types of divergence.

3.1.4 Specialized Bias Detection

Our multi-agent bias detection system identified potential bias in approximately 48% of the analyzed reviews, with varying confidence levels across different bias categories. This near-even split between biased and unbiased reviews suggests that reviewer

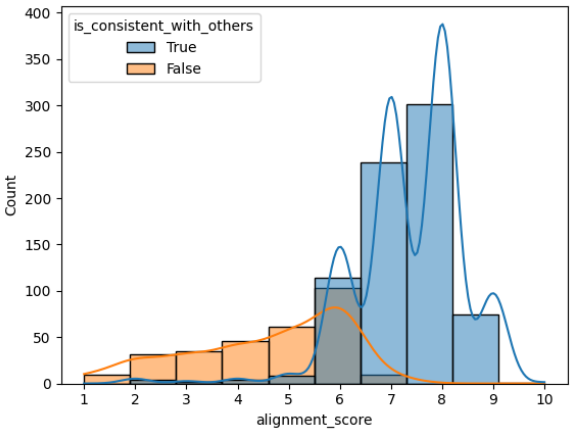

Figure 2: Alignment Score KDE Plot

bias may be more prevalent in academic peer review than previously acknowledged in the literature. Analysis of confidence scores (refer Fig. 3) showed an interesting asymmetry: our agents expressed higher confidence (mean: 9.4, range: 9-10) when classifying reviews as "unbiased" compared to when identifying specific bias types (mean: 7.8, range: 7-8). This pattern suggests that recognizing the absence of bias represents a simpler classification task than identifying specific bias subtypes, which require more nuanced analysis of the review text and context.

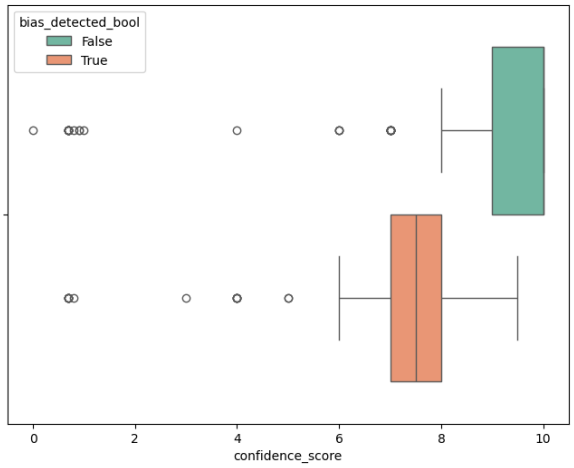

Figure 3: Confidence Score Box Plot

3.2 Qualitative Analysis

The comparison between two human experts and our agent reveals significant variations in bias detection sensitivity and interpretation. Human Expert 1 identified no bias across all six reviews, while Human Expert 2 detected bias in four reviews (67%), and the agent found bias in three reviews (50%). This disparity highlights fundamental disagreements about what constitutes bias versus legitimate scholarly critique.

The agent demonstrated systematic evidence extraction and reasoning comparable to the more sensitive human expert, providing specific quotes and detailed justifications for bias classifications. However, it showed some inconsistency in bias type categorization, classifying reviewer preferences for alternative methodological approaches as "confirmation bias" rather than "methodology bias" as identified by the human expert. The agent's intermediate sensitivity suggests it may serve as a useful screening tool, though the wide variation in human expert judgments (0% to 67% bias detection rates) underscores the inherent subjectivity in distinguishing between legitimate academic criticism and inappropriate reviewer bias.

4 Future Scope and Conclusion

Our work establishes a foundational framework for automated bias detection in academic peer review, yet several promising research directions emerge that warrant future investigation. The most immediate priority involves establishing robust validation methodologies through comprehensive human expert annotation studies and creating benchmark datasets with established ground truth labels. Additionally, expanding beyond our current five bias types to include demographic preferences, institutional prestige effects, and cognitive biases would provide more comprehensive coverage of reviewer fairness issues.

The integration of multi-modal analysis presents significant opportunities by incorporating reviewer expertise profiles, citation patterns, institutional affiliations, and historical interaction patterns for more nuanced bias assessment. This work demonstrates the feasibility of automated bias detection as a scalable alternative to resource-intensive manual audits, revealing bias in approximately 48% of 1,200 analyzed reviews, with novelty bias (29%) and confirmation bias (22%) most prevalent. Our hierarchical agent architecture successfully provides comprehensive fairness assessment through specialized modules, establishing a foundation for systematic bias monitoring that could enhance scholarly communication integrity and provide actionable insights for improving review fairness.

Limitations

While our LLM-driven framework demonstrates promising capabilities for automated bias detection in academic peer review, several important limitations must be acknowledged to properly contextualize our findings and guide future research directions.

Dataset Scale and Representativeness

Our analysis was constrained to 200 papers from the PeerSum dataset, yielding approximately 1,200 reviews. This limited sample size raises questions about generalizability across different academic disciplines, conference venues, and reviewer populations, as the dataset represents only a fraction of the broader academic peer review landscape.

LLM-Inherent Biases and Limitations

Our reliance on large language models introduces potential systematic biases from training data and may exhibit inconsistent performance across different writing styles or cultural contexts. The black-box nature of LLMs limits interpretability and validation of bias classifications, potentially hindering adoption by conference organizers requiring transparent evaluation tools.

Ground Truth and Validation Challenges

Our most significant limitation lies in the absence of established ground truth labels for bias in peer review, as bias detection represents subjective, contextually dependent judgment. Without inter-annotator agreement studies or validation against human expert judgments, we cannot establish precision and recall metrics or appropriately calibrate confidence thresholds.

Bias Type Coverage and Granularity

Our framework focuses on five specific bias types and employs binary classification (biased vs. unbiased), which may oversimplify bias that often exists on a spectrum. This approach may fail to capture mild bias cases that, while present, may not significantly impact overall evaluation quality.

5 Ethical Considerations

This work is not intended to replace human judgment in the peer review process. Instead, it provides an additional layer of confidence for bias detection. The study does not involve human subjects or private data; all analyses were conducted on publicly available, anonymized peer review datasets. The model outputs are interpretable and include confidence estimates to ensure transparency. All large language model (LLM) components were evaluated for fairness and alignment to minimize the risk of unintended bias.

References

Gwendolyn B Emerson, Winston J Warme, Fredric M Wolf, James D Heckman, Richard A Brand, and Seth S Leopold. 2010. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Archives of internal medicine*, 170(21):1934–1939.

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic AI for scientific discovery: A survey of progress and challenges. *arXiv preprint arXiv:2503.08979*.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with LLM agents. *arXiv preprint arXiv:2406.12708*.

Chhavi Kirtani, Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Murari Mandal, and Dhruv Kumar. 2025. Revieweval: An evaluation framework for AI-generated reviews. *arXiv preprint arXiv:2502.11736*.

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

Michael J Mahoney. 1977. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175.

Emaad Manzoor and Nihar B. Shah. 2020. Uncovering latent biases in text: Method and application to peer review. *arXiv preprint arXiv:2010.15300*.

Eduard Hovy Miao Li and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of EMNLP 2023*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Stephen Politzer-Ahles, Teresa Girolamo, and Samantha Ghali. 2020. Preliminary evidence of linguistic bias in academic reviewing. *Journal of English for academic purposes*, 47:100895.

Karanbir Singh and William Ngu. 2025. Bias-aware agent: Enhancing fairness in AI-driven knowledge retrieval. *arXiv preprint arXiv:2503.21237*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Andrew Tomkins, Min Zhang, and William D. Heavlin. 2017. Reviewer bias in single-blind peer review. *Proceedings of the National Academy of Sciences*, 114(2):127–132.

Jian Wang, Reinhilde Veugelers, and Paula Stephan. 2017. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Le, Michiel Bosma, Fei Xia, and Dale Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Christine Wennerås and Agnes Wold. 1997. Nepotism and sexism in peer-review. *Nature*, 387(6631):341–343.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *arXiv preprint arXiv:2501.10326*.