MEASURING LLM-ASSOCIATED SIGNATURES IN PEER REVIEW: A COMPARATIVE STUDY OF ICLR 2025 AND 2026

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly used to support academic peer review through linguistic refinement and content generation. However, detectors based primarily on surface-level writing patterns may conflate language editing with automated scientific judgment. This paper presents a comparative empirical study of 300 ICLR 2025–2026 submissions and 1,201 linked official reviews, including 75 accepted and 75 rejected papers from each year. SemDetect, a claim-level semantic detection framework, is used to distinguish human-associated, rewrite-associated, and AI-associated signatures. Of 1,178 scorable reviews, 776 (65.9%) are classified as rewrite associated, 401 (34.0%) as human associated, and one (0.1%) as AI associated; 23 reviews are retained separately as unscorable. Human-associated and rewrite-associated reviews have similar mean ratings (5.11 versus 5.06) and reviewer-reported confidence (3.58 versus 3.57), but rewrite-associated reviews exhibit lower raw rating dispersion (1.69 versus 1.89). At the paper level, rewrite proportion is negatively correlated with within-paper rating variance, although the association does not meet the conventional significance threshold after adjustment for mean rating, review count, and conference cohort ($p = 0.067$). Rewrite prevalence is also similar for accepted and rejected papers (66.3% versus 65.5%) and is not independently associated with final decisions. The findings indicate that the dominant detected signature is refinement rather than complete review generation. They also show that rating dispersion provides complementary information to average ratings, although the adjusted evidence for variance compression remains inconclusive. Because detection cannot establish authorship, all outputs are interpreted as probabilistic indicators rather than proof of individual LLM use.

1 INTRODUCTION

Peer review plays a central role in determining which scientific contributions are communicated and published. The growing availability of Large Language Models (LLMs) is changing how reviews are prepared. Reviewers may use these systems to improve clarity, summarize manuscripts, identify weaknesses, or generate substantial portions of an evaluation. A review refined using an LLM may preserve the reviewer's scientific judgment, whereas a substantially generated review may delegate part of the evaluation process to an automated system. Treating these practices as equivalent limits understanding of how LLM assistance relates to conference evaluation.

Existing research presents a mixed picture of LLMs as scientific reviewers. LLMs can produce fluent and organized evaluations, but their reviews may lack technical verification, manuscript specific reasoning, and sufficiently critical feedback (Zhou et al., 2024). They can assist with identifying weaknesses and improving critiques, although expert verification remains necessary (Du et al., 2024; Garg et al., 2025). Research based on scientific claims further shows that a plausible critique is not necessarily grounded in the evaluated manuscript (Ou et al., 2025). Linguistic quality therefore cannot be treated as evidence of reliable scientific judgment.

Several studies have examined LLM associated language in peer review. Liang et al. (2024) estimated the prevalence of substantially modified review text across major artificial intelligence con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ferences. Other approaches use linguistic features, token distributions, or comparisons with reviews generated from the same manuscript (Kumar et al., 2024). Detection benchmarks have also examined performance across domains and generation settings (Agrahari et al., 2025). However, many existing methods estimate prevalence at the collection level or divide reviews into only two categories: human and AI generated.

A binary distinction does not adequately represent cases in which a reviewer develops the scientific evaluation and subsequently uses an LLM to revise its presentation. Detection methods based on style, perplexity, or token distributions are also sensitive to paraphrasing, model changes, and domain variation (Sadasivan et al., 2023; Weber-Wulff et al., 2023). Their outputs cannot establish authorship or prove that a reviewer used an LLM. The present study therefore uses SemDetect, a semantic framework that compares a target review with reference reviews generated from the same manuscript (Duarte et al., 2026). It distinguishes human associated, rewrite associated, and AI associated signatures. All classifications are interpreted as probabilistic indicators rather than verified descriptions of reviewer behaviour.

The research question extends beyond the prevalence of these signatures. Reviews with similar average ratings may differ in how broadly they use the rating scale. A higher proportion of rewrite associated reviews could correspond to lower score dispersion even when average ratings and reviewer confidence remain unchanged. Such a pattern could appear as stronger reviewer agreement, particularly for papers near a decision boundary. Differences in dispersion may also reflect manuscript quality, conference cohort, review count, or dependence among reviews of the same paper. Analysis at both the review and paper levels is therefore required.

This study examines 300 ICLR 2025 and 2026 submissions linked with 1,201 official reviews. It evaluates the prevalence of human associated, rewrite associated, and AI associated signatures and examines their relationships with ratings, reviewer confidence, within paper score variance, and final decisions. Unscorable reviews are retained as a separate processing outcome rather than being assigned to the human category. Because the design is observational and the classifications are probabilistic, all findings are interpreted as associations rather than causal effects.

The main contributions are as follows:

- A verified dataset links 300 ICLR 2025 and 2026 submissions with 1,201 official reviews and final decisions.
- The analysis separates human associated, rewrite associated, and AI associated signatures while retaining unscorable reviews as a distinct outcome.
- Review level and paper level analyses examine the relationships among rewrite associated signatures, rating dispersion, reviewer confidence, and final decisions while accounting for review dependence, conference cohort, and manuscript provenance.

2 Related Work

Research on LLMs in peer review has examined the availability of review datasets, the quality of automatically generated reviews, and the detection of LLM associated text. Existing work provides evidence about the prevalence and quality of AI assisted reviewing, but gives less attention to reviews that may have been refined rather than fully generated. It also remains unclear whether these intermediate signatures correspond to differences in rating dispersion or conference decisions.

2.1 Peer Review Data and LLM Assistance

Open peer review datasets established the foundation for computational research on scientific evaluation. Kang et al. (2018) introduced PeerRead for tasks such as review analysis and decision prediction. Dycke et al. (2023) later developed NLPeer, which contains manuscripts, reviews, and revisions from multiple computer science venues. These resources support the study of review language and outcomes, but were not designed to distinguish different levels of LLM assistance or connect such distinctions with reviewer disagreement.

Research evaluating LLMs as scientific reviewers presents mixed results. LLMs can generate fluent and structured evaluations, but remain limited in technical assessment, score prediction, and specific

criticism (Zhou et al., 2024). They can assist with reviewing and critique revision, although expert verification remains necessary (Du et al., 2024). Evaluation frameworks further show that apparent fluency does not guarantee sufficient coverage, specificity, or usefulness (Garg et al., 2025). Similarly, a plausible critique may not be grounded in the scientific claims of the evaluated manuscript (Ou et al., 2025). These findings distinguish linguistic quality from substantive evaluation quality.

Most of this literature examines reviews generated substantially or entirely by an LLM. Less attention has been given to reviews in which the scientific assessment is written by a reviewer and subsequently refined using an LLM. Such refinement may preserve the reviewer's main position while changing how criticism, confidence, and numerical judgment are expressed.

2.2 Detection of LLM Associated Review Text

Approaches for identifying LLM associated review text operate at different levels. Liang et al. (2024) estimated the prevalence of AI modified content using changes in word frequency distributions. This approach supports collection level estimation but does not produce verified classifications for individual reviews. Other methods use linguistic features or comparisons with reviews generated from the same manuscript (Kumar et al., 2024). Detection benchmarks have also examined generalization across domains and generation settings (Agrahari et al., 2025).

Reliable detection remains difficult because performance can change across models, prompts, domains, and paraphrasing strategies (Sadasivan et al., 2023; Weber-Wulff et al., 2023). Detector outputs therefore cannot establish authorship or prove that a reviewer used an LLM. The present study uses SemDetect, which compares a target review with reference reviews generated from the same manuscript and constructs claim level semantic features (Duarte et al., 2026). Its outputs distinguish human associated, rewrite associated, and AI associated signatures. Reviews lacking sufficient semantic information are retained separately as unscorable.

2.3 Research Gap and Study Goal

Existing studies primarily examine the prevalence, linguistic characteristics, and quality of LLM generated or modified reviews. Limited evidence is available on whether rewrite associated signatures correspond to changes in numerical rating distributions. This is an important gap because two groups of reviews may have similar mean ratings while differing in rating dispersion. At the paper level, lower dispersion could appear as stronger reviewer agreement even when the average recommendation remains unchanged.

The present study addresses this gap by connecting SemDetect classifications with review ratings, reviewer reported confidence, within paper rating variance, and final conference decisions. The analysis separates classified and unscorable reviews, accounts for reviews clustered within papers, and evaluates whether rewrite prevalence is associated with rating dispersion or final decisions after considering mean rating, review count, and conference year. Because the design is observational, the results are interpreted as associations rather than evidence that LLM assistance caused changes in reviewer scores or decisions.

3 Methodology

3.1 Study Design

This study examines whether LLM associated signatures in peer review text correspond to differences in numerical ratings, reviewer confidence, rating dispersion, and final conference decisions. Official reviews from ICLR 2025 and 2026 are linked with numerical evaluations, manuscript metadata, and final decisions. Figure 1 summarizes the analytical pipeline.

The analysis is conducted at two levels. At the review level, canonical review text is processed using SemDetect (Duarte et al., 2026). The detector assigns one of three predicted categories: human associated, rewrite associated, or AI associated. Reviews without sufficient semantic information are recorded separately as unscorable. Unscorable is treated as a processing outcome rather than a fourth predicted category.

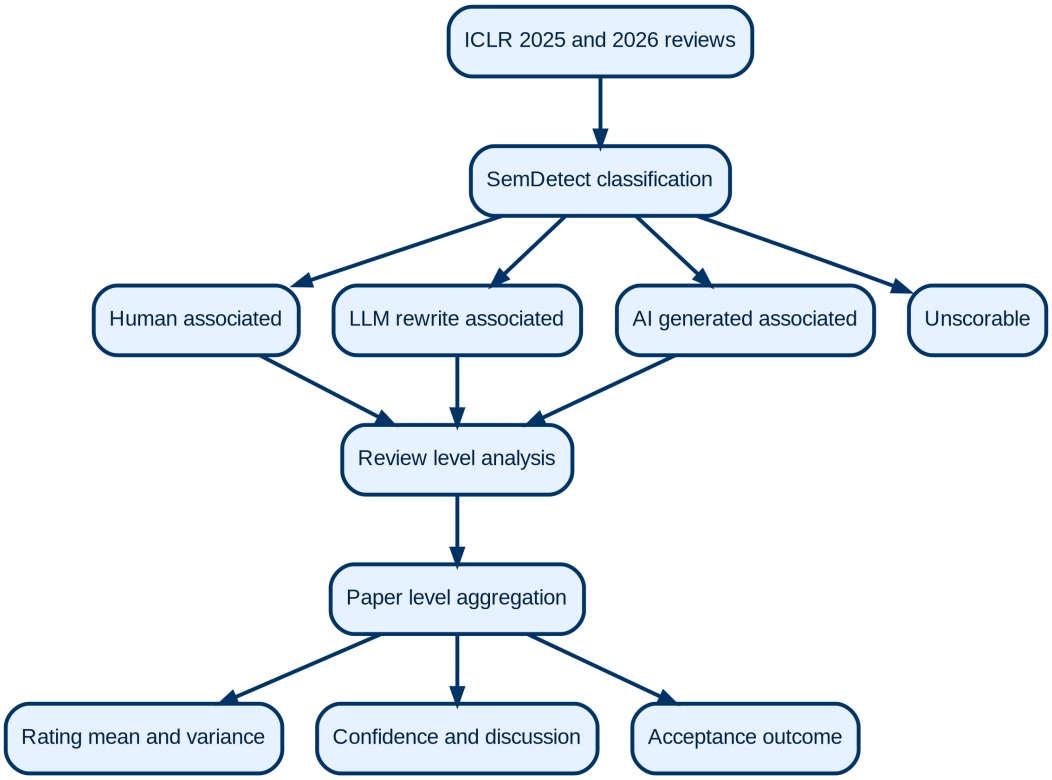


Figure 1: **Overview of the analytical pipeline.** Official ICLR reviews are processed using SEMDETECT and recorded as human associated, rewrite associated, AI associated, or unscorable. Review level measures are then aggregated by paper to examine rating means, rating variance, reviewer confidence, rewrite prevalence, and final decisions. Unscorable reviews are retained separately rather than assigned to the human category.

At the paper level, reviews belonging to the same submission are aggregated before relationships among rewrite prevalence, rating variance, confidence, and final decisions are evaluated. This structure accounts for dependence among reviews of the same paper and prevents papers with more reviews from receiving disproportionate weight.

3.2 DATASET CONSTRUCTION

The dataset was constructed from publicly available ICLR submissions and review records obtained from a pinned release of an OpenReview based benchmark.[1] Candidate papers were placed in a deterministic order for each conference year before eligibility screening. Papers were retained only when their titles, manuscript text, anonymity status, official reviews, and final decisions could be verified.

The final dataset contains 300 papers and 1,201 official reviews. It includes 75 accepted and 75 rejected papers from each conference year. The 2025 accepted and rejected groups contain 310 and 307 reviews, while the corresponding 2026 groups contain 291 and 293 reviews. Each paper has between three and six reviews, with an average of approximately four. All retained reviews include valid ratings and confidence scores.

Two manuscript verification tiers were used. Tier A contains 150 ICLR 2025 papers whose public PDFs were downloaded, hashed, parsed, matched by title, and confirmed anonymous. Tier B contains 150 ICLR 2026 papers whose text was obtained from the pinned and hash verified bench-

[1] https://github.com/qhjqhj00/iclr-openreview-reviews/releases/tag/v1.0

mark. These records passed title and anonymity checks, although their source PDFs were not stored locally.

Verification tier is perfectly aligned with conference year, so their contributions cannot be estimated separately. Differences between the 2025 and 2026 samples are therefore interpreted as cohort differences rather than isolated year or verification effects. The exact revision available at the original submission deadline could not be established for every paper. The analysis consequently uses verified anonymous review stage manuscripts without assuming that every artifact represents the initial submission version.

The sample was intentionally balanced by final decision and does not represent the natural ICLR acceptance rate. Descriptive acceptance proportions and model intercepts are therefore not interpreted as population level acceptance probabilities.

3.3 Review Processing and Classification

Each canonical review combines the available summary, strengths, weaknesses, and questions sections in their original order. SHA-256 hashes verify that review text matches the frozen source release. Review identifiers and paper forum identifiers provide deterministic links among reviews, manuscripts, and decisions.

For review i of paper p, $r_{ip} \in [1, 10]$ denotes the numerical rating and $c_{ip} \in [1, 5]$ denotes reviewer confidence. Both values were extracted from official records without imputation.

SemDetect returns a predicted category $\mathcal{C}_{ip}$ and a corresponding probability vector $\mathbf{p}_{ip}$. Predicted categories are used to define the primary paper level exposure. Detector probabilities are retained in the frozen outputs but are not interpreted as verified authorship probabilities. The single AI associated review is reported descriptively and excluded from human versus rewrite comparisons because one observation cannot support category level inference.

3.4 Paper Level Measures

For each paper, $\bar{r}_p$ and $\bar{c}_p$ denote the mean rating and mean confidence across its n_p reviews. Within paper rating variance is calculated as

$$s_p^2 = \frac{1}{n_p - 1} \sum_{i=1}^{n_p} (r_{ip} - \bar{r}_p)^2 . \tag{1}$$

This measure represents rating dispersion among reviewers of the same paper and is not treated as a direct measure of review quality.

The primary exposure is the paper level rewrite proportion:

$$q_p = \frac{n_{\mathrm{rewrite},p}}{n_{\mathrm{rewrite},p} + n_{\mathrm{human},p}}. \tag{2}$$

Unscorable and AI associated reviews are excluded from this denominator and tracked separately. This definition prevents unscorable reviews from being implicitly treated as human associated.

The final decision is represented by Y_p, where $Y_p = 1$ denotes acceptance and $Y_p = 0$ denotes rejection. These measures support the analyses described in Section 4.

All analyses use publicly available anonymous conference records. SemDetect outputs are interpreted as probabilistic indicators of textual patterns rather than proof of LLM use, reviewer identity, or policy violation. Because the design is observational, estimated relationships are reported as associations rather than causal effects.

4 EXPERIMENTAL SETUP

4.1 SEMDETECT CONFIGURATION

The frozen SemDetect pipeline uses `gpt-5.4-nano` to generate reference reviews and `gpt-4.1-mini` for claim cleaning and normalization. The pretrained classifier produces probabilities for human associated, rewrite associated, and AI associated signatures.

Valid predictions were obtained for 1,178 of the 1,201 reviews. The remaining 23 reviews lacked sufficient semantic information and were retained as unscorable rather than reassigned to the human category. Human versus rewrite comparisons exclude both unscorable reviews and the single AI associated review.

4.2 STATISTICAL ANALYSIS

Review ratings and confidence scores are summarized using means, standard deviations, medians, and interquartile ranges. Differences between human associated and rewrite associated reviews are estimated as

$$z_{ip} = \beta_0 + \beta_1 I(\mathcal{C}_{ip} = \text{rewrite}) + \beta_2 I(\text{Cohort}_p = 2026) + \epsilon_{ip}, \tag{3}$$

where z_{ip} represents either rating r_{ip} or confidence c_{ip}. Standard errors are clustered by paper to account for dependence among reviews of the same manuscript. The coefficient β_1 represents the adjusted difference between rewrite associated and human associated reviews. Differences in raw rating dispersion are additionally examined using the Brown–Forsythe test.

The primary paper level analysis estimates

$$s_p^2 = \alpha + \beta_1 q_p + \beta_2 \bar{r}_p + \beta_3 I(\text{Cohort}_p = 2026) + \beta_4 n_p + e_p, \tag{4}$$

where q_p is the rewrite proportion, $\bar{r}_p$ is the mean rating, and n_p is the review count. Heteroskedasticity robust HC_3 standard errors are used. Spearman correlation and a descriptive comparison above and below the median rewrite proportion provide complementary analyses. The continuous specification remains primary because median splitting discards information.

Final decisions are examined using logistic regression:

$$\text{logit}[P(Y_p = 1)] = \alpha + \beta_1 \bar{r}_p + \beta_2 s_p^2 + \beta_3 q_p + \beta_4 I(\text{Cohort}_p = 2026). \tag{5}$$

Coefficients, odds ratios, and p-values are reported. Because the sample contains equal numbers of accepted and rejected papers by design, the model intercept does not estimate the natural ICLR acceptance rate.

Robustness checks compare the continuous rewrite proportion with the median split groups and exclude the single AI associated observation from human versus rewrite comparisons. Conference cohort is included in the models, while verification tier is omitted because it is perfectly aligned with year. All estimates are interpreted as associations rather than causal effects.

5 RESULTS

5.1 PREDICTED REVIEW CATEGORIES

SemDetect produced valid classifications for 1,178 of the 1,201 reviews. Among the scorable reviews, 776 (65.9%) were classified as rewrite associated, 401 (34.0%) as human associated, and one (0.1%) as AI associated. The remaining 23 reviews (1.9% of the complete sample) were retained separately as unscorable. Rewrite associated reviews represented 66.4% of scorable reviews in 2025 and 65.3% in 2026.

Table 1: Review level rating and confidence statistics.

Category	N	Rating	SD	Confidence	SD
Human associated	401	5.11	1.89	3.58	0.77
Rewrite associated	776	5.06	1.69	3.57	0.78
AI associated	1	8.00	N/A	4.00	N/A

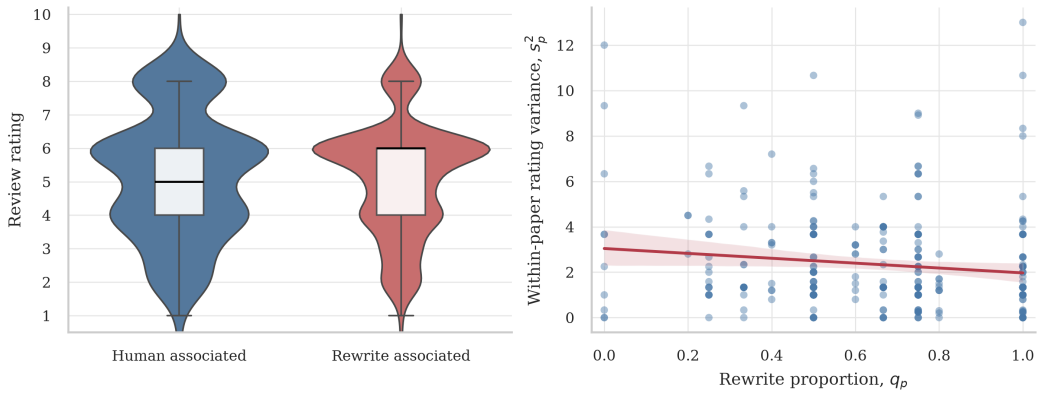


Figure 2: **Rating distributions and paper-level rating dispersion.** The left panel compares ratings for human-associated and rewrite-associated reviews. The right panel shows paper-level rewrite proportion q_p against within-paper rating variance s_p^2. The fitted line and shaded 95% confidence interval summarize the descriptive linear relationship. Adjusted regression results are reported below.

5.2 Ratings and Reviewer Confidence

Human associated and rewrite associated reviews had similar mean ratings, at 5.11 and 5.06, respectively. Their mean confidence scores were also nearly identical, at 3.58 and 3.57. The main descriptive difference concerned rating dispersion: the rating standard deviation was 1.89 for human associated reviews and 1.69 for rewrite associated reviews. Confidence dispersion was similar across the two categories.

The Brown–Forsythe test indicated a difference in raw rating dispersion between human associated and rewrite associated reviews ($p = 0.004$). However, this comparison does not account for reviews clustered within papers or differences between conference cohorts.

In the cohort adjusted model with paper clustered standard errors, rewrite classification was not associated with mean rating ($\hat{\beta} = -0.064$, SE $= 0.115$, $p = 0.576$). The corresponding confidence difference was also small and nonsignificant ($\hat{\beta} = -0.008$, SE $= 0.047$, $p = 0.860$). The two categories therefore remained similar in average rating and reported confidence.

5.3 Paper Level Rating Variance

Rewrite proportion was weakly negatively correlated with within paper rating variance ($\rho = -0.134$, $p = 0.020$). Papers below the median rewrite proportion had a mean rating variance of 2.61 and a median of 2.00, compared with a mean of 2.14 and a median of 1.33 among papers at or above the median. The rank based group comparison yielded $p = 0.048$.

Figure 2 shows similar rating centres for the two predicted categories but lower raw dispersion among rewrite associated reviews. At the paper level, the descriptive relationship between rewrite proportion and rating variance is weakly negative.

After adjustment for mean rating, review count, and conference cohort, the rewrite proportion coefficient remained negative but did not meet the conventional 0.05 significance threshold ($\hat{\beta} = -1.08$, HC_3 SE $= 0.59$, $p = 0.067$). The descriptive results are therefore consistent with lower rating

Table 2: Selected estimates from the acceptance model.

Predictor	Coefficient	Odds ratio	p-value
Mean rating	3.088	21.93	< 0.001
Rating variance	0.033	1.03	0.728
Rewrite proportion	0.460	1.58	0.467

dispersion among papers with more rewrite associated reviews, but the adjusted evidence remains suggestive rather than conclusive.

5.4 Association with Final Decisions

Rewrite prevalence was similar across decision groups. Rewrite associated reviews represented 66.3% of scorable reviews for accepted papers and 65.5% for rejected papers.

Mean paper rating was strongly associated with acceptance ($\hat{\beta} = 3.088$, OR $= 21.93$, $p < 0.001$). Neither rating variance ($\hat{\beta} = 0.033$, OR $= 1.03$, $p = 0.728$) nor rewrite proportion ($\hat{\beta} = 0.460$, OR $= 1.58$, $p = 0.467$) was independently associated with the final decision after adjustment for mean rating and conference cohort.

These results do not provide evidence that rewrite associated signatures confer an acceptance advantage. Because the sample is balanced by decision, the fitted model should not be used to estimate the natural probability of acceptance at ICLR.

6 Discussion and Conclusion

The results reveal a distinction between the prevalence of LLM associated signatures and their relationship with conference evaluation. Rewrite associated signatures appear in approximately two thirds of the scorable reviews, while only one review is classified as AI associated. This suggests that the dominant detected pattern is linguistic refinement rather than complete review generation. However, the detector identifies textual patterns and cannot establish how an individual review was produced.

Human associated and rewrite associated reviews have nearly identical mean ratings and reviewer confidence. The main descriptive difference concerns rating dispersion. Rewrite associated reviews have a lower raw rating standard deviation, and papers with higher rewrite proportions exhibit lower within paper variance in the unadjusted analyses. This relationship weakens after adjustment, with the primary variance coefficient reaching $p = 0.067$. The evidence is therefore consistent with possible variance compression but is not sufficient to establish a reliable adjusted association. Rewrite proportion is also not independently associated with final decisions after accounting for mean rating and conference cohort.

Several limitations affect interpretation. SemDetect classifications are probabilistic and may be sensitive to domain, model, and paraphrasing changes. Conference year is perfectly aligned with manuscript verification tier, so their contributions cannot be separated. The balanced decision sample does not represent the natural ICLR acceptance rate, and the precise submission cutoff revision could not be verified for every manuscript. The analysis is also limited to two conference cohorts, while the single AI associated review prevents meaningful inference about fully generated reviews.

Overall, the findings support cautious, distribution level auditing rather than deterministic judgments about individual reviewers. Average ratings alone may not capture all patterns associated with LLM supported reviewing, but the current evidence does not demonstrate that rewrite associated signatures alter final acceptance decisions.

6.1 Future Work

Future research should extend the analysis to additional conferences, years, and review scales. Voluntary disclosure or controlled experiments could provide ground truth for distinguishing language refinement from substantial generation. Larger studies could also evaluate calibrated detector prob-

abilities, alternative semantic detection methods, rebuttal activity, meta review recommendations, and changes in ratings over the review cycle. Such evidence would help determine whether the observed dispersion pattern generalizes and whether it affects how disagreement is interpreted during conference decision making.

AI USE STATEMENT

Generative AI tools were utilized to generate reference reviews, support claim normalization, assist with methodology refinement, assist with analysis script construction, and support LaTeX document formatting. AI tools were not utilized to generate, alter, or synthesize the official conference reviews, numerical ratings, confidence scores, manuscript metadata, or decision records analyzed in this study. All AI-assisted outputs were rigorously verified against primary sources and frozen dataset releases. The authors assume full responsibility for the integrity and accuracy of this manuscript.

REPRODUCIBILITY STATEMENT

Reproducibility is supported through frozen dataset releases, cryptographic SHA-256 hashes, deterministic processing pipelines, and automated verification scripts. The supplementary materials include the execution notebook, SEMDETECT configuration, data extraction and verification scripts, and complete analytical outputs. The execution environment records the reference generation, claim processing, embedding, feature extraction, and classification dependencies. Verification scripts confirm coverage across all $N = 300$ papers and $M = 1{,}201$ reviews and validate relational joins, input hashes, probability vectors, and feature matrices. API keys are omitted, and precomputed classifier outputs are provided to support reproduction without additional API access.

ETHICS STATEMENT

This study relies exclusively on publicly available, anonymized peer-review records retrieved from OpenReview. No attempt was made to deanonymize, re-identify, or contact reviewers, authors, or area chairs. All analytical findings are reported in aggregate. SEMDETECT classifications represent probabilistic indicators of textual and semantic signatures rather than deterministic proof of undisclosed AI usage or institutional policy violations. Unscorable boundary cases are explicitly isolated to prevent systematic classification bias.

AUTHOR CONTRIBUTIONS

Authors 1 and 2 jointly conceived and led the study, formulated the research questions, developed the methodology and statistical analysis plan, interpreted the findings, and led the preparation and revision of the manuscript. Author 3 provided project supervision, methodological guidance, theoretical feedback, and critical review of the manuscript. Author 4 constructed and validated the dataset, performed cryptographic verification, executed the SEMDETECT classification pipeline, and prepared the data processing and verification scripts under the supervision of Authors 1, 2, and 3. All authors reviewed and approved the final manuscript and take responsibility for its content.

REFERENCES

Shifali Agrahari, Sujit Kumar, and Ranbir Singh Sanasam. Can you really trust that review? ProtoFewRoBERTa and DetectAIRev: A prototypical few shot method and multi domain benchmark for detecting AI generated reviews. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia Pacific Chapter of the Association for Computational Linguistics*, pp. 2118–2140, Mumbai, India, dec 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-ijcnlp.132. URL https://aclanthology.org/2025.findings-ijcnlp.132/.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta,

Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5081–5099, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.292. URL https://aclanthology.org/2024.emnlp-main.292/.

André V. Duarte, Brian Tufts, Aditya Oke, Fei Fang, Arlindo L. Oliveira, and Lei Li. Sem-Detect: Semantic level detection of AI generated peer-reviews. *arXiv preprint arXiv:2605.21713*, 2026. doi: 10.48550/arXiv.2605.21713. URL https://arxiv.org/abs/2605.21713.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 5049–5073, 2023. doi: 10.18653/v1/2023.acl-long.277. URL https://aclanthology.org/2023.acl-long.277/.

Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. ReviewEval: An evaluation framework for AI generated reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20542–20564, Suzhou, China, nov 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1120. URL https://aclanthology.org/2025.findings-emnlp.1120/.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1647–1661, 2018. doi: 10.18653/v1/N18-1149. URL https://aclanthology.org/N18-1149/.

Sandeep Kumar, Mohit Sahu, Vardhan Gacche, Tirthankar Ghosal, and Asif Ekbal. 'Quis custodiet ipsos custodes?' who will watch the watchmen? on detecting AI-generated peer-reviews. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22663–22679, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1262. URL https://aclanthology.org/2024.emnlp-main.1262/.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024. doi: 10.48550/arXiv.2403.07183. URL https://arxiv.org/abs/2403.07183.

Jiefu Ou, William Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. CLAIMCHECK: How grounded are LLM critiques of scientific papers? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, nov 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1185. URL https://aclanthology.org/2025.findings-emnlp.1185/.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023. doi: 10.48550/arXiv.2303.11156. URL https://arxiv.org/abs/2303.11156.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *arXiv preprint arXiv:2306.15666*, 2023. doi: 10.48550/arXiv.2306.15666. URL https://arxiv.org/abs/2306.15666.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? A comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351, Torino, Italia, may 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.816/.