FROM FLUENCY TO ACCOUNTABILITY: A CLAIM-LIFECYCLE ANALYSIS OF AI-LIKE PEER REVIEWS AT ICLR 2026

Junjie Nian
College of Computer Science and Artificial Intelligence
Fudan University
Shanghai, China
github.com/JunjieNian/ICLR26-review-analysis

ABSTRACT

Large language models are rapidly becoming part of the ordinary infrastructure of peer review: they can polish prose, organize criticism, and draft substantial portions of a report. This shift creates a governance problem that detection alone cannot answer. A review matters not only because its sentences appear specific or well supported, but because its concerns can be answered, pursued, and ultimately connected to a decision. We study that chain in the public ICLR 2026 review record, covering 13,738 accept/reject papers, 53,463 reviews, and 416,862 extracted claims. Because individual AI-use labels are unavailable and unreliable, we construct a continuous unsupervised AI-likeness style proxy rather than classifying reviews as human- or AI-written. Reviews with higher scores show no broad deficit in specificity, evidential support, redundancy, or marginal decision information. Their claims nevertheless follow a different downstream trajectory. Within the same paper, a full-range increase in the 0–1 proxy is associated with a lower probability of author response ($\beta = -0.066$), a small increase in reviewer follow-up ($\beta = 0.006$), and lower meta-review uptake ($\beta = -0.017$). These observational patterns suggest that fluent review text can still lose continuity as it moves through deliberation. We therefore propose claim-level traceability, rather than stylistic detection alone, as a central criterion for evaluating AI-assisted peer review.

1 INTRODUCTION

Peer review was already under pressure before generative AI became widely available. Submission volumes have grown, relevant expertise has become more specialized, and reviewers are asked to produce careful judgments under severe time constraints. Large language models offer an immediate response to that pressure: they can turn notes into polished prose, expand a short concern into a structured critique, or draft a report from a manuscript. Their use is no longer merely hypothetical. Corpus-level work estimates that 6.5–16.9% of review text at several post-ChatGPT AI conferences was substantially modified by an LLM (Liang et al., 2024); a separate study places a lower bound of 15.8% on AI-assisted reviews at ICLR 2024 (Latona et al., 2024).

This growing presence has made detection the most visible policy question: which reviews were written with AI, and can they be reliably identified? Yet detection provides only a partial account of risk. A review is not a self-contained writing sample. It is a move in a deliberative process. Its criticisms consume rebuttal space, shape what reviewers discuss, and furnish the area chair (AC) with reasons for a final recommendation. The central institutional question is therefore not simply whether a report *sounds* machine-written. It is whether the concerns it raises can be located, answered, contested, and carried into the decision record.

Consider two equally fluent reviews. One offers a long catalogue of locally plausible issues, each weakly prioritized; the other identifies a small number of consequential weaknesses and explains why they matter. Both may look specific when read in isolation. Under a tight rebuttal budget, however, authors may triage the first, its reviewer may continue to press unresolved details, and the AC may

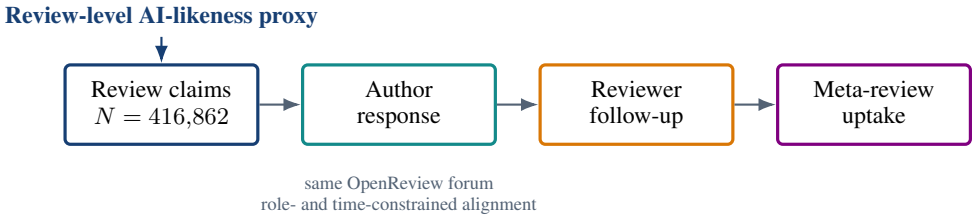

Figure 1: The claim-lifecycle framework. A review-level style proxy is linked to claim-level outcomes as criticisms move from the original review through rebuttal, follow-up, and institutional synthesis.

find little that can be compactly synthesized. The second review may leave a clearer institutional trace. Text quality and process usefulness can therefore diverge.

We study that divergence by asking: *How do claims from more AI-like reviews travel through peer-review deliberation?* The public ICLR 2026 record lets us follow the same concern from the original review into author responses, reviewer follow-up, and the meta-review. We first ask whether reviews with more AI-like style actually show the familiar textual deficiencies—vagueness, weak grounding, repetition, or little decision information. We then shift the unit of analysis from the review to the claim and measure whether each concern remains visible at later stages.

The resulting picture is not a simple indictment of AI-like writing. Across 53,463 reviews, higher AI-likeness is not accompanied by a general collapse in specificity, support, originality, or marginal predictive information. The clearer separation emerges downstream: within the same paper, claims from more AI-like reviews are less likely to receive an author response and less likely to appear in the meta-review, even though reviewers are slightly more likely to follow up on them. We call this separation between textual adequacy and downstream persistence *process-level discontinuity*.

This paper contributes a way to evaluate AI-assisted reviewing without pretending that style reveals authorship. Its claim-lifecycle representation connects review content to the social and institutional work that follows. Empirically, it identifies where high-AI-likeness reviews do and do not differ. Conceptually, it shifts the governance target from policing prose to maintaining traceable reasons: a useful review should leave concerns that participants can answer and decision makers can account for.

2 Related Work

AI use in peer review. The first wave of empirical work established that LLM assistance is already visible in scientific reviewing. Distributional methods estimate its prevalence at the corpus level (Liang et al., 2024), while quasi-experimental work relates likely assistance to ratings and acceptance (Latona et al., 2024). Moving from a corpus trend to an individual accusation is much harder. Reviews are short, technical, frequently edited, and written by authors with diverse linguistic backgrounds; detectors also shift as models and writing practices change. Recent ICLR 2026 benchmarking accordingly treats reliable attribution as unresolved (Yu et al., 2026). Our continuous proxy is designed for studying aggregate variation, not for deciding who used a model.

Generated reviews and review assistance. Work on generated reviews raises a different concern. Models readily produce organized, professional-looking reports, yet can struggle to identify consequential defects, calibrate scrutiny to the paper, or turn criticism into actionable requests (Li et al., 2025). Assistance need not replace the reviewer: the ICLR 2025 Review Feedback Agent instead prompted human reviewers to revise vague or unprofessional passages and increased discussion engagement in a randomized deployment (Thakkar et al., 2025). Together, these studies suggest that "AI review" is not a single treatment. Generation, editing, and feedback support can affect different parts of the process, making prose quality an incomplete outcome.

Rebuttal and deliberation. Research on rebuttals shows that outcomes depend not only on what reviewers initially write, but also on how authors respond and how participants interact afterward (Huang et al., 2023; Kargaran et al., 2025). This is also the process imagined by conference policy:

ICLR asks reviewers to engage in discussion and asks ACs to synthesize reviews together with author responses (International Conference on Learning Representations, 2026b;a). We connect the emerging literature on AI-assisted reviewing to this older deliberative perspective. The claim, rather than the whole report or its numerical score, is the smallest unit that can visibly persist, change, or disappear across stages.

3 Data and Setting

3.1 Corpus

ICLR is especially suitable for studying deliberation because OpenReview exposes more than a final score. The public forum links the original reviews to rebuttals, later comments, and the AC's meta-review, allowing a concern to be followed through time without joining separate datasets. We retrieve the ICLR 2026 forums and retain submissions with final accept or reject decisions, excluding withdrawals and desk rejections. This produces 13,738 papers and 53,463 official reviews. Segmenting the weakness and question fields yields 416,862 claims—the units whose later trajectories we study. The much larger collection of response, follow-up, and meta-review passages in Table 1 supplies the candidate text to which those claims are aligned.

Table 1: Corpus scale. Counts refer to the final aligned analysis release.

Unit	Count
Accept/reject papers	13,738
Official reviews	53,463
Extracted review claims	416,862
Author-response units	612,428
Reviewer follow-up units	62,003
Meta-review units	206,637

3.2 The ICLR 2026 security incident

The 2026 cycle was also exceptional. An OpenReview security incident interrupted the discussion period; the program chairs froze discussion, restored review text and scores to their pre-discussion state, and reassigned ACs, who were then asked to infer the expected outcome from the available record (ICLR 2026 Program Chairs, 2025). This disruption makes score change a poor measure of rebuttal success. It also makes the surviving textual record unusually important: what authors addressed, what reviewers pursued, and what the newly assigned AC ultimately wrote are the observable traces left by deliberation. We use those traces while treating the incident as a serious boundary on generalization, not as a source of causal leverage.

4 Methods

4.1 A continuous AI-likeness proxy

Because the forums do not contain reliable ground-truth labels for model use, a binary detector would introduce false certainty at the first step of the analysis. We instead describe how closely each review resembles the learned style profile associated with machine-polished text. Let review i contain tokens $w_{i1}, \ldots, w_{iT_i}$. We compute average token log-probability under GPT-2,

$$\bar{\ell}_i = \frac{1}{T_i - 1} \sum_{t=2}^{T_i} \log p_\theta(w_{it} \mid w_{i,<t}), \tag{1}$$

and derive distributional features such as perplexity and variation in token surprisal. These signals are combined with review length, lexical diversity, hedging, discourse transitions, sentence-length variation, list formatting, and passive voice. After standardization and principal-component compression, an Isolation Forest assigns an anomaly score (Liu et al., 2008). We orient and rescale this score

to $s_i \in [0, 1]$, where larger values indicate greater AI-likeness in the learned style space. Terciles make the descriptive comparisons readable, while all regressions retain the continuous score.

The name of the construct is intentionally cautious. A high score may arise from model assistance, but it may also reflect disciplined structure, non-native-language editing, field convention, or reviewer experience. Nothing in the analysis licenses an authorship judgment about a particular review.

4.2 TEXT-LEVEL OUTCOMES

Before asking what happens to a claim, we test the more familiar possibility that high-AI-likeness reviews are simply worse pieces of text. The outcomes cover concreteness, grounding, duplication, substantive coverage, and decision information; no single metric is asked to stand in for review quality.

Specificity. A rule-based index captures whether criticism is anchored to the object under review. It counts references to sections, figures, tables, equations, datasets, baselines, ablations, metrics, and numbers, normalizes by length, and caps extreme values at five.

Claim mode and evidence support. A stratified full-text sample of 3,000 reviews provides a closer test of grounding. From these reviews we extract 23,461 claims, then use a local language-model annotator to distinguish factual-supportable statements from requests, expectations, subjective judgments, and fragments. Only factual-supportable claims proceed to a second stage that retrieves paper passages and judges the claim supported, contradicted, or insufficiently evidenced. The separation matters: "add an ablation" may be a sensible recommendation, but it is not a factual statement that the submitted paper must already entail.

Redundancy and issue uniqueness. To distinguish repeated boilerplate from genuinely different coverage, we compute embedding similarity between reviews of the same paper. We also cluster weakness claims into issues and divide the number of distinct issue clusters contributed by a review by its number of extracted claims. These measures capture complementary forms of redundancy: similarity in whole-review language and repetition in the concerns themselves.

Incremental decision contribution. A review may be original yet irrelevant to the eventual recommendation. We therefore train a LightGBM model to predict acceptance from the complete set of reviews for each paper (Ke et al., 2017), then remove one review at a time. For review r on paper p, incremental information contribution is

$$\mathrm{IIC}_{r,p} = \left| \widehat{P}(A_p = 1 \mid R_p) - \widehat{P}(A_p = 1 \mid R_p \setminus \{r\}) \right|. \tag{2}$$

The absolute change asks how much predictive information disappears with that review. We compare paper-level means for low- and high-AI-likeness reviews and estimate a controlled review-level model.

4.3 CLAIM-LIFECYCLE ALIGNMENT

The lifecycle begins by turning each weakness or question field into claims using numbered lists, bullets, headings, and sentence boundaries. For a claim c, only text from the same paper forum and an eligible later stage can become a match: author comments can indicate response, later comments by the originating reviewer can indicate follow-up, and the AC's meta-review can indicate institutional uptake. We combine TF–IDF cosine similarity with token overlap,

$$M(c, u) = \lambda \cos(\mathrm{tfidf}(c), \mathrm{tfidf}(u)) + (1 - \lambda)\mathrm{overlap}(c, u). \tag{3}$$

The highest-scoring eligible passage is then classified with stage-specific thresholds as a direct match, a weaker or aggregated match, or unmatched. The binary outcomes count direct and partial response, claim-level and weak follow-up, and explicit and aggregated meta-review uptake. This design favors visible textual continuity; implicit acknowledgment that shares no language with the original concern can be missed.

4.4 Statistical analysis

The descriptive analysis uses nonparametric tests because most review outcomes are skewed and several are sparse. We report Kruskal–Wallis and Mann–Whitney tests with Bonferroni correction, Cliff's δ, bootstrap confidence intervals, a 10,000-draw permutation test for redundancy, and a paper-paired Wilcoxon test for IIC. Effect sizes and paired comparisons matter more here than significance alone: at this scale, very small differences can produce very small p-values.

For lifecycle outcome $k \in \{\text{response}, \text{follow-up}, \text{uptake}\}$, we estimate a paper fixed-effects linear probability model:

$$Y_{pim}^{(k)} = \alpha_p + \beta_k s_{pi} + \gamma^\top Z_{pim} + \epsilon_{pim}, \tag{4}$$

where p indexes papers, i reviews, and m claims. Controls include review rating and confidence, claim token count, question status, and answerability. Follow-up additionally controls for prior response; uptake controls for response and follow-up. Estimation uses within-paper demeaning and HC1 heteroskedasticity-robust standard errors.

Comparing claims within the same paper removes stable differences in paper topic and overall quality, while the controls absorb observed variation in the review and claim. It does not make reviewer assignment random. Moreover, s_{pi} remains on its original 0–1 scale: β_k is the model-implied probability difference across the full range, not a one-standard-deviation effect. Equation 4 therefore describes conditional association rather than a causal effect of AI use.

5 Results

5.1 Fluency is not the fault line

If AI-like style were merely a marker for poor reviewing, the high-score tier should be consistently vaguer, less grounded, more repetitive, and less consequential. That pattern does not appear. High-AI-likeness reviews are slightly more specific, but the difference is substantively negligible. After separating factual statements from requests and opinions, the two ends of the distribution contain almost the same share of factual-supportable claims. The supported share among adjudicated factual claims is directionally lower in the high tier, yet the review-level comparison is too uncertain to distinguish it from sampling variation (Table 2).

Nor do more AI-like reviews collapse onto the same generic script. Reviews in high–high pairs are, on average, less similar to one another than reviews in low–low pairs, and their extracted claims cover marginally more distinct issues. Removing a high-AI-likeness review from the acceptance-prediction model also changes the prediction about as much as removing a low-AI-likeness review. These measures capture different aspects of quality and none proves equivalence, but together they make a broad text-deficit account difficult to sustain.

Table 2: Text-level results. Tests compare low and high AI-likeness unless otherwise stated. "Supported" is restricted to factual-supportable claims with valid judgments.

Outcome	Low	High	Δ	Test	Reading
Specificity	3.280	3.341	+0.061	$p = 0.0013$, $\delta = -0.022$	Statistically detectable, substantively negligible
Factual-claim share	0.3467	0.3442	-0.0025	$p = 0.584$	Nearly identical claim composition
Supported-factual share	0.5895	0.5511	-0.0384	$p = 0.340$	Directional, not statistically separated
Review-pair similarity	0.462	0.416	-0.046	permutation $p < 10^{-4}$	High–high pairs are less similar
Unique-claim ratio	0.818	0.831	$+0.014$	MW $p = 3.3 \times 10^{-6}$	Small increase in issue uniqueness
Paired IIC	–	–	-0.0019	Wilcoxon $p = 0.114$	No reliable low–high difference

The relevant contrast therefore lies less in what the review looks like on first reading than in what happens after it enters the forum.

5.2 The divergence appears after the review is posted

The unadjusted lifecycle rates reveal the first sign of separation. Authors address a slightly smaller share of claims in high-AI-likeness reviews, while the reviewers who wrote those claims return to them slightly more often. At first glance, meta-review uptake appears nearly unchanged across terciles. The apparent stability masks a change in the unmatched tail, however: 66.38% of high-tier claims leave no detectable trace in the meta-review, compared with 62.90% in the low tier. In other words, more follow-up conversation does not necessarily translate into more visible influence on the final rationale. These comparisons remain descriptive because papers and reviewers differ across tiers, so we next compare claims within the same paper.

Table 3: Unadjusted lifecycle rates by review AI-likeness tier. "No uptake" is a claim-level label; other rows are review-level mean claim rates.

Outcome	Low	Medium	High	High–Low
Author response	34.50%	34.80%	33.47%	$-1.02\,pp$
Reviewer follow-up	0.62%	0.66%	0.77%	$+0.15\,pp$
Any meta-review uptake	7.09%	7.31%	7.27%	$+0.18\,pp$
No meta-review uptake	62.90%	63.42%	66.38%	$+3.48\,pp$

5.3 Within papers, interaction and uptake move apart

The within-paper models turn the descriptive hint into a coherent trajectory (Table 4; Figure 2). Among claims about the same submission, higher review AI-likeness is associated with a lower probability of author response and a lower probability of meta-review uptake. Reviewer follow-up moves in the opposite direction, but by much less. The result is not simple disengagement: high-AI-likeness claims can remain active enough for their reviewers to revisit them while still being less likely to enter the AC's synthesis.

The coefficients quantify this separation. Moving across the full 0–1 score range corresponds to a 6.63 percentage-point decrease in author-response probability, a 0.56-point increase in follow-up, and a 1.67-point decrease in meta-review uptake. Response and follow-up are themselves positively associated with later uptake in the final model, which is consistent with a general lifecycle intuition: a concern has a better chance of becoming a decision reason when participants keep it visible. The high-AI-likeness pattern is distinctive because additional reviewer activity does not fully restore that continuity.

Table 4: Coefficient on the 0–1 AI-likeness score in paper fixed-effects linear probability models ($N = 416{,}862$ claims). Confidence intervals use HC1 robust standard errors.

Outcome	β	SE	95% CI	p
Author response	-0.0663	0.0043	$[-0.0748, -0.0578]$	6.16×10^{-53}
Reviewer follow-up	$+0.0056$	0.0012	$[+0.0033, +0.0080]$	3.15×10^{-6}
Meta-review uptake	-0.0167	0.0030	$[-0.0227, -0.0107]$	4.12×10^{-8}

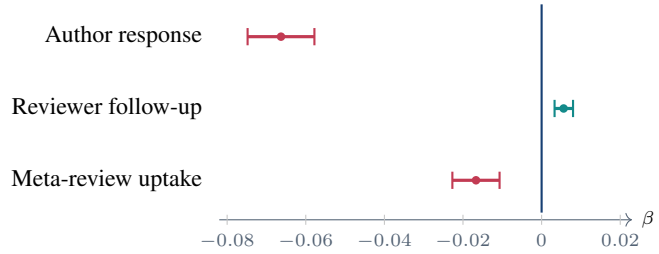

Figure 2: Within-paper fixed-effects coefficients and 95% confidence intervals. The horizontal scale is the coefficient for a full-range increase in the original 0–1 AI-likeness proxy; it is not a standardized-effect axis.

The full-range coefficients are deliberately easy to compare across outcomes, but they are not typical case contrasts: most reviews occupy a much narrower part of the score distribution. The low model R^2 values (0.003–0.012) are equally important. Style explains only a small fraction of what happens to any individual claim. The evidence supports a precise average trajectory, not a rule for predicting or judging a particular reviewer.

5.4 What two discussion threads reveal

Two high-AI-likeness threads make the difference between interaction and uptake concrete. A review of *Not How You Think, It's What You See* yielded 40 extracted claims. Authors responded extensively and the reviewer followed up at an unusually high rate, producing a busy discussion. Yet none of those claims matched the meta-review rationale. The conversation was active without becoming part of the institutional summary.

The review of *SWE-fficiency* followed a different path. It contained only six extracted claims, centered on benchmark soundness and workload coverage. Those concerns were answered, pursued, and still recognizable in the final rationale. The comparison does not establish a mechanism, but it shows why counts of comments or polished sentences are insufficient. A long checklist can generate substantial conversational work while offering the AC no compact decision reason; a smaller set of prioritized concerns can remain traceable from criticism to outcome.

6 Discussion

6.1 Review quality is a property of a process

The central result is a decoupling between a review's apparent adequacy and the persistence of its claims. High-AI-likeness reviews are not broadly more generic, unsupported, repetitive, or marginal to decision prediction. Yet their concerns follow a different average path once other participants must work with them. Authors respond less often, reviewers return slightly more often, and ACs carry less of the original criticism into the final rationale. We use *process-level discontinuity* to name this gap. It is not evidence that AI assistance causes claims to fail; it is evidence that an audit confined to the submitted review can miss a consequential form of friction.

Attentional dilution offers one plausible interpretation. Generative systems make it cheap to expand a few observations into a comprehensive-looking report. The resulting concerns can each be locally reasonable while the document as a whole lacks a hierarchy of importance. Authors working within a rebuttal limit must triage; reviewers may continue to press the points left unanswered; and ACs must compress several reports and a discussion into a short justification. Under those constraints, a claim can remain conversationally active while disappearing from the institutional record. The case comparison is consistent with this mechanism, although claim count, severity, and prioritization still need to be tested directly as mediators.

6.2 Design implications

If the problem is traceability rather than fluency, a warning that text "looks AI-generated" is poorly matched to the task. Review platforms could instead give important claims persistent identifiers, allowing authors to answer them in place and reviewers to mark them resolved, still open, or superseded. A rebuttal interface could reveal which major concerns remain uncovered without treating every sentence as equally important. When the AC writes the meta-review, lightweight provenance links could connect each decision reason to the claims and exchanges it summarizes. These features would not require the platform to know whether AI was used, and they would improve accountability for human-written reviews as well.

This reframing also changes what evaluation should reward. A useful review is not merely one that contains many specific observations. It is one whose consequential claims are prioritized clearly enough to survive limited attention and whose resolution can be reconstructed afterward. Future review assistants should therefore be evaluated on whether they help reviewers formulate a small number of answerable, decision-relevant concerns, not only on fluency, coverage, or similarity to human reports.

6.3 Limitations

Construct validity. The AI-likeness score is an unsupervised style proxy, not a validated detector or record of model use. Its orientation and composition may capture field, language background, reviewer experience, or formatting conventions. Conclusions must therefore remain about *AI-like style*, not AI authorship.

Alignment validity. TF–IDF and lexical overlap favor explicit restatement. They can miss implicit response, merged rebuttal paragraphs, and AC synthesis that abstracts away from the original wording. Threshold error may also vary with claim length or writing style. Manual validation and embedding-based sensitivity analyses are needed.

Inference. The design is observational. Paper fixed effects absorb stable paper-level properties, but do not solve reviewer assignment, within-paper role differences, or unobserved claim severity. Claims from the same review are dependent, while current inference uses HC1 rather than review-clustered standard errors. The large sample makes very small associations statistically detectable, so confidence in the direction of an average pattern should not be confused with a large practical effect.

External validity. ICLR 2026 is one venue and an exceptional year. The security incident changed discussion, review states, and AC assignment. Cross-year replication at ICLR 2024–2025 and other conferences is necessary before generalizing the lifecycle pattern.

7 Ethics and Responsible Interpretation

The corpus consists of publicly accessible OpenReview material, but public availability does not eliminate ethical obligations. The analysis is aggregate and should not be used to accuse individual reviewers of undisclosed AI use. Case studies are linked to public forums to support auditability, yet they illustrate process types rather than reviewer intent. Raw data and large text caches are excluded from the public repository; only code and compact result summaries are released. The ICLR 2026 incident further requires care: leaked identities and harassment are part of the conference context but are not reproduced or analyzed here.

8 Reproducibility Statement

The public repository is the canonical companion to this paper. It includes scripts for corpus preparation, AI-likeness scoring, text-quality analysis, lifecycle alignment, statistical modeling, and figure generation, together with compact Markdown and JSON result summaries. Raw OpenReview dumps, paper full text, and large intermediate artifacts are excluded. The manuscript source and the ICLR 2026 style dependencies are bundled with the release. The numerical claims were checked against

`outputs/results/main_findings.json`, `rq1_refined_vllm.json`, and `claim_lifecycle_findings.json`.

9 Conclusion

As AI assistance becomes ordinary in scientific writing, the consequential question is not whether reviewers can still produce polished reports. They can. The harder question is whether the resulting concerns can support a deliberative institution. In ICLR 2026, more AI-like reviews do not show a general collapse in textual quality, but their claims are less likely to be answered and less likely to survive into the meta-review, even as they prompt slightly more follow-up. That separation turns AI-assisted reviewing from a problem of prose authenticity into a problem of accountable reasons. Establishing how general the pattern is will require manual alignment audits, inference that respects review-level dependence, direct measures of claim priority, and replication across less exceptional conference years. The broader standard is already clear: whatever tools reviewers use, the concerns that shape scientific decisions should remain visible from the moment they are raised to the moment a decision is justified.

References

Jun Huang, Wei Bo Huang, Yi Bu, et al. What makes a successful rebuttal in computer science conferences? a perspective on social interaction. *Journal of Informetrics*, 17(3):101427, 2023. doi: 10.1016/j.joi.2023.101427.

ICLR 2026 Program Chairs. ICLR 2026 response to security incident. https://blog.iclr.cc/2025/12/03/iclr-2026-response-to-security-incident/, December 2025. Accessed 2026-07-20.

International Conference on Learning Representations. ICLR 2026 area chair guide. https://iclr.cc/Conferences/2026/AreaChairGuide, 2026a. Accessed 2026-07-20.

International Conference on Learning Representations. ICLR 2026 reviewer guide. https://iclr.cc/Conferences/2026/ReviewerGuide, 2026b. Accessed 2026-07-20.

Amir Hossein Kargaran, Nafiseh Nikeghbal, Jing Yang, and Nedjma Ousidhoum. Insights from the ICLR peer review and rebuttal process. *arXiv preprint arXiv:2511.15462*, 2025. URL https://arxiv.org/abs/2511.15462.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*, 2024. URL https://arxiv.org/abs/2405.02150.

Ruochi Li, Haoxuan Zhang, Edward Gehringer, Ting Xiao, Junhua Ding, and Haihua Chen. Unveiling the merits and defects of LLMs in automatic review generation for scientific papers. In *2025 IEEE International Conference on Data Mining*, pp. 1370–1379, 2025. doi: 10.1109/ICDM65498.2025.00146.

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. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2403.07183.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can LLM feedback enhance review quality? a randomized study of 20k reviews at ICLR 2025. *arXiv preprint arXiv:2504.09737*, 2025. URL https://arxiv.org/abs/2504.09737.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an LLM? benchmarking AI text detection in peer review. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HyZwf1rt4s.

A Outcome Definitions

Table 5: Lifecycle label aggregation used in the binary models.

Binary outcome	Positive labels
Author response	directly addressed; partially addressed
Reviewer follow-up	claim follow-up; weak follow-up
Meta-review uptake	explicit uptake; aggregated uptake

Claims labeled "resolved then dropped" are not counted as meta-review uptake. The label records a concern that appears to have been addressed in discussion but is absent from the final rationale. "No uptake" denotes no eligible alignment above the stage threshold.

B Illustrative OpenReview Cases

Table 6: Process cases used for qualitative interpretation. Rates are generated by the alignment pipeline and should be treated as approximate.

Type	Forum	Claims	Response	Follow-up	Uptake
Checklist-heavy	https://openreview.net/forum?id=WAQIxi7ifb	50	38.5%	1.8%	1.8%
Focused concern	https://openreview.net/forum?id=vClBDezZUo	2	100.0%	–	25.0%
Discussion-heavy, dropped	https://openreview.net/forum?id=i0zjotaTnv	40	77.8%	43.2%	0.0%
Institutionally retained	https://openreview.net/forum?id=JllgJyP4Tm	6	80.0%	28.6%	28.6%

C LLM Usage Disclosure

AI writing and coding assistants were used for organization, terminology consistency, prose revision, LaTeX authoring, and portions of plotting code. The analysis data, sample sizes, coefficients, test statistics, and effect sizes originate from the author's analysis pipeline and tracked result files. The author verified the reported evidence and remains responsible for the study design, interpretation, and final manuscript.