

Insights from the ICLR Peer Review and Rebuttal Process

Amir Hossein Kargaran^{1,3*} Nafiseh Nikeghbal^{2,3*} Jing Yang^{4,5} Nedjma Ousidhoum⁶
amir@cis.lmu.de, nafiseh.nikeghbal@tum.de

¹LMU Munich

²Technical University of Munich

³Munich Center for Machine Learning

⁴University of Southern California

⁵Paper Copilot

⁶Cardiff University

Abstract

Peer review is a cornerstone of scientific publishing, including at premier machine learning conferences such as ICLR. As submission volumes increase, understanding the nature and dynamics of the review process is crucial for improving its efficiency, effectiveness, and the quality of published papers. We present a large-scale analysis of the ICLR 2024 and 2025 peer review processes, focusing on before- and after-rebuttal scores and reviewer–author interactions. We examine review scores, author–reviewer engagement, temporal patterns in review submissions, and co-reviewer influence effects. Combining quantitative analyses with LLM-based categorization of review texts and rebuttal discussions, we identify common strengths and weaknesses for each rating group, as well as trends in rebuttal strategies that are most strongly associated with score changes. Our findings show that initial scores and the ratings of co-reviewers are the strongest predictors of score changes during the rebuttal, pointing to a degree of reviewer influence. Rebuttals play a valuable role in improving outcomes for borderline papers, where thoughtful author responses can meaningfully shift reviewer perspectives. More broadly, our study offers evidence-based insights to improve the peer review process, guiding authors on effective rebuttal strategies and helping the community design fairer and more efficient review processes. Our code and score changes data are available at github.com/papercopilot/iclr-insights.

1 Introduction

Peer review has been central to scientific publishing since 1665 (Kachooei & Ebrahimzadeh, 2022), with formal oversight beginning in 1752 (Kronick, 1990). The process has long been regarded as a cornerstone of academic integrity (Shah, 2022; Price & Flach, 2017), but faces increasing challenges due to rapid scientific growth. In computer science, for example, conference publications carry disproportionate weight compared to journals (Vrettas & Sanderson, 2015; Tomkins et al., 2017; Kim, 2019; Meho, 2019).¹ Both academic pressures, such as “publish or perish,” (Van Dalen & Henkens, 2012; Grimes et al., 2018) and technical advances, including LLM-assistance (Liang et al., 2024), have contributed to increased submission numbers.

Rebuttals, or formal responses from authors to reviewers, are a routine part of conference peer review, particularly in machine learning (ML) conferences. They allow authors to clarify misunderstandings, provide additional evidence, and support for author–reviewer engagement. While rebuttals may improve paper quality, prior research suggests that author responses generally have only a marginal impact on final scores (Gao et al., 2019). Building on these observations, and considering the additional effort required from authors,

*Equal contribution.

¹Top-tier venues such as NeurIPS, ICML, and ICLR now receive tens of thousands of submissions annually, with ICLR 2026 reviewing over 19,000 individual papers.

Table 1: Statistics of papers, reviews, and score-change records in our ICLR 2024 and 2025 corpus.

Year	#Paper	#Review	#Score Change Record
2025	11672	46748	46353
2024	7405	28028	26878

reviewers, and meta-reviewers, it is important to understand: 1) what proportion of papers are affected by rebuttal score changes, 2) when a conference should implement a rebuttal stage, 3) the association between rebuttals and score changes, 4) the association between reviewer comments and ratings, and 5) how authors can maximize the impact of their rebuttals. In this work, we investigate these questions to provide a clearer understanding of rebuttal effectiveness in peer review. To this end, we use a large-scale open dataset collected using the OpenReview API covering ICLR 2024 and 2025. We supplement this dataset with previous review scores—i.e., the original scores that were later overwritten by updated scores—to analyze how rebuttals influence score changes (see our dataset details in Table 1). We examine not only the frequency of rebuttal score changes but also temporal patterns of reviewer and author activity , as well as the relationship between textual content and the changes in scores after rebuttals. Our analysis leverages both traditional statistical methods and LLM-based techniques to identify what factors—such as evidence-backed clarification, reviewer engagement, and deviations from co-reviewers—contribute the most to effective rebuttals. Our analyses yield several consistent insights into the dynamics of ICLR peer review and rebuttals:

1. Rebuttals primarily affect borderline papers, with score changes concentrated in the mid-range (5–6), where even small shifts can change outcomes. We estimate that rebuttals influence roughly one in five top papers (based on the acceptance threshold) at ICLR through rank improvements.
2. Late reviewer submissions may correlate with slightly higher scores, though the effect is minimal. Rebuttals submitted in the middle of the rebuttal period may be the most effective, while very early or last-minute rebuttals have less impact on engagement and score changes.
3. Reviewer disagreements decrease by roughly 9–10% after rebuttals, with the strongest convergence observed for high-quality papers (oral/spotlight) and minimal effect for low-scoring or rejected papers.
4. When comparing common strengths and weaknesses across rating categories, low-rated papers tend to be criticized for writing, experimental, and methodological flaws, while high-rated papers are recognized for novelty and methodological soundness.
5. Initial scores and other reviewers’ scores strongly predict rebuttal score changes, but rebuttals that provide evidence-backed clarifications and avoid generic or vague defenses are more likely to result in positive score changes.

2 Corpus

ICLR is one of the few venues that fully publishes its peer review process (Yang, 2025; Wang et al., 2023), licensed under CC BY 4.0² and hosted on OpenReview (Soergel et al., 2013). All stages of the review process—including reviews, author responses, and final decisions—are publicly accessible, even for papers that were ultimately rejected. This transparency ensures that studies on peer review using ICLR data are not subject to biases such as survivorship (Brown et al., 1992).

Each submission is typically reviewed by three to five reviewers, who provide structured written feedback along with numerical scores. The review format includes a summary of the paper, identified strengths and weaknesses, and questions for the authors. Numerical scores are provided for soundness, presentation, contribution, confidence, and overall rating. The overall rating for ICLR 2024 and 2025 is a discrete number in {1, 3, 5, 6, 8, 10}.³ After the initial reviews are released, authors can submit a rebuttal, and reviewers are encouraged to read them, engage in discussions, and revise their assessments accordingly, before the final decisions are made. Our corpus consists of data from ICLR 2024 and 2025, covering the full review cycle. To gather this dataset, we used the OpenReview API, which provides structured access to all submission-related records, including reviews, rebuttals, decisions, and submission metadata. Although OpenReview retains

²The license metadata for each note from authors and reviewers in ICLR OpenReview is under CC BY 4.0.

³The overall rating for ICLR 2026, however, is mapped to {0, 2, 4, 6, 8, 10}.

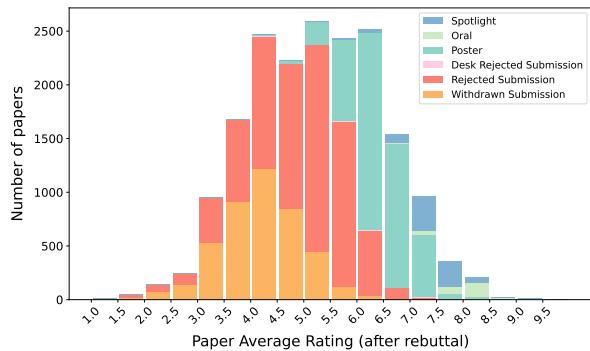


Figure 1: Distribution of submitted papers by average overall rating score and final decision.

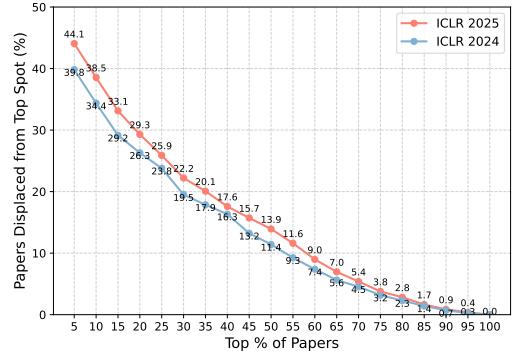


Figure 2: Percentage of papers displaced in top % threshold in 2024 and 2025 (before → after rebuttal).

archived versions, it only displays the final versions⁴, meaning that preliminary reviews submitted during the rebuttal phase are overwritten. To track score changes accurately, data must therefore be collected both immediately after the initial reviews and again after the rebuttal period ends. We archived the dataset at the time of review releases both in 2024 and 2025, containing both before- and after-rebuttal scores, enabling the analysis of how reviewer assessments change. We release the data through the Paper Copilot platform (Yang, 2025; Yang et al., 2025) and our GitHub repository.⁵ Table 1 provides basic statistics of our ICLR 2024 and 2025 corpus. The dataset includes over 19,000 papers and over 74,000 reviews. Due to technical issues at the time of review release, some papers (less than 5%) had no recorded scores before the rebuttal; therefore, the number of score change records is less than the total number of reviews. Figure 1 shows the distribution of submitted papers based on their average review ratings (after rebuttal) and their final decision outcomes.

3 Empirical Analysis of Review and Rebuttal Dynamics

Table 2: Paper- and review-level statistics for our ICLR 2024–2025 corpus, showing score changes (Δ) after rebuttal, acceptance rates (Acpt.%), and discussion metrics: author/reviewer participation percentage (Part%) and conversation turns between authors and reviewers (ConvTurn).

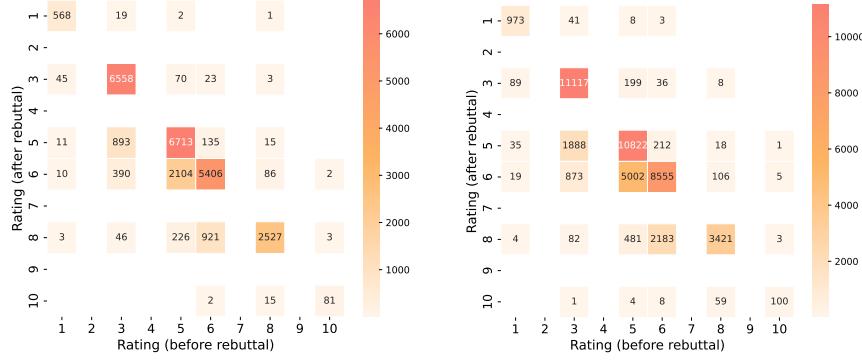
Year	Type	Per Paper			Per Review				
		#Paper	Acpt.%	Δ Rating	#Reviews	ConvTurn	AuthPart%	RevPart%	Δ Rating
2025	Increase	5807	55.7	5.21 → 5.97	10728	2.21 ± 0.83	95.65	86.88	4.64 → 6.34
	Decrease	377	8.0	4.88 → 4.41	640	2.04 ± 1.13	83.75	66.41	5.95 → 4.10
	Keep	5247	7.8	4.30 → 4.30	34985	1.47 ± 0.68	65.92	39.07	4.81 → 4.81
	Total	11431	32.1	4.78 → 5.15	46353	1.65 ± 0.79	73.05	50.51	4.78 → 5.15
2024	Increase	2930	57.6	5.31 → 6.01	4666	2.04 ± 0.75	99.79	80.73	4.58 → 6.30
	Decrease	251	6.4	4.91 → 4.44	359	1.71 ± 0.89	90.25	50.7	6.28 → 4.49
	Keep	3792	12.4	4.51 → 4.51	21853	1.35 ± 0.57	75.26	30.65	4.91 → 4.91
	Total	6973	31.2	4.86 → 5.14	26878	1.47 ± 0.67	79.72	39.61	4.86 → 5.14

3.1 Effect of Rebuttals

RQ: How many papers or reviews are affected by the rebuttal process? Table 2 shows paper- and review-level statistics for ICLR 2024 and 2025, grouped by whether the overall rating (per paper) or review score (per review) increased, decreased, or remained unchanged after rebuttal. In both years, most review scores remained unchanged (2024: 81%, 2025: 75%), followed by increases (2024: 17%, 2025: 23%), with decreases being the least frequent outcome (2024: 1%, 2025: 1%). Among papers with increased scores, 42%

⁴Except for the abstract, PDF submission, and supplementary materials, whose archived versions are also publicly available.

⁵<https://github.com/papercopilot/iclr-insights>



(a) ICLR 2024

(b) ICLR 2025

Figure 3: Change in rating scores from before (x-axis) to after (y-axis) the rebuttal.

in 2024 and 44% in 2025 were still not accepted despite improved reviewer assessments. However, papers with increased scores were far more likely to be accepted (57.6% and 55.7%) than those with unchanged (12.4% and 7.8%) or decreased (6.4% and 8.0%) scores. Most score updates occurred in borderline cases, with the most frequent changes in this order: $5 \rightarrow 6$, $6 \rightarrow 8$, and $3 \rightarrow 5$, as shown in the Figure 3. Reviews with increased scores show higher engagement, including more conversation turns and stronger author-reviewer participation than in the *keep* or *decrease* score cases, indicating that active round-trip rebuttal discussions are often correlated with positive score changes. Interestingly, decreases also occur in cases with high conversation turns, showing greater variance than in the *increase* category, suggesting that both lack of reviewer responses and active discussions can also be correlated with score reductions.

RQ: Does score change affect paper ranking (proxy for acceptance)? If all review scores shift by a constant amount, paper rankings remain unchanged, making the effect of score changes neutral. We show in Figure 2 the percentage of papers displaced from the top ranks after rebuttal-induced score changes, measured as a function of the proportion of top papers considered. The displacement is most evident among the highest-ranked papers: in both years, over 40% of papers in the top 5% before rebuttal are replaced after score updates. This may occur because authors of papers potentially assured of acceptance at the top have little incentive to pursue higher scores, whereas those with lower—but still acceptable—scores may actively seek improvements. This results in large shifts within the top-ranked. The overall patterns for 2024 and 2025 are consistent, with slightly higher displacement in 2025 across most thresholds. These results highlight that even small score updates during rebuttal can substantially affect the relative ranking of papers, particularly among top-ranked submissions where acceptance decisions are most competitive. For ICLR 2024 and 2025, the acceptance rates are around 30% (30.81%, and 31.75% respectively). As shown in Figure 2, nearly 20% of papers lost their position in the top 30% after rebuttal. This displacement drops to about half when the acceptance rate is set at 50%. This indicates that rebuttals may matter more when acceptance rates are low, whereas for conferences (or workshops) with higher acceptance rates, rebuttals may have limited impact. Note that ranking alone does not determine acceptance, as other factors—such as meta-review assessments based on paper quality, topic interest, or the quality of reviews—can also determine the fate of a paper. Here, we use ranking as a proxy for acceptance and do not consider other positive effects of rebuttals that may exist, such as improvements in quality or changes in a paper’s ultimate outcome. However, the impact of these factors are difficult to measure, as we cannot determine whether improvements in a revised paper are due to the rebuttal itself, the review, or other aspects.

3.2 Temporal Dynamics of Reviews and Rebuttals

RQ: Does Reviewer 2 exist? There is an inside joke in the ML (and other) communities about reviewer ordering; for example, Reviewer 2 (or 3) symbolizes the peer reviewer who writes vague or unhelpful reviews, assigns low scores, and refuses to budge during the rebuttal (Watling et al., 2021; Worsham et al., 2022; Jin, 2023; Kinnear et al., 2025; Tardy, 2018; Lundy & Stalford, 2022). Some studies have investigated whether such reviewers truly exist or if the perception arises solely from their assigned numbers (Peterson, 2020).

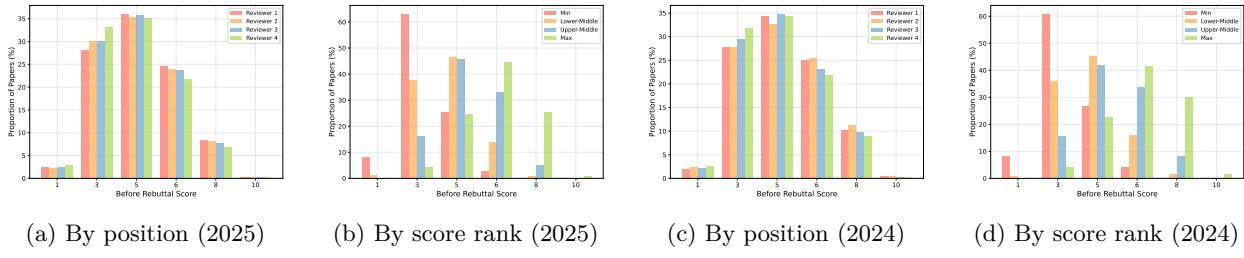


Figure 4: Distribution of before-rebuttal reviewer scores across papers, grouped either by reviewer position or score rank for ICLR 2025 and ICLR 2024.

When examining the order of review submissions based on their creation timestamps, we find that the review submitted first appears last in the public display order; consequently, the most recently submitted review is shown first to all viewers. Most of the ICLR papers had four reviews (2024: 65%, 2025: 67%). In this part, we only consider papers with four reviews and assign review numbers 1 to 4. As shown in Figure 4a, the distribution of scores, regardless of the papers they are assigned to, appears similar across reviewers. However, Reviewers 4, 3, 2, and 1, in order, tend to show increasing generosity-, which means that the reviewer who submits last (Reviewer 1) tends to submit higher scores. The result is the same for ICLR 2024 in Figure 4c. Even though the reviews in Figure 4a appear relatively similar, ordering the review scores per paper reveals a different picture in Figure 4b. Every paper consistently receives both high and low scores. Figure 4b also shows that the most frequently assigned low scores are 3, middle scores are 5, and the highest scores are 6.

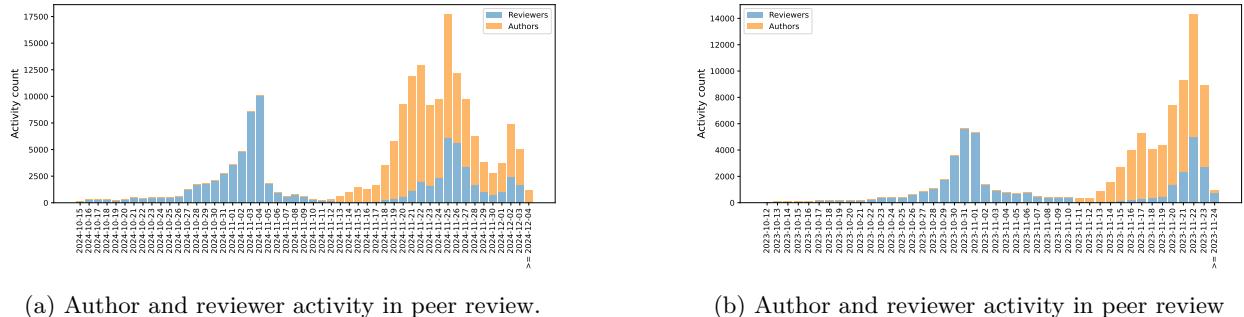
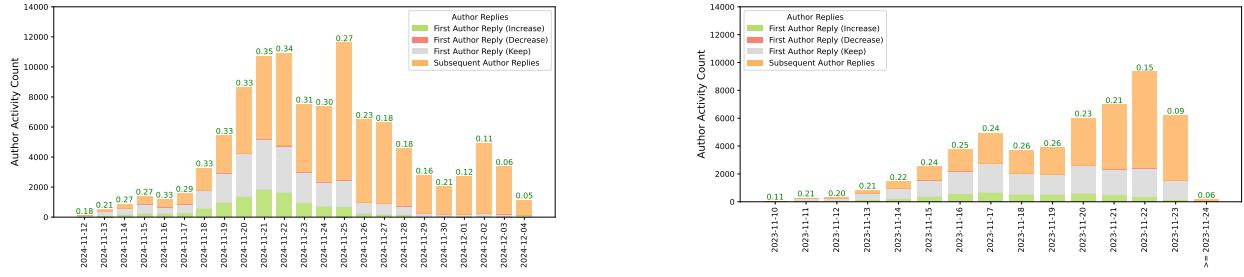


Figure 5: Author and reviewer activity in ICLR peer review (2025, left; 2024, right).

RQ: When are reviews and rebuttals typically submitted during the peer review process? Figure 5a shows the daily activity of reviewers and authors in ICLR 2025, measured by the number of submitted messages across the review and rebuttal phases. ICLR deadlines use AoE, but we use UTC here, in conformity with the default on OpenReview. During the review-writing phase (October 15 – November 4), activity is exclusively from reviewers, peaking on November 3–4, immediately before the review deadline. This indicates that reviewer activity is highly concentrated around the deadline. After the submission of reviews, activity drops for emergency reviews until the reviews are released, marking the start of the rebuttal phase on November 12. In the early days of the rebuttal period, author contributions are limited yet more prominent compared to reviewers. Reviewer and author activities rise later, peaking on November 25–26 for reviewers, near the official rebuttal deadline. Even after the rebuttal period is extended, overall activity remains low. Interestingly, a similar pattern occurred in ICLR 2024 (Figure 5b), but with a difference: reviewers were the most active on the last day, whereas in 2025 authors were more active—likely due to a message from the committee or area chairs prompting reviewer action. This pattern shows that the activities are mostly shaped by the deadlines of each group rather than by steady interaction.

RQ: When is the best time for authors to start the rebuttal? The rebuttal starts with the first message; usually, authors prepare their entire rebuttal in this initial interaction. Figure 6a shows daily author activities in ICLR 2025. We color the first messages based on whether the reviewer later changed



(a) Author activity in rebuttal.

(b) Author activity in rebuttal

Figure 6: Author activity in rebuttal (2025, left; 2024, right). The figure show author first messages (colored by whether reviewers later changed their scores) versus other messages. Numbers above bars indicate the percentage of first messages leading to score increases later.

their score. Numbers above each bar indicate the percentage of first messages that led to a later review score increase. As expected, messages submitted late—after or near the original rebuttal deadline—are less often associated with score increases. Interestingly, messages submitted very early were also less successful, which may be attributed to the rebuttal being rushed or to the papers having received very low scores. During the period November 18–24, nearly one-third of first messages later led to a review score increase. The same pattern, with some differences, also holds for ICLR 2024 (Figure 6b).

3.3 Co-Reviewers Influence

We use the term “co-reviewers influence” to refer to the incentive for reviewers to update their scores and reach a consensus. Such consensus is often explicitly encouraged by area chairs, particularly when there is high deviation among review scores and new reviewers cannot be assigned. Co-reviewers influence has also been described in other contexts as peer pressure (Gao et al., 2019), herd behavior (Banerjee, 1992), conformity bias (Buechel et al., 2015), or balance between reviewers (Huang et al., 2023); we do not differentiate between these forms. Since we do not have access to controlled experimental data (i.e., a control group of reviewers without exposure to others’ reviews), our analysis cannot fully attribute the observed convergence of scores solely to co-reviewers influence. We formalize co-reviewers influence as the extent to which pairs of scores move closer together. This can be computed per review pair or per paper across all reviewer pairs. For reviewers i and j with scores s_i and s_j , disagreement is $|s_i - s_j|$. For a paper with n reviewers, average disagreement is $\frac{1}{\binom{n}{2}} \sum_{i < j} |s_i - s_j|$.

We compute reviewer disagreement before and after rebuttals and calculate the normalized difference (after – before)/9, where 9 is the score range. A larger decrease indicates a stronger co-reviewers influence.

RQ: From when can we expect co-reviewers influence to take effect? Influence can occur as soon as reviews are released. When examining the last modification timestamp of the reviews upon release; initially, each timestamp is set to the release time. We observe that many changes—mostly increases in scores—occur within the first hour of release (see Figure 7).

RQ: Do reviewers’ scores diverge or converge? Table 3 presents the average reviewer disagreement across different submission categories and in total for ICLR 2024 and 2025. Disagreements are reported for the before and after rebuttal, with both absolute (Δ) and relative (Rel. $\Delta\%$) changes also shown. The results demonstrate a consistent reduction in disagreements after rebuttal in both years, indicating that reviewer evaluations become more balanced and aligned during this phase. The effect, however, is not uniform across categories. For spotlight and oral submissions, reductions are largest: 29% & 26% for spotlight and 41% & 48% for oral in 2024 and 2025, respectively. This indicates that potentially high-quality submissions are more strongly affected by co-reviewers influence, leading to greater reviewer convergence. Poster submissions show more moderate but consistent decreases. By contrast, rejected and withdrawn submissions show only small reductions (generally below 7%), indicating limited rebuttal influence when scores are low. Overall, average distances drop by 9–10%. This shows that rebuttals can act as a balancing step, consistently reducing

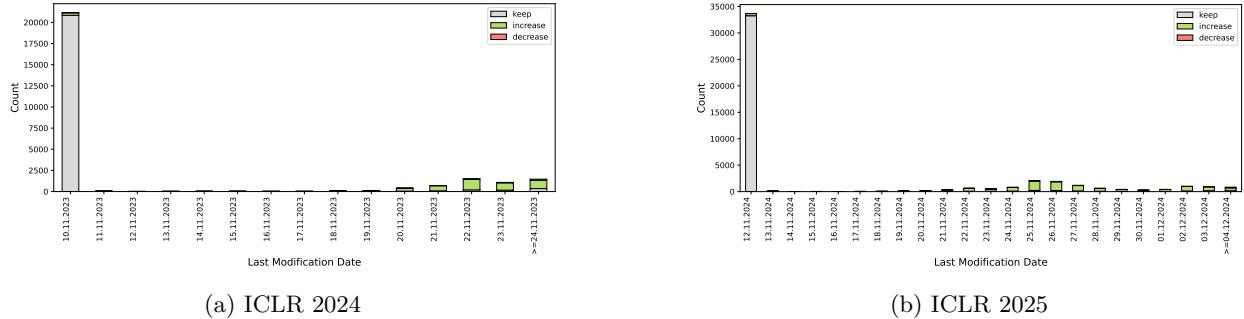


Figure 7: Last modification date of reviews, with the first day corresponding to the release day of the reviews. The colors indicate the review score changes.

Table 3: Average reviewer disagreement before and after rebuttals, by submission category and year. The table reports absolute (Δ) and relative (Rel. $\Delta\%$) changes in disagreement, with lower values after rebuttals indicating higher peer pressure.

Year	Phase	Submission Category						Total
		Spotlight	Oral	Poster	Rejected	Withdrawn	Desk Rejected	
2025	Before	0.1888	0.1715	0.1645	0.1677	0.1502	0.0199	0.1623
	After	0.1398↓	0.0889↓	0.1387↓	0.1579↓	0.1489↓	0.0166↓	0.1478↓
	Δ	-0.0490	-0.0826	-0.0258	-0.0098	-0.0013	-0.0033	-0.0145
	Rel. $\Delta\%$	-25.96	-48.16	-15.68	-5.86	-0.88	-16.49	-8.92
2024	Before	0.1899	0.1690	0.1660	0.1651	0.1480	0.0636	0.1620
	After	0.1357↓	0.0992↓	0.1383↓	0.1535↓	0.1449↓	0.0522↓	0.1456↓
	Δ	-0.0542	-0.0699	-0.0277	-0.0116	-0.0031	-0.0115	-0.0164
	Rel. $\Delta\%$	-28.54	-41.33	-16.69	-7.00	-2.09	-18.00	-10.12

reviewer divergence, with the effect most pronounced in top-tier categories, which likely would have been accepted at least as a poster even without this change.

4 LLM-Based Analysis of Review and Rebuttal Dynamics

4.1 Reviewer Evaluation Patterns

RQ: What are the most common strengths and weaknesses per paper and overall rating group? We show the common strengths and weaknesses per rating group in Appendix Figures 9-10. Our analysis reveals a clear gradient: as ratings increase, the perception of weaknesses decreases, while the recognition of strengths grows. Low-rated papers are dominated by writing flaws, such as unclear wording, and experimental flaws, such as weak baselines, whereas high-rated papers are highlighted for novelty, such as original ideas, and methodology, such as elegant and efficient models. Mid-range ratings reflect a mixed evaluation, where reviewers balance both weaknesses and strengths, often emphasizing promising contributions that are still underdeveloped.

We extract this information by prompting GPT-4o to identify strengths and weaknesses in each review. To ensure consistency, we provide GPT-4o with a structured taxonomy of high-level categories and subcategories that the model must use when labeling the text. For weaknesses, the categories cover aspects such as Novelty & Contribution (e.g., lack of originality, incremental improvement), Motivation (e.g., weak justification of the problem), Methodology & Technical Soundness (e.g., unrealistic assumptions, unclear algorithmic description), Experiments & Evaluation (e.g., insufficient baselines, missing ablations), Results (e.g., marginal gains), Data (e.g., poor data quality), Writing & Presentation (e.g., unclear wording), Broader Impact & Ethics, Related Work, and Venue Fit. The strength taxonomy mirrors this structure, including subcategories such as high novelty, strong theoretical justification, broad and realistic experimental coverage, substantial

Table 4: Top feature importances. Weakness features are marked with \times , strengths with \checkmark . The results are the mean_(std) over 10 independent runs.

Type	Feature	Avg. Coef.
\times	Novelty & Contribution	0.47(0.01)
\times	Experiments & Evaluation	0.32(0.01)
\times	Word Count	0.25(0.01)
\checkmark	Writing & Presentation	0.23(0.01)
\times	Methodology & Technical Soundness	0.23(0.01)
\times	Motivation	0.20(0.01)
\checkmark	Methodology & Technical Soundness	0.20(0.01)
\checkmark	Word Count	0.19(0.01)
\checkmark	Results	0.19(0.01)
\checkmark	Experiments & Evaluation	0.18(0.01)
\checkmark	Novelty & Contribution	0.18(0.01)

gains, high-quality data, and clear presentation. The categories were iteratively refined with input from two academic experts, based on over 100 review statements from ICLR 2024–2025 and the seed weakness categories of Gao et al. (2019) (see Appendix A.2). The full prompt templates and complete category definitions—with all subcategories—are provided in the Appendix Figures 11–12. To reduce computational load, we select a sample of 4,000 papers from both ICLR 2024 and 2025, evenly distributed across different paper outcomes during the rebuttal phase.

RQ: Which strengths and weaknesses influence the overall score the most? We show the top categories associated with strengths and weaknesses in Table 4. To compute these categories, we use word counts as a strong feature (see next RQ) and category appearance vectors (from the previous RQ) for ‘strengths’ and ‘weaknesses’ as additional features. The output labels represent the overall rating (before rebuttal) for each review. They are discrete and carry meaningful information, so we treat them as distinct classes and perform classification. For scores 1 and 3, we assign the category *Low*, and similarly assign the category *High* to scores 8 and 10 due to their low frequency of occurrence. The borderline scores 5 (Low-B) and 6 (High-B) are treated as separate classes. We use multinomial logistic regression, splitting the data into training (80%), validation (10%), and test (10%) sets, and normalize all features using a scaler fitted on the training data. We perform a grid search over class weights based on the validation set. We choose multinomial logistic regression for its interpretability as it allows us to identify important features. When using all features, the macro-f1 score is 0.49, compared to 0.43 when only using word counts. The top-ranked features, according to their average absolute coefficient (Avg. |Coef.|), are shown in Table 4. A closer look at the subcategories of the top two features (see Figure 9) shows that for ‘Novelty & Contribution’, the most frequent issues are overlap with prior work, lack of originality, and lack of clear contribution. For ‘Experiments & Evaluation’, the most common concerns are insufficient or weak baselines, too few datasets or limited domain, missing ablation tests, and reproducibility issues.

RQ: What further insights can be derived from the relationship between review scores and textual features, such as word count? Figure 8 shows the correlations between score-based criteria (SC) and word count features (WC). All non-zero correlations have p -values below 0.001. We find that the overall rating correlates most strongly with contribution, then soundness. The overall rating shows a positive correlation with the word count of the strengths section and a negative correlation with the word count of the weaknesses section. Interestingly, reviews with higher confidence scores tend to submit longer weaknesses sections and are associated with lower overall ratings. Table 5 supports these findings by breaking down average scores and word counts across submission categories. Accepted papers (oral, spotlight, and poster) generally receive higher scores and longer reviewer text in the summary, strengths, and questions sections, whereas rejected submissions (desk-rejected, withdrawn, and rejected) receive lower scores and longer weaknesses sections. This pattern suggests that reviewers provide more strengths in high-quality papers, and more extensive critical or constructive feedback for weaker papers.

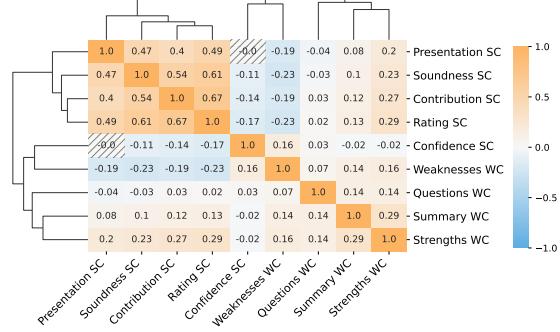


Figure 8: Correlation heatmap of after rebuttal score (SC) and word count (WC). Non-hatched cells indicate $p < 0.001$.

Table 5: Average reviewer scores and word counts by submission category. The abbreviations are: Sound. (Soundness), Present. (Presentation), Contrib. (Contribution), Confid. (Confidence), and Summ. (Summary). Bold numbers indicate the highest value in each column.

Category	Scores					Word Counts			
	Sound.	Present.	Contrib.	Rating	Confid.	Summ.	Strengths	Weaknesses	Questions
Oral	3.21	3.19	3.13	7.74	3.58	103.86	90.32	167.50	104.67
Spotlight	3.09	3.07	2.95	7.24	3.56	98.74	80.54	160.91	92.59
Poster	2.87	2.87	2.67	6.26	3.57	92.80	72.15	174.53	88.03
Desk Rejected	2.57	2.58	2.40	5.14	3.72	80.27	52.02	182.75	78.28
Rejected	2.51	2.55	2.27	4.79	3.64	86.62	60.39	205.63	88.36
Withdrawn	2.34	2.40	2.08	4.07	3.81	82.00	55.73	220.34	82.83

Table 6: Score changes in ICLR 2025 reviews across different score fields based on the type of overall score change. ‘Any’ indicates any change, while ‘Any Except Rating’ excludes changes to the overall rating.

Type	# No Score Change	# Score Change						
		Any	Rating	Confidence	Soundness	Contribution	Presentation	Any Except Rating
Decrease	0	640	640	116	114	116	77	267
Increase	0	10728	10728	717	1325	1303	935	3021
Keep	33913	1072	0	579	265	214	281	1072

4.2 Common Rebuttal Strategies and Outcomes

RQ: Which rebuttal strategies are associated with changes in reviewer scores? To answer this question, we only consider data for papers whose authors participated in the rebuttal, as a lack of author participation could result in reviewers also not participating. Here, we only consider the overall rating score change, as most reviews (71% of cases, see Table 6) reflect changes only in the overall score and do not update other aspects, such as soundness. We aim to identify which strategies, used to convince the reviewer to change their scores, are more likely to lead to success or failure. Hence, we prompt GPT-4 to annotate strategies based on the given categories for the same batch of sampled papers in §4.1. The categories were iteratively refined with input from two academic experts, based on over 100 review statements from ICLR 2024–2025 and the seed starting with the strategies suggested by Noble (2017); Kennard et al. (2022); Huang et al. (2023); Gao et al. (2019); Li et al. (2025a) (see Appendix A.2). The final categories are presented in Appendix Figure 13. We use multinomial logistic regression to predict score changes for three classes—“increase”, “decrease”, and “keep”—because it offers interpretability, allowing us to identify important features. The data is split into training (80%), validation (10%), and test (10%) sets, and features are normalized using a scaler fitted on the training data. We perform a grid search over class weights based on performance on the validation set.

We show the results of training and testing on both three-class and two-class settings—following the approach suggested by Huang et al. (2023), who merge “decrease” and “keep” into a single group—in Table 7. In the three-class setting, “keep” has the largest class size, while “decrease” has the smallest, and the macro F1-score for each class correlates with these sizes. The overall macro F1 is 0.52 for the three-class task and 0.71 for the two-class task. This prediction task is challenging because several factors, including paper quality and reviewer characteristics, are not considered in the model. We also show the top features and their coefficients, which indicate the importance of each strategy in the 3-class classification in Table 8. Some of the most important features are aspects over which the author has limited control in the rebuttal phase, such as the initial overall rating, the contribution and soundness scores, and the average of other reviewers’ scores. A notable feature is reviewer engagement, which can lead to either an increase or a decrease in scores. Although authors can respond to engagement, the effect ultimately depends on the reviewers’ willingness to participate. Other important features are more strategy-based. For example, a clearly “evasive stance” or a “generic/vague defense” strategy tends to only help reviewers maintain their original scores, while “Bare agreement/disagreement” and providing “evidence-backed clarification” can help increase the scores.

Table 7: F1-scores per class and macro F1 for 2- and 3-class classification. The results are the mean_(std) over 10 independent runs.

Setting	Class	F1-score	Macro F1	Random Macro F1
2-class (Decrease/Keep vs Increase)	Decrease/Keep	0.80 _(0.01)	0.69 _(0.02)	0.50
	Increase	0.57 _(0.03)		
3-class (Decrease, Keep, Increase)	Decrease	0.18 _(0.05)		
	Increase	0.59 _(0.02)	0.51 _(0.02)	0.33
	Keep	0.76 _(0.02)		

Table 8: Top features and coefficients for DEC, INC, and KEEP classes. For All it shows the mean absolute coefficient across classes. The results are the mean_(std) of 10 independent runs.

Feature	All (Avg. Coef.)	DEC	INC	KEEP
Overall rating score (before the rebuttal)	0.67 _(0.03)	0.92 _(0.04)	-1.01 _(0.03)	0.08 _(0.02)
Mean overall rating score of other reviewers (before the rebuttal)	0.37 _(0.02)	-0.56 _(0.03)	0.55 _(0.01)	0.01 _(0.02)
Reviewer engagement (number of notes)	0.25 _(0.02)	0.17 _(0.03)	0.20 _(0.02)	-0.37 _(0.02)
Generic/vague defense (strategy)	0.19 _(0.15)	0.03 _(0.17)	-0.29 _(0.15)	0.26 _(0.12)
Contribution score (before the rebuttal)	0.19 _(0.02)	-0.27 _(0.02)	0.28 _(0.01)	-0.01 _(0.01)
Evidence backed clarification (strategy)	0.15 _(0.10)	-0.03 _(0.12)	0.23 _(0.10)	-0.20 _(0.09)
Bare agreement/disagreement (strategy subcategory)	0.13 _(0.09)	-0.05 _(0.10)	0.19 _(0.09)	-0.14 _(0.07)
Evasion (stance subcategory)	0.10 _(0.04)	-0.06 _(0.03)	-0.10 _(0.04)	0.15 _(0.07)
Disagree (stance)	0.10 _(0.12)	0.02 _(0.05)	-0.15 _(0.13)	0.13 _(0.18)
Method details (strategy subcategory)	0.10 _(0.06)	0.01 _(0.07)	-0.15 _(0.06)	0.14 _(0.05)
Future promise (strategy subcategory)	0.10 _(0.08)	0.05 _(0.09)	0.09 _(0.08)	-0.15 _(0.07)
Soundness score (before the rebuttal)	0.10 _(0.02)	-0.12 _(0.03)	0.14 _(0.02)	-0.03 _(0.02)
Broad assertion (strategy subcategory)	0.09 _(0.08)	0.04 _(0.10)	0.09 _(0.08)	-0.14 _(0.07)
Reviewer confidence score (before the rebuttal)	0.09 _(0.02)	0.12 _(0.02)	-0.13 _(0.01)	0.01 _(0.01)
Evasion (strategy subcategory)	0.08 _(0.03)	0.12 _(0.04)	-0.06 _(0.03)	-0.07 _(0.03)

5 Related Work

The peer review process has been the focus of multiple research efforts on various NLP tasks (Lin et al., 2023a; Drori & Te’eni, 2024; Kuznetsov et al., 2024; Staudinger et al., 2024), including review and rebuttal analysis, among others (see Appendix A.1 for more).

Review Analysis. Prior work has examined the content and characteristics of reviews, including their quality and tone. For example, studies have measured review length and overall quality (Geldsetzer et al., 2023), evaluated politeness or harshness in peer reviews (Verma et al., 2022; Bharti et al., 2024), detected misinformed or deficient review points (Ryu et al., 2025; Zhang et al., 2025b), assessed the utility of reviews for authors (Sadallah et al., 2025), explored argumentative perspectives to identify disagreements between reviewers and scoring discrepancies across submission versions (Gao et al., 2019; Chakraborty et al., 2020), and analyzed reviewer confidence in specific sections or aspects of a paper to derive insights into overall review quality. Since reviewers are typically asked to provide numerical scores alongside their textual feedback, some studies have focused on predicting scores or acceptance decisions from textual features (Kang et al., 2018; Fernandes & Vaz-de Melo, 2022; 2024); for instance, Ghosal et al. (2019); Ribeiro et al. (2021) applied sentiment analysis techniques to estimate acceptance likelihood based on review language.

In our work, we similarly focus on the content of reviews, but identify common categories of strengths and weaknesses, examine their relationship with different rating groups, and use them to model scoring patterns. The interpretability of this framework further enables us to determine which review attributes most strongly influence reviewer decisions.

Rebuttal Analysis. Some peer review processes include a conversation between authors and reviewers, conducted to resolve misunderstandings, address comments, and provide additional experiments. The effects of this engagement can be reflected in the rating score and final version of a publication. Therefore, analyzing the differences between successive submission versions, as well as interactions between reviewers and authors and corresponding score changes, is relevant. Several datasets have been developed to study these

phenomena (Gao et al., 2019; Hua et al., 2019; Cheng et al., 2020; Choudhary et al., 2021; Kennard et al., 2022; Huang et al., 2023; Purkayastha et al., 2023; Ruggeri et al., 2023; Bharti et al., 2024).

Among these, the most closely related papers are by Gao et al. (2019) and Huang et al. (2023). The former introduced an open corpus from ACL 2018 and analyzed before- and after-rebuttal score changes, though their dataset may be biased since 69% of authors withheld consent to share responses. The latter examined over 3,000 ICLR 2022 papers, treating rebuttals as social interactions between authors and reviewers and identifying common rebuttal strategies, but did not release the underlying data or score changes.

In our work, we extend this line of research by analyzing a larger corpus of ICLR submissions and exploring new angles, including when rebuttals are most effective, how timing is associated with their impact, and which factors, identified through LLMs, are associated with reviewer scores and rebuttal outcomes.⁶

6 Conclusion

We aim to provide a systematic account of how rebuttals and reviewer dynamics shape outcomes in large-scale conference peer review. To this end, we analyze ICLR 2024 and 2025 review data, combining statistical methods with LLM-based categorization of review texts and rebuttal exchanges. Our study reveals that rebuttals primarily affect borderline papers, with approximately 20% of accepted submissions likely benefiting from rebuttal-driven score increases. We further find that initial reviewer scores and co-reviewer ratings are the strongest predictors of rebuttal score changes, indicating substantial peer influence. Reviewer disagreements narrow after rebuttals, especially for high-scoring paper submissions, while low-scoring ones remain largely unaffected. Textual analysis highlights that evidence-backed clarifications and precise responses are most strongly associated with positive changes, whereas vague or defensive rebuttals have little effect. These results offer practical guidance for authors in crafting effective rebuttals and shed light on the role of reviewer interactions in shaping final outcomes. More broadly, they inform program chairs seeking to design review processes that balance fairness and efficiency amid increasing submission volumes.

Limitations

- 1) Our findings should be interpreted with caution. Although our analyses reveal clear statistical patterns, the results should be interpreted as correlational and descriptive, not causal.
- 2) While the findings of this paper are specific to ICLR, they provide valuable insights tailored to this venue, which is the primary focus of our study.
- 3) Although a small subset of the data may be missing—for example, in cases where the meta-review references an author-reviewer discussion that is not available or due to a small portion of lost score change records from technical issues—the overall scale of our dataset ensures that these gaps are unlikely to meaningfully impact the findings.
- 4) We rely in part on LLM-based categorization of review and rebuttal texts to identify strengths, weaknesses, and strategies and LLM judgments remain imperfect. However, we iteratively refined the categories and prompts with expert input and achieved relatively high pairwise agreement with human annotations.
- 5) We predict review scores and rebuttal outcomes only to identify features most strongly correlated with them, not to set predictive performance benchmarks. Accordingly, we used an interpretable multinomial logistic regression model, and the results should be read as exploratory.
- 6) We primarily analyze the review and rebuttal stages, without considering the meta-review stage, which may affect borderline papers. Since meta-reviews largely depend on the quality of the reviews and the perceived expertise of the reviewers, focusing on the review and rebuttal stages allows us to more clearly isolate the dynamics that directly drive score changes. We leave the meta-review analysis for future work.

⁶Our concurrent work (Jung et al., 2025), focuses on acceptance decisions and does not include experiments on score changes, which is the main focus of our study. In contrast, we analyze score dynamics and do not examine acceptance decisions, in order to avoid potential incentives for gaming or reverse-engineering the factors that influence paper acceptance.

Reproducibility Statement

We do not publish any OpenReview data, as it is already publicly available and can be accessed at any time through the official OpenReview API to reproduce our results. We share the ICLR 2024 and 2025 score changes we obtained—which can no longer be accessed through standard means—under the same OpenReview license, CC BY 4.0. All experimental code is open source. The prompts are provided in the Appendix and will also be shared alongside the code. The exact GPT-4o model used in these experiments is gpt-4o-2024-08-06, with `top_p=0` and `temperature=0`. The total computation cost for the experiments was \$500 in OpenAI credits.

Ethics statement

This study investigates the peer review and rebuttal dynamics of ICLR 2024 and 2025 using publicly available data from OpenReview, which is licensed under CC BY 4.0. We did not collect any private or confidential data. All reviews, rebuttals, and decisions are already openly accessible as part of ICLR’s transparent peer review policy. To respect the integrity of reviewers and authors, we focus on aggregated analyses and refrain from attributing any results to specific individuals.

Although peer reviews may contain strong critiques, disagreements, or subjective opinions, our analysis treats them as research artifacts for understanding trends in review and rebuttal processes, not for evaluating individual reviewers or authors. Although we acknowledge that automated text analysis with LLMs may introduce limitations or biases in categorization, these methods were only used to identify general patterns of strengths, weaknesses, and rebuttal strategies. Human experts guided the design of annotation categories to ensure the extracted insights were meaningful and appropriate.

The purpose of this work is to provide constructive insights for authors, reviewers, and program chairs, with the overarching goal of improving fairness, efficiency, and transparency in scientific peer review. We refrain from making any causal claims, as our analyses are based on observational data and cannot disentangle causal effects from underlying confounding factors. Consequently, all reported relationships should be interpreted as correlational rather than causal. To facilitate transparency and reproducibility, we release our code, prompts, and derived score-change data under the same CC BY 4.0 license as the original OpenReview records.

References

- Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.
- Prabhat Kumar Bharti, Meith Navlakha, Mayank Agarwal, and Asif Ekbal. Politepeer: does peer review hurt? a dataset to gauge politeness intensity in the peer reviews. *Language Resources and Evaluation*, 58(4):1291–1313, 2024.
- Stephen J Brown, William Goetzmann, Roger G Ibbotson, and Stephen A Ross. Survivorship bias in performance studies. *The Review of Financial Studies*, 5(4):553–580, 1992.
- Berno Buechel, Tim Hellmann, and Stefan Klößner. Opinion dynamics and wisdom under conformity. *Journal of Economic Dynamics and Control*, 52:240–257, 2015.
- Souvik Chakraborty, Pawan Goyal, and Animesh Mukherjee. Aspect-based sentiment analysis of scientific reviews. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pp. 207–216, 2020.
- Nuo Chen, Momeng Duan, Andre Huikai Lin, Qian Wang, Jiaying Wu, and Bingsheng He. Position: The current ai conference model is unsustainable! diagnosing the crisis of centralized ai conference, 2025. URL <https://arxiv.org/abs/2508.04586>.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7000–7011, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.569. URL <https://aclanthology.org/2020.emnlp-main.569/>.

Juhwan Choi, JungMin Yun, Changhun Kim, and YoungBin Kim. Position paper: How should we responsibly adopt llms in the peer review process? *OpenReview Archive Direct Upload*, 2025. URL <https://openreview.net/forum?id=KZ3NspcpLN>.

Gautam Choudhary, Natwar Modani, and Nitish Maurya. React: A re view comment dataset for act ionability (and more). In *International Conference on Web Information Systems Engineering*, pp. 336–343. Springer, 2021.

Iddo Drori and Dov Te’eni. Human-in-the-loop ai reviewing: feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109, 2024.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. NLPeer: A unified resource for the computational study of peer review. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5049–5073, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.277. URL <https://aclanthology.org/2023.acl-long.277/>.

Gustavo Lúcius Fernandes and Pedro OS Vaz-de Melo. Between acceptance and rejection: challenges for an automatic peer review process. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pp. 1–12, 2022.

Gustavo Lúcius Fernandes and Pedro OS Vaz-de Melo. Enhancing the examination of obstacles in an automated peer review system. *International Journal on Digital Libraries*, 25(2):341–364, 2024.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1274–1290, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1129. URL <https://aclanthology.org/N19-1129/>.

Pascal Geldsetzer, Markus Heemann, Pauli Tikka, Grace Wang, Marika Mae Cusick, Ali Lenjani, and Nandita Krishnan. Prevalence of short peer reviews in 3 leading general medical journals. *JAMA Network Open*, 6(12):e2347607–e2347607, 2023.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In Anna Korhonen, David Traum, and Lluís Márquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1120–1130, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1106. URL <https://aclanthology.org/P19-1106/>.

David Robert Grimes, Chris T Bauch, and John PA Ioannidis. Modelling science trustworthiness under publish or perish pressure. *Royal Society open science*, 5(1):171511, 2018.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2131–2137, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1219. URL <https://aclanthology.org/N19-1219/>.

Junjie Huang, Win-bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427, 2023.

-
- Steven Jecmen, Minji Yoon, Vincent Conitzer, Nihar B Shah, and Fei Fang. A dataset on malicious paper bidding in peer review. In *Proceedings of the ACM Web Conference 2023*, pp. 3816–3826, 2023.
- Steven Jecmen, Nihar B Shah, Fei Fang, and Leman Akoglu. On the detection of reviewer-author collusion rings from paper bidding. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=o58uy91V2V>.
- Song Jin. How not to be the dreaded reviewer# 2, 2023.
- Sangkeun Jung, Goun Pyeon, Inbum Heo, and Hyungjin Ahn. What drives paper acceptance? a process-centric analysis of modern peer review, 2025. URL <https://arxiv.org/abs/2509.25701>.
- Amir Kachooei and Mohammad H Ebrahimzadeh. What is peer review? *The Archives of Bone and Joint Surgery*, 10(1):1–2, 2022.
- 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 Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1149. URL <https://aclanthology.org/N18-1149/>.
- Neha Nayak Kennard, Tim O’Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. DISAPERE: A dataset for discourse structure in peer review discussions. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1234–1249, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.89. URL <https://aclanthology.org/2022.naacl-main.89/>.
- Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=18QemUZaIA>.
- Jinseok Kim. Author-based analysis of conference versus journal publication in computer science. *Journal of the Association for Information Science and Technology*, 70(1):71–82, 2019.
- Benjamin Kinnear, Lynelle Govender, and Helen R Church. How to be reviewer 2: Lessons in academic curmudgeonry. *Medical Education*, 2025.
- David A. Kronick. Peer review in 18th-century scientific journalism. *JAMA*, 263(10):1321–1322, 03 1990. ISSN 0098-7484. doi: 10.1001/jama.1990.03440100021002. URL <https://doi.org/10.1001/jama.1990.03440100021002>.
- Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummersfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, and Iryna Gurevych. What can natural language processing do for peer review?, 2024. URL <https://arxiv.org/abs/2405.06563>.
- Loka Li, Ibrahim Aldarmaki, Minghao Fu, Wong Yu Kang, Yunlong Deng, Qiang Huang, Jing Yang, Jin Tian, Guangyi Chen, and Kun Zhang. How effective is your rebuttal? identifying causal models from the openreview system. In *NeurIPS 2025 Workshop on CauScienc: Uncovering Causality in Science*, 2025a. URL <https://openreview.net/forum?id=NM6Vv15FoJ>.
- Miao Li, Eduard Hovy, and Jey Lau. Summarizing multiple documents with conversational structure for meta-review generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7089–7112, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.472. URL <https://aclanthology.org/2023.findings-emnlp.472/>.

Rui Li, Jia-Chen Gu, Po-Nien Kung, Heming Xia, Junfeng Liu, Xiangwen Kong, Zhifang Sui, and Nanyun Peng. Llm-reval: Can we trust llm reviewers yet?, 2025b. URL <https://arxiv.org/abs/2510.12367>.

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, 2025c. URL <https://arxiv.org/abs/2509.19326>.

Siqing Li, Wayne Xin Zhao, Eddy Jing Yin, and Ji-Rong Wen. A neural citation count prediction model based on peer review text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4914–4924, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1497. URL <https://aclanthology.org/D19-1497/>.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Dify Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=YX7QnhxESU>.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Automated scholarly paper review: Concepts, technologies, and challenges. *Information fusion*, 98:101830, 2023a.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Moprd: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35(34):24191–24206, 2023b.

Ryan Liu and Nihar B. Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing, 2023. URL <https://arxiv.org/abs/2306.00622>.

Laura Lundy and Helen Stalford. In praise of reviewer 2. *The International Journal of Children's Rights*, 30(3):615–616, 2022.

Lokman I Meho. Using scopus's citescore for assessing the quality of computer science conferences. *Journal of Informetrics*, 13(1):419–433, 2019.

William Stafford Noble. Ten simple rules for writing a response to reviewers, 2017.

David AM Peterson. Dear reviewer 2: Go f'yourself. *Social Science Quarterly*, 101(4):1648–1652, 2020.

Barbara Plank and Reinard van Dalen. Citetracked: A longitudinal dataset of peer reviews and citations. In *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pp. 116–122. CEUR Workshop Proceedings (CEUR-WS. org), 2019.

Simon Price and Peter A Flach. Computational support for academic peer review: a perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79, 2017.

Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14479–14495, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.894. URL <https://aclanthology.org/2023.emnlp-main.894/>.

Ana Carolina Ribeiro, Amanda Sizo, Henrique Lopes Cardoso, and Luís Paulo Reis. Acceptance decision prediction in peer-review through sentiment analysis. In *EPIA Conference on Artificial Intelligence*, pp. 766–777. Springer, 2021.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. A dataset of argumentative dialogues on scientific papers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7684–7699, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.425. URL <https://aclanthology.org/2023.acl-long.425/>.

Hyun Ryu, Doohyuk Jang, Hyemin S. Lee, Joonhyun Jeong, Gyeongman Kim, Donghyeon Cho, Gyouk Chu, Minyeong Hwang, Hyeongwon Jang, Changhun Kim, Haechan Kim, Jina Kim, Joowon Kim, Yoonjeon Kim, Kwanhyung Lee, Chanjae Park, Heecheol Yun, Gregor Betz, and Eunho Yang. Reviewscore: Misinformed peer review detection with large language models, 2025. URL <https://arxiv.org/abs/2509.21679>.

Abdelrahman Sadallah, Tim Baumgärtner, Iryna Gurevych, and Ted Briscoe. The good, the bad and the constructive: Automatically measuring peer review’s utility for authors, 2025. URL <https://arxiv.org/abs/2509.04484>.

Rylan Schaeffer, Joshua Kazdan, Yegor Denisov-Blanch, Brando Miranda, Matthias Gerstgrasser, Susan Zhang, Andreas Haupt, Isha Gupta, Elyas Obbad, Jesse Dodge, Jessica Zosa Forde, Francesco Orabona, Sanmi Koyejo, and David Donoho. Position: Machine learning conferences should establish a "refutations and critiques" track, 2025. URL <https://arxiv.org/abs/2506.19882>.

Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. MReD: A meta-review dataset for structure-controllable text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2521–2535, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.198. URL [https://aclanthology.org/2022.findings-acl.198/](https://aclanthology.org/2022.findings-acl.198).

David Soergel, Adam Saunders, and Andrew McCallum. Open scholarship and peer review: a time for experimentation. *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. An analysis of tasks and datasets in peer reviewing. In Tirthankar Ghosal, Amanpreet Singh, Anita Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin (eds.), *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pp. 257–268, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.sdp-1.24/>.

Ivan Stelmakh, Nihar B Shah, and Aarti Singh. Catch me if i can: Detecting strategic behaviour in peer assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4794–4802, 2021.

Ivan Stelmakh, John Frederick Wieting, Yang Xi, Graham Neubig, and Nihar B Shah. A gold standard dataset for the reviewer assignment problem. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=XofMH05yVY>.

Hao Sun, Yunyi Shen, and Mihaela van der Schaar. Openreview should be protected and leveraged as a community asset for research in the era of large language models, 2025. URL <https://arxiv.org/abs/2505.21537>.

Lu Sun, Stone Tao, Junjie Hu, and Steven P Dow. Metawriter: Exploring the potential and perils of ai writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction*, 8 (CSCW1):1–32, 2024.

Jaroslaw Szumega, Lamine Bougueroua, Blerina Gkotse, Pierre Jouvelot, and Federico Ravotti. The open review-based (orb) dataset: Towards automatic assessment of scientific papers and experiment proposals in high-energy physics, 2023. URL <https://arxiv.org/abs/2312.04576>.

Christine M Tardy. We are all reviewer# 2: A window into the secret world of peer review. In *Novice writers and scholarly publication: Authors, mentors, gatekeepers*, pp. 271–289. Springer, 2018.

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, 2025. URL <https://arxiv.org/abs/2504.09737>.

Andrew Tomkins, Min Zhang, and William D Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

Hendrik P Van Dalen and Kène Henkens. Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology*, 63(7):1282–1293, 2012.

Rajeev Verma, Rajarshi Roychoudhury, and Tirthankar Ghosal. The lack of theory is painful: Modeling harshness in peer review comments. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 925–935, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.67. URL <https://aclanthology.org/2022.aacl-main.67/>.

George Vrettas and Mark Sanderson. Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology*, 66(12):2674–2684, 2015.

Gang Wang, Qi Peng, Yanfeng Zhang, and Mingyang Zhang. What have we learned from openreview? *World Wide Web*, 26(2):683–708, 2023.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In Brian Davis, Yvette Graham, John Kelleher, and Yaji Sripada (eds.), *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 384–397, Dublin, Ireland, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.inlg-1.44. URL <https://aclanthology.org/2020.inlg-1.44/>.

Chris Watling, Shiphra Ginsburg, and Lorelei Lingard. Don’t be reviewer 2! reflections on writing effective peer review comments. *Perspectives on Medical Education*, 10(5):299–303, 2021.

Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. The ai imperative: Scaling high-quality peer review in machine learning, 2025. URL <https://arxiv.org/abs/2506.08134>.

Christopher Worsham, Jaemin Woo, André Zimmerman, Charles F Bray, and Anupam B Jena. An empirical assessment of reviewer 2. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 59:00469580221090393, 2022.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2189–2198, 2022.

Jing Yang. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=gnyqRarPzW>.

Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper copilot: Tracking the evolution of peer review in ai conferences, 2025. URL <https://arxiv.org/abs/2510.13201>.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. Meta-review generation with checklist-guided iterative introspection. *Computation and Language Repository*, 2023.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. Re²: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions, 2025a. URL <https://arxiv.org/abs/2505.07920>.

Haoxuan Zhang, Ruochi Li, Sarthak Shrestha, Shree Harshini Mamidala, Revanth Putta, Arka Krishan Aggarwal, Ting Xiao, Junhua Ding, and Haihua Chen. Reviewguard: Enhancing deficient peer review detection via llm-driven data augmentation, 2025b. URL <https://arxiv.org/abs/2510.16549>.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *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/>.

Changjia Zhu, Junjie Xiong, Renkai Ma, Zhicong Lu, Yao Liu, and Lingyao Li. When your reviewer is an llm: Biases, divergence, and prompt injection risks in peer review, 2025. URL <https://arxiv.org/abs/2509.09912>.

A Appendix

A.1 Additional Related Work

In NLP research on peer review, beyond analyzing reviews and rebuttals, several other main tasks have been studied, including reviewer assignment (Stelmakh et al., 2021; Jecmen et al., 2023; 2025; Stelmakh et al., 2025), review generation (Wang et al., 2020; Yuan et al., 2022; Liu & Shah, 2023; Szumega et al., 2023; Zhou et al., 2024; Zhang et al., 2025a), meta-review generation (Shen et al., 2022; Wu et al., 2022; Li et al., 2023; Lin et al., 2023b; Zeng et al., 2023; Sun et al., 2024), guided skimming (Dycke et al., 2023), and citation prediction (Li et al., 2019; Plank & van Dalen, 2019). More recently, new tasks have emerged, including the use of LLMs as tools for reviewing, which raises concerns about quality and fairness (Choi et al., 2025; Sun et al., 2025; Zhu et al., 2025; Thakkar et al., 2025; Wei et al., 2025; Li et al., 2025c;b). These developments have sparked debates around the so-called “AI conference peer review crisis” (Kim et al., 2025; Chen et al., 2025), highlighting the need to rethink peer review management in computer science, including proposals to establish dedicated tracks for refutations and critiques in ML conferences (Schaeffer et al., 2025).

A.2 Design of Prompts

Weakness/Strength Categories. We adopt weakness categories from Gao et al. (2019) as a starting point and use them, to also define corresponding strength categories, in order to classify both weaknesses and strengths. To refine the categories, two academic experts (authors of this paper; compensated according to their employment contract) independently reviewed 20 papers selected from different decision outcomes (reject, oral, spotlight, etc.) and annotated the weaknesses and strengths in their reviews. Weaknesses and strengths were either assigned to existing categories or, when necessary, placed into newly proposed categories or subcategories. After this initial round, the authors met to compare results and reached consensus on a preliminary taxonomy. The goals of this meeting were (a) to merge overlapping categories and (b) to validate the newly proposed categories. To test the stability of the taxonomy, the authors then reviewed an additional 10 papers (later also used as a validation set) and compared their results. Since no new categories or subcategories emerged, the taxonomy was finalized.

Based on the outcomes of these two rounds, we established a stable set of categories and subcategories. Using this refined taxonomy, we designed prompts for GPT-4o to classify reviews at scale. We experimented with different prompt designs on the validation set, enabling comparison between GPT-4o outputs and our manual annotations. We then selected the prompt that achieved the highest alignment with the manual labels (84% pairwise human agreement) and applied it to a larger set of review statements (see the weakness prompt in Figure 11 and the strength prompt in Figure 12).

Figure 9 and Figure 10 present the main weaknesses and strengths highlighted by reviewers at ICLR 2024 and 2025, organized by paper rating scores before rebuttal. For each rating group (1, 3, 5, 6, 8, and 10), the most common categories of criticism and praise are identified, along with their top three subcategories. Low-rated papers are often criticized for unclear writing, weak baselines, limited datasets, poor unclear algorithmic description, while high-rated papers are recognized for novelty, methodological soundness, efficient model or approach, and clear presentation.

Strategy Categories. We adopt strategy categories from Noble (2017); Gao et al. (2019); Kennard et al. (2022); Huang et al. (2023); Li et al. (2025a) as a starting point. Like the weakness and strength category design, the same two academic experts independently selected 10 papers from each category (increase, decrease, keep) to develop more fine-grained strategies by defining subcategories. They further discussed these strategies in a meeting and finalized a stable set of categories and subcategories. Using this refined taxonomy, we designed prompts for GPT-4o to classify reviews at scale. We validated different prompt designs against manual annotations of additional 10 papers and selected the one with the highest alignment (81% pairwise human agreement). Our taxonomy consists of three main fields: *coverage*, *stance*, and *strategy*. Coverage indicates whether the author addresses a reviewer’s concern. Stance reflects the author’s position—agreement or disagreement with the reviewer’s point—thereby revealing alignment or conflict. Strategy characterizes how the response is formulated, distinguishing evidence-backed clarifications from vague or generic defenses, which allows us to assess the substance and persuasiveness of rebuttals. Each of these fields contains a set of subcategories, which are detailed in the Figure 13.

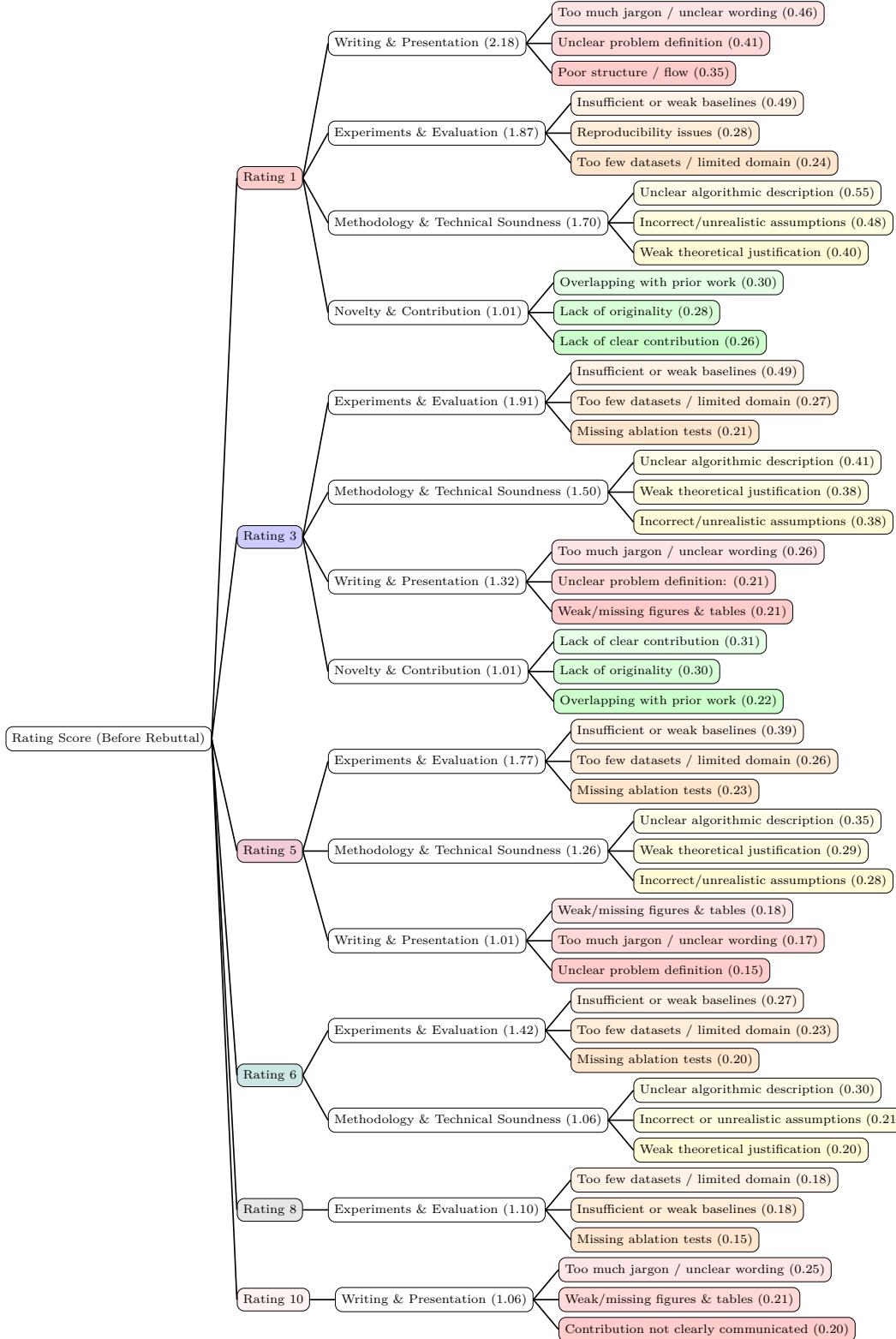


Figure 9: Taxonomy of weaknesses raised by reviewers at ICLR 2024 and 2025, categorized by rating scores. Values in parentheses represent the number of mentions per review. We report only categories with values above 1.0 and, for each, select the top three subcategories.

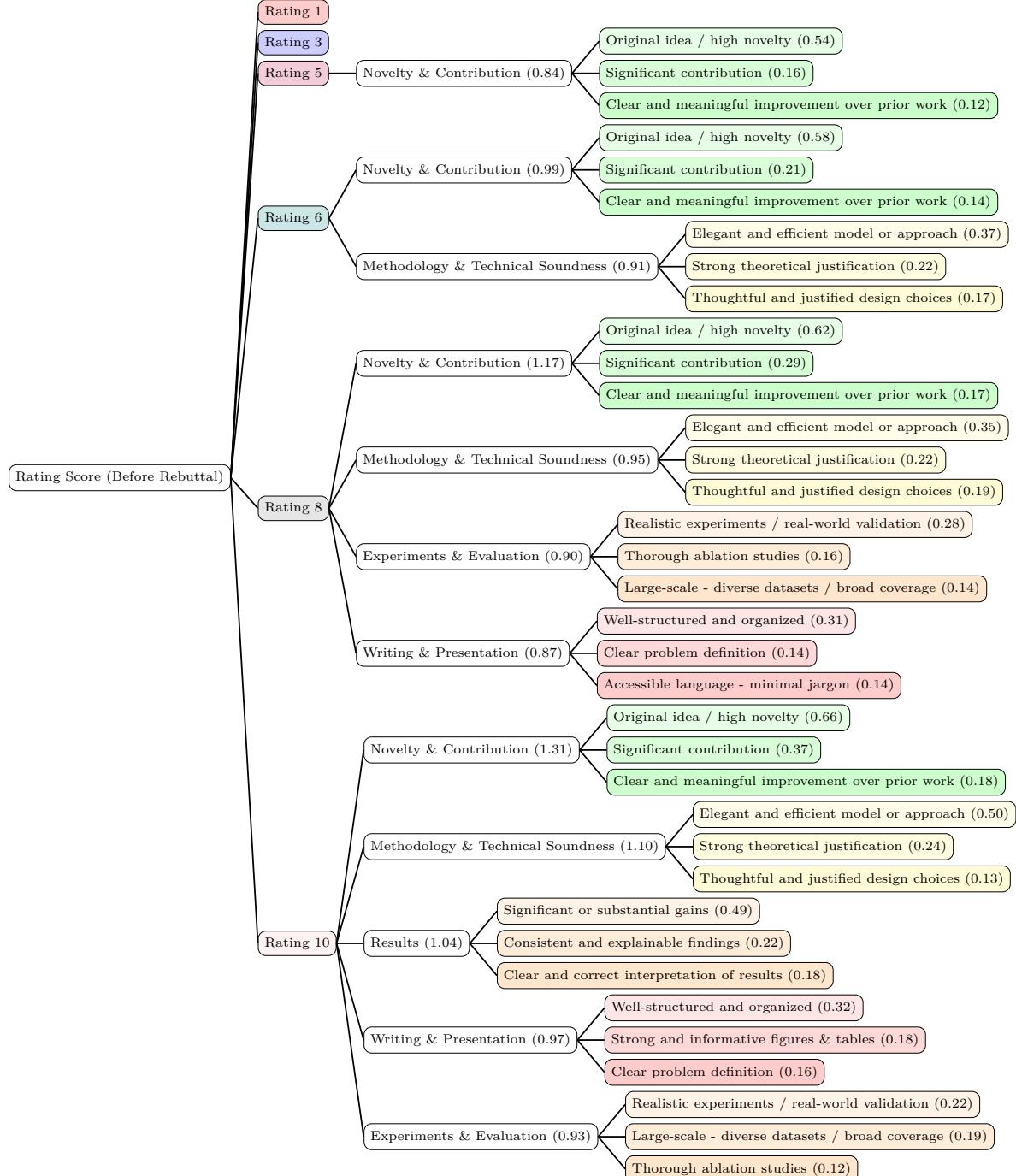


Figure 10: Taxonomy of strengths raised by reviewers at ICLR 2024 and 2025, categorized by rating scores. Values in parentheses represent the number of mentions per review. We report only categories with values above 0.8 and, for each, select the top three subcategories.

Weakness Annotation Prompt

Task Description: You are an annotator. Your task is to analyze the weaknesses in a computer science paper review. For each paragraph in the text, produce a list of annotations. For each annotation:

1. Assign one of the main categories from the list below. You must not use “Other” unless it is absolutely impossible to map the weakness to any category. Even a very weak or partial connection (as little as 1%) is enough to assign an existing category.
2. Assign a subcategory: use one of the subcategories defined under the chosen main category. You must always choose a defined subcategory if there is any possible relation at all, no matter how small. Only if it is truly impossible to relate the weakness to any defined subcategory, then use “Other”.
3. Extract the exact text span that expresses the weakness.
4. If multiple weaknesses exist in one paragraph, annotate each separately in annotation list.

Always prioritize using the predefined categories and subcategories over “Other”.

Main categories and their subcategories:

1. **Novelty & Contribution:** Lack of originality; Incremental improvement; Lack of clear contribution; Overclaiming novelty; Overlapping with prior work; Work not mature enough for publication; Other.
2. **Motivation:** Weak or missing motivation; Problem not well justified as important; No clear real-world or theoretical relevance; Other.
3. **Methodology & Technical Soundness:** Weak theoretical justification; Incorrect or unrealistic assumptions; Overly complicated model; Cherry-picked design choices; Unclear algorithmic description; Scalability or computational impracticality; Lack of consideration of established improvement techniques; Other.
4. **Experiments & Evaluation:** Insufficient or weak baselines; Too few datasets / limited domain or language coverage; Small-scale experiments / lack of real-world validation; Poor generalizability across settings; Missing error or failure analysis; Missing ablation tests; Lack of statistical significance tests; No human evaluation when needed; Evaluation metrics not well justified; Unfair comparisons (own method tuned, others not); Reproducibility issues (missing details, code not available); Other.
5. **Results:** Marginal gains / not significant; Overinterpretation of results; Contradictory or unexplained findings; Weak interpretability / lack of explanation; No qualitative examples or case studies; Other.
6. **Data:** Dataset not publicly available; Poor data quality / noise not addressed; Missing dataset documentation or statistics; Missing inter-annotator agreement scores; Synthetic or toy datasets only; Other.
7. **Writing & Presentation:** Unclear problem definition; Poor structure / flow; Too much jargon / unclear wording; Weak or missing figures & tables; Inconsistent terminology; Grammar, typos, formatting issues; Contribution not clearly communicated; Other.
8. **Broader Impact, Ethics & Relevance:** No discussion of societal risks or ethical implications; Unrealistic claims of impact; Ignoring potential biases, fairness, or safety issues; Other.
9. **References & Related Work:** Missing related work; Missing important prior work; Outdated references; Other.
10. **Fit & Scope for Venue:** Mismatch with venue; Other.
11. **Other:** Other.

Output Format: Return the annotations as a JSON list, where each element corresponds to a paragraph. Put the JSON list in the <START_LIST> <END_LIST>.

```
{  
  "paragraph": "<original paragraph>",  
  "annotation_list": [  
    {  
      "main_category": "<one of the 11 main categories>",  
      "subcategory": "<one of the subcategories for the selected main category>",  
      "source": "<exact phrase/sentence/paragraph that expresses the weakness>"  
    }  
  ]  
}
```

Figure 11: Weakness annotation prompt

Strength Annotation Prompt

Task Description: You are an annotator. Your task is to analyze the strengths in a computer science paper review. For each paragraph in the text, produce a list of annotations. For each annotation:

1. Assign one of the main categories from the list below. You must not use “Other” unless it is absolutely impossible to map the strength to any category. Even a very weak or partial connection (as little as 1%) is enough to assign an existing category.
2. Assign a subcategory: use one of the subcategories defined under the chosen main category. You must always choose a defined subcategory if there is any possible relation at all, no matter how small. Only if it is truly impossible to relate the strength to any defined subcategory, then use “Other”.
3. Extract the exact text span that expresses the strength.
4. If multiple strengths exist in one paragraph, annotate each separately in annotation list.

Always prioritize using the predefined categories and subcategories over “Other”.

Main categories and their subcategories:

1. **Novelty & Contribution:** Original idea / high novelty; Significant contribution; Clear and meaningful improvement over prior work; Strong theoretical or practical impact; Well-scoped and mature work; Other.
2. **Motivation:** Well-motivated problem; Clearly important problem; Strong real-world or theoretical relevance; Other.
3. **Methodology & Technical Soundness:** Strong theoretical justification; Realistic and valid assumptions; Elegant and efficient model or approach; Thoughtful and justified design choices; Clear and reproducible algorithmic description; Scalable and computationally practical; Builds on established techniques effectively; Other.
4. **Experiments & Evaluation:** Strong baselines; Large-scale, diverse datasets / broad coverage; Realistic experiments / real-world validation; Good generalizability across settings; Detailed error or failure analysis; Thorough ablation studies; Statistical significance tested; Human evaluation included when needed; Well-justified evaluation metrics; Fair comparisons; Reproducible (code and details provided); Other.
5. **Results:** Significant or substantial gains; Clear and correct interpretation of results; Consistent and explainable findings; Strong interpretability / well-explained results; Includes qualitative examples or case studies; Other.
6. **Data:** High-quality datasets; Publicly available data; Detailed dataset documentation and statistics; Inter-annotator agreement provided; Realistic datasets (not toy); Other.
7. **Writing & Presentation:** Clear problem definition; Well-structured and organized; Accessible language, minimal jargon; Strong and informative figures & tables; Consistent terminology; Grammatically correct, well-formatted; Contribution clearly communicated; Other.
8. **Broader Impact, Ethics & Relevance:** Thoughtful discussion of societal risks or ethical implications; Realistic and meaningful claims of impact; Awareness of biases, fairness, or safety issues; Other.
9. **References & Related Work:** Comprehensive related work; Covers important prior work; Up-to-date references; Other.
10. **Fit & Scope for Venue:** Strong fit with venue; Other.
11. **Other:** Other.

Output Format: Return the annotations as a JSON list, where each element corresponds to a paragraph. Put the JSON list in the <START_LIST> <END_LIST>.

```
{  
  "paragraph": "<original paragraph>",  
  "annotation_list": [  
    {  
      "main_category": "<one of the 11 main categories>",  
      "subcategory": "<one of the subcategories for the selected main category>",  
      "source": "<exact phrase/sentence/paragraph that expresses the strength>"  
    }  
  ]  
}
```

Figure 12: Strength annotation prompt

Reviewer–Author Stance & Strategy Annotation Prompt

Task Description: You are an annotator.

You are given a multi-turn conversation between reviewers and authors, starting with the reviewer. The reviewer message contains a set of weaknesses and questions. Your task is to analyze each reviewer weakness and question and the corresponding author response(s).

For each reviewer weakness/question:

1. Determine if the author answered it, and record this in the `coverage` field. If not answered, set `coverage` to `Not Answered` and leave all other fields empty.
2. Identify the author's stance towards the reviewer point in the `stance` field (`Disagree` or `Agree`).

Agree/Disagree stance subcategories:

- `reject_validity`: Reject the validity of the question or weakness
- `evasion`: Mitigate the importance of the question or weakness
- `reject_request`: Reject a request from the reviewer
- `contradict_statement`: Contradict a statement presented as a fact in the question or weakness
- `completion_claim`: Claim that a requested task has been completed
- `concede_point`: Concede the validity of a weakness or question
- `Promise`: Promise a change by camera-ready deadline
- `future_work`: Express approval for a suggestion, but for future work

3. Determine the author's strategy in answering (ignoring stance):

evidence_backed_clarification subcategories:

- `method_details`: gives technical clarifications (formulas, hyperparameters, algorithm steps)
- `new_table_figures`: provides new tables or figures
- `analysis`: provides deeper breakdowns or error analysis
- `experiments_in_paper`: refers to figures, tables, ablations, significance tests, baselines, or human evaluations in the original paper
- `new_experiments`: provides new ablations, tests, baselines, or evaluations done to answer the reviewer in rebuttal
- `citation`: references prior work or other papers
- `other`: evidence-backed but none of the above

generic_defense_vague subcategories:

- `repetition`: repeats earlier claims without new support
- `broad_assertion`: general phrases like "our method is strong" without details
- `evasion`: avoids addressing the reviewer's concern directly
- `future_promise`: vague improvements promised without specifics
- `bare_agreement_or_disagreement`: only says "we agree" or "we disagree" without justification
- `other`: vague or defensive but not fitting above

Output Format: Return the annotations as a JSON list inside `<START_LIST>` `<END_LIST>` tags. Each element corresponds to a reviewer weakness/question:

```
{  
  "reviewer_point": "<exact reviewer weakness/question>",  
  "author_response": "<summary or quote of author response>",  
  "coverage": "Answered / Not Answered",  
  "stance": "<Disagree/Agree>",  
  "stance_subcategory": "<stance subcategory>",  
  "strategy": "<evidence_backed_clarification/generic_defense_vague>",  
  "strategy_subcategory": "<strategy subcategory>"  
}
```

Figure 13: Reviewer–Author dialogue annotation prompt