Judging a Review by its Cover: A Reliability Analysis of LLM-based Peer Review Evaluation Metrics

Abstract

Peer-review evaluation is increasingly being automated with LLM-as-a-judge metrics, but this creates a measurement risk. A review may receive a high score because it is fluent, organized, and polished, rather than because it provides a strong evaluation of the paper. This risk is especially important in AI-assisted reviewing, where reviewers may use LLMs to improve clarity or presentation while preserving the underlying judgments. We propose a statistical framework for testing whether peer-review evaluation metrics capture substantive review quality beyond surface-level linguistic form. The framework compares original human reviews with faithful LLM rewrites that preserve the same evaluative content while changing wording and presentation. Using 4,044 meaning-preserving rewrites derived from 674 human reviews from ICLR and NeurIPS, we evaluate 29 content-oriented peer-review evaluation metrics drawn from four prior works through complementary tests of *surface sensitivity* and *robustness*. Although these metrics are intended to capture review properties beyond surface-level, writing-dependent characteristics, we find that sensitivity to rewriting is widespread. Twenty-three metrics assign significantly different scores to reviews whose evaluative content is preserved, while only seven satisfy our robustness criterion. The patterns are consistent across two LLM judge models, suggesting that the issue is not specific to a single judge. These findings show that many peer-review evaluation metrics partially conflate review quality with linguistic presentation, and indicate that robustness to meaning-preserving rewriting should be validated before such metrics are used to compare human-written, AI-assisted, and AI-generated reviews. All evaluation results, prompts, and implementation code are publicly available on our GitHub repository at https://anonymous.4open.science/r/judging_reviews_by_cover-BB10/

1 Introduction

Peer review plays a central role in scientific publishing, shaping publication decisions, funding outcomes, and academic careers [8, 9, 23]. Yet this process is increasingly strained by growing submission volumes, which place substantial pressure on reviewers, editors, and conference organizers. At the same time, large language models (LLMs) are beginning to change how reviews are written. LLMs are now used both to assist reviewers and to generate review text [12–14]. This shift creates a parallel need for scalable methods that can evaluate peer-review quality without relying exclusively on costly human assessment.

Peer-review-specific LLM-as-a-judge metrics have become a prominent approach for automated peer review evaluation, scoring reviews along dimensions such as analytical depth [6], structured review component coverage [22], holistic review quality [4], and human-anchored comparative judgment [7], with recent benchmarks such as PRISM [15] and PRAIB [5] further expanding evaluation criteria. These developments make peer review evaluation more scalable, but they also raise a basic measurement question. A review can be well written without being a high-quality review, fluent, organized, and polished, while offering shallow analysis, weak evidence, or unhelpful recommendations, or containing a substantive and useful evaluation while being less polished in presentation. For LLM-as-a-judge metrics to support reliable peer review evaluation, they must distinguish the *quality of the review as an evaluation* from the *quality of the review as a piece of writing* [21]. Scores for substantive properties such as actionability, constructiveness, and depth of technical analysis should not change merely because the same review is expressed in a more polished style.

This paper studies a central reliability problem in LLM-based peer-review evaluation. We ask whether metrics intended to measure substantive review quality instead reward surface form. This is especially important in AI-assisted reviewing, where an LLM may improve clarity, fluency, or organization while preserving the reviewer's underlying judgments. If a metric assigns different scores to the original and rewritten versions of the same review, it may be responding to linguistic presentation rather than evaluative content, making polished reviews appear better even when their substantive assessment has not changed.

The concern is consistent with broader findings on LLM-based evaluation, including prompt sensitivity, position bias, judge-model disagreement, and preferences for particular writing styles [17, 19, 24, 26], and connects to work on language invariance, where meaning-preserving transformations should not change properties intended to remain constant under the surface variation [1]. Similar questions have been studied in machine translation and summarization, where automatic metrics are expected to remain reliable under meaning-preserving transformations [10, 16]. Despite the growing use of LLM-as-a-judge metrics for peer-review evaluation, their reliability has not been systematically examined to date.

We study this problem from two distinct complementary perspectives. *Surface sensitivity* asks whether meaning-preserving rewriting induces a statistically detectable change in metric scores, identifying metrics that systematically respond to changes in wording or presentation despite preserved evaluative content. *Robustness* asks whether any score change is small enough to be practically negligible [11]. These two perspectives capture different forms of metric reliability. *Surface sensitivity* identifies whether rewriting changes a metric's scores in a systematic way, whereas *robustness* determines whether the size of that change is small enough to be acceptable for practical evaluation. A metric can therefore be sensitive without being practically unreliable, if the shift is very small, and it can also appear non-sensitive without being demonstrably robust, if equivalence has not been established. Considering both perspectives allows us to distinguish metrics that merely show no detectable shift from metrics that provide positive evidence of stability under meaning-preserving rewriting. We operationalize our work using 4,044 LLM-generated rewrites derived from 674 human reviews from ICLR and NeurIPS submissions from 2021 to 2024 from the PeerPrism dataset [18]. The rewritten reviews are designed to preserve the original review's judgments, arguments,

and recommendations while altering wording and presentation. We evaluate 29 content-oriented metrics drawn from four prior works, ReviewEval [6], REMOR [22], RottenReviews [4], and ScholarPeer [7]. We assess *surface sensitivity* using paired significance testing and *robustness* using equivalence testing, and we replicate the analysis with two different SOTA LLMs as judge models.

Among the 29 examined metrics, 23 assign significantly different scores to rewritten reviews whose evaluative content remains unchanged, while only 7 satisfy our robustness criterion. The finding is largely consistent across judge models, with 24 of 29 metrics receiving the same classification under both LLMs. The results suggest that many peer-review evaluation metrics partially conflate substantive review quality with linguistic presentation. Our findings suggest that robustness to meaning-preserving rewriting should be validated before LLM-as-a-judge metrics are used to compare human, AI-assisted, and AI-generated reviews.

More concretely, this paper makes the following contributions: **(1)** We formulate *surface sensitivity* and *robustness* as distinct but complementary criteria for evaluating peer-review metrics, separating statistically detectable score shifts from practically meaningful instability under meaning-preserving rewriting; **(2)** We present a systematic analysis of *surface sensitivity* and *robustness* in LLM-as-a-judge metrics for peer-review evaluation, covering 29 content-oriented metrics drawn from four prior works and 4,044 meaning-preserving rewrites of 674 human reviews from ICLR and NeurIPS; and **(3)** We show that sensitivity to surface-level rewriting is pervasive while robustness is comparatively rare, with results that are largely consistent across two different LLMs, highlighting the need to validate metric stability before drawing conclusions about human and AI-assisted review quality.

2 Analytical Framework

Our objective is to evaluate whether peer review evaluation metrics measure the substantive assessment expressed in a review or also respond to its surface form. We consider pairs of reviews in which an original and its meaning-preserving rewrite express the same evaluative content but differ in wording, phrasing, organization, or style. A reliable content-oriented metric should remain stable under such transformations, since the quality of the underlying review as an evaluation has not changed.

Let R denote an original peer review and R' a meaning-preserving rewrite that preserves the evaluative content of R while modifying only its surface realization. A metric designed to measure properties such as analytical depth, constructiveness, or aspect coverage should assign similar scores to R and R', since the underlying assessment has not changed.

This setting motivates two complementary notions of metric reliability. *Surface sensitivity* asks whether meaning-preserving rewriting induces a systematic change in metric scores. *Robustness* asks whether any such change is small enough to be practically negligible. Surface sensitivity identifies systematic score shifts, whereas robustness provides positive evidence that their magnitude remains within an acceptable range.

For each metric m and paired instance (R_i, R_i'), we compute the paired score difference $d_i = S_m(R_i') - S_m(R_i)$, where $S_m(\cdot)$ denotes the score assigned by metric m.

Let $D_m = \{d_i^{(m)}\}_{i=1}^n$ denote the collection of paired differences for metric m, which forms the basis of both analyses. To assess surface sensitivity, we test whether rewriting induces a systematic shift using the Wilcoxon signed-rank test [25]:

$$H_0 : \text{median}(D_m) = 0 \qquad \text{vs.} \qquad H_1 : \text{median}(D_m) \neq 0. \quad (1)$$

Rejecting H_0 indicates that the metric assigns systematically different scores to original and rewritten reviews. Because the rewrite is intended to preserve evaluative content, such a shift is interpreted as evidence that the metric is affected by surface-level presentation. We assess surface sensitivity using the Wilcoxon signed-rank test [25], a non-parametric procedure for paired comparisons that does not require normally distributed differences.

To assess *robustness*, we use the Two One-Sided Tests (TOST) procedure [11]. Whereas the Wilcoxon test asks whether a non-zero shift can be detected, equivalence testing asks whether the shift is sufficiently small to be considered practically negligible. For each metric m, we define an equivalence bound $\delta_m > 0$ and test:

$$H_0 : |\mu_m| \geq \delta_m \qquad \text{vs.} \qquad H_1 : |\mu_m| < \delta_m, \quad (2)$$

where $\mu_m = \mathbb{E}[D_m]$ is the mean paired difference. A metric is considered robust if the TOST procedure rejects the null hypothesis, providing positive evidence that the true mean shift lies within the equivalence interval $[-\delta_m, +\delta_m]$. Following the conventional small-effect threshold of $d = 0.2$ [3] and recommendations for equivalence testing when no external criterion for practical significance is available [11], we set $\delta_m = 0.2\sigma_m$, where σ_m is the standard deviation of the paired differences for metric m. This scale-normalized bound allows the same robustness criterion to be applied across metrics with different scoring ranges.

Both tests are applied independently to all evaluated metrics. To control the family-wise error rate, we apply Bonferroni correction [2] $\alpha_{\text{corrected}} = \frac{0.05}{M}$, where M is the number of evaluated metrics. The same corrected threshold is used for sensitivity and robustness testing. Combining the Wilcoxon and TOST outcomes allows us to distinguish four forms of metric behavior. A metric is classified as **sensitive only** when the Wilcoxon test rejects the null hypothesis, but the TOST procedure does not, meaning that rewriting induces a systematic score shift whose magnitude cannot be treated as practically negligible. A metric is classified as **robust only** when the Wilcoxon test does not reject the null hypothesis and the TOST procedure does, meaning that no systematic shift is detected and equivalence is positively established. A metric falls into the **sensitive ∩ robust** category when both tests reject their null hypotheses, indicating that rewriting produces a statistically detectable shift, but one whose magnitude remains within the equivalence bound. Finally, a metric is classified as **inconclusive** when neither test rejects its null hypothesis, meaning that no systematic shift is detected, but equivalence is also not established.

3 Experimental Setup

Peer Review Evaluation Metrics. We evaluate metrics introduced in four prior works: ReviewEval [6], REMOR [22], RottenReviews [4], and ScholarPeer [7]. Together, these metrics cover analytical depth, actionability, review-component coverage, holistic quality, and human-anchored comparative judgment. Metrics from ReviewEval focus on depth and actionability, REMOR measures explicit review components, RottenReviews evaluates broad

Table 1: Per-metric surface sensitivity and robustness using GPT-5-mini as the judge. Metrics are classified with Bonferroni-corrected Wilcoxon and TOST tests ($\alpha = 0.05/29$). σ_m is the standard deviation of paired differences, $\delta_m = 0.2\sigma_m$ is the equivalence bound, and the final column reports whether the classification agrees with Gemini-2.5-Flash.

Source	Metric	Scale	σ_m	δ_m	Mean Δ	Wilcoxon p	Sensitive?	TOST p	Robust?	Gemini-2.5-Flash Agreement
ReviewEval	Literature Comparison	[0,1]	0.451	0.090	+0.051	6.35×10^{-4}	✓	0.004	–	✓
	Methodological Scrutiny	[0,1]	0.250	0.050	+0.060	1.75×10^{-15}	✓	0.893	–	✓
	Results Interpretation	[0,1]	0.277	0.055	+0.082	2.21×10^{-17}	✓	0.998	–	✓
	Theoretical Contributions	[0,1]	0.410	0.082	+0.044	1.99×10^{-4}	✓	0.002	–	✓
	Logical Gaps Identification	[0,1]	0.275	0.055	+0.053	4.45×10^{-9}	✓	0.413	–	✓
	Overall Depth	[0,1]	0.205	0.041	+0.058	1.71×10^{-18}	✓	0.995	–	✓
	Total Insights	[0,∞)	5.292	1.058	-0.356	0.048	–	2.05×10^{-5}	✓	✓
	Actionable Insights	[0,∞)	5.769	1.154	-0.319	0.085	–	3.87×10^{-6}	✓	✓
	Actionability Score	[0,1]	0.070	0.014	+0.001	0.814	–	1.00×10^{-8}	✓	✓
REMOR	Criticism	[0,1]	0.193	0.039	+0.017	0.004	–	3.08×10^{-4}	✓	–
	Example	[0,1]	0.117	0.023	+0.018	2.20×10^{-5}	✓	0.074	–	✓
	Importance & Relevance	[0,1]	0.152	0.030	+0.048	2.15×10^{-19}	✓	1.000	–	✓
	Materials & Methods	[0,1]	0.198	0.040	-0.005	0.328	–	2.94×10^{-8}	✓	–
	Praise	[0,1]	0.191	0.038	+0.032	4.51×10^{-8}	✓	0.146	–	✓
	Results & Discussion	[0,1]	0.150	0.030	-0.006	0.099	–	3.85×10^{-7}	✓	–
	Suggestion & Solution	[0,1]	0.156	0.031	+0.028	8.69×10^{-7}	✓	0.237	–	✓
RottenReviews	Comprehensiveness	[0,5]	0.411	0.082	+0.383	1.59×10^{-103}	✓	1.000	–	✓
	Objectivity	[0,5]	0.384	0.077	+0.372	1.94×10^{-104}	✓	1.000	–	✓
	Fairness	[0,5]	0.380	0.076	+0.310	2.87×10^{-86}	✓	1.000	–	✓
	Actionability	[0,5]	0.590	0.118	+0.291	9.90×10^{-45}	✓	1.000	–	✓
	Constructiveness	[0,5]	0.524	0.105	+0.311	5.76×10^{-60}	✓	1.000	–	✓
	Relevance Alignment	[0,5]	0.129	0.026	+0.012	2.64×10^{-5}	✓	3.34×10^{-4}	✓	–
	Overall Quality	[0,5]	13.286	2.657	+3.887	1.39×10^{-28}	✓	0.998	–	✓
	Overall Score	[0,100]	5.151	1.030	+5.573	5.06×10^{-128}	✓	1.000	–	✓
ScholarPeer	Technical Accuracy	[1,10]	0.881	0.176	-0.419	5.97×10^{-161}	✓	1.000	–	✓
	Constructive Value	[1,10]	0.891	0.178	-0.707	$<10^{-300}$	✓	1.000	–	–
	Analytical Depth	[1,10]	0.782	0.156	-1.192	$<10^{-300}$	✓	1.000	–	✓
	Novelty & Significance	[1,10]	1.203	0.241	-0.265	3.44×10^{-41}	✓	0.894	–	✓
	Overall	[1,10]	0.737	0.147	-0.595	$<10^{-300}$	✓	1.000	–	–

quality dimensions such as fairness and constructiveness, and ScholarPeer compares reviews against a paired human-written review, allowing us to examine whether human anchoring affects sensitivity to rewriting. Because our goal is to test whether substantive metrics are affected by surface-level variation, we retain only content-oriented metrics, including those for analytical depth, constructiveness, actionability, relevance, and aspect coverage, as reported in Table 1. We exclude metrics that explicitly target style or presentation, namely *Presentation & Reporting* from REMOR and *Usage of Technical Terms* and *Clarity & Readability* from RottenReviews. The resulting analysis contains $M = 29$ metrics, giving $\alpha_{\text{corrected}} = 0.05/29 \approx 0.00172$.

Peer Review Dataset. We use PeerPrism [18], a benchmark designed to separate evaluative content from surface realization in peer reviews. PeerPrism contains 160 ICLR and NeurIPS papers from 2021–2024. Our analysis uses only the *rewritten* partition, which contains 4,044 faithful rewrites of 674 original human reviews. These rewrites preserve each review's judgments, arguments, strengths, weaknesses, and recommendations while changing wording, phrasing, and style. We pair each original review with its rewritten variants, yielding $n = 4{,}044$ paired comparisons per metric, pooled across reviews, papers, years, venues, and rewriting models.

Human Validation of Rewrites. Our framework assumes that R' preserves the evaluative content of R. To validate this, we conducted a human annotation study on 50 randomly sampled original–rewrite pairs. Three computer science graduate students annotated the pairs, with each pair evaluated by two annotators. Annotators rated how faithfully the rewritten review preserved the original review's content on an overall assessment on a 1–5 Likert scale. Additionally they indicated whether the original and rewritten reviews would provide essentially equivalent feedback to a paper author using a binary Yes or No judgment.

Faithfulness scores were high across annotators, with mean ratings of 4.28, 4.96, and 4.28 out of 5. Agreement was also strong: 90% within-one-point agreement for faithfulness, 92% exact agreement for binary equivalence, and 94% of annotations indicating equivalent feedback. Overall, this validation supports the content faithfulness of the rewrites and their use for testing surface sensitivity and robustness. Detailed annotation guidelines are provided in the project repository.

Judge Models and Scoring Protocol. All evaluations are performed using GPT-5-mini [20] as the judge model. For each metric, we use the original prompt and scoring rubric from the work that introduced the metric without modification. To assess whether the findings depend on the judge model, we repeat the full analysis using Gemini-2.5-Flash and compare the resulting sensitivity and robustness classifications against the GPT-5-mini results.

4 Results and Findings

Table 1 reports the per-metric results, and Figure 1 summarizes the joint evidence from the Wilcoxon and TOST procedures. The question is whether metrics intended to evaluate substantive review quality remain stable when the same content is expressed in a different linguistic form. The overall observed pattern is that many LLM-as-a-judge metrics respond significantly to how that evaluation is written. More detailed observations are as follows:

(1) Surface sensitivity is a broad measurement problem. Rewriting reviews while preserving their content still causes many peer-review evaluation metrics to assign systematically different scores. This is shown in Table 1 by the *Sensitive?* column and the

Bonferroni-significant Wilcoxon p-values. Since the rewrites preserve the original review content, these shifts cannot be interpreted as ordinary differences in review quality; they indicate sensitivity to linguistic realization. This is especially concerning for metrics that require broad interpretive judgment, such as analytical depth, comprehensiveness, objectivity, fairness, constructiveness, and overall quality. Such judgments may be affected by fluency, organization, explicitness, and tone, even when the underlying critique is unchanged. The pattern appears across metrics: ReviewEval's depth-oriented metrics, RottenReviews' broad quality metrics, and ScholarPeer's human-anchored dimensions are especially sensitive, while several REMOR dimensions also show sensitivity. This suggests that surface sensitivity is not an artifact of a single metric source or scoring scale, but a broader challenge in LLM-as-a-judge peer-review evaluation.

(2) Reliable metrics tend to be anchored in concrete review content. The metrics that appear reliable under the robustness criterion share an important property. They evaluate relatively concrete, localized, or explicitly identifiable aspects of a review rather than broad qualitative impressions. Under the TOST criterion, the most reliable metrics are ReviewEval's *Total Insights*, *Actionable Insights*, and *Actionability Score*, REMOR's *Criticism*, *Materials & Methods*, and *Results & Discussion*, and RottenReviews' *Relevance Alignment*. These metrics are less tied to overall polish and more tied to identifiable evaluative content. For metric design, this suggests that concrete content-based judgments are more stable under rewriting than global quality assessments. This does not mean component-level metrics are sufficient, since high-quality reviewing also requires synthesis and prioritization, but it suggests that broad constructs may need to be decomposed into explicit subproperties or validated against meaning-preserving rewrites.

(3) Sensitivity and robustness identify different kinds of reliability. The joint view in Figure 1 shows why the Wilcoxon and TOST tests should be interpreted together. A metric can show a statistically detectable shift under rewriting while still remaining practically stable if the shift is small. Conversely, a metric can fail to show a significant shift without providing positive evidence that the effect is negligible. The joint classification therefore separates metrics that merely lack evidence of sensitivity from metrics that provide affirmative evidence of robustness. Most metrics fall into the *Sensitive only* region, meaning that rewriting induces detectable score shifts that exceed the equivalence bound. These metrics should be treated as unreliable for evaluating review substance under surface-level rewriting, because their scores change in ways that are too large to be considered practically negligible. By contrast, *Total Insights*, *Actionable Insights*, *Actionability Score*, *Criticism*, *Materials & Methods*, and *Results & Discussion* fall into the *Robust only* region, while *Relevance Alignment* is *Sensitive* ∩ *Robust*, showing a small but practically negligible shift. Reliability also varies across metric sources. ReviewEval includes both sensitive depth metrics and robust insight-based metrics, REMOR contains several robust component-level metrics, RottenReviews is largely sensitive except for *Relevance Alignment*, and ScholarPeer has no metric that satisfies the robustness criterion.

(4) Directionality reveals metric-specific presentation effects. The Mean Δ column in Table 1 shows that rewriting affects different metric sources in different directions. For ReviewEval,

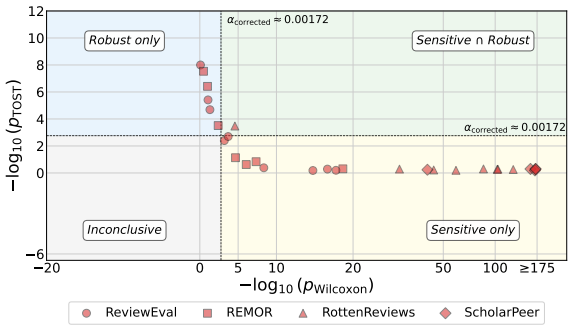

Figure 1: Sensitivity and robustness evidence for all metrics. The four shaded areas correspond to *Robust only*, *Sensitive only*, *Sensitive* ∩ *Robust*, and *Inconclusive* regions.

REMOR, and RottenReviews, rewritten reviews generally receive higher scores than the original human reviews, with especially large positive shifts for RottenReviews metrics such as *Comprehensiveness* (+0.383) and *Overall Score* (+5.573). This suggests that LLM judges often treat improved fluency, organization, or explicitness as evidence of stronger review quality. ScholarPeer shows the opposite pattern as its metrics assign lower scores to rewritten reviews, even though the evaluative content is preserved. A plausible explanation is its human-anchored scoring design, where reviews are evaluated relative to a paired human-written reference. Under this setup, rewriting may change the perceived relationship between the rewritten review and the human reference, reducing the score even when the underlying evaluation remains aligned. This indicates that human anchoring does not eliminate surface sensitivity. It may instead introduce a different presentation-dependent behavior in which the metric becomes sensitive to perceived similarity to the reference rather than to evaluative substance alone.

(5) The findings are not judge-model artifacts. The replication with Gemini-2.5-Flash indicates that the main conclusions are not specific to GPT-5-mini. As shown in the right-most column of Table 1, most metric classifications remain unchanged across judge models, and the few disagreements occur close to the statistical decision boundary. This consistency suggests that the observed behavior reflects a broader property of LLM-as-a-judge evaluation rather than an idiosyncrasy of a single judge model. Replacing one strong judge model with another is therefore unlikely to resolve the issue on its own. The more fundamental problem lies in how peer-review quality metrics are specified, how much they rely on holistic judgment, and whether they are validated against meaning-preserving changes in surface form.

5 Conclusion

We presented a robustness analysis of LLM-as-a-judge metrics for peer-review evaluation using meaning-preserving rewrites. Many metrics intended to assess substantive review quality were sensitive to changes in wording and presentation despite preserved evaluative content, while robust metrics tended to focus on concrete review content. These findings suggest that robustness to meaning-preserving rewriting should be validated before such metrics are used to compare human-written, AI-assisted, and AI-generated reviews.

GenAI Usage Disclosure

Large language models were used in the preparation of this manuscript for grammar correction, prose polishing, and coding support in the implementation of the experimental pipeline. All research design, analysis, interpretation of results, and scholarly conclusions are the work of the authors.

References

[1] Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. Language invariant properties in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*. 84–92.

[2] Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze* 8 (1936), 3–62.

[3] Jacob Cohen. 1992. Statistical power analysis. *Current directions in psychological science* 1, 3 (1992), 98–101.

[4] Sajad Ebrahimi, Soroush Sadeghian, Ali Ghorbanpour, Negar Arabzadeh, Sara Salamat, Muhan Li, Hai Son Le, Mahdi Bashari, and Ebrahim Bagheri. 2025. RottenReviews: Benchmarking Review Quality with Human and LLM-Based Judgments. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 5642–5649.

[5] Julia Farganus, Tomasz Jan Kajdanowicz, et al. 2026. PRAIB: Peer Review AI Benchmark of Behaviour of LLM-Assisted Reviewing. *arXiv preprint arXiv:2605.29815* (2026).

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

[7] Palash Goyal, Mihir Parmar, Yiwen Song, Hamid Palangi, Tomas Pfister, and Jinsung Yoon. 2026. ScholarPeer: A Context-Aware Multi-Agent Framework for Automated Peer Review. *arXiv preprint arXiv:2601.22638* (2026).

[8] Serge PJM Horbach and Willem Halffman. 2018. The changing forms and expectations of peer review. *Research integrity and peer review* 3, 1 (2018), 8.

[9] Jacalyn Kelly, Tara Sadeghieh, and Khosrow Adeli. 2014. Peer review in scientific publications: benefits, critiques, & a survival guide. *Ejifcc* 25, 3 (2014), 227.

[10] Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the sixth conference on machine translation*. 478–494.

[11] Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science* 8, 4 (2017), 355–362.

[12] Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert West. 2024. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150* (2024).

[13] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183* (2024).

[14] Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622* (2023).

[15] Ngoc Phan Phuoc Loc, La Viet, Toan Huynh, Thanh Tran Khanh, Duy A Nguyen, Tuan Anh Nguyen Pham, Thanh Nguyen, Nitesh V Chawla, Wray Buntine, Kok-Seng Wong, et al. 2026. PRISM: A Multi-Dimensional Benchmark for Evaluating LLM Peer Reviewers. *arXiv preprint arXiv:2605.26730* (2026).

[16] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 7297–7306.

[17] Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37 (2024), 68772–68802.

[18] Soroush Sadeghian, Alireza Daqiq, Radin Cheraghi, Sajad Ebrahimi, Negar Arabzadeh, and Ebrahim Bagheri. 2026. PeerPrism: Peer Evaluation Expertise vs Review-writing AI. *arXiv preprint arXiv:2604.14513* (2026).

[19] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076* (2023).

[20] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267* (2025).

[21] Amanda Sizo, Adriano Lino, Álvaro Rocha, and Luís Paulo Reis. 2025. Defining quality in peer review reports: a scoping review. *Knowledge and Information Systems* 67, 8 (2025), 6413–6460.

[22] Pawin Taechoyotin and Daniel Acuna. 2025. Remor: Automated peer review generation with llm reasoning and multi-objective reinforcement learning. *arXiv preprint arXiv:2505.11718* (2025).

[23] Jonathan P Tennant, Jonathan M Dugan, Daniel Graziotin, Damien C Jacques, François Waldner, Daniel Mietchen, Yehia Elkhatib, Lauren B Collister, Christina K Pikas, Tom Crick, et al. 2017. A multi-disciplinary perspective on emergent and future innovations in peer review. *F1000Research* 6 (2017), 1151.

[24] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9440–9450.

[25] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*. Springer, 196–202.

[26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.