A Report on “A Large-Scale Randomized Study of Large Language Model Feedback in Peer Review” by Thakkar et al. (2026)

Reviewer 2

February 24, 2026

v1

isitcredible.com

Disclaimer

This report was generated by large language models, overseen by a human editor. It represents the honest opinion of The Catalogue of Errors Ltd, but its accuracy should be verified by a qualified expert. Comments can be made here. Any errors in the report will be corrected in future revisions.

> I am wiser than this person; for it is likely that neither of us knows anything fine and good, but he thinks he knows something when he does not know it, whereas I, just as I do not know, do not think I know, either. I seem, then, to be wiser than him in this small way, at least: that what I do not know, I do not think I know, either.
>
> Plato, *The Apology of Socrates*, 21d

> To err is human. All human knowledge is fallible and therefore uncertain. It follows that we must distinguish sharply between truth and certainty. That to err is human means not only that we must constantly struggle against error, but also that, even when we have taken the greatest care, we cannot be completely certain that we have not made a mistake.
>
> Karl Popper, 'Knowledge and the Shaping of Reality'

Overview

Citation: Thakkar, N., Yuksekgonul, M., Silberg, J., Garg, A., Peng, N., Sha, F., Yu, R., Vondrick, C., & Zou, J. (2026). A Large-Scale Randomized Study of Large Language Model Feedback in Peer Review. *Nature Machine Intelligence*

URL: https://doi.org/10.1038/s42256-026-01188-x

Abstract Summary: This study introduces Review Feedback Agent, a system leveraging multiple large language models to improve peer review quality by providing automated feedback to reviewers. A randomized controlled study at ICLR 2025 with over 20,000 reviews showed that 27% of reviewers who received automated feedback updated their reviews, incorporating over 12,000 suggestions, leading to more informative reviews and increased engagement during rebuttals.

Key Methodology: A large-scale randomized controlled study was conducted at ICLR 2025, involving over 20,000 reviews. The Review Feedback Agent, a multi-LLM system, provided automated feedback to reviewers, and the impact on review updates, length, engagement, and scores was analyzed.

Research Question: Can large language models (LLMs) enhance peer review quality by providing automated feedback to reviewers?

Summary

Is It Credible?

Thakkar et al. present a large-scale randomized controlled trial conducted during the ICLR 2025 conference to test whether an automated system can improve the quality of peer reviews. The authors claim that deploying a multi-LLM system to flag vague, misunderstood, or unprofessional comments causally improves review clarity, specificity, and actionability. The headline findings suggest that 26.6% of reviewers who received feedback updated their reviews, incorporating over 12,000 suggestions, and that the intervention led to "substantially longer reviews (80 additional words among updaters)" and increased engagement during the rebuttal phase (p. 1).

The most prominent effect size highlighted in the abstract—the 80-word increase in review length—is an observational comparison between reviewers who self-selected to update their reviews and those who did not. Because this comparison breaks the initial randomization, it cannot be causally attributed solely to the feedback; it is highly likely that more conscientious reviewers were both more likely to update their text and more likely to write longer revisions. The true causal intent-to-treat effect reported in the main text is a much more modest, though still statistically significant, increase of 14 words (p. 4). Presenting the larger observational figure so prominently risks inflating the perceived impact of the intervention for casual readers. While the article does not explicitly list the baseline average length of an initial review in the main table, Extended Data Fig. 2 visualizes length changes relative to the baseline, providing some context for these increases.

The study design also struggles to fully isolate the content of the LLM feedback from the psychological impact of being monitored. Reviewers received notifications under the official heading "Review Feedback from Associate Program Chairs" (p. 7).

This official scrutiny alone could trigger a Hawthorne effect, motivating reviewers to put more effort into their work regardless of the specific AI-generated suggestions. The authors attempt to counter this by analyzing incorporation rates, showing that specific feedback types (e.g., requests for analysis) were incorporated more often than others, and verifying via LLM that 17.7% of specific suggestions were enacted (p. 5). While this suggests the content was relevant, without a placebo control group receiving a generic quality-check notification, it remains difficult to definitively separate the behavioral impact of the specific semantic content from the social prompt of being observed.

Furthermore, the study's definition of review quality is largely limited to stylistic and structural improvements. The system was deliberately designed to avoid evaluating the substantive scientific correctness of a reviewer's critique (p. 3). While the "Actor" prompts did include instructions to check if a reviewer requested something "obviously present" in the paper (p. 27), the system generally cannot detect if a reviewer's core theoretical premise is flawed. By pushing reviewers to make their comments more specific and actionable, the agent risks encouraging them to elaborate on scientifically invalid criticisms, potentially making a bad review more specifically wrong. The reliance on a self-policing architecture, where the same underlying base model (Claude 3.5 Sonnet) generates the feedback and runs the reliability tests (p. 9), further compounds the risk of systemic blind spots, though the authors mitigate this by using distinct agents with different system prompts.

The evidence supporting the claim that the revised reviews were actually better relies heavily on a blinded human evaluation of 100 reviews. However, the authors omit any measure of inter-annotator agreement for this highly subjective preference task, which is a standard requirement for establishing the reliability of qualitative judgments. Additionally, the statistical analysis uses standard t-tests that do not appear to account for the cluster-randomized design, where all reviews for a single paper were assigned to the same condition (pp. 4, 8). While the authors performed a robustness check across subfields (p. 13) to ensure consistency, the primary statistical tests likely overstate the precision of the findings by treating clustered reviews as independent samples.

Despite these methodological limitations, the study successfully demonstrates that automated systems can be deployed at scale to nudge reviewer behavior. The high incorporation rates and the increased engagement during the rebuttal period indicate that reviewers found the structural feedback useful. However, the study ultimately provides stronger evidence that LLMs can effectively enforce formatting and specificity guidelines, rather than demonstrating they can elevate the underlying scientific rigor of the peer review process.

The Bottom Line

Thakkar et al. provide compelling evidence that automated, LLM-generated nudges can prompt peer reviewers to write longer, more specific critiques and engage more deeply during author rebuttals. However, the headline effect sizes are somewhat inflated by observational comparisons, and the study design cannot fully separate the value of the AI's specific advice from the psychological effect of being monitored by conference organizers. Furthermore, because the intervention primarily targets the structural form of reviews rather than their scientific validity, the results should be interpreted as a successful behavioral nudge rather than a comprehensive upgrade to scientific evaluation.

Potential Issues

Potential for misinterpretation of causal claims based on post-treatment subgroup analysis: The article's abstract gives prominence to a finding that may be misinterpreted as a direct causal effect of the intervention. The abstract states the intervention "led to substantially longer reviews (80 additional words among updaters)" (p. 1). This 80-word figure is derived from an observational comparison between reviewers in the treatment group who self-selected to update their reviews and those who did not. Because this comparison is not randomized, the observed difference cannot be causally attributed to the feedback alone and is likely confounded by pre-existing reviewer characteristics. The authors are transparent about this, acknowledging that more engaged reviewers were more likely to revise their reviews and clearly reporting the much smaller, methodologically sound intent-to-treat effect of 14 words in the main text (p. 4). However, the prominent placement of the larger 80-word figure in the abstract, even with the qualifier "among updaters," may inflate the perceived impact of the intervention for readers who do not engage with the full text.

Potential for uncontrolled demand characteristics: The study's design cannot fully disentangle the effect of the LLM feedback's content from the psychological effect of being monitored. Reviewers were informed that the feedback was AI-generated and delivered under the official heading "Review Feedback from Associate Program Chairs" (p. 7). This intervention signals that a reviewer's work is being scrutinized, which on its own could motivate them to be more diligent, a phenomenon known as a Hawthorne effect. The authors provide counter-evidence, noting that specific feedback types had higher incorporation rates (p. 5), suggesting the content itself was influential. However, the absence of a placebo control group—for example, one receiving a generic notification of a quality check without specific suggestions—makes it difficult to isolate the impact of the feedback's content from the act of intervention itself.

Limited scope of "quality" improvement: The study's claim of enhancing "review quality" is largely confined to stylistic and structural aspects, as the intervention was deliberately designed to avoid evaluating a review's substantive scientific correctness (pp. 3, 9). This creates a significant blind spot: the system generally cannot determine if a reviewer's core criticism is scientifically valid. By encouraging reviewers to make comments more specific and actionable, the system risks prompting them to elaborate on a flawed premise, potentially making a review more specifically wrong and thus more harmful. The authors are transparent about this, describing it as a "conservative design approach" chosen due to the limitations of current LLMs (p. 8). While the prompts did include instructions to check for obvious factual errors regarding the article's content (p. 27), these guardrails do not address the underlying scientific validity of the reviewer's theoretical arguments, meaning the study's findings on quality improvement apply primarily to form rather than scientific substance.

Statistical analysis may not fully account for cluster randomization: The study's statistical analysis may not fully account for the experimental design, potentially affecting the precision of its statistical claims. The experiment randomized at the paper level, assigning all reviews for a given paper to the same condition and thus creating clusters of non-independent data (p. 8). However, the analysis of review-level outcomes, such as the change in review length, uses standard two-sided t-tests for independent samples (p. 4). This approach does not account for the intra-cluster correlation—the fact that reviews of the same paper are likely to be more similar to each other than to reviews of different papers. While more robust methods like mixed-effects models or clustered standard errors were not used for the primary analysis, the authors did perform a robustness check by analyzing the effect across different subfields, which showed consistent trends (p. 13). Given the large reported effect size on review length, the main finding is likely to be robust, but the reported p-values may overstate the statistical significance.

Missing reliability metrics for blinded qualitative evaluation: The primary evidence for an improvement in review quality relies on a blinded human evaluation that lacks a key metric needed to assess its reliability. The study used a sample of 50 reviews from the treatment group and 50 from the control group for this evaluation, conducted by "experienced AI PhD students" (pp. 5, 25). However, the article does not report any measure of inter-annotator agreement, such as Fleiss' kappa or Krippendorff's alpha, for this preference task. While the authors did validate their automated incorporation metric against human labels (reporting 92% accuracy), the subjective human preference study lacks this standard confirmation of consistency (p. 25).

Potentially biased method for measuring feedback incorporation: The automated method for measuring feedback "incorporation" was validated in a way that may have introduced confirmation bias. This key metric was measured using an LLM to determine if feedback generated by a similar LLM was integrated into revised reviews (p. 4). During the validation of this pipeline against 222 human-labeled items, the authors performed a post-hoc re-evaluation of the human labels and concluded that most of the model's "errors" were in fact human errors (p. 25). The authors describe this as an error analysis based on verifiable content, but this re-adjudication was performed with knowledge of the model's predictions. This process risks biasing the evaluation in favor of the model, potentially leading to an overly optimistic assessment of the measurement tool's accuracy.

Self-policing system architecture: The quality control mechanism for the feedback-generating agent relies on a self-policing architecture that may be vulnerable to systemic blind spots. The "reliability tests" used as guardrails were executed by the same underlying base model (Claude Sonnet 3.5) that generated the feedback (p. 9). While the system uses distinct agents with different prompts for generation and checking (p. 6), the fundamental logic and capabilities are shared. If the base model has an inherent failure mode or bias, using that same model to check its own output is unlikely to detect that failure. The authors are transparent about this design, but the lack of an independent verification model means the high reported pass rate (96%) could reflect genuine quality or the checker's inability to detect the generator's flaws.

Debatable design limitations: The study's design includes several trade-offs that may limit the interpretation of its findings. First, randomizing at the paper level means a single reviewer could have been assigned to review papers in both the treatment and control groups. The experience of receiving AI feedback on one review could alter their behavior on subsequent control-group reviews, potentially contaminating the control group and biasing the estimated treatment effect downwards. Second, the clustering analysis used to categorize the types of feedback provided was performed entirely by LLMs without human validation (p. 6). The resulting clusters therefore represent how the model categorizes its own output, which may not align with human-centric categories. Finally, the study was conducted at a single AI conference (ICLR), and its findings may not generalize to other academic disciplines with different peer review cultures. The authors are largely transparent about these issues, but they represent important context for interpreting the results.

Omission of contextual information and sensitivity analyses: The article omits some details that would help in assessing the practical significance and robustness of the findings. For instance, a key result is that reviewers who updated their reviews increased their length by an average of 80 words, but the article does not report the baseline average length of an initial review in the main table, though it is visualized in Extended Data Fig. 2 (p. 4). Additionally, the definition of an "updated" review relies on specific thresholds (an edit distance greater than 5 and, for the control group, a delay of at least 1 hour). While these thresholds are justified, the article does not include a sensitivity analysis to show that the main findings are robust to these specific choices (p. 4).

Presentation and clerical issues: The article contains several minor inconsistencies in its reporting. The summary of findings states that annotators preferred revised reviews “17 percentage points more often in the feedback group than in the control group” (p. 2), but the data presented in the results show a 24 percentage point difference (68% preference in the feedback group vs. 44% in the control group) (p. 5). The 17 percentage point figure likely refers to the difference in update rates (26.6% vs. 9.4%), not the preference study, creating confusion. The start date of the study is reported as both October 14 (p. 2) and October 15 (p. 8). The exclusion rate for reviews selected for feedback is stated as “Less than 8%” in the text (p. 8), but the data in Figure 1a imply a rate of 15.7% (3,521 out of 22,467 selected reviews). Finally, for several statistical tests, the article reports p-values but omits the corresponding test statistics and degrees of freedom (p. 4, Table 1).

Future Research

Placebo-controlled interventions: To isolate the effect of the LLM feedback content from the Hawthorne effect of being monitored, future experimental designs should include an active placebo control group. This group could receive a generic notification from the program chairs reminding them of review guidelines without providing specific, AI-generated semantic feedback on their text.

Rigorous cluster-robust statistical analysis: Because peer review interventions are often randomized at the paper level, creating clusters of non-independent reviews, future evaluations should employ mixed-effects models or clustered standard errors. This would properly account for intra-cluster correlation and provide more accurate estimates of statistical significance for review-level outcomes.

Inter-annotator reliability reporting: Future studies relying on human evaluations of review quality, clarity, or actionability must report standard metrics of inter-annotator agreement, such as Fleiss' kappa or Krippendorff's alpha, to establish the consistency and reliability of the subjective judgments underpinning their conclusions.

isitcredible.com