nature portfolio

Peer Review File

Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025

Corresponding Author: Dr James Zou

A version of this paper was originally rejected for publication by Nature Machine Intelligence, however that decision was reconsidered after appeal by the authors.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
SUMMARY
This papers reports the results of a randomized controlled trial testing the effects of providing reviewers with feedback on their reviews. The study has many merits: the trial is well designed, it's large-scale, deployed in an excellent real-world setting, and the question of how much LLMs can improve reviews (and, more generally, augment humans in intellectually demanding tasks) is very timely. I see one substantial issue and a few medium/minor ones.

SUBSTANTIAL ISSUE
The blinded evaluation of review quality appears to suffer from "selection on the dependent variable" (i.e. quality). For the set of treatment reviews, the authors selected only those where the "proportion of incorporated feedback exceeded 0.60". If I'm understanding correctly, this isn't a random set of treatment reviews but a (random?) set of those that received feedback *AND* the reviewers found the feedback valuable enough to incorporate, i.e. they thought the feedback was relatively high quality. So, this comparison, unsurprisingly, goes on to show that reviews revised after high-quality feedback are perceived better than those without feedback, but it tells us almost nothing about feedback quality on average. At best it tells us that *sometimes* the feedback must have been good because it is associated with higher post-feedback quality. I say "at best" because even that more limited conclusion is not guaranteed: the treatment reviews you (non-randomly) selected may have been better *pre-feedback* than randomly selected control reviews. So, a more fair comparison is between a random sample of reviews from the control group vs. a random sample of reviews from the treatment group (with no other conditioning).

So, I'd strongly urge redoing this analysis with a more fair comparison. Another option is to just remove this quality evaluation component from the paper, although it would weaken it. Yet another option is to acknowledge that the comparison isn't very informative, and certainly not feature it in the abstract.

I would also like to see more detail in this section: the exact instructions given to the annotators, were the reviews rated independently or side-by-side, how many of each type, the experience level of the annotators, etc.

MEDIUM
1. The current Related Works section is a bit too focused on uncritical summary and is placed in an odd position. Putting it after the Introduction would be more typical. More importantly, I think this section could better motivate your research choices. The elephant in the room is that this study is designed around a relatively "easy" task --- identifying vague statements, those that are already addressed, and unprofessional ones. Notably absent are more intellectually challenging tasks like identifying weaknesses, making novel connections, providing constructive and non-obvious suggestions. This "elephant" is currently downplayed, as I believe there's just one sentence in the Discussion addressing it: "Additionally, it would be interesting to explore ... to generate more nuanced feedback for complex issues." In my view, this is not a weakness of the study but a strength --- you go after a task that's more achievable with current technology and still valuable. This (smart) research choice can flow from your Related Works section: you could give a even-handed/critical overview, noting that current research does not provide support for LLMs performing well on the more intellectual review tasks, and therefore you make the very reasonable move to test them on easier tasks. Without such a context, a reader is left asking, "Well, if LLMs are so great, why are the authors going after the relatively low-hanging fruit of 'vagueness' and 'unprofessionalism'"?

2. Going back to my "elephant in the room," it was interesting to see that the treatment caused only 0.6% more changes in review scores (relative to baseline). This paints a picture where reviewers incorporate the feedback to improve the tone/specificity of their reviews, but ultimately don't change their opinions/scores much and so the direction of science does not change much. In the current version, this key result (and the most plausible interpretation of it) is not in the abstract or the summary of findings but is "hidden" on page 8. I think the paper should be much more upfront about this result, i.e. abstract and summary of findings. If such results are downplayed, there's a very big risk that the more casual reader comes away with an overly hyped sense of LLM capabilities relative to what this study (very excellently) demonstrates. (And, in my view, the capabilities demonstrated in the paper are already substantial!)

MINOR
1. In Section 3.4, it was odd to see feedback gets clustered into 5 *new* topics when it was, by design, generated around 3 topics. Why not just go with the original 3?

I hope these comments help move this excellent project forward!

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Summary:
This paper develops a Review Feedback Agent that provides feedback to human reviewers. Their study found that 26.6% of reviewers changed their reviews after reading the AI-generated feedback. The length of the reviews increased by 80 words, and these reviewers were also more engaged during the rebuttal phase.
Strengths:
1. The tool has significant social impact, and the study is based on a real-world use case; therefore, the results effectively demonstrate the usefulness of the tool.
2. The result analysis is thorough and insightful.
Weaknesses:
1. My major concern is that the study only includes ICLR conference papers and reviews. Since it is a heavily engineered system, I wonder how well it generalizes to other conferences and what would need to be changed to adapt it.
2. The Critic module and Reality Check module are still LLMs-based prompts. Are there any human studies evaluating the quality of the feedback, or surveys asking reviewers how useful they found the feedback?
3. There are more details or discussion can be included in the paper to make the study more comprehensive and insightful, please check the questions/suggestions.
Questions/Suggestions:
1. Are there any cases where the two Actors do not agree with each other? What is the percentage, and how are these cases handled?
2. How can more reviewers be encouraged to update their reviews based on the feedback?
3. The clustering experiments are interesting. I can see that the clustering is more fine-grained compared to the three-target feedback in the Actor module prompt. I still wonder about the percentage of the three-target feedback.
4. In my opinion, the "Addressing Misunderstandings" feedback is especially powerful. But I wonder if LLMs are intelligent enough to generate such feedback, and whether simple few-shot prompt engineering is sufficient.
5. An ablation study where the system is only given the review (without the paper PDF) and the resulting differences in feedback would reveal how much the LLM relies on the paper's content. This could serve as an indirect indicator of the LLM's trustworthiness.
6. What are the limitations of the current feedback system? Is there room for further improvement?

(Remarks on code availability)
N/A

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Great work by the authors in addressing the first round comments. Here are some outstanding issues:

COMPARISONS
However, the blind evaluation piece is still not quite the right comparison, in my view. In the first round, I was proposing you evaluate a sample of final reviews (whether revised or not) taken from each condition. That would let you straightforwardly measure the treatment effect of feedback on quality, which, in the big picture, is what we ultimately want the feedback to do. This would give us an intent-to-treat effect. I suppose another approach would be to use instrumental variables, with condition as the instrument, feedback-taking as the treatment, and quality as the outcome; this would give us the LATE

effect.

All the other comparisons involve selection bias and cannot be straightforwardly interpreted. For example, in the current set-up you're conditioning on *revising* in both groups. That is a *choice* reviewers made and, indeed, we know that the choice was made much more often in the feedback condition. So, the initial randomization gave you two very similar groups, but after you condition on revising, we lose any guarantees that the groups are similar and cannot make causal statements. So, I again would urge either doing a clean test, with no extra conditioning, or interpreting these results very tentatively, something like, "Among those choosing to revise their initial reviews, the increase in quality was higher for those who received feedback."

QUALITY VS. ACTIONABILITY
Thanks also for showing the instructions. After seeing these, I don't think we should call the outcome "review quality." Instead, the construct is arguably one dimension of quality, perhaps something like, "Clarity" or "Actionability".

INTERPRETATION AND ABSTRACT
The revised part of the abstract is too strong an interpretation of the data:
"This suggests that many reviewers found the AI-generated feedback sufficiently helpful to merit updating their reviews, although the intervention had minimal impact on final scores (only 0.6% increase in score changes), suggesting that AI feedback primarily enhanced review quality rather than systematically changing the final score in one direction." The last part may be true, or not; we don't quite know whether reviewers perceived review quality (rather, actionability/clarity) to be enhanced, or these are demand effects because participants didn't want to disappoint the experimenters who clearly put a lot of time into engineering this system, etc. If you remove the last part of the sentence then it's more defensible and fine by me: "This suggests that many reviewers found the AI-generated feedback sufficiently helpful to merit updating their reviews, although the intervention had minimal impact on final scores (only 0.6% increase in score changes)."

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
1.My original weakness 1 concerns generalization. The authors state that the review pipeline is almost identical across other venues, therefore they expect their method to apply to other venues as well. Meanwhile, the authors also acknowledge that the system prompt will need to be adapted to specific domains. The reviewer agrees with their first point. Regarding the second point, the reviewer suggests that in the supplementary material the authors highlight which parts of the prompts should be modified when applying the method to other venues.
2.For all the weakness, questions and suggestions the reviewer raised, the authors have provided either experiments or explanations and discussions. I believe these responses are sufficient. One additional proposal the reviewer has regarding "How can more reviewers be encouraged to update their reviews based on the feedback?" is the following. Based on the authors' analysis that late submission reviewers are more likely to submit revisions than those who submit their reviews close to the deadline, possibly because the latter do not have enough time to revise, would it be helpful to integrate a "review revision phase" after the initial review submission in the submission review pipeline? Maybe could be added as part of the discussion.

In summary, the reviewer is content with the current manuscript.

(Remarks on code availability)

Manuscript NATMACHINTELL-A25041472
Response to reviewers

Overall comments to reviewers

We sincerely thank the reviewers for their thorough evaluation of our manuscript and their constructive insights on our AI feedback system.

We are encouraged that all reviewers recognized the merit and potential impact of our work. Reviewer 1 described our study as "excellent" multiple times and praised that our "trial is well designed, it's large-scale, [and] deployed in an excellent real-world setting." We appreciate Reviewer 1's acknowledgment of the significant practical value demonstrated by our system. Reviewer 2 highlighted the "significant social impact" of our tool and found our analysis "thorough and insightful." We are grateful that Reviewer 2 recognized the real-world applicability of our approach and the comprehensive nature of our evaluation.

Most importantly, neither reviewer identified fundamental flaws in our methodology or conclusions. The primary concerns raised were Reviewer 1's suggestions on evaluation design and Reviewer 2's questions about generalizability and implementation details, all of which we have comprehensively addressed. The reviewers' insights have been invaluable in strengthening our work. We have carefully incorporated all feedback in our revised manuscript, including re-running key analyses with improved methodologies and providing additional detail on system implementation and broader applicability. Please see below, in blue, for a point-by-point response to the reviewers' comments and concerns. In our revised manuscript, we have also highlighted key changes in blue.

Summary of Major Revisions

These revisions significantly strengthened our manuscript and carefully addressed all the reviewers' comments. Below is a high-level summary of the key changes we made:

1. We expanded the blinded evaluation to assess the improvements in review quality due to LLM feedback (see section 4.2.1 of the revised manuscript)
    a. We re-conducted the blinded evaluation analysis using random samples from both control and treatment groups with no other conditioning (addressing Reviewer 1's concern about selection bias). We found that feedback significantly improved both the quality and detectability of revisions, with annotators preferring modified reviews from the feedback group over the control group (68% vs 44%) and marking fewer ties (20% vs 48%).
    b. We added detailed annotation instructions and annotator information for transparency.
2. We strengthened the generalizability claims of our system
    a. We provided evidence for consistent effectiveness across diverse research areas within ICLR by re-running key results stratified by paper subcategory.

    b. We addressed concerns about venue-specific implementation.
    c. We discussed transferability to other conferences and required adaptations.
3. We reframed the related works and key conclusions
    a. We substantially revised the related works section to better motivate our focus on achievable and low-risk AI tasks: vague statements, unprofessional language, and content misunderstandings.
    b. We emphasized the key conclusion that this intervention had small impact on final scores (only a 0.6% increase in score changes), suggesting that AI feedback primarily enhanced review quality rather than changing scientific evaluations.
4. We provided a thorough discussion of the limitations of our system and the top priority next steps.

Reviewer #1 (Remarks to the Author):

SUMMARY
This papers reports the results of a randomized controlled trial testing the effects of providing reviewers with feedback on their reviews. The study has many merits: the trial is well designed, it's large-scale, deployed in an excellent real-world setting, and the question of how much LLMs can improve reviews (and, more generally, augment humans in intellectually demanding tasks) is very timely. I see one substantial issue and a few medium/minor ones.

We sincerely thank the reviewer for their thorough evaluation of our manuscript! We appreciate their recognition of our ‘excellent’ study as ‘well-designed’ and ‘timely.’ We have carefully reviewed their insightful comments and responded to each in detail below.

SUBSTANTIAL ISSUE
The blinded evaluation of review quality appears to suffer from "selection on the dependent variable" (i.e. quality). For the set of treatment reviews, the authors selected only those where the "proportion of incorporated feedback exceeded 0.60". If I'm understanding correctly, this isn't a random set of treatment reviews but a (random?) set of those that received feedback *AND* the reviewers found the feedback valuable enough to incorporate, i.e. they thought the feedback was relatively high quality. So, this comparison, unsurprisingly, goes on to show that reviews revised after high-quality feedback are perceived better than those without feedback, but it tells us almost nothing about feedback quality on average. At best it tells us that *sometimes* the feedback must have been good because it is associated with higher post-feedback quality. I say "at best" because even that more limited conclusion is not guaranteed: the treatment reviews you (non-randomly) selected may have been better *pre-feedback* than randomly selected control reviews. So, a more fair comparison is between a random sample of reviews from the control group vs. a random sample of reviews from the treatment group (with no other conditioning).

So, I'd strongly urge redoing this analysis with a more fair comparison. Another option is to just remove this quality evaluation component from the paper, although it would weaken it. Yet another option is to acknowledge that the comparison isn't very informative, and certainly not feature it in the abstract.

I would also like to see more detail in this section: the exact instructions given to the annotators, were the reviews rated independently or side-by-side, how many of each type, the experience level of the annotators, etc.

Thank you for this feedback. We have redone this experiment in line with the reviewer's comments, where now we compare a random sample of 50 reviews from the control group, which did not receive LLM feedback, and a random sample of 50 reviews from the feedback group. In this analysis, we compare initial and modified pre-rebuttal reviews. The only condition we impose is ensuring these two versions of a review are not identical; this is so we do not waste our annotators' time by telling them there is a difference when there actually is not. From our previous analysis, we know that the edit rate is quite low to begin with, and much lower in the control group compared to the feedback group (around a 17 percentage point difference).

Here are the instructions given to the annotators: For each pair of reviews, select which version is clearer, more specific, and more actionable for authors. Consider whether the review provides sufficient detail for authors to understand and address the concerns raised. Focus on the specificity of feedback, clarity of explanations, and how effectively the review guides authors toward concrete improvements. If you prefer review version A, select 'A'. If you prefer review version B, select 'B'. If the reviews are of equal quality (or there are minimal/insignificant differences between the two versions), you can select 'tie.'

Annotators were provided with a spreadsheet, where each row corresponded to one review. There was a column representing review version A, review version B, and then an empty column to put their preference. For each review, we shuffled whether A/B was the original or revised review. The annotators were experienced AI PhD students. We have included these instructions in Appendix F of our revised manuscript.

The annotation study revealed a clear preference for revised reviews in both groups, with the feedback group showing a substantially stronger effect. In the feedback group (n=50), annotators preferred the revised version in 68% of cases, the original version in 12% of cases, and rated them as tied in 20% of cases. This represents a strong preference for the post-feedback revisions. In contrast, the control group (n=50) showed a more modest preference pattern. Annotators preferred the revised version in 44% of cases, the original version in 8% of cases, and rated them as tied in 48% of cases. While revised reviews were still preferred when annotators detected differences, nearly half of the review pairs were judged as equivalent quality.

The difference in preference patterns between groups was substantial and statistically significant. Annotators preferred revised reviews from the feedback group 24 percentage points

more often than those from the control group (68% vs 44%, p = 0.027). Annotators were also 28 percentage points more likely to mark reviews from the control group as ties (48% vs 20%, p = 0.001). This demonstrates that feedback significantly increases both the rate and detectability of improvement, helping reviewers make better revisions that are substantial enough to be noticed by human evaluators. These results provide statistically significant evidence that feedback leads to more substantive improvements rather than superficial changes that go unnoticed.

We have added these new results and the details on how this evaluation was carried out to section 4.2.1 of the revised manuscript.

MEDIUM

1. The current Related Works section is a bit too focused on uncritical summary and is placed in an odd position. Putting it after the Introduction would be more typical. More importantly, I think this section could better motivate your research choices. The elephant in the room is that this study is designed around a relatively "easy" task --- identifying vague statements, those that are already addressed, and unprofessional ones. Notably absent are more intellectually challenging tasks like identifying weaknesses, making novel connections, providing constructive and non-obvious suggestions. This "elephant" is currently downplayed, as I believe there's just one sentence in the Discussion addressing it: "Additionally, it would be interesting to explore ... to generate more nuanced feedback for complex issues." In my view, this is not a weakness of the study but a strength --- you go after a task that's more achievable with current technology and still valuable. This (smart) research choice can flow from your Related Works section: you could give a even-handed/critical overview, noting that current research does not provide support for LLMs performing well on the more intellectual review tasks, and therefore you make the very reasonable move to test them on easier tasks. Without such a context, a reader is left asking, "Well, if LLMs are so great, why are the authors going after the relatively low-hanging fruit of 'vagueness' and 'unprofessionalism'"?

We thank the reviewer for this insightful feedback. We have substantially revised the related works section, which we moved to directly follow the introduction, to address these concerns:

“Previous research has found that evaluating peer reviews is a highly challenging task, even for human experts. One study found that human reviewers are often influenced by superficial cues such as review length or the associated paper score, and that subjectivity among human evaluators remains substantial [33]. These difficulties highlight why current LLMs are unlikely to succeed at the nuanced intellectual tasks central to generating high-quality peer reviews. Empirical evidence shows that LLM-generated reviews tend to be generic, frequently overlooking novelty, failing at balanced methodological critique, and diverging significantly from expert judgments [34, 35, 36]. Given these documented limitations of current LLMs in complex review evaluation, our study deliberately focuses on well-defined, tractable tasks: identifying vague statements, unprofessional language, and issues already addressed in the manuscript. These tasks leverage LLMs' strengths in text analysis and semantic understanding while avoiding the need for novel insight generation or deep domain expertise. This strategic choice

allows us to test whether LLMs can provide immediate value in areas where their current capabilities are better suited."

We reframed our related work to provide empirical context for LLM limitations in peer review tasks and used this evidence to justify our strategic focus on well-defined, tractable tasks like identifying vague statements, unprofessional language, and previously addressed issues. These areas leverage LLMs' text analysis strengths while avoiding complex evaluation requiring domain expertise.

The reviewer was correct that this represents a strength of our study. The revised section now frames our approach as a thoughtful response to current technological capabilities, emphasizing immediate practical value while acknowledging where LLMs are not yet ready to replace human expertise.

References:

33. Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B. Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments, 2024.
34. Burak Kocak, Mehmet Ruhi Onur, Seong Ho Park, Pascal Baltzer, and Matthias Dietzel. Ensuring peer review integrity in the era of large language models: A critical stocktaking of challenges, red flags, and recommendations. European Journal of Radiology Artificial Intelligence, 2:100018, 2025.
35. Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review, 2024.
36. Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. Mind the blind spots: A focus-level evaluation framework for llm reviews, 2025.

2. Going back to my "elephant in the room," it was interesting to see that the treatment caused only 0.6% more changes in review scores (relative to baseline). This paints a picture where reviewers incorporate the feedback to improve the tone/specificity of their reviews, but ultimately don't change their opinions/scores much and so the direction of science does not change much. In the current version, this key result (and the most plausible interpretation of it) is not in the abstract or the summary of findings but is "hidden" on page 8. I think the paper should be much more upfront about this result, i.e. abstract and summary of findings. If such results are downplayed, there's a very big risk that the more casual reader comes away with an overly hyped sense of LLM capabilities relative to what this study (very excellently) demonstrates. (And, in my view, the capabilities demonstrated in the paper are already substantial!)

This is a very insightful comment. We have revised the paper to address this concern by prominently featuring the minimal score change finding (0.6% increase) in the abstract as requested. We now state: "the intervention had minimal impact on final scores (only 0.6%

increase in score changes), suggesting that AI feedback primarily enhanced review quality rather than systematically changing the final score in one direction."

We also highlighted this finding in our summary of main findings on page 2. We agree that this demonstrates our system improves review quality and engagement without influencing scientific judgment, which is a strength of our approach that prevents misinterpretation of our results.

MINOR
1. In Section 3.4, it was odd to see feedback gets clustered into 5 *new* topics when it was, by design, generated around 3 topics. Why not just go with the original 3?

Thank you for this question. While our system was designed around 3 categories, we conducted additional analysis that revealed why clustering into 5 categories was more appropriate. We randomly sampled 5,000 feedback items and used LLM-based prompting to determine which of the three original categories each belonged to. We found that only 2 items addressed unprofessional comments and 49 addressed content misunderstandings, while ~99% were related to addressing vague or unjustified statements. We have updated section 4.4 of our revised manuscript to include these results.

This distribution occurred because our agent rarely commented on content misunderstandings - it had to be absolutely certain there was an error and provide a direct quote from the paper, as we did not tolerate hallucinations. Given that the vast majority of feedback fell into the "specificity" category, the 5-cluster analysis provides meaningful subcategories within this dominant type, offering more granular insights into the different ways our system helped reviewers improve specificity rather than forcing the sparse data from the other categories into an artificial framework.

I hope these comments help move this excellent project forward!

We thank the reviewer for their helpful suggestions!

Reviewer #2 (Remarks to the Author):

Summary:
This paper develops a Review Feedback Agent that provides feedback to human reviewers. Their study found that 26.6% of reviewers changed their reviews after reading the AI-generated feedback. The length of the reviews increased by 80 words, and these reviewers were also more engaged during the rebuttal phase.

Strengths:
1. The tool has significant social impact, and the study is based on a real-world use case; therefore, the results effectively demonstrate the usefulness of the tool.
2. The result analysis is thorough and insightful.

We sincerely thank the reviewer for their comprehensive summary of our work! We appreciate their recognition of our system's 'significant social impact' and their praise that our 'result analysis is thorough and insightful.' We have carefully reviewed their insightful comments and responded to each in detail below.

Weaknesses:
1. My major concern is that the study only includes ICLR conference papers and reviews. Since it is a heavily engineered system, I wonder how well it generalizes to other conferences and what would need to be changed to adapt it.

Thank you for this comment. We focused our experiment on ICLR because it is a large and popular AI conference that covers a diverse range of research topics. Due to the considerable time and cost in setting up the partnership with a major conference and executing a study with tens of thousands of reviews, it was not feasible for us to conduct the same experiment with multiple conferences. ICLR's setup is representative of most AI conferences because it uses the same core workflow: paper submission, assignment to multiple reviewers, author rebuttal, meta-review by area chairs, and final decisions by program chairs. This structure is nearly identical across major AI venues like NeurIPS and ICML, making ICLR's format typical in how research is evaluated and selected.

Our dataset encompasses a substantial variety within ICLR's diverse research landscape, spanning applications (26.4% of papers), foundation models and generative AI (20.1%), responsible AI and infrastructure (20.9%), learning methods (16.4%), and theory and optimization (11.3%). The 20,000+ reviews we gave feedback to represent diverse writing styles, technical approaches, and reviewer backgrounds from institutions worldwide. ICLR's open nature reduced ethical concerns about sharing review data with LLMs and enabled the large-scale validation that makes it the logical starting point for such research.

As shown in our analysis across five major AI subfields (see Supplementary Figures S1-3), our system demonstrates consistent effectiveness regardless of research area. These fields were generated by first looking at the distribution of all submissions across the 21 primary areas. Then, we grouped relevant categories such that we had 5 main categories. The categories are:

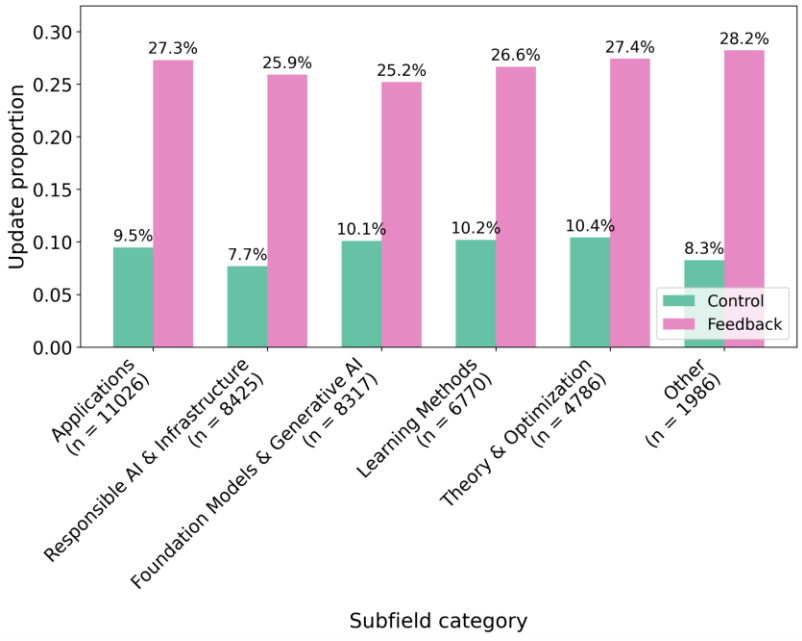
1. Applications (26.4% of papers) - Computer vision, NLP, robotics, physical sciences, etc.
    a. 2,062 feedback vs 2,000 control papers
2. Foundation Models & Generative AI (20.1%) - LLMs, foundation models, generative models
    a. 1,586 feedback vs 1,625 control papers
3. Responsible AI & Infrastructure (20.9%) - Safety, fairness, interpretability, datasets, infrastructure
    a. 1,541 feedback vs 1,541 control papers
4. Learning Methods (16.4%) - Representation learning, RL, transfer learning, etc.
    a. 1,240 feedback vs 1,277 control papers

5. Theory & Optimization (11.3%) - Learning theory, optimization, probabilistic methods
    a. 881 feedback vs 853 control papers

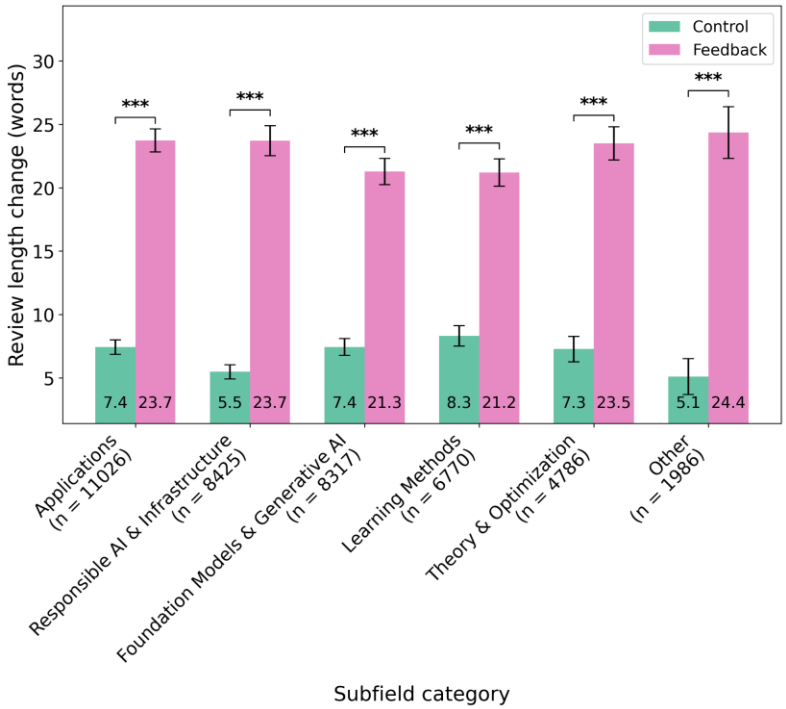
We combined all other papers that belonged to a category not included above in the 'Other' category. We see that feedback consistently produces significant increases in update rates (Supplementary Figure S1) and review length (Supplementary Figure S2) across all subfields. We also observe that incorporation rates remain stable (63-69%) across all subfields (Supplementary Figure S3). This cross-subfield consistency suggests that the underlying reviewer challenges (vague comments, content misunderstandings, and unprofessional remarks) are universal rather than domain-specific. We have added all of these results to Appendix C of our revised manuscript.

While our system is heavily engineered, much of this complexity represents transferable infrastructure rather than venue-specific customization. The multi-LLM pipeline architecture and reliability testing framework, as well as our thorough analysis, would transfer directly to other venues. Most major CS conferences share the fundamental elements our system leverages: organized review sections, author rebuttals, and reviewer-author dialogue. Since most AI conferences use OpenReview, technical deployment would be seamless.

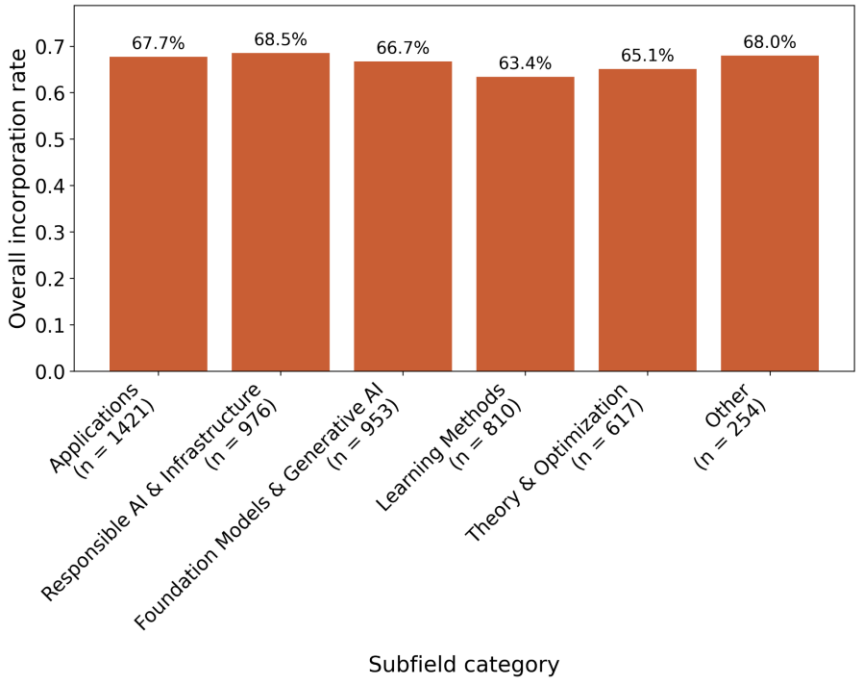
While implementation at other venues may require domain-specific prompt adaptation and field terminology updates, this work establishes the first large-scale evidence for LLM-enhanced peer review with validated methodology and reusable infrastructure for extension to other academic venues.



Supplementary Figure S1: Review update proportions (based on the analysis in Figure 3B) are consistent when stratified by subfield category. In other words, reviewers are as likely to update their review regardless of what category the paper falls into. The n values represent the number of reviews for papers categorized within each specific subfield.



Supplementary Figure S2: Based on the analysis in Figure 3C, feedback consistently results in similar review length increases across all subfields. This trend is replicated in the control group as well. The n values represent the number of reviews for papers categorized within each specific subfield.



Supplementary Figure S3: Incorporation rates remain stable (63-69%) across all subfields (based on the analysis in Figure 4A). This suggests that reviewers were as likely to incorporate

feedback into their review regardless of what subfield category the paper fell into. The n values represent the number of updated reviews for papers categorized within each specific subfield.

2. The Critic module and Reality Check module are still LLMs-based prompts. Are there any human studies evaluating the quality of the feedback, or surveys asking reviewers how useful they found the feedback?

Thank you for this question. In line with comments we received from another reviewer, we re-conducted our human study, which was a blind preference evaluation of review quality. We compared a random sample of 50 reviews from the control group, which did not receive LLM feedback, and a random sample of 50 reviews from the feedback group. In this analysis, we compared initial and modified pre-rebuttal reviews. We provide the exact details of this study in section 4.2.1 of the revised manuscript, and we also provide the instructions given to the annotators in Appendix F.

The annotation study revealed a clear preference for revised reviews in both groups, with the feedback group showing a substantially stronger effect. In the feedback group (n=50), annotators preferred the revised version in 68% of cases, the original version in 12% of cases, and rated them as tied in 20% of cases. This represents a strong preference for the post-feedback revisions. In contrast, the control group (n=50) showed a more modest preference pattern. Annotators preferred the revised version in 44% of cases, the original version in 8% of cases, and rated them as tied in 48% of cases. While revised reviews were still preferred when annotators detected differences, nearly half of the review pairs were judged as equivalent quality.

The difference in preference patterns between groups was substantial and statistically significant. Annotators preferred revised reviews from the feedback group 24 percentage points more often than those from the control group (68% vs 44%, $p = 0.027$). Annotators were also 28 percentage points more likely to mark reviews from the control group as ties (48% vs 20%, $p = 0.001$). This demonstrates that feedback significantly increases both the rate and detectability of improvement, helping reviewers make better revisions that are substantial enough to be noticed by human evaluators. These results provide statistically significant evidence that feedback leads to more substantive improvements rather than superficial changes that go unnoticed. We have added these new results and the details on how this evaluation was carried out to section 4.2.1 of the revised manuscript.

While this experiment measures review quality rather than feedback quality directly, we use review quality as a proxy for feedback quality. The substantial improvements in review quality observed in the feedback group compared to the control group indicate that the feedback provided was of high quality, as poor-quality feedback would likely not have driven such meaningful improvements.

3. There are more details or discussion can be included in the paper to make the study more comprehensive and insightful, please check the questions/suggestions.

Thank you for the questions and suggestions! We address each question individually below.

Questions/Suggestions:
1. Are there any cases where the two Actors do not agree with each other? What is the percentage, and how are these cases handled?

We would like to clarify that the Actors could not explicitly disagree with each other. The Actors simply produced initial lists of feedback regarding the same feedback. For all our LLMs, we used a high temperature to encourage more diverse outputs. If one Actor provided feedback on a review comment and the other didn't, that can be interpreted as a disagreement, as they clearly had different opinions about the quality of the specific review comment. However, that disagreement would not be flagged by the system as the Aggregator module would simply combine these lists and remove any duplicate cases (times the Actors agreed). The Critic module would remove any feedback it felt was unnecessary or incorrect, in other words, times it disagreed with the Actors.

From our development testing, we found that the two Actors provided different initial feedback on 15 out of 50 reviews (30%). However, these disagreements were minimal in scope, averaging only 0.5 different comments per review and representing just 5.6% of all feedback comments. The vast majority (70%) of reviews received identical feedback from both Actors, with 14 reviews having only 1-2 different comments and just 1 review having more substantial differences. We hope this clarifies the reviewer's understanding and answers their question.

2. How can more reviewers be encouraged to update their reviews based on the feedback?

This is a great question, and one we extensively discussed amongst ourselves while planning the experimental setup. While we decided that we wanted to alter the review process as little as possible so we could most accurately understand how this singular review feedback intervention impacted the review process, there were many ideas we discussed.

Our data showed that reviewers who submitted early were significantly more likely to update their reviews, suggesting that encouraging earlier submissions could naturally increase responsiveness by providing more time for reflection and revision. We also found higher incorporation rates when reviewers received fewer, more targeted feedback items rather than comprehensive lists. This suggests that we should focus on giving only the most impactful suggestions and maintaining rigorous quality standards through reliability testing.

Conference organizers could implement recognition for reviewers who demonstrate engagement with feedback, or provide reviewers with statistics showing how their updated reviews changed (for example, review length metrics, as our study found an average 80-word increase among those who updated). Since our system was optional, making feedback incorporation more explicitly valued in reviewer guidelines and sharing anonymized examples of successful improvements could demonstrate value to skeptical reviewers.

The 27% voluntary update rate we observed represents meaningful behavioral change, but these approaches could potentially increase engagement while maintaining the quality improvements we saw in review specificity and author-reviewer dialogue.

3. The clustering experiments are interesting. I can see that the clustering is more fine-grained compared to the three-target feedback in the Actor module prompt. I still wonder about the percentage of the three-target feedback.

Thank you for the suggestion. We have updated the paper to include information about the percentage of the three-target feedback. While our system was designed around 3 categories, we conducted additional analysis that revealed why clustering into 5 categories was more appropriate. We randomly sampled 5,000 feedback items and used LLM-based prompting to determine which of the three original categories each belonged to. We found that only 2 items addressed unprofessional comments and 49 addressed content misunderstandings, while ~99% were related to addressing vague or unjustified statements. We have updated section 4.4 of our revised manuscript to include these results.

This distribution occurred because our agent rarely commented on content misunderstandings - it had to be absolutely certain there was an error and provide a direct quote from the paper, as we did not tolerate hallucinations. Given that the vast majority of feedback fell into the "specificity" category, the 5-cluster analysis provides meaningful subcategories within this dominant type, offering more granular insights into the different ways our system helped reviewers improve specificity rather than forcing the sparse data from the other categories into an artificial framework.

4. In my opinion, the "Addressing Misunderstandings" feedback is especially powerful. But I wonder if LLMs are intelligent enough to generate such feedback, and whether simple few-shot prompt engineering is sufficient.

We appreciate this comment and agree with the reviewer. At the time when we ran this experiment, we found that LLMs were often not confident enough to generate such feedback hallucination-free, especially when only given simple few-shot prompts. Therefore, our system rarely gave feedback in that category (~1% of feedback given), even though we agree that this category of feedback has the potential to be the most important and impactful in improving review quality. This is a very clear next step we would like to take as LLMs and prompting techniques have already improved drastically over the past year, and we anticipate will continue to improve over time.

5. An ablation study where the system is only given the review (without the paper PDF) and the resulting differences in feedback would reveal how much the LLM relies on the paper's content. This could serve as an indirect indicator of the LLM's trustworthiness.

We appreciate this suggestion. For two of our three feedback categories (improving specificity and addressing unprofessional remarks), the system primarily analyzes review text patterns and does not require paper content. However, our system only provides "misunderstanding" feedback when it can identify specific paper content that addresses the reviewer's concern, and we require verbatim quotes from the paper using exact quote tags. We validated this approach extensively on our development set to ensure zero hallucination; when the system references paper content, those references are always accurate. Without access to the paper PDF, the system simply cannot generate this type of feedback, as it would have no basis for claiming that reviewer concerns are addressed elsewhere in the manuscript.

As we noted in our clustering analysis, only ~1% of feedback addressed misunderstandings, likely because our stringent requirements (absolute certainty plus verbatim quotes) caused the model to err on the side of caution. An ablation removing paper access would eliminate this small but important category of feedback, rather than revealing differences in LLM trustworthiness.

6. What are the limitations of the current feedback system? Is there room for further improvement?

We thank the reviewer for prompting this important discussion. There are several limitations to our current system. Most significantly, our conservative design approach may have been overly restrictive. We implemented stringent requirements to avoid hallucinations, particularly requiring absolute certainty and verbatim quotes for addressing misunderstandings. This resulted in only ~1% of feedback addressing content misunderstandings, despite this category having the greatest potential impact, as the reviewer astutely mentioned. This suggests we may have missed opportunities to catch genuine reviewer oversights that could have substantially improved review quality.

Right now, the system only addresses three types of review issues: vague comments, content misunderstandings, and unprofessional remarks. This does not capture all aspects of review quality that could benefit from feedback. Additionally, our reliability testing approach, while preventing errors, may have created an overly conservative system that limited higher-impact interventions.

Over the course of the past year, since we ran this experiment, LLM capabilities have drastically improved. Reasoning models such as GPT-o1 didn't come out until after we had finalized our system for deployment. We anticipate that many gains can be achieved by testing more advanced models that could better handle the nuanced task of identifying content misunderstandings without hallucinations.

Finally, running this system across diverse AI conferences would certainly help us assess its robustness and effectiveness across different domains, and we need to evaluate whether our approach would generalize beyond AI to other fields with different reviewing cultures. We have updated our discussion to better reflect these limitations and areas for improvement.

Reviewer #2 (Remarks on code availability):

N/A

Manuscript NATMACHINTELL-A25041472A-Z
Response to reviewers

Overall comments to reviewers

We sincerely thank the reviewers for their thoughtful feedback on our revised manuscript. We are pleased that Reviewer #1 acknowledged our "great work" addressing their initial comments and that Reviewer #3 was "content with the current manuscript." We have carefully addressed all remaining concerns, and all major revisions are detailed below. Please see below, in blue, for a point-by-point response to the reviewers' comments and concerns.

Summary of Major Revisions

1. We clarified the setup of the blinded evaluation to specify that the reviewers are those who chose to revise their initial review, and removed unfounded causal statements (see section 3.2.1 of the revised manuscript).
    a. We also changed "review quality" to "review clarity, specificity, and actionability" throughout the analysis to accurately reflect what was measured.
2. In the abstract, we removed conclusions about impact on scores that are not fully supported by the results.
3. We specified which system prompt elements need modification for other venues (see Supplementary Note 4).
4. We added recommendations for a dedicated review revision phase to give all reviewers adequate time to incorporate feedback (see Discussion).

Reviewer #1:
Remarks to the Author:
Great work by the authors in addressing the first round comments.

We sincerely appreciate the reviewer's recognition of our revisions!

Here are some outstanding issues:

COMPARISONS
However, the blind evaluation piece is still not quite the right comparison, in my view. In the first round, I was proposing you evaluate a sample of final reviews (whether revised or not) taken from each condition. That would let you straightforwardly measure the treatment effect of feedback on quality, which, in the big picture, is what we ultimately want the feedback to do. This would give us an intent-to-treat effect. I suppose another approach would be to use instrumental variables, with condition as the instrument, feedback-taking as the treatment, and quality as the outcome; this would give us the LATE effect.

All the other comparisons involve selection bias and cannot be straightforwardly interpreted. For example, in the current set-up you're conditioning on *revising* in both groups. That is a *choice* reviewers made and, indeed, we know that the choice was made much more often in

the feedback condition. So, the initial randomization gave you two very similar groups, but after you condition on revising, we lose any guarantees that the groups are similar and cannot make causal statements.
So, I again would urge either doing a clean test, with no extra conditioning, or interpreting these results very tentatively, something like, “Among those choosing to revise their initial reviews, the increase in quality was higher for those who received feedback.”

We appreciate the reviewer's careful attention to the interpretation of our blind evaluation. We agree that we need to clarify the wording of how we present these results. We have updated the wording in the final paragraph of section 3.2.1 to the following:

"The difference in preference patterns between groups was substantial and statistically significant. Among reviewers who chose to revise their initial reviews, annotators preferred revised reviews from the feedback group 24 percentage points more often than those from the control group (68% vs 44%, p = 2.7e-2). Annotators were also 28 percentage points more likely to mark reviews from the control group as ties (48% vs 20%, p = 1.0e-3). This demonstrates that among reviewers who chose to revise their initial review, feedback significantly increases both the rate and detectability of improvement, helping reviewers make better revisions that are substantial enough to be noticed by human evaluators. These results provide complementary evidence that feedback not only increases the rates of revision (Figure 1b), but also improves the substance and detectability of those revisions when reviewers choose to make them."

QUALITY VS. ACTIONABILITY
Thanks also for showing the instructions. After seeing these, I don’t think we should call the outcome “review quality.” Instead, the construct is arguably one dimension of quality, perhaps something like, “Clarity” or “Actionability”.

We thank the reviewer for catching this misleading phrasing. We agree, and have updated the text related to this analysis (section 3.2.1 and Supplementary Note 2) to explicitly say ‘review clarity, specificity, and actionability’ rather than just ‘review quality.’

INTERPRETATION AND ABSTRACT
The revised part of the abstract is too strong an interpretation of the data:
“This suggests that many reviewers found the AI-generated feedback sufficiently helpful to merit updating their reviews, although the intervention had minimal impact on final scores (only 0.6% increase in score changes), suggesting that AI feedback primarily enhanced review quality rather than systematically changing the final score in one direction.” The last part may be true, or not; we don’t quite know whether reviewers perceived review quality (rather, actionability/clarity) to be enhanced, or these are demand effects because participants didn’t want to disappoint the experimenters who clearly put a lot of time into engineering this system, etc. If you remove the last part of the sentence then it’s more defensible and fine by me: “This suggests that many reviewers found the AI-generated feedback sufficiently helpful to merit

updating their reviews, although the intervention had minimal impact on final scores (only 0.6% increase in score changes)."

Thank you for this feedback. We agree with the reviewer and appreciate their insight. We have removed the last part of the sentence in the abstract in line with the reviewer's request.

Reviewer #3:
Remarks to the Author:
1.My original weakness 1 concerns generalization. The authors state that the review pipeline is almost identical across other venues, therefore they expect their method to apply to other venues as well. Meanwhile, the authors also acknowledge that the system prompt will need to be adapted to specific domains. The reviewer agrees with their first point. Regarding the second point, the reviewer suggests that in the supplementary material the authors highlight which parts of the prompts should be modified when applying the method to other venues.

We appreciate this constructive suggestion. We agree that providing explicit guidance on prompt adaptation would facilitate adoption at other venues. We have added the following paragraph to the end of Supplementary Note 4:

"To adapt these system prompts for use in other conference venues, we recommend considering a few key points. First, our prompts explicitly reference ICLR's standardized review sections (summary, strengths, weaknesses, and questions) and scoring system, which may differ at other venues and therefore need to be updated. Second, the in-context examples we provide reflect ICLR-style machine learning papers and typical ML review concerns (e.g., experimental baselines, novelty claims, methodological details); for non-ML venues, these examples could be tailored to that field's typical concerns and conventions. Finally, while our Actor system prompt references common issues identified in AI conference reviewer guidelines and would likely generalize, incorporating venue-specific reviewer guidelines could further improve alignment with that venue's quality standards and expectations."

2.For all the weakness, questions and suggestions the reviewer raised, the authors have provided either experiments or explanations and discussions. I believe these responses are sufficient. One additional proposal the reviewer has regarding "How can more reviewers be encouraged to update their reviews based on the feedback?" is the following. Based on the authors' analysis that late submission reviewers are more likely to submit revisions than those who submit their reviews close to the deadline, possibly because the latter do not have enough time to revise, would it be helpful to integrate a "review revision phase" after the initial review submission in the submission review pipeline? Maybe could be added as part of the discussion.

We thank the reviewer for acknowledging our thorough effort to address all of their questions and suggestions. This is a wonderful insight, and we have added the following to the 3rd paragraph of our discussion:

“Finally, from a practical implementation perspective, we noticed that reviewers who submitted early were more likely to revise their review than those who submitted late. Therefore, it could be helpful to introduce a dedicated review revision phase after the initial review submission, allowing all reviewers to have adequate time to update their review based on feedback if desired.”

In summary, the reviewer is content with the current manuscript.

We sincerely appreciate the reviewer's recognition of our revisions!