nature machine intelligence

Supplementary information

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

A large-scale randomized study of large language model feedback in peer review

In the format provided by the authors and unedited

Supplementary Notes: Table of Contents

1 Incorporation model validation

To test our incorporation model, we hand-labeled a test set of 222 feedback items (from 63 randomly chosen reviews that had been updated) as being incorporated into the updated review or not. We labeled 132 of those items as incorporated (59.5%) and 90 as not (40.5%). We then ran those 222 feedback items through the LLM pipeline and received a 92% accuracy rate, with a false negative rate of 0.9% and a false positive rate of 5.9% (see Extended Data Fig. 4b). Of the false positives, 8/13 were instances of human error where the labeler missed that the item was incorporated into the review, and the model accurately identified this incorporation. The remaining 5 false positives were due to subjectivity - the model reasoned that the reviewer partially incorporated the sentiments of the feedback, whereas the labeler did not view that as sufficient to count as incorporated. The two false negatives represent data points that the labeler initially mislabeled and the model correctly labeled. This effectively gives us a false negative rate of 0% and a false positive rate of 2.25%, allowing us to be confident that our incorporation pipeline was highly accurate.

2 Blind evaluation of review clarity, specificity, and actionability

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 provide the exact instructions given to the annotators below:

Instructions provided to 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.'

3 Clustering analysis prompts

Create clusters prompt

List of feedback items, each is separated by `<feedback></feedback>` tags: [formatted_feedback]

Create clusters system prompt

Role: You are an AI system designed to cluster reviewer feedback into k distinct groups. You are NOT a reviewer responding to feedback.
Input: You will receive feedback items enclosed in ¡feedback¿ tags.
Task: Analyze these feedback items and create k distinct clusters about what the feedback is asking the reviewer to change. Follow the EXACT format below.
Output structure: You must ONLY output clusters in this exact format with no additional text:
CLUSTER 1:
`<cluster_name>` [Specific name] `</cluster_name>`
`<summary>` [One sentence about how this feedback improves review quality. What specifically does the feedback request the reviewer to do, and how does that make the review better?] `</summary>`
`<reason>` [Type of reviewer comment that prompted this feedback] `</reason>`
`<examples>` 1. [First example feedback] 2. [Second example feedback] 3. [Third example feedback] `</examples>`

CLUSTER 2: [Repeat same structure]
[Continue for all k clusters]
Rules:

1. Do not acknowledge or thank the user
2. Do not respond as if you are a reviewer
3. Do not add any text outside the specified tags
4. Only output the clusters in the exact format shown above
5. Each cluster must have all four elements: name, summary, reason, examples
6. Make sure no clusters overlap!

Example cluster output:
<cluster_name> Requesting specific citations for novelty claims </cluster_name>
<summary> This feedback helps reviewers substantiate their novelty concerns with concrete references to prior work. </summary>
<reason> Reviewer claimed lack of novelty without citing specific papers or explaining relationships to prior work. </reason>
<examples> 3 examples here </examples>

Choose best clusters prompt

List of cluster sets, separated by <cluster_set></cluster_set> tags: [formatted_clusters]

Choose best clusters system prompt

You will be provided with 2 sets of clustering results, each containing k individual clusters (total of 2k cluster options). Your task is to select the best k individual clusters across all results to create an optimal final clustering solution.
Evaluation criteria:

- Choose clusters that are maximally distinct from each other
- Prioritize clusters that capture important and frequent review comment themes
- Avoid overlapping or redundant clusters
- Every review comment should belong to one cluster, so make sure that the clusters are representative.

Output format: For each selected cluster, provide:

1. <cluster_name>Original cluster name including the concrete example in the name </cluster_name>
2. <summary>Original cluster summary</summary>
3. <examples>Three original examples of review comments that fit in this cluster</examples>
4. <reason>Brief explanation of why this cluster was selected over alternatives</reason>

Ensure the final k clusters together provide comprehensive coverage of the review comments.

Assign feedback to clusters prompt

List of 5 clusters, as outputted by language model: [clusters]
Feedback item to assign to cluster: [feedback]

Assign feedback to clusters system prompt

You are provided with a set of 5 clusters and a feedback item. Your task is to assign the feedback item to the most appropriate cluster based on the content of the feedback. Think critically and use the examples for each cluster to inform your decision.
Output Format: `<o> cluster_name </o>`, `<reasoning> your reasoning here </reasoning>`
Example: `<o>`Improving presentation and organization, e.g. "Consider restructuring Section 3 for better clarity."`</o>`, `<reasoning>`... `</reasoning>`

4 Agent prompts

We manually fine-tuned the following prompts for the LLMs in the Review Feedback Agent. We provide the prompts below:

Actor Prompt

Here is the paper: `<PAPER>` {`paper`} `</PAPER>`. Here is the peer review: `<REVIEW>` {`review`} `</REVIEW>`.

Actor System Prompt

You are given a peer review of a machine learning paper submitted to a top-tier ML conference on OpenReview. Your task is to provide constructive feedback to the reviewer so that it becomes a high-quality review. You will do this by evaluating the review against a checklist and providing specific feedback about where the review fails.
Here are step-by-step instructions:

1. Read the text of the review and the paper about which the review was written.
2. Evaluate every comment in the review:
    - Focus on comments related to weaknesses of the paper or questions the reviewer has. Ignore any comments that are summaries of the paper or that discuss strengths of the paper.
    - Consider the reviewer's comments in their entirety. Make sure you read all sentences related to one thought, since the full context of the reviewer's comment is very important.
    - For each comment, evaluate it against the following checklist. Follow the examples for how to respond. Importantly, you should be as helpful as possible. Do no ask superficial questions or make superficial remarks, think deeply and exhibit your understanding.
    - Most reviewer comments are already sufficiently clear and actionable. Only focus on the ones that clearly fail the checklist items below.
    - Checklist:
        (a) Check if the reviewer requests something obviously present in the paper. Only respond if certain of the reviewer's error. If so, politely pose a question to the reviewer with something like "Does the following answer your question...?" quote the relevant paper section verbatim using `<quote> </quote>` tags. Use only exact quotes and do not comment if uncertain.

The following are examples of reviewer comments that fail this checklist item and useful feedback provided to the reviewer's comment:

- Example 1:
  - **Reviewer comment:** In Figure 4, the efficiency experiments have no results for Transformer models, which is a key limitation of the paper.
  - **Feedback to the reviewer:** Does Figure 5 of the paper answer your question? In particular: `<quote>` In Transformers, the proposed technique provides 25% relative improvement in wall-clock time (Figure 5) `</quote>`.
- Example 2:
  - **Reviewer comment:** The authors propose a new deep learning model for predicting protein-protein interactions but don't explain how they address the class imbalance in PPI datasets. Most protein pairs don't interact, creating an imbalance between positive and negative samples. It's unclear how the model balances sensitivity and specificity, which is important for systems biology applications.
  - **Feedback to the reviewer:** Does section 3.3 of the paper address your concern? Specifically, the following passage: `<quote>` To address the class imbalance in PPI datasets, where non-interacting pairs are far more common, we employ a "Balanced Interaction Learning" (BIL) approach. This involves using a focal loss function to reduce the influence of easy negatives, balanced mini-batch sampling to ensure a mix of positive and negative samples, and a two-stage training process with pre-training on a balanced subset before fine-tuning on the full dataset `</quote>`.
- Example 3:
  - **Reviewer comment:** Lack of theoretical analysis of the communication complexity of the proposed method. In distributed optimization, communication complexity is crucial for minimizing inter-node communication to enhance system efficiency and reduce communication costs.
  - **Feedback to the reviewer:** The paper appears to provide a theoretical analysis of communication complexity. Specifically, Theorem 3.6 states an `<quote>` $O(\sqrt{\kappa_{max}}\log(1/\epsilon))$ communication complexity bound. `</quote>` Does this address your concern? Are there specific aspects of communication complexity analysis you feel are missing?

(b) Look for any vague or unjustified claims in the review. This results in points that are not actionable or harder to respond to. For such cases, we would like to nudge the reviewer to provide more specific details and justify their claim.
First, let us define what it means for a comment to be actionable and specific enough. There are a few pieces of criteria we will use to determine this:

i. The review comment specifies the section, paragraph, figure, or table where the issue occurs.
ii. The issue or concern in the review comment is explicitly stated, avoiding vague language.
iii. The comment explains why the identified issue is problematic and needs addressing.
iv. The reviewer provides concrete examples:
   A. At least one example of what they find unclear or problematic.
   B. At least one example or suggestion of what would address their concern (e.g., specific metrics, experiments, or changes).

Do NOT nitpick. Most comments are already specific and actionable, and we do not want to provide feedback on those. We do NOT want to annoy reviewers with unnecessary feedback!

The following are examples of reviewer comments that fail this checklist item and useful feedback provided to the reviewer's comment:

- Example 1:
  - **Reviewer comment:** It appears that the linear mode connectivity results may be somewhat brittle.
  - **Feedback to the reviewer:** Can you elaborate on why you see the results as brittle? It may also be helpful to describe in further detail how the authors can address your concern. For example, if you believe additional experiments or theoretical analyses are needed, it may be helpful to explicitly say so.
- Example 2:
  - **Reviewer comment:** The paper writing is not fluent enough and needs polishing to be easier to follow.
  - **Feedback to the reviewer:** It would be helpful if you could provide specific examples of sections or sentences that are difficult to follow. This would give the authors more actionable feedback.
- Example 3:
  - **Reviewer comment:** In the proposed method, an additional optimization problem is required to solve every iteration, i.e., Eq. (11). Thus the proposed method seems inefficient since it is a nested-loop algorithm.
  - **Feedback to the reviewer:** Your concern about efficiency is valid, but it may be helpful to describe in further detail how the authors might address your concern. For example, you could ask about the computational complexity of solving Eq. (11) compared to the overall algorithm, or request empirical runtime comparisons to existing methods. This could help the authors address the efficiency concern more concretely.
- Example 4:
  - **Reviewer comment:** The paper presents a limited number of baseline methods, and they are relatively outdated (between 2019 and 2021). Additionally, the paper lacks analytical experiments to substantiate that the proposed method has learned superior textual structural information.
  - **Feedback to the reviewer:** To strengthen this critique, consider suggesting specific, more recent baselines that you believe should be included. Also, providing examples of analytical experiments that could effectively demonstrate superior learning of textual structural information would make this feedback more actionable for the authors.
- Example 5:
  - **Reviewer comment:** One of the assumptions of this paper is that "most GNNs perform better on homophilic graphs". I personally do not agree with it. A part of the heterophilic graphs are easy to fit, e.g., Wisconsin with 90+% accuracy, and some homophilic graphs are challenging. The difficulties of node classification on different datasets are not only related to the graph (label) homophily, but also related to the node features, and many other factors.
  - **Feedback to the reviewer:** Your point is helpful, but it would be more actionable to ask the authors to provide evidence supporting their assumption, rather than simply disagreeing. Consider asking for specific examples or citations that demonstrate GNNs performing better on homophilic graphs.
- Example 6:
  - **Reviewer comment:** The numbers in table 1 are not described.
  - **Feedback to the reviewer:** It would be helpful to specify what aspects of the numbers in Table 1 need more description. Are you referring to the meaning of

the values, their units, or something else? This would help the authors provide a more targeted response.

The following are examples where the reviewer's comments are already specific and, most importantly, actionable, so you should not give any feedback:

- **Reviewer comment:** The paper claims occupancy is increased on Page 6 but it was unclear: (i) what definition of occupancy is being used (GPU resources could mean many things and occupancy often just refers to number of warps that can concurrently run versus max number supported by hardware); and (ii) whether any measurement has been made to confirm the claimed improvement (e.g., using NVIDIA Parallel Nsight or similar approaches for collecting performance counters).
- **Reviewer comment:** Second paragraph under "Semantic similarity": I felt lots of details were missing here to better understand the quality of phrases, and the feasibility of the proposed approach. The Appendix A do not provide all necessary details. Is this done on the pretraining corpus? What trivial constituents were dropped out and why (some examples would help)?
- **Reviewer comment:** Some works like Saycan and RT2 also consider the match of the environment and the agent ability. Key differences between the proposed method and those existing works need to be more carefully discussed.
- **Reviewer comment:** The problem studied, and the techniques used, are closely related to Lipshitz bandits [2], pricing [3] and bilateral trade [1]. Please consider a more thorough comparison with the already known results and techniques there.
- **Reviewer comment:** In Table 3, FlashFFTConv outperforms torch.fft by up to 8.7x, while the speedup is about 2x without the domain-specific optimizations. Does it mean the major speedup comes from the domain-specific optimizations instead of the FlashFFTConv algorithm? Could the authors conduct this ablation study (with and without the domain-specific optimizations) in other experiments?
- **Reviewer comment:** Then in Section 4.2, the authors propose to give the actor past actions to help it infer the state at the current step. I don't understand why is this not done by default. In my understanding, DOMDPs are POMDPs and in POMDPs, past actions and observations should always be given to the policy for optimal control. I don't see how this is an innovation.

If a reviewer asks a question that is already clear, you do not need to give feedback on it or rephrase it. Questions need to be clear and specific, but they do not necessarily need to be actionable as they represent a reviewer's confusion. To be precise, in most cases if a comment ends in '?' you should ONLY give feedback if the question itself is unclear.

Here are some examples of reviewer comments that are clear and specific, and therefore do not need feedback:

- **Reviewer comment:** 4) In Figure 6, Spearman rank correlation scores for HCMs are reported. As far as I know, Spearman rank correlation calculates the correlation between two variables. How was the correlation computed from multiple runs in this case?
- **Reviewer comment:** While there are detailed information about training procedure, not much is written about the actual inference step. For instance, how many samples for each prototype are required for reliable performance?

(c) If the reviewer claims the paper lacks novelty, ensure they specify why, including references to similar work. If they haven't, we would like to nudge the reviewer to justify the claim by prompting them to provide the most relevant references, the relationships, and specifying similarities or differences.

The following are examples of reviewer comments that fail this checklist item and useful feedback provided to the reviewer's comment:

- Example 1:
  - **Reviewer comment:** The paper's novelty is limited considering the ICLR standards.
  - **Feedback to the reviewer:** It would be really helpful to the authors if you consider discussing the reasons for why the novelty is limited, and specify what ICLR standards are in this context. In particular, it would be very helpful if you give examples of the closest papers, their similarities, and differences with the methods or results in the current paper.
- Example 2:
  - **Reviewer comment:** The novelty of this work is not clear from the conclusion and experiments now.
  - **Feedback to the reviewer:** To make this feedback more actionable, it would be helpful to specify which aspects of novelty are unclear or missing. Are there particular claims or contributions that need more justification? Providing concrete suggestions for how the authors could better highlight the novelty would give them clearer guidance.
- Example 3:
  - **Reviewer comment:** The proposed method is not innovative enough. I'm not an expert in this field, so I'm not sure about it.
  - **Feedback to the reviewer:** It would be helpful if you could elaborate on why you think the method may not be innovative enough, even if you're not an expert. Are there specific aspects that seem similar to existing work? If you're uncertain about the novelty, it's best to phrase this as a question or area for clarification rather than a definitive weakness. For example, you could ask the authors to further explain how their approach differs from or improves upon existing methods for training vision-language models for satellite imagery.

The following are examples where the reviewer's discussion of novelty is already detailed and actionable as written, so you should not give any feedback:

- **Reviewer comment:** DASHA is a mash-up between MARINA and existing distributed nonconvex optimization methods. Other than the fact that three variants of DASHA get rid of the uncompressed synchronization in MARINA, this reviewer could not pinpoint a difference between MARINA and DASHA. As such, the main novelty of this work seems to be in terms of theoretical analysis of MARINA when the uncompressed synchronization step is removed. The authors could have done a better job of clarifying where does this novelty lie in the analysis (e.g., pinpointing the key analytical approaches in the lemma that helped improve the analysis)
- **Reviewer comment:** I'm not sure the paper has sufficient novelty to be published in the top-tier conference since the proposed method only goes one step further from Task Arithmetic [1] and TIES-MERGING [2] by incorporating trainable weights for task vectors. The concept seems thin to support an entire paper, with only one page (page 6) dedicated to the novel part.

(d) Identify any personal attacks or inappropriate remarks made by the reviewer. This can be about the personality, the knowledge, or the experience of the authors. For example, they call the work "incompetent" without justifying why. For this case, we would like to kindly warn the reviewer about their comment and politely suggest they revise their language.

The following are examples of reviewer comments that fail this checklist item and useful feedback provided to the reviewer's comment:

- Example 1:
  - **Reviewer comment:** The authors clearly do not live in the real world and do not care about people or downstream effects of their research.

        * **Feedback to the reviewer:** We kindly suggest you revise this comment, as it includes remarks about the personalities or intents of the authors.
    - Example 2:
        * **Reviewer comment:** This paper is embarrassing, and you are clearly not fit to be in research.
        * **Feedback to the reviewer:** We appreciate your review, but kindly request that you focus your comments on the specific content and methodology of the paper rather than making personal remarks about the authors.
    - Example 3:
        * **Reviewer comment:** This MC-IS method for estimating the score will NEVER work well in high dimensions due to variance and thus why works such as [1,2,3,4] which are clearly aware of this formulation (as they either state it in their appendices or use it for subsequent calculation) pursue an optimization alternative to estimating the drift.
        * **Feedback to the reviewer:** Consider revising this comment to avoid absolute statements like "NEVER". Instead, you could phrase it as a concern about scalability to high dimensions, and ask the authors to address this limitation or provide evidence that it can work in higher dimensions.

3. Provide feedback:

    - For each comment that fails according to the checklist, write concise feedback in the following format:
        - Comment: the verbatim comment of interest
        - Feedback: your concise feedback
    - If you do not identify any issues with a comment, do not include it in your feedback list.
    - If you find no issues in the review at all, respond with: 'Thanks for your hard work!'

Remember:

- Be concise, limiting your feedback for each comment to 1-2 sentences.
- Do not summarize your feedback at the end or include a preamble at the beginning.
- Do not repeat anything the reviewer already included in their review, and do not praise anything the reviewer wrote as we want to provide constructive feedback.
- Your feedback will be sent to reviewers. Do not mention that you are using a checklist or guidelines.
- Do not address the authors at all or provide suggestions to the authors. You are only giving feedback to the reviewer.
- Do not provide feedback to any comments that mention a score or rating. You do not care about the reviewer's score or rating for this paper.
- Do not provide feedback to any comments that discuss typos.

Aggregator Prompt

Here is the paper: `<PAPER>` {`paper`}`</PAPER>`.
Here are the lists of feedback: `<FEEDBACK_LIST>` {`feedbacks`} `</FEEDBACK_LIST>`.
Here is the peer review: `<REVIEW>` {`review`} `</REVIEW>`.

Aggregator System Prompt

You will be given multiple lists of feedback about a peer review of a machine learning paper submitted to a top-tier ML conference. The aim of the feedback is to guide a reviewer to make the review high-quality. Your task is to aggregate the lists of feedback into one list.

Here are the guidelines that were followed to generate the feedback lists originally: <ORIGINAL_GUIDELINES> {ACTOR_SYSTEM_PROMPT} </ORIGINAL_GUIDELINES>
Here are step-by-step instructions:

1. Read the multiple feedback lists provided for that review, the text of the review, and the paper about which the review was written.
2. For all feedback lists, aggregate them into one list with the best comment-feedback pairs from each list:
    - For each comment-feedback pair in the multiple lists that are similar, determine which provides the best feedback and keep only that pair.
    - If there are unique comment-feedback pairs in the multiple lists, critically determine if it is an essential piece of feedback needed to improve the review. If it is unnecessary or redundant, remove the comment-feedback pair.
    - You should end up with one feedback list that has no repeated comments from the review and that is high quality.
    - Return the feedback list in the format you received it in, where the pairs are formatted as:
        - **Comment:** {{the verbatim comment of interest}}
        - **Feedback:** {{your concise feedback}}

Critic Prompt

Here is the paper: <PAPER> {paper}</PAPER>.
Here is the feedback: <FEEDBACK> {feedback} </FEEDBACK>.
Here is the peer review: <REVIEW> {review} </REVIEW>.

Remember:

- You are a critic that will help reviewers improve their comments and reviews. Your valuable feedback will help improve their review.
- Do not address the authors at all or provide suggestions to the authors. You are only giving feedback to the reviewer.

Critic System Prompt

You are a critic that will help reviewers improve their reviews. You are given a list of feedback to the reviewer comments of a machine learning paper submitted to a top-tier ML conference on OpenReview. The aim of the feedback is to guide a reviewer to improve their comments and review as a whole. Your task is to edit the feedback to the reviewer comments for correctness and clarity.

Here, feedback means the feedback given to the reviewer comments to improve them, so the feedback will be given to the reviewer.

Here are the guidelines that were followed to generate the feedback to the reviewer comments originally: <ORIGINAL_GUIDELINES> {ACTOR_SYSTEM_PROMPT} </ORIGINAL_GUIDELINES>. You

should keep in mind to adhere to the above guidelines.

Here are step-by-step instructions:

1. Read the feedback list provided for reviewer comments, the full text of the review itself, and the paper about which the review was written.
2. Evaluate every piece of feedback in the feedback list:
   - For each feedback item, it is imperative that you evaluate the correctness of the feedback. If there is a quote in the feedback, ensure that the quote appears **verbatim** in the paper. You need to check every quote and factual claim in the feedback and edit for correctness. If the feedback is not correct, edit it so it is or if you cannot then remove it.
   - For each feedback item, evaluate if it is clear. You should make sure it would not confuse or frustrate the reviewer who reads it.
3. Remove comment-feedback pairs that are too nitpicky, unnecessary, or superficial. Also remove comment-feedback pairs that do not actually provide suggestions to the reviewer or address an issue with the review, but rather just praise and agree with their comment; the feedback should lead to the reviewer changing their comment.

   Here are some examples of comment-feedback pairs that should be entirely removed from the final feedback list:

   (a) **Reviewer comment:** The novelty remains concerned. It seems that the algorithm combines MLR + a bisimulation loss.

   Feedback to the reviewer: It would be helpful if you could elaborate on why you see the novelty as limited. In particular, it would be very helpful if you could discuss the key differences between simply combining MLR and bisimulation loss versus the approach proposed in this paper. This would give the authors a better chance to clarify the novelty of their work.

   (b) **Reviewer comment:** The paper lacks a proper related work section, which makes it challenging for readers to quickly grasp the background and understand the previous works. It is crucial to include a comprehensive discussion on related works, especially regarding the variance-reduced ZO hard-thresholding algorithm and the variance reduction aspect.

   Feedback to the reviewer: To make this comment more actionable, consider suggesting specific areas of related work that should be covered, such as key papers on ZO optimization, hard-thresholding methods, and variance reduction techniques in optimization. This will give the authors clearer guidance on what to include.

   (c) **Reviewer comment:** The paper is not very well-presented and is hard to follow. First of all, it is unclear in the hybrid setting considered, what are the relative relations of the guest parties? In the introduction, it appears that they share the same feature space but have different sample IDs, however, in 3.1 they appear to have different dimensions and unclear alignment. It is suggested that the paper properly define the problem setting. A figure on how data is partitioned by different parties would also help.

   Feedback to the reviewer: Your suggestion for a clearer definition of the problem setting and a visual representation of data partitioning is excellent. This would significantly improve the paper's clarity and readability.

   (d) **Reviewer comment:** 3) the model performance of the proposed methods still appear to be a little inferior to the centralized setting, not exactly "comparable" as claimed. It is important to understand whether the proposed method is "lossless" or "lossy" and why. I think more detailed examinations and explanations are needed here.

Feedback to the reviewer: Your observation about the performance gap between the proposed method and the centralized setting is insightful. Requesting a more detailed analysis of whether the method is lossless or lossy, along with explanations for any performance differences, would significantly enhance the paper's contribution.

(e) **Reviewer comment:** Q2: It appears that the introduced projection loss can be directly optimized with respect to the trigger T. What's the rationale behind setting an upper bound and optimizing the projection loss through this bound? Does this approach offer computational benefits?

Feedback to the reviewer: This question effectively probes the authors' methodological choices. It's a clear and concise query that could lead to valuable insights about the paper's approach. The authors' response could provide important context about the trade-offs involved in their method.

4. Edit comments based on evaluations:
   - Do not add any new points unless the previous feedback obviously missed something important.
   - If you do not identify any issues with a comment-feedback pair, do not edit it.
5. The feedback will be shared with the reviewers for them to improve their comments. Address the reviewer in the second person (e.g., "you") and do not refer to them as "the reviewer."
6. Return the feedback list in the format you received it in, where the pairs are formatted as:
   - **Comment:** {{the verbatim comment of interest}}
   - **Feedback:** {{your concise feedback}}

Remember:

- You are a critic that will help reviewers improve their comments and reviews.
- Be concise, limiting your feedback for each reviewer comment to 1-2 sentences.
- Do not summarize your feedback at the end or include a preamble at the beginning.
- Do not repeat anything the reviewer already included in their review.
- Do not mention that you are using a checklist or guidelines.
- Do not address the authors at all or provide suggestions to the authors. You are only giving feedback to the reviewer.

Formatter Prompt

Here is the feedback for you to format: {`feedback`}

Formatter System Prompt

You will be given a set of feedback given to various reviewer comments in a peer review of a machine learning paper. Your response, which will be the list of reviewer comments and feedback to them, will be shared with the reviewers who wrote the review, so that they can improve their reviews and the peer review cycle.

Your task is to format the feedback into a structured format. You should format the feedback as a list of comment-feedback pairs:

- **Reviewer comment:** {{a comment}}
- **Feedback to the reviewer:** {{feedback to the comment}}
- **Reviewer comment:** {{another comment}}
- **Feedback to the reviewer:** {{feedback to the comment}}

  ...

Your goal is to only keep feedback to the reviewers that can help them improve their comments. You should only pay attention to lines that start with "Comment" or "Feedback".

- Only keep the comment-feedback pairs where the feedback can help improve the reviewer. If there is no suggestion for improvement, remove the entire comment-feedback pair.
  - Here is an example of a comment-feedback pair that should be removed from the final feedback list:
    - **Reviewer comment:** Section 2.2. "It independently formulates new approaches" → Is it a hallucination or a feature? It looks like a hallucination to me. If this is important for achieving good performance, can you provide an ablation study based on whether to allow new approaches or not?
    - **Feedback to the reviewer:** This is a thoughtful question about an important aspect of the methodology. Your suggestion for an ablation study is particularly valuable and could provide insights into the method's effectiveness.
  - If the feedback says "No changes needed" or something with a similar meaning, remove the entire comment-feedback pair.
- Do not modify the content of the feedback at all, only format it into the bullet point format described above.
- The response you send will be immediately shared with the reviewers. Thus, there should be NO OTHER TEXT in the output, for example no preamble or conclusion sentences. Only respond with the list of feedback & reviewer comment bullets, and no other text.
- Since your response will immediately be sent to the reviewers, if there is no feedback, just say "Thanks for your hard work!".

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.

We also provide the prompt used for the incorporation analysis:

Incorporation Analysis Prompt

Task: Determine if the following feedback suggestion was incorporated into the modified version of a review. Also, categorize the given feedback into exactly one of these three categories:

1. **ACTIONABLE_VAGUE:** Encouraging reviewers to rephrase vague review comments, making them more actionable for the authors. *For example, the feedback says:* "It would be helpful to suggest specific baselines that you think must be included. Are there particular methods you feel are missing from the current comparison? Could you elaborate why?"
2. **CONTENT_CLARIFY:** Highlighting sections of the paper that may already address some of the reviewer's questions (clarifying content). *For example, the feedback says:* "Does Figure 5 of the paper answer your question? In particular: 'In Transformers, the proposed technique provides 25% relative improvement in wall-clock time (Figure 5)'."
3. **ADDRESS_UNPROFESSIONAL:** Identifying and addressing unprofessional or inappropriate remarks in the review. *For example, the feedback says:* "We appreciate your review, but kindly request that you focus your comments on the specific content and methodology of the paper rather than making personal remarks about the authors."

Instructions:

1. Read the original review and modified review.
2. Read the reviewer's original comment and the feedback given to the reviewer.
3. Determine if the changes suggested in the feedback were incorporated into the modified review as compared to the original review. If the reviewer's original comment appears *verbatim* in the modified review still, you should return `FALSE` for the incorporation. The incorporations should be clear and quite explicit. Think critically about if the incorporation is significant enough to count.
4. Determine which of the three categories best describes the primary purpose of the feedback.
5. Think step by step and explain your reasoning.

Output Format: Please provide your final answer as two comma-separated values between `<OUTPUT>` tags, where:

- The first boolean is `TRUE` or `FALSE` depending on whether the feedback was incorporated.
- The second string is one of these three options: `ACTIONABLE_VAGUE`, `CONTENT_CLARIFY`, or `ADDRESS_UNPROFESSIONAL`.

Example: `<OUTPUT>TRUE, ACTIONABLE_VAGUE</OUTPUT>`

5 Reliability test prompts

We generated the following reliability tests to be run in real-time after feedback was generated. For each reliability test, we provide examples of feedback that would fail it within the prompt as in-context examples.

Test 1 Prompt: Does feedback only praise reviewer?

Task: You are given feedback to a review and potentially the review itself. Your task is to identify problematic feedback that lacks actionable content. In particular, our goal is to only have feedback

such that upon reading the feedback, the reviewer would want to change their comments.
Reviewer comments are marked with **Reviewer Comment:** and feedback to the reviewer is marked with **Feedback to the reviewer:**.
We are specifically looking for **Feedback to the reviewer** items that fail to provide actionable guidance. A feedback item is considered non-actionable if it ONLY does one or more of the following without any additional constructive content:

- Praises the reviewer's comment
- Agrees with the reviewer's point
- Only affirms the reviewer's observation
- Restates or summarizes the reviewer's comment
- Makes a general statement about the importance of the topic

None of these would lead to the reviewer changing their comment.
Actionable feedback MUST include at least one of the following:

- Specific suggestions for improving the review comment
- Guidance on how to make the comment more constructive or helpful for the authors
- Upon reading the feedback, the reviewer should find something to change in their review.

Is there a feedback item starting with '**Feedback to the reviewer:**' that fails to provide actionable feedback as defined above? If so, return True.
Example:

- **Reviewer Comment:** "in heterophilic networks, vertices with high structural and semantic similarities are generally farther away from each other" Any evidence for this claim?
- **Feedback to the reviewer:** This is a good question that challenges a key assumption of the paper. The feedback appropriately suggests asking for evidence or citations to support this claim.

Example:

- **Reviewer Comment:** Researchers already find that heterophily is not always harmful and homophily assumption is not always necessary for GNNs [2,3,4,5]. How does this paper align with these works?
- **Feedback to the reviewer:** This is an excellent point that highlights recent developments in the field. The feedback correctly suggests asking the authors to discuss how their work relates to and differs from these existing findings.

Example:

- **Reviewer Comment:** How do you select spectral nodes and why? Why do you want to connect the spectral nodes?
- **Feedback to the reviewer:** These are important technical questions about the proposed method. The feedback appropriately suggests asking for more details and justification for these design choices.

For such cases, you should return True.

- Analyze each **Feedback to the reviewer** item individually.

- Determine if the feedback item contains any part such that upon reading the feedback, the reviewer would want to change their comments.
- If a feedback item contains only praise, agreement, restatement, or general statements without any actionable content that would actually help the reviewer change their comment, consider it non-actionable.

First, think step by step and send your reasoning between `<REASONING>` {{your reasoning}} `</REASONING>` tags. Then, respond with TRUE or FALSE between `<o>` tags, such as `<o> TRUE </o>` or `<o> FALSE </o>`.
Return True if there exists at least one non-actionable feedback item as defined above. Return False only if all feedback items include actionable content.

Test 2 Prompt: Does feedback address author?

Task: Identify feedback that incorrectly addresses the paper authors instead of the reviewer.
Input:

- A set of reviewer comments and corresponding feedback items
- Reviewer comments are marked with **Reviewer comment:**
- Feedback is marked with **Feedback to the reviewer:**

Instructions:

1. Analyze each piece of feedback marked with "**Feedback to the reviewer:**".
2. Identify any feedback that appears to be addressing the paper authors rather than the reviewer.
3. Look for key indicators that the feedback is meant for authors, such as:
    - References to "your method," or "your approach"
    - Mentions of "readers" or how to improve the paper for the audience
    - Advice on how to modify experiments, results, or conclusions
    - Any language that assumes the recipient can make direct changes to the paper

Output:

1. Provide your reasoning within `<REASONING>` tags.
2. Return your final decision within `<OUTPUT>` tags:
    - TRUE if any feedback item incorrectly addresses the authors
    - FALSE if all feedback items correctly address the reviewer

Examples of incorrect feedback (addressing authors) that should return True:
Example 1:

- **Reviewer Comment:** The authors argued that "previous works mostly focus on addressing specific types of distribution shifts", which seems improper and incorrect. E.g., the typical works for graph OOD learning EERM [1] and DIR [2] do not assume the type of distribution shifts in their problem formulation.
- **Feedback to the reviewer:** It would be helpful to clarify the distinction between the approach in this paper and previous works like EERM and DIR. Specifically, how does modeling distribution shifts as a mixture of components differ from not assuming specific types of shifts? This would help readers better understand the novelty and contribution.

Example 2:

- **Reviewer Comment:** The main critique of the paper is that it is quite elaborate, and there is no comparison to a far simpler available baseline: send the answer to GPT-J and accept sufficiently high probability answers (this is a simplified version of their L = 2 setup, but requiring training only a single scalar for the classification case).
- **Feedback to the reviewer:** Your suggestion for a simpler baseline comparison is valuable. It would be helpful to include this comparison in the paper, especially for APIs that provide log probabilities. This could demonstrate the advantages of your more complex approach while acknowledging simpler alternatives. Consider discussing why your method might be preferable even when log probabilities are available.

Example 3:

- **Reviewer Comment:** The 5% quality improvement on some datasets seems likely to be a real ensembling effect, but it should be presented as such (that it likely has the same cause as ensembling methods).
- **Feedback to the reviewer:** To strengthen your paper, consider discussing the relationship between FrugalGPT and traditional ensembling techniques. Highlight both similarities and differences, and explain how this relates to the observed quality improvements. This would provide more context for your results and situate your work within the broader field of machine learning.

Example of correct feedback (addressing the reviewer) that should return False:

- **Reviewer comment:** The proposed solution could benefit from more details. Exploring the soft-prompt size (i.e., the variable N) would be nice. The N used is not specified in the main text either, which I believe is an important detail.
- **Feedback to the reviewer:** To make this feedback more actionable, you could ask the authors to include the value of N used in their experiments in the main text, and suggest adding an ablation study exploring different soft-prompt sizes and their impact on performance and robustness.

Note: The key is to identify feedback that directly instructs or suggests changes to the paper, rather than feedback that helps the reviewer improve their review or reviewing skills.

Test 3 Prompt: Does feedback restate review?

Task: You are given feedback to a review and potentially the review itself. Analyze each '- **Reviewer Comment**:' '- **Feedback to the reviewer**:' pair. Do any of the feedback items simply restate what the comment says without providing any new meaningful and unique suggestions?
Example of feedback that restates the comment without adding value:

- **Reviewer comment:** Can examples or further clarification be given for the 3.1 sentence "enhancing the accountability of the output"? This isn't clear, at least to me.
- **Feedback to the reviewer:** This is a good point that could lead to improved clarity in the paper. To make your comment more actionable, you could ask the authors to provide examples or further clarification for the sentence "enhancing the accountability of the output".

In this example, the feedback essentially repeats the reviewer's request for examples and clarification without providing any new insights or meaningful and unique suggestions on how to improve the comment.
Output:

1. Provide your reasoning within `<REASONING>` tags.
2. Return your final decision within `<OUTPUT>` tags:
    - TRUE if you find any feedback items that restate what the comment says
    - FALSE if no feedback items simply restate the comment

Test 4 Prompt: Ensure comments appear in review

Task: You are given feedback to a review and potentially the review itself. Do all the items starting with '**Review Comment:**' appear verbatim in the original review?

6 Main results stratified by ICLR primary areas

We focused our experiment on ICLR because it is a large and popular AI conference that covers a diverse range of research topics. Furthermore, ICLR's workflow is representative of most AI conferences because it uses the same core process: 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%). Our system demonstrates consistent effectiveness regardless of primary research area. The five major AI subfields we stratified our results by were generated by first looking at the distribution of all submissions across the 21 primary areas. Then, we grouped relevant categories together 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 (Extended Data Fig. 1) and review length (Extended Data Fig. 2) across applications, generative AI, learning methods, theory, responsible AI, and infrastructure domains. We also observe that incorporation rates remain stable (63-69%) across all subfields (Extended Data Fig. 5). This cross-subfield consistency suggests that the underlying reviewer challenges (vague comments, content misunderstandings, and unprofessional remarks) are universal rather than domain-specific.

7 Average score changes during review and rebuttal periods

In Extended Data Fig. 3a, we examined the potential change in review scores (soundness, presentation, contribution, rating, and confidence) between the initial and modified reviews across the groups during the review period. We found that reviewers who were selected to receive feedback did not change their scores more than those in the control group (top panel). We also saw that of reviewers who received feedback, reviewers who updated their review were significantly more likely to decrease their soundness score and increase their confidence score at the end of the review period (before the rebuttal period began) compared to those who did not update their review. This suggests that reviewers who updated their reviews became more confident in their assessments.

In Extended Data Fig. 3b, we conducted the same analysis during the rebuttal period. Similar to the review period, we found that reviewers who were selected to receive feedback did not change their scores more than those in the control group (top panel). Of reviewers who received feedback, those who updated their reviews significantly increased all scores except confidence compared to those who did not update their reviews. From this, we see that reviewers who updated their reviews were much more engaged in the rebuttal process.