Responsible NLP Checklist

Paper title: *Evaluating the Impact of Reviewer Guideline Design on LLM-Based Automated Peer Review*
Authors: *Haowen Li, Yoichi Ishibashi, Masafumi Oyamada*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☒ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
Our dataset contains no PII or offensive content. We only use publicly available ICLR 2024 papers and reviews that are already anonymized and filtered to remove names, affiliations, and other identifiers (Section 3).

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
We report data statistics in Section 3 and Appendix A, D, E(e.g., 7,262 papers and 28,028 reviews from ICLR 2024, 1,500 sampled for evaluation, 3 models 5 guideline conditions).

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The setup, including prompt templates (Appendix C), data sampling, token limits, and evaluation procedure (RMSE computation), is described in Sections 34 and Appendix A.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Privately revealed to ACL ARR 2026 January Program Chairs, ACL ARR 2026 January Submission9217 Area Chairs, ACL ARR 2026 January Submission9217 Authors, ACL ARR 2026 January Submission9217 Reviewers, ACL ARR 2026 January Submission9217 Senior Area Chairs All quantitative results include RMSE and correlation metrics (Table 1 and Table 2) with comparative averages across models and conditions (Section 34).

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☒ N/A D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
(left blank)

☒ N/A D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
(left blank)

☒ N/A D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
(left blank)

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☒ E1. If you used AI assistants, did you include information about their use?
AI tools (e.g., ChatGPT) were used only for phrasing and writing suggestions, not for generating research content or analysis.