Responsible NLP Checklist

Paper title: *Putting HUMANS first: Efficient LAM Evaluation with Human Preference Alignment*
Authors: *Woody Haosheng Gan, William Barr Held, Diyi Yang*

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.

☑ A2. Did you discuss any potential risks of your work?
Section "Ethical Considerations" (appears after Conclusion and Limitations section, before Reference)

☑ 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?
Section 3 and Appendix B: Benchmark subsets derived from 5 publicly available datasets that have undergone their own review processes. Section Ethical Considerations: Personal information automatically filtered from text feedback before analysis (Appendix H.1). Audio recordings stored securely with restricted access and will be noise-masked before any sharing.

☑ 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?
Section 3.1.2 and Appendix B: 40 tasks with detailed specifications in their sizes. Section 3.3: 3-fold CV with 12 source/6 held-out models, 100 repeats. Section 4.2: 776 human evaluations across 7 models with numbers on each model specified. Section 5: Released benchmark subsets at sizes 10, 20, 30, 50, 100, 200 with regression and benchmark weights.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.2 and Appendix C: Hyperparameters for all methods including IRT (5-dim, lr=0.1, 500 epochs), Ridge regression (via CV over {0.001,...,100}), clustering (K-Means, PCA dims). Section 4.1.1 and Appendix F: VAD configuration parameters. Section 4.4: Ridge {10^-4,...,10^4} via nested leave-one-out CV.

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.

☑ 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?
Tables 1 (Section 3.3), 6, 7 (Appendix D.2): Mean correlation SEM across 300 evaluations (3-fold CV, 100 repeats). Table 2 (Section 4.2): Mean ratings SEM across conversations. Figures 2 (Section 3.3), 3 (Section 4.3), 4 (Section 4.4), 5-13 (Appendix D): Shaded regions/error bars show confidence intervals. Table 11 (Appendix I): Mean SEM for pairwise ranking accuracy.

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

☑ 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.?
Appendix G: Full participant instructions provided, consent form (Figure 16), scenario-specific instructions for three conversation types (Figures 17-19), conversation interface details (Figure 20), and post-conversation evaluation interface (Figures 21, 22).

☑ 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)?
Section 4.1.3 and Appendix G.2-G.3: Participants recruited via Prolific (native English speakers, US-based, 18+, balanced gender). Compensation: $\backslash 0.25 base + \$0.25/minute conversation(\backslash 2.50$ *for 10 minutes) + bonuses ($1 task completion,* $\backslash 0.25 feedback), totaling \$2.75 - \backslash 4$ *for 10-13 minutes (minimum $15/hour rate).*

☑ 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)?
Appendix G.4: Informed consent procedures detailed (Figure 16), including study workflow, voice recording disclosure, simulated environment clarification, data usage explanation, privacy protections, and right to withdraw without penalty.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 4.1.3 and Appendix G.1: Study approved by authors' institution's Institutional Review Board (IRB) prior to data collection. 776 participants recruited through Prolific with informed consent.

☑ **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?
Appendix B.1: LLM/LAM-as-a-judge for benchmark evaluation (GPT-4o, GPT-4o-audio, GPT-4o-mini, Gemini-2.5-Flash). Appendix G.8.1: GPT-4.1 and o4-mini for scenario generation. Appendix H.1: GPT-5.2 for automated user feedback analysis.