Responsible NLP Checklist

Paper title: *LongMP-Bench: A Benchmark for Multimodal Persona Understanding in Long-Term Dialogues*
Authors: *Zhuoqun Li, Zhaopei Huang, Wenxuan Wang, Qin Jin*

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?
Potential risks associated with the work are discussed in the Ethical Statement section.

☑ **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?
We discuss data containing personally identifying or offensive content in the Ethical Statement section.

☑ 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 reported basic statistics for LongMP-Bench, including the number of examples, in Appendix B.1.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We report the key hyperparameters used during inference, including generation settings such as temperature, top-p, and max tokens, in Section 4.3. Since we only perform inference and do not conduct any training, we do not provide details on hyperparameter search or the best-found hyperparameter values.

☒ 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?
As all models were evaluated with a temperature of 0, the generation process was deterministic, minimizing randomness and maximizing reproducibility. Therefore, we report results from a single run for each configuration.

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?**

☑ 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.?
The instructions provided to participants are included in Appendix A.4.3.

☑ 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)?
Information regarding participant recruitment and compensation is provided in Appendix A.4.3.

☑ 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)?
This information is detailed in Appendix A.4.3.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No formal ethics review board approval was obtained, as the study does not involve minors, medical data, or any sensitive personal information

☑ **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?
We include information about the use of AI assistants in Appendix C.2, where we discuss the use of an AI assistant (Gemini-1.5-pro) as a judge and analyze its correlation with human ratings.