Responsible NLP Checklist

Paper title: *PsychEval: A Multi-Session and Multi-Therapy Benchmark for High-Realism AI Psychological Counselor*
Authors: *Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, JingYuan Zhao, Qin Chen, Ningning Zhou, Kai Chen, Liang He*

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?
See Limitations Statement. This article explores potential hazards such as the lack of detailed information other than the text and the lack of high-risk scenarios, which make the scenarios insufficiently representative of extreme or high-risk situations.

☑ **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?
The data used in this work does not contain any personally identifiable information. All data is either synthetic, publicly available, or properly anonymized, and does not include offensive content.

☑ 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?
See Section 3.1 (Data Collection), Analysis of Benchmark, Table 1, and Table 3, et al. The paper reports detailed dataset statistics, including the number of cases and sessions, skill and stage distributions, and benchmark configurations necessary to understand and reproduce the experiments.

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We utilized closed-source pre-trained LLMs via API for data generation and evaluation. We did not perform model training or hyperparameter search. Experimental configurations and sampling details are described in Appendix E.1.

☑ 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?

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.

Section 5.1 (Table 1) reports descriptive statistics of the constructed dataset. Section 5.3 (Tables 4, 5) and Appendix D (Table 6, 7) report the average evaluation results and structural characteristics across the sampled instances.

☑ **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.?
see Appendix F

☑ 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)?
see Appendix F

☑ 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)?
See Appendix F

☒ 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?
During the preparation of this work, the author(s) used chatgpt in order to improve the language and readability of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.