Responsible NLP Checklist

Paper title: *The Pursuit of Empathy: Evaluating Small Language Models for PTSD Dialogue Support*
Authors: *Suhas BN, Yash Mahajan, Dominik O. Mattioli, Andrew M. Sherrill, Rosa I. Arriaga, Christopher Wiese, Saeed Abdullah*

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?
The paper discusses potential risks in Section 11 ("Limitations") , which covers the limitations of synthetic data , the lack of adversarial testing for harmful advice or crisis mishandling , and issues of generalizability. Further discussion on the risks of context-inappropriate responses is found in Section 6 ("Discussion").

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

☑ B1. Did you cite the creators of artifacts you used?
The creators of all artifacts used are cited throughout the paper and listed in the "References" section (starting on page 10). This includes citations for the language models , datasets, and evaluation metrics such as BERTScore , ROUGE-L , and METEOR.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The license for the created TIDE dataset is specified in Section 10 ("Data Availability"), which states the dataset is publicly available under the CC BY-NC 4.0 license.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
TIDE dataset is intended for research to support broader empathetic AI research, consistent with non-commercial use. The ethical considerations for using such models in sensitive contexts are discussed in Section 3.6 ("Ethical Considerations and Safety").

☑ 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?
In Section 4.2, we discuss that the TIDE dataset is fully synthetic, derived from 500 client personas, and does not contain real personally identifiable information. All scenarios were clinically reviewed

by psychology experts to ensure trauma sensitivity and emotional plausibility, thereby mitigating offensive or harmful content. The synthetic nature is reiterated in the Limitations section (Section 11).

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Documentation and details of the created TIDE dataset, including its composition, structure, and generation methodology, are provided in Section 4.1 and 4.2. Demographic characteristics of the human evaluators are documented in Section 6.5 and Table 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?
Key statistics for the TIDE dataset (10,000 instances, 500 personas, 20 interactions per persona) are reported in Section 4.2. Descriptive statistics are in Table 3.

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The parameter sizes of the models used (0.5B to 5B) are stated in the Abstract and Introduction. Details of the computing setup, including training sample sizes, epochs, and LoRA parameters, are provided in the caption for Table 1 on page 6.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup and all relevant hyperparameters for fine-tuning (epochs, gradient accumulation, learning rate, LoRA rank and alpha, max length) are listed in the caption of Table 1 on page 6.

☑ 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?
Descriptive statistics (Mean Std. Dev.) are reported for all automatic evaluation metrics in Table 1. For human evaluations, mean empathy ratings are annotated on distributions in Figure 2 , and means, standard deviations, and p-values are reported in the demographic analysis in Section 6.5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
The specific automatic metrics and embeddings used are reported in Section 4.3.1, with citations to the original implementation papers (e.g., BERTScore, ROUGE-L, METEOR, and all-MiniLM-L6-v2).

☑ **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 full text of the survey scenarios and instructions given to participants for the empathy rating task are provided in Appendix A. The complete experimental protocol is detailed in Appendix B.

☑ 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 about participant recruitment via the Prolific platform is detailed in Appendix B and Appendix C. Participants were restricted to first-language English speakers in the United States. The use of Prolific implies compensation standards that are fair and above minimum wage for U.S. participants.

☑ 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)?
Human evaluation study was IRB-approved (see Abstract) and administered on the Qualtrics platform, where participant consent was obtained.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We were approved by Penn State IRB for the human evaluation.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Basic demographic characteristics of the final participant pool (N=116), including age, gender, race, education, and employment status, are reported in Table 3 in Appendix D. The study's inclusion criteria (first-language English speakers in the U.S.) are described in Appendix C.

☑ **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?
Yes, the paper details the use of LLM's. Specifically, Claude Sonnet 3.5 was used to generate the reference responses for the synthetic TIDE dataset, which is a core part of the methodology. This process is described in Section 4.2 ("Generation Methodology"). The language models were used for data generation and as experimental subjects, not for writing the manuscript.