

Responsible NLP Checklist

Paper title: *How Real Are Synthetic Therapy Conversations? Evaluating Fidelity in Prolonged Exposure Dialogues*

Authors: *Suhas BN, 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'
- ☐ 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 Section 9 (Ethical Concerns) and the Limitations section.

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

- ☒ B1. Did you cite the creators of artifacts you used?

Citations for artifacts used are provided throughout the paper, particularly in Section 3 (Related Work) and Section 4 (Methodology)

- ☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?

he license for the created artifact is discussed in Section 10 (Data Availability).

- ☒ 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)?

The use of artifacts is consistent with research purposes as detailed in Section 4 (Methodology) and within the ethical framework described in Section 9 (Ethical Concerns).

- ☒ 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?

This is discussed in Section 4 (Methodology) and Section 9 (Ethical Concerns). The real-world dataset, which contains sensitive information, is not released and is protected by IRB protocols. The synthetic dataset is publicly available and does not contain real personally identifiable information.

- ☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

Documentation is provided in Section 4 (Methodology) and Appendix A, which defines key terms and metrics.

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.

- ☒ 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?
Statistics for the data are reported in Section 4 (Methodology) and presented in detail in Table 2 and Table 5.

☒ **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 study did not involve training or fine-tuning models. The work focuses on the statistical analysis of existing real-world data and pre-generated synthetic data.
- ☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup for the comparative analysis is detailed in Section 4 (Methodology) 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?
Descriptive statistics, including means, standard deviations, and results from statistical tests, are reported in Section 5 (Findings) and specifically in Tables 2, 3, 4, and 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 packages and tools used for analysis are described in Section 4 (Methodology).

☒ **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.?
This information is not included due to clinical data protocols. Data collection and annotation were conducted under IRB-approved procedures, as mentioned in Section 4 (Methodology) and Section 9 (Ethical Concerns).
- ☒ 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)?
Recruitment and payment details are not reported. Clinical participants were recruited under an IRB protocol, and the expert annotators were affiliated with the research project.
- ☒ 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)?
Consent is discussed in Section 4 (Methodology) and Section 9 (Ethical Concerns). All data from human subjects was collected with participant consent under IRB-approved protocols.
- ☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics review board approval is confirmed in Section 4 (Methodology) and Section 9 (Ethical Concerns).
- ☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Detailed characteristics are not reported to protect the privacy of the small expert annotator pool. It is noted in Section 4.1.5 and Section 9 that annotators had clinical expertise.

☒ **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 assistants were not used in the research or writing process.