Responsible NLP Checklist

Paper title: *Empathy Applicability Modeling for General Health Queries*
Authors: *Shan Randhawa, Agha Ali Raza, Kentaro Toyama, Julie Hui, Mustafa Naseem*

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 9

☑ **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 answered 'No' because the source datasets were already anonymized and cleaned; therefore, we did not explore handling of personally identifying information or offensive content in the study.

☑ 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?
Sections4.1, 4.2.1, and 4.2.2 describe the sampled dataset of9,500 patient queries and its 1,296 dualannotated/8,000GPTonly split, while Section5.5 details the traindevtest partitions for both the Human and Autonomous experiments used in model training and evaluation.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.5

☒ 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?
Section6.1 (Table3) reports accuracy, macroF1, and weightedF1 from a single run on the test set for each training set and baseline.

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?**

N/A 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.?
Our study did not have any human subjects as we use a publicly available anonymized dataset. However, we did use human annotators and we have provided the detailed instructions used in Appendix B referred in Section 4.2.1

☑ 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 4.2.1

☑ 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)?
Section 4.2.1

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 4.1

☑ **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?
Sections 4.2.2 (annotation) and 5.4 (evaluation baselines) describe our use of GPT4o and o1 for labeling and zeroshot baseline evaluation; additionally, ChatGPT was used only for grammatical review of manuscript text.