Responsible NLP Checklist

Paper title: *MentalSeek-Dx: Towards Progressive Hypothetico-Deductive Reasoning for Real-world Psychiatric Diagnosis*
Authors: *Xiao Sun, ymyang, Xinyi Jiang, Yu Tian, Junnan Zhu, Jiang Zhong, Qin Lei, Jingwang Huang, Haoyang Zeng, xinyu zhou, Xin Xiao, Kaiwen Wei*

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 are discussed in detail in Section 5 Experiments on MentalDx Bench and the Ethics 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 described the checking and anonymization procedures in Section 3 MentalDx Bench, and verified that the benchmark contains no personally identifying or offensive content with detailed statistics in Section 5 Experiments on MentalDx Bench.

☑ 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?
Detailed statistics, inputs, outputs and data distributions are presented in Section 3 MentalDx Bench. The dataset is used for testing only, so no train/test/dev split is needed.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental setup is detailed in Section 7 MENTALSEEK-DX Evaluation, and all hyperparameters are fully specified in Section 9.10 Hyperparameter Settings.

☑ 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?
Experimental setup, hyperparameter search and optimal hyperparameters are fully discussed in Section 5 Experiments on MentalDx Bench.

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.?
Details including the number of annotation experts, affiliated hospitals, and risk disclaimers for annotators are fully documented in Section 3 MentalDx Bench.

☑ 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)?
Participant recruitment, payment, and the reasonableness of compensation are fully described in Section 3 MentalDx Bench.

☑ 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)?
As stated in Section 3 MentalDx Bench, all data collection and labeling followed the ethical principles of the Declaration of Helsinki. All records were fully deidentified before processing, so individual informed consent was waived.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics review board approval is detailed in Section 3 MentalDx Bench and Section Appendix.

☑ **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 used large language models for data cleaning, and disclosed the usage in Section 6.1 Pretraining Setup: "The entire automatic procedure is executed onsite within the medical center using an on-premise deployment of DeepSeek-R1, ensuring that all data remains local and under institutional governance."