Responsible NLP Checklist

Paper title: *MedEinst: Benchmarking the Einstellung Effect in Medical LLMs through Counterfactual Differential Diagnosis*
Authors: *Wenting Chen, Guolin Huang, Wenxuan Wang, Zhongrui Zhu*

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?
We discuss limitations of coverage and clinical scope in the Limitations section, but we do not include a dedicated discussion of broader potential risks or misuse risks of the work in the current manuscript.

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

N/A 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 benchmark is constructed from DDXPlus-derived clinical cases and rewritten narratives rather than real identified patient records. The current manuscript does not involve collection or release of personally identifying information, so anonymization procedures are not applicable.

☑ 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?
Section 3.3 and Appendix B.3

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 and Appendix C

☑ 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?
Section 5

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.?
The manuscript reports the physician quality-control protocol and evaluation dimensions, but it does not include the full text of instructions given to annotators.

☒ 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)?
The current manuscript states that four board-certified physicians with over 8 years of clinical experience participated in quality control, but it does not provide recruitment or compensation details.

N/A 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)?
The work does not curate personal data from human participants or patient-subject records collected by the authors. The benchmark is derived from an existing source dataset and transformed into paired clinical narratives.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The current manuscript does not report ethics board approval or exemption. The human involvement is limited to expert quality control by physicians rather than collection of participant data for a human-subject study.

☑ **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?
The manuscript describes the use of AI systems in the research pipeline, including GPT-5, DeepSeek-R1, and Gemini-2.5-Pro as an LLM-as-a-Judge committee for pair verification, GPT-5 as the critic model in CGME, and LLM-based rewriting/evidence substitution during benchmark construction. In addition, AI-based writing assistants were used for language polishing and grammar checking during manuscript preparation; all technical content, claims, and final wording were verified by the authors.