Responsible NLP Checklist

Paper title: *Patient-Similarity Cohort Reasoning in Clinical Text-to-SQL*
Authors: *Yifei Shen, Yilun Zhao, Justice Ou, Tinglin Huang, Arman Cohan*

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?
No. The paper focuses on benchmark construction and evaluation. We did not include a dedicated discussion of potential societal/clinical risks in the current version.

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

☑ B1. Did you cite the creators of artifacts you used?
Sections: 1 and 3 (and References). We cite the creators of MIMIC-IV and prior clinical text-to-SQL datasets/benchmarks used as baselines or background.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section: 3

☑ 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)?
Appendix D (SQL generation/refinement prompts specify BigQuery and the official MIMIC-IV datasets).

☑ 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?
Sections 1 and 3. We use only de-identified identifiers (e.g., optional de-identified anchor patient identifiers such as subject_id/hadm_id) and build the benchmark on the credentialed MIMIC-IV dataset.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix N (Annotation Guideline), including validation criteria and benchmark item structure.

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?
Table 3 (dataset size and splits).

☑ **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?
Appendix E

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1

☒ 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?
No. Main results are reported as single-run benchmark scores without error bars or multi-seed statistics. We provide diagnostic statistics such as a humanGPT agreement study (Appendix K) and correlation analyses between scoring signals (Appendix M), but not variability estimates for the model results.

N/A 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?
N/A. Our evaluation is based on executing SQL in Google BigQuery and rubric-based judging, and we do not rely on external NLP metric packages with tunable parameters (e.g., ROUGE/Spacy).

☑ **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.?
Appendix N (Annotation Guideline) and Appendix I (Annotation Interface).

☒ 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)?
No. We describe the annotator pool at a high level (Appendix A, Table 6), but we do not report recruitment procedures and compensation/payment details.

☒ 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)?
No. The manuscript does not discuss whether/how consent was obtained for the underlying patient data. The study uses the existing MIMIC-IV research resource, for which the dataset creators report an IRB waiver of informed consent.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. The manuscript does not report IRB approval/exemption for the underlying patient data. The MIMIC-IV resource itself is reported by its maintainers to have IRB approval and a waiver of informed consent.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No. We describe annotator expertise/training (Appendix A, Table 6), but we do not report demographic or geographic characteristics of the annotator population.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)