Responsible NLP Checklist

Paper title: *CLEAR: A Clinically Grounded Tabular Framework for Radiology Report Evaluation*
Authors: *Yuyang Jiang, Chacha Chen, Shengyuan Wang, Feng Li, Zecong Tang, Benjamin M. Mervak, Lydia chelala, Christopher M Straus, Reve Chahine, Samuel G. Armato III, Chenhao Tan*

> 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?
See 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?
See Appendix A and B.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
See Appendix A.

☑ 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)?
See Appendix A and B.

☒ 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?
Source dataset of CLEAR-Bench is already de-identified and didn't include any offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
See Section 3 and Appendix A.

☑ 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?
See Section 3 and Appendix A.

☑ **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?
See Appendix D.

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

☑ 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?
See Section 4.

☑ 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?
See Appendix A and D.

☑ **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.?
See Appendix E.

☒ 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)?
All of our human annotators are listed as our co-authors for this project and they're five board-certified radiologists.

☒ 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)?
All of our human annotators are listed as our co-authors for this project and they're five board-certified radiologists.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
All of our human annotators are listed as our co-authors for this project and they're five board-certified radiologists.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
All of our human annotators are listed as our co-authors for this project and they're five board-certified radiologists.

☑ **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?
See Appendix B. Mostly, we only use AI assistants to polish our wordings.