Responsible NLP Checklist

Paper title: *CREPE: Rapid Chest X-ray Report Evaluation by Predicting Multi-category Error Counts*
Authors: *Gihun Cho, Seunghyun Jang, Hanbin Ko, Inhyeok Baek, Chang Min Park*

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?
The work proposes an automatic evaluation metric for chest X-ray report generation. It does not involve deployment in clinical practice, patient interaction, or decision-making, so potential risks are minimal.

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

☑ B1. Did you cite the creators of artifacts you used?
We cited the creators of datasets (MIMIC-CXR, ReXVal, ReFiSco, RadEvalX, RaTE-Eval) and models (CheXagent, BiomedBERT).

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We used only publicly available datasets under their data use agreements.

☑ 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)?
All datasets were used strictly under their intended research licenses and access conditions.

☑ 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?
All datasets are de-identified (e.g., MIMIC-CXR). No personally identifiable or offensive content is included.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We described datasets, annotation protocols, and evaluation settings.

☑ 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?
We reported dataset sizes, error distributions, and evaluation splits.

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.

☑ **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?
We reported model (BiomedBERT), training setup (10 epochs, batch size 64, max 512 tokens), and GPU (NVIDIA A6000). Training used moderate GPU resources.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We described hyperparameter search and best configuration.

☑ 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?
We reported correlations, error bars, and confidence intervals.

☑ 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?
We specified model backbones (BiomedBERT, ClinicalBERT, Bio_ClinicalBERT) and generation methods (CheXagent).

☒ **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.?
We used existing datasets with predefined annotation protocols; we did not recruit new 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)?
We did not conduct new annotation; we used publicly available datasets.

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)?
All datasets are de-identified and governed by institutional data use agreements.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
MIMIC-CXR and other datasets were used under approved data use agreements.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We relied on existing datasets with expert annotations; demographic details were not reported.

☑ **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 have used AI assistant as grammar checker, spell checker, etc.