Responsible NLP Checklist

Paper title: *MedLayBench-V: A Large-Scale Benchmark for Expert-Lay Semantic Alignment in Medical Vision Language Models*
Authors: *Han Jang, Junhyeok Lee, Heeseong Eum, Kyu Sung Choi*

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?
Section 'Limitations' discusses risks including synthetic data reliance, English-only restriction, and modality imbalances

☑ **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?
Section 3.1. All data are derived from ROCOv2, which sources from PubMed Central Open Access articles containing de-identified medical images.

☑ 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 4.2, Tables 12, and Appendix B (Table A1, A2).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.4 (Experimental Setup paragraph) and Appendix A (Algorithm 1, prompt template). Section 3.3 details the LLM choice (Llama-3.1-8B-Instruct).

☑ 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?
Appendix D (Table A3, bootstrap significance testing, n=1,000). Tables 12 report full statistics across train/val/test 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. 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.?
Section 4.3. Evaluation followed the protocol of Jeblick et al. (2024) using a 5-point Likert scale across four criteria.

☑ 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)?
Section 4.3. Two board-certified radiologists and one lay reader were recruited from collaborating institutions on a voluntary basis.

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 data are sourced from publicly available datasets (ROCOv2/PMC-OA) under open access licenses.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No new data was collected from human subjects. All data are from publicly available open-access sources.

☑ **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?
Section 3.3 and Appendix A. Llama-3.1-8B-Instruct was used for constrained text refinement in the SCGR pipeline. The prompt template is provided in Figure A1.