Responsible NLP Checklist

Paper title: *CliniCAST: Benchmarking Acoustic Grounding and Text Dominance in Medical Triage*
Authors: *Kyusik Kim, Hyunwoo Yoo, Jaehoon Choi, Kitae Kim, Gail Rosen, Bongwon Suh*

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?
Limitations and Ethical Considerations

☑ **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?
Ethical Considerations

☑ 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

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2 and 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?
Section 5

☑ **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 F

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.

☒ N/A 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 two respiratory-medicine specialists who performed clinical validation participated as domain-expert collaborators in a quality-assurance role rather than as recruited external annotators or crowdworkers. No crowdsourcing platform was used, and no separate per-task compensation scheme applied. No other human annotators participated in data collection or validation.

☒ 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)?
CliniCAST contains no human-subject data. All audio is synthetic, generated by ElevenLabs v3 TTS from pre-shipped commercial voice identities (licensed by the TTS provider for research use). All textual scripts are generated by GPT-5.2 from researcher-authored prompts. No data whose use would require human-subject consent is collected, curated, or released.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human-subject data collection was performed; CliniCAST consists entirely of synthetically generated text and audio. The clinical validation by two respiratory-medicine specialists (Appendix F) evaluated AI-generated scenarios and audio outputs as a professional quality check rather than as research involving identifiable human subjects. Ethics review board approval was therefore not applicable to this 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?
Fully documented in the paper. Section 3 describes the use of GPT-5.2 (thinking mode) to generate both the 25 Task 1 patient-utterance scripts per disease and the 11 Task 2 reassurance paraphrases, and the use of ElevenLabs v3 as the TTS synthesis engine for all 5,856 audio samples. The exact prompts used for script generation and for reassurance-text generation are provided in Figures 2 and 3 of Appendix D. All evaluated audio-language models listed in Section 4.2 are the object of study rather than authoring assistants. AI assistants were not used to generate paper text beyond routine proofreading under full author review.