Responsible NLP Checklist

Paper title: *Evaluating Robustness of Large Audio Language Models to Audio Injection: An Empirical Study*
Authors: *Guanyu Hou, Jiaming He, Yinhang Zhou, Ji Guo, Yitong Qiao, Rui Zhang, Wenbo Jiang*

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?
Ethical Concerns section.

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

☑ B1. Did you cite the creators of artifacts you used?
Sections 3.2 and 4.1

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The focus is on the empirical study, and the utilized artifacts are standard research tools whose license information can be found in their original cited sources.

☒ 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)?
The use of standard artifacts (e.g., using evaluation benchmarks for evaluation) is conventional and aligns with their intended purpose, making an explicit discussion redundant for the expert audience.

☒ 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?
The study uses public research datasets or synthetic data not expected to contain PII. The ethical focus is on the potential misuse of the attack methods, not the data content, where risk is considered negligible.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.2

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?
Section 3.2

☑ **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?
While some model names indicate their parameter count (e.g., Salmonn-7B), the paper does not detail the computational budget or infrastructure. The study focuses on evaluating the robustness of pre-trained models via inference, where such details are often considered secondary to the experimental outcomes.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The study evaluates pre-trained models through inference, which typically does not involve hyperparameter tuning. The experimental setup is detailed in terms of tasks, models, and datasets, but not model-specific hyperparameters, as they are not relevant to the core investigation.

☑ 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 3.3

☒ 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?
The paper identifies the models and versions used, but omits detailed parameters as standard settings were employed and these values are not central to the main conclusions on LALM robustness

☑ **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 D

☒ 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 study refers to the participants as "human volunteers", which implies they were not compensated. This human assessment was a small-scale validation (100 items per task) of the primary automated evaluation metrics. In the context of such a validation performed by a small number of expert evaluators, a formal discussion of recruitment and payment is often omitted as it is not a large-scale crowdsourcing or formal human-subject study.

☒ 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)?
The study uses public datasets where consent is presumed to have been handled by the original creators. For the evaluation conducted by "human volunteers", a formal consent discussion is not included, as this was a small-scale validation task integral to the research process, not a formal human-subjects study.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This type of computational research, which uses public data and involves expert author evaluation, typically does not require a formal ethics review board approval as it poses minimal risk to human subjects.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The human assessment was conducted by "two human volunteers" to validate the automated metrics. As this was a small-scale check rather than a large-scale data annotation effort, details about the annotators were not included.

☒ **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)