Responsible NLP Checklist

Paper title: *Jamendo-MT-QA: A Benchmark for Multi-Track Comparative Music Question Answering*
Authors: *Junyoung Koh, Jaeyun Lee, Soo Yong Kim, GYU HYEONG CHOI, Jung In Koh, Jordan Phillips, Yeonjin Lee, Min Song*

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 "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?
Section "Ethical Considerations". We derive the dataset from Creative Commons-licensed Jamendo audio and release only annotations and metadata. We monitor/filter potentially sensitive or biased content and do not collect personal or sensitive information in the human evaluation.

☑ 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 3.5; Appendix E (Statistical Analysis of the Dataset), including Tables 1016.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 4.14.2; Appendix H (Baseline Inference Settings), especially Tables 1718 and prompt templates in H.3.

☑ 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 4.3; Table 1; Appendix A.1 and Appendix E. We report aggregate human means/MAD, benchmark scores across models, and descriptive dataset statistics.

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.?
We provide the evaluation criteria and rating scales in Appendix A.1, but not the full verbatim participant instruction text or screenshots due to space.

☒ 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)?
Annotators were recruited internally for research quality control. We do not report payment details because no crowdsourcing platform or external paid recruitment was used.

☒ 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)?
(left blank)

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section "Ethical Considerations" (Human Subjects): the study was determined to be exempt from formal IRB review under institutional guidelines.

☑ **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?
Sections 3 and 4; Appendix C (Prompt Template for Multi-Track QA Generation); Appendix D (LLM-as-a-Judge); Appendix H.