Responsible NLP Checklist

Paper title: *Musical Score Understanding Benchmark: Evaluating Large Language Models' Comprehension of Complete Musical Scores*
Authors: *Congren Dai, Yue Yang, Krinos Li, Huichi Zhou, Shijie Liang, Zhang Bo, Enyang Liu, Ge Jin, Hongran An, Haosen Zhang, Peiyuan Jing, KinHei Lee, Zhenxuan Zhang, Xiaobing Li, Maosong Sun*

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?
No dedicated potential risks section is included. The paper focuses on benchmark construction and evaluation for musical score understanding, and does not involve deployment in safety-critical settings.

☑ 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?
The benchmark is constructed from musical scores and question–answer annotations rather than personal data. Annotator recruitment and study context are described in Appendix: Recruitment and Payment.

☑ 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?
See Section 3 and Section 4.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 5.1.

☑ 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?
See Section 5.

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.?
See Appendix: Instructions Given to 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)?
See Appendix: Recruitment and Payment.

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)?
The study uses public musical scores and expert annotation of benchmark answers rather than personal data collected from individuals for downstream release.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No formal ethics review board approval was obtained. The study involved voluntary expert annotators, no sensitive personal data, and the procedure followed the host institution's ethical guidelines as described in Appendix: Recruitment and Payment.

☑ **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?
See Appendix: Use of LLMs.