Responsible NLP Checklist

Paper title: *MARS-Bench: A Multi-turn Athletic Real-world Scenario Benchmark for Dialogue Evaluation*
Authors: *Chenghao Yang, Yinbo Luo, Zhoufutu Wen, Qi Chu, Tao Gong, Longxiang Liu, Kaiyuan Zhang, Jianpeng Jiao, Ge Zhang, Wenhao Huang, Nenghai Yu*

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?
We note that domain specificity and automatic evaluation biases may limit generalizability and reliability. These concerns are discussed in the limitations.

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

☑ B1. Did you cite the creators of artifacts you used?
See Table 1 for the MathQA dataset and Table 3 for the evaluation models used. Metric settings are described in Section 4.1.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
All datasets used are open-source and properly cited. We access LLMs via paid APIs under their respective terms of service.

☑ 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)?
We used all datasets and APIs in accordance with their intended research use. Our benchmark is for research use only.

☒ 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 data comes from public sports sources and was manually reviewed to ensure it contains no personally identifying information or offensive content.

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

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?
See Section 3.3 for dataset statistics and task composition.

☑ **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?
The evaluations were conducted via third-party paid APIs, and the providers do not disclose model sizes or compute infrastructure. We therefore do not have access to model size or compute budget details.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 4.1

☑ 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 4.2

N/A 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?
(left blank)

☑ **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.?
Annotators were instructed to verify the factual correctness of generated questions. No subjective labeling or sensitive content was involved.

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)?
(left blank)

☑ 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 data is collected from publicly available sports records that do not involve personal user data, so individual consent was not required.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The data was collected from publicly available sources and does not involve human subjects, so ethics review board approval was not required.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
All annotation was performed by the authors. No additional demographic data was collected. See Section 3.1

☑ **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?
AI assistants are used to help coding.