Responsible NLP Checklist

Paper title: *League of LLMs: A Benchmark-Free Paradigm for Mutual Evaluation of Large Language Models*
Authors: *Qianhong Guo, Wei Xie, Xiaofang Cai, Enze Wang, Shuoyoucheng Ma, Xiaobing Sun, Tian Xia, Kai Chen, Xiaofeng Wang, Baosheng Wang*

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?
Yes. We discuss potential risks in the Limitations section (10).

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

N/A 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?
N/A. Our artifacts are model-generated questions/answers/scores (math/programming) and do not involve human-subject data or personally identifying information; we only perform lightweight human sanity checks of generated content.

☑ 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?
Yes. Data scale and statistics are reported in Section 3.3.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. We describe the experimental setup and key hyperparameters in Section 3.3.

☑ 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?
We report descriptive statistics in Sections 46 and Appendices AE (e.g., mean/SD/95% CIs across runs, and significance tests where applicable).

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.?
No. The paper only reports a sanity check: the generated questions and reference answers were independently validated by at least three human annotators; no detailed, standardized instruction protocol is described.

☒ 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)?
No. The paper does not report recruitment or payment details; annotators only participated in an internal sanity check.

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)?
N/A. We did not collect or use any personal/human-subject data; annotators only validated the generated questions and reference answers as a sanity check.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. The annotators only conducted a minimal-risk internal sanity check and no personal/sensitive data were collected.

☑ **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?
Yes. We used AI assistants (e.g., ChatGPT) for writing assistance (e.g., language polishing); all experiments, analyses, and conclusions were produced and verified by the authors.