## Responsible NLP Checklist

Paper title: *Challenging the Boundaries of Reasoning: An Olympiad-Level Math Benchmark for Large Language Models*
Authors: *Haoxiang Sun, Yingqian Min, Zhipeng Chen, Xin Zhao, Ji-Rong Wen*

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.*

N/A A2. Did you discuss any potential risks of your work?
*The work introduces a mathematical reasoning benchmark and does not pose direct potential risks. The dataset consists of Olympiad-level math problems sourced from printed publications and does not involve sensitive, harmful, or dual-use content.*

☑ **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?
*The dataset consists entirely of Olympiad-level mathematical problems with numerical or formal answers. It does not contain any personally identifying information or offensive 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?
*Detailed statistics including problem counts, category distributions, answer format specifications, and contamination analysis are reported in Section 3 (Tables 1, 2, 3, and Table 11 in Appendix).*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Experimental setup and hyperparameters are described in Section 4.1.1 (natural language evaluation) and Section 4.2.1 (formal language evaluation).*

☑ 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?
*Multiple metrics are reported across all experiments, broken down by category, language, and difficulty level in Section 4 (Tables 410). Models with fewer samples are explicitly marked to indicate potential instability.*

---

*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.?
*The roles and tasks of human annotators are described in Section 3.1 (verification) and Section 3.4 (bilingual translation).*

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)?
*The annotators are academic collaborators involved in the research project, not crowdworkers or paid external participants.*

☑ 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)?
*We have received agreement from human annotators.*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*The work involves constructing a mathematics benchmark and does not require ethics board approval. Human annotators participated as expert collaborators.*

☑ **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?
*The paper fully discloses how LLMs were utilized in the dataset construction process in Section 3.1, Section 3.4, and Appendix A.2. Additionally, AI assistants were employed for language polishing and refinement of the paper.*