Responsible NLP Checklist

Paper title: *IRIS: Interleaved Reinforcement with Incremental Staged Curriculum for Cross-Lingual Mathematical Reasoning*
Authors: *Navya Gupta, Rishitej Reddy Vyalla, Avinash Anand, Chhavi Kirtani, Erik Cambria, Zhengchen Zhang, Zhengkui Wang, Timothy Liu, Aik Beng Ng, Simon See, Rajiv Ratn Shah*

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?
(left blank)

☑ **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?
(left blank)

☑ 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 1 and 3.1 report dataset statistics including total size (29k examples), difficulty splits, and train/test split methodology across all three languages.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 and Appendix report full hardware specifications, hyperparameters including learning rates, optimizer, warmup ratio, and batch size.

☑ 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?
Appendix reports additional training reward dynamics and curriculum configuration comparisons across all experimental settings.

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.?
Annotators were given task-specific guidelines to verify step-wise correctness of high-schoollevel mathematical solutions. These instructions were internal, non-sensitive, and focused on mathematical validity; full instruction text was not included in the paper due to space constraints. A summary of the annotation process and validation criteria is provided in Section 3.1 and Appendix 12.4.

☒ 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 undergraduate and master's students with technical backgrounds who participated as part of an academic research setting. The annotation task involved validating mathematical reasoning and did not involve sensitive content. Detailed recruitment and compensation information was not included, as participation followed standard institutional research practices.

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)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **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?
Section 3.1 discloses use of Llama-3.3-70B for generating step-wise reasoning annotations and IndicTrans2 for multilingual translation of problems and solutions.