Responsible NLP Checklist

Paper title: *On the Step Length Confounding in LLM Reasoning Data Selection*
Authors: *Bing Wang, Rui Miao, Chen Shen, Shaotian Yan, kaiyuan liu, Ximing Li, Xiaosong Yuan, Sinan Fan, Jun Zhang, Jieping Ye*

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?
This paper is an experimental analysis and methodological study that focuses on general reasoning tasks of large language models, and involves no potential risks.

☑ 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 training dataset we used was derived from publicly available mathematical reasoning datasets, and the evaluation benchmarks are also widely adopted datasets. As a result, our data contain no information that identifies individual people or any offensive content. In addition, our code has been carefully anonymized.

☑ 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?
In Section2.2 and AppendixB, we provide a detailed description of the training data used in our study, including the number of questions, the number of responses, and other relevant statistics. The evaluation benchmarks we employed are identical to those used in the original work.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section2.2 and AppendixB provide a complete description of our experimental setup, including training parameters, data scale, and model cards.

☑ 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?
Tables1 and3 clearly show that the metric we used is simple accuracy.

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?**

☐ N/A 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.?
(left blank)

☐ 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)

☐ 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?
We do not use the AI assistants.