Responsible NLP Checklist

Paper title: *a1: Steep Test-time Scaling Law via Environment Augmented Generation*
Authors: *Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Yuyao Ge, Jun Wan, Yurong Wu, Xueqi Cheng*

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?
Section: Ethics Statement. We discuss dual-use potential of enhanced reasoning capabilities and computational resource considerations.

☑ **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?
Our EAG-2K dataset consists of mathematical and scientific reasoning problems derived from the s1 dataset. It contains no personally identifying information or offensive content as it focuses on abstract mathematical and logical reasoning tasks.

☑ 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 3 (Dataset). We report dataset size (2,000 samples), composition breakdown (Raw: 200, Iterative-Refinement: 800, Direct-Execution: 1,000), token length statistics, and success rates across retry attempts (Table 1).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.4 (Training Details). We report training configuration including: 8 epochs, batch size 8, learning rate 8e-6, warmup steps, optimizer settings (AdamW with =0.9, =0.95), sequence length (12K tokens), and hardware (8 NVIDIA A100 GPUs, ~12 hours).

☑ 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?
Section 5 (Experiments). We report Pass@1 metrics for all models across multiple benchmarks

(AIME24, AIME25, MATH500, GPQA). Results show mean performance; we use greedy decoding for reproducibility. Ablation study (Table 2) provides comparative statistics across model variants.

☒ **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?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)