Responsible NLP Checklist

Paper title: *Thinking Traps in Long Chain-of-Thought: A Measurable Study and Trap-Aware Adaptive Restart*
Authors: *chenkang, Fan Yu, Junjie Nian, SIHAN ZHAO, Zhuoka Feng, Zijun Yao, wang heng, Yu Minshen, Yixin Cao*

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 4.2 and Appendices FH report the data construction statistics, including 6,000 raw trajectories from 1,500 problems, 3,661 labeled trajectories after filtering, difficulty/pattern breakdowns, and the 80/10/10 train/dev/test split.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 4.2, 4.4, and 5.1, together with Appendices CD, report the experimental setup and the hyperparameters actually used, including decoding settings, the resampling budget N=36, restart thresholds, and policy training settings (e.g., learning rate 1e-5, 8H20 GPUs, epoch=1, maximum sequence length 36k). We did not perform an extensive hyperparameter search beyond these reported settings.

☑ 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?
Sections 5.2 and 6, together with Tables 16 and Appendices KO, report descriptive statistics including

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.

trap ratios, counts, average accuracies, escape rates, AUC-ROC, token overhead, correlations, and intervention transition statistics. The paper is explicit about summary metrics such as AVG@4, VOTE@16, and Pass@16. We do not report multi-seed error bars.

☑ **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.?
Section 4.2 and Appendix I describe the manual audit protocol, but the paper does not provide the full text of annotator instructions, screenshots, or participant-facing risk disclosures.

☐ 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?
language polishing, coding assistance