Responsible NLP Checklist

Paper title: *One Refiner to Unlock Them All: Inference-Time Reasoning Elicitation via Reinforcement Query Refinement*
Authors: *Yixiao Zhou, Dongzhou Cheng, zhiliang wu, Yi Yang, Yu Cheng, Hehe Fan*

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 study is restricted to mathematical and logical reasoning tasks using public benchmarks. The proposed framework does not interact with human users or generate sensitive content, thus posing no direct societal risk.

☑ **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?
We exclusively utilize standard, publicly available mathematical and logical reasoning benchmarks (e.g., GSM8K, MATH, MMLU-Pro). These datasets are widely used in the NLP community and are not known to contain personally identifying information or offensive content. The curated SFT data is derived from these same vetted sources.

☑ 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 4.1, Appendix C.1, Appendix C.2, Appendix E.1, and Appendix E.2

☑ **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 C.4

☑ 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 4.1, Section 4.2, Appendix D

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.?
This research does not involve human subjects, annotators, or crowdworkers. All experiments were conducted using automated large language models and established public benchmarks.

☒ 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)?
All data curation and evaluation were performed by the authors or through automated processes using large language models. No external human participants or crowdworkers were recruited.

☒ 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)?
This research exclusively utilizes standard, publicly available reasoning benchmarks (e.g., GSM8K, MATH, MMLU-Pro) and data curated internally by the authors. No new data from human participants was collected for this study.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This study is a purely computational research project that utilizes publicly available, anonymized datasets (e.g., GSM8K, MATH, MMLU-Pro). Since it does not involve human participants, clinical trials, or the collection of sensitive personal data, ethics review board approval was not required.

☑ **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?
Appendix H