Responsible NLP Checklist

Paper title: *CoTJudger: A Graph-Driven Framework for Automatic Evaluation of Chain-of-Thought Efficiency and Redundancy in LRMs*
Authors: *Siyi Li, Jiajun Shi, Shiwen Ni, Ge Zhang, Shuaimin Li, Shijian Wang, Zhoufutu Wen, Yizhi LI, Hamid Alinejad-Rokny, Jiaheng Liu, Min Yang, Wenhao Huang*

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?
The work focuses on automated evaluation of Chain-of-Thought (CoT) efficiency and redundancy in Large Reasoning Models (LRMs) via a graph-driven framework. It does not involve human subjects, sensitive data handling, or deployment of models in real-world scenarios that would introduce direct risks. The research is purely analytical and diagnostic, with no inherent risks to individuals, systems, or society.

☑ **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 dataset (896 queries) is constructed from open-source benchmarks (OmniMath, HumanEval, GPQA, Big-Bench Hard) that are publicly available and designed for research purposes. These benchmarks do not contain personally identifying information (PII) or offensive content by design. The paper also specifies data cleaning steps (Appendix A) to remove open-ended queries and excessively long answers, but PII/offensive content was not a concern given the source of the data.

☑ 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 (Data Description) and Appendix A report detailed statistics: total query count (896), domain distribution (Math: 360, General Reasoning: 270, Programming: 164, PCB: 98), subdomain breakdowns (e.g., Math subdomains include Logic, Algebra, Calculus), and core development set details (2,688 CoTs from 3 models). Appendix A also provides subdomain distribution percentages.

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

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.

The experiments involve evaluating 21 LRMs using greedy decoding without system instructions (Section 5.1) to standardize generation. Since the work is an evaluation framework (not model training or fine-tuning), there are no hyperparameters to search or tune. Model-specific adaptations (e.g., CoT extraction from API fields) are detailed in Appendix E, but no hyperparameter optimization is required for the frameworks modules.

☑ 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 5.2 (Main Evaluation Results)

☒ **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?
The framework (CoTJudger) uses GPT-5 for key modules, including step atomization (Section 4.1), atomic node classification (Section 4.2), answer node detection/verification (Section 4.3), graph construction (Section 4.4), and path validation (Section 4.5). Detailed prompts for these GPT-5-assisted modules are provided in Appendix D (Figures 915), including structured output schemas to ensure reliability.