Responsible NLP Checklist

Paper title: *Conflicts Make Large Reasoning Models Vulnerable to Attacks*
Authors: *Honghao Liu, Chengjin Xu, Xuhui Jiang, Cehao Yang, Shengming Yin, Zhengwu Ma, Lionel Ni, Jian Guo*

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?
Our method can expose vulnerabilities in large reasoning models (LRMs) by injecting conflicting prompts, as illustrated in Figure 1, where all harmful details are abstracted to prevent misuse. While this approach advances research on model robustness, it also presents dual-use risks, potentially enabling jailbreaks if misapplied.

☑ **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 use third-party benchmarks like HarmfulQ and HarmBench rather than collecting original data, no new personally identifying information or unintended offensive content was gathered.

☑ 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?
We provided the number of queries in each benchmark in Appendix B (Details of Benchmarks).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup and hyperparameters are discussed in Section 4.1 (Experimental setup) and Appendix A (Inference Model Configurations).

☑ 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?
We provided the bar plot with error bars (Figure 6) and Tables of averages and variances (Table 10, Table 11, Table 12) in Appendix D (Additional Experiments).

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 mentioned the use of Ai Assistant in a section after ethical consideration.