Responsible NLP Checklist

Paper title: *AutoRAN: Automated Hijacking of Safety Reasoning in Large Reasoning Models*
Authors: *Jiacheng Liang, Tanqiu Jiang, Yuhui Wang, Rongyi Zhu, Fenglong Ma, Ting Wang*

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?
Potential risks are discussed in the "Ethics Statement" section.

☑ **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?
Discussed in the Ethics Statement. We used established public safety benchmarks (AdvBench, HarmBench, StrongReject, XSTest) curated for adversarial research. Harmful outputs were used strictly for analytical purposes and not redistributed.

☑ 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?
Dataset statistics are provided in Section 4.1: AdvBench (50 prompts), HarmBench (50 prompts), StrongReject (54 prompts). Defensive alignment dataset details (500 attack-response pairs) are in Section 5.4.1.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experimental setup and hyperparameters are described in Section 4.1. Key hyperparameters: maximum iterations nturn=10n_turn=10 nturn=10, helpfulness threshold h=7h^*=7 h=7. Full algorithm is in Appendix C.

☑ 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?
Descriptive statistics including ASR and ANQ are reported in Tables 110. Re-sampling stability analysis (mean std across two runs) is provided in Appendix A.5, Table 15.

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?**

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