Responsible NLP Checklist

Paper title: *ASTRA: An Automated Framework for Strategy Discovery, Retrieval, and Evolution for Jailbreaking LLMs*
Authors: *Xu Liu, Yan Chen, Kan Ling, Yichi Zhu, Hengrun Zhang, Guisheng Fan, Huiqun Yu*

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?
See Section Ethical Considerations.

☑ **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?
See Ethical Considerations. We utilize standard red-teaming datasets (HarmBench, AdvBench) which intentionally contain harmful/offensive queries for the purpose of robustness evaluation. These datasets do not contain personally identifiable information. We explicitly acknowledge the offensive nature of the data and discuss steps to prevent misuse, such as committing not to disseminate the generated harmful content.

☑ 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?
See Section 4.1 Experimental Setup.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 4.1 and 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?
See Section 4.1 Experimental Setup.

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.?
Not applicable. No human subjects or annotators were involved.

☒ 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)?
Not applicable. No human subjects or annotators were involved.

☒ 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)?
Not applicable. This study utilizes standard, publicly available open-source datasets.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. This research is purely computational and utilizes existing publicly available datasets. It does not involve human subjects, user studies, or primary data collection from individuals, and thus falls outside the scope of Ethics Review Board oversight regarding human participants.

☑ **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 utilized AI assistants for grammatical polishing, sentence refinement, and generating LaTeX formatting code. All scientific claims and the final manuscript were verified by the authors. Separately, Large Language Models (e.g., GPT-4o, Llama-3) were employed as the experimental subjects and core components (Attacker, Judge, Target) of the proposed framework, as detailed in Section 3 (Methodology) and Section 4 (Experiments).