

## Responsible NLP Checklist

Paper title: *Jailbreaking Attacks vs. Content Safety Filters: How Far Are We in the LLM Safety Arms Race?*

Authors: *Yuan Xin, Dingfan Chen, Linyi Yang, Michael Backes, Xiao Zhang*

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?

*Yes. We include a content warning below the title and discuss risks in the Ethical Considerations 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?

*Yes, we discuss this in the Ethical Considerations section.*

- 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 3.3. We report dataset size (417 harmful and 417 benign prompts), category distribution, and data sources.*

### C. Did you run computational experiments?

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

*Section 3.3 and Appendix A.1. We describe all models, datasets, attack configurations, filter thresholds, and hyperparameters.*

- 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?

*No. Each attack-filter-model combination is evaluated in a single run following the official implementations. The evaluation is deterministic given fixed model outputs and filter predictions.*

---

*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.?

*N/A. No human subjects or annotators were involved in this work.*

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)?

*N/A. No human participants 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)?

*N/A. All data is derived from previously published benchmarks.*

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*N/A. No human subjects research was conducted.*

**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?

*Yes. See the Acknowledgments section. We used AI assistants for proofreading and editing.*