Responsible NLP Checklist

Paper title: *RAP-ID: Mechanistic Prompt Injection Detection via Impostor Behavior Analysis*
Authors: *Yuchen Yang, Lei Peng, Yujie He, yang yu, Zhongxin Wu, Yanlei Shi*

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?
Section 6 (Limitations) and Ethical Considerations. We discuss potential dual-use risks of the risk vocabulary and clarify that RAP-ID is intended as a complementary safety layer rather than a standalone safeguard.

☑ **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?
Appendix B.2 and Ethical Considerations. The paper discusses the construction of the risk vocabulary used by the PC module, which includes sensitive or offensive terms necessary for safety detection. Our evaluation uses public benchmarks, and we do not collect personal identifying information.

☑ 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 4.1 (Experimental Setup) describes the attack and benign datasets used in evaluation, and Appendix B.1/B.2 further details the system-prompt pool and risk vocabulary construction.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 describes the experimental setup and hyperparameter calibration procedure, and Appendix A.1 (Table 6) reports the exact hyperparameter settings used for the Qwen3-8B and Qwen2-1.5B variants.

☑ 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 4 reports the main evaluation metrics (F1, TPR, FPR), while Tables 25 provide ablations, threshold sensitivity, and cross-architecture results. Appendix A.2 reports efficiency statistics including latency, throughput, and peak memory.

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.?
N/A. This work does not involve human subjects or paid annotators.

☒ 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)?
N/A. This work does not involve human subjects or paid annotators.

☒ 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)?
N/A. This work does not involve human subjects or paid annotators.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. This work does not involve human subjects or paid annotators.

☑ **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?
AI assistants were used only for limited grammatical polishing and formatting assistance. They were not used to generate core scientific ideas, experimental results, or research conclusions.