Responsible NLP Checklist

Paper title: *PARASITE: Conditional System Prompt Poisoning to Hijack LLMs*
Authors: *Viet Pham, Thai Le*

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 discussed in the "Broader Impacts and 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?
No, our paper does not explicitly discuss whether the datasets contain personally identifying information or offensive content. We used publicly available datasets (TriviaQA, TruthfulQA) commonly used in prior work, and we assumed these had been vetted accordingly.

☑ 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?
Yes, we report the number of examples, the train/test splits, and paraphrasing strategies used to construct both the benign and malicious subsets in Sections 5 and 6, with additional statistics provided in Appendix B

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, we describe the Section 5, 6, and other hyperparameters in Appendix B.

☑ 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?
Yes, we report descriptive statistics including F1 and EM in Section 5 , 6, 7, 8, and Appendix D

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.?
No, we did not use human subjects or annotators in any part of our study. All data was collected from publicly available sources, and all evaluations were conducted automatically.

☒ 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 participants were recruited or paid, as the study did not involve any human annotation or human-subject interaction.

☒ 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)?
No, we did not discuss data consent because we exclusively used publicly available datasets (TriviaQA and TruthfulQA) that are commonly used in academic research.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No, ethics review board approval was not obtained because our study used publicly available datasets and did not involve human subjects, private data, or sensitive information.

☑ **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, we used AI assistants (ChatGPT) to support writing (grammar checking)