Responsible NLP Checklist

Paper title: *Proxy Barrier: A Hidden Repeater Layer Defense Against System Prompt Leakage and Jailbreaking*
Authors: *Pedro Schindler Freire Brasil Ribeiro, Iago Alves Brito, Rafael Teixeira Sousa, Fernanda Bufon Frber, Julia Soares Dollis, Arlindo Rodrigues Galvo Filho*

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?
This work focuses on a defensive mechanism to prevent misuse of LLMs. The primary goal is to reduce risks, and the methods described do not introduce new potential risks

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
Sec 4

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The artifacts used are well-known public datasets and models (e.g., Llama, GPT, Unnatural Instructions) cited according to academic standards. A detailed discussion of their licenses was not included as their use falls under standard research purposes

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
All datasets and models were used for their intended research purposes of evaluating LLM performance and security, but their intended use was not specifically discussed within the paper

☑ 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?
The datasets used are standard academic benchmarks that do not contain personally identifying information. The jailbreak datasets contain potentially offensive content by design, but it is used strictly for evaluating the model's safety defenses and is not reproduced in the paper

---

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.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
As we used existing, publicly available artifacts, we refer to the original papers for their detailed documentation

☑ 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 and the Appendix details with precision the number of prompts examples used for each experiment, along with their sources.

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and the Appendix specify the different model families and sizes used in the experiments, but computing infrastructure was not specified

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The Appendix provides detailed hyperparameters for the fine-tuning experiment

☑ 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?
Tables in Section 4 and the Appendix report results, clearly states as being from a single run.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
The Appendix describes the formulas and thresholds used for evaluation metrics employed: ROUGE-L

☒ **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)

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(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)