Responsible NLP Checklist

Paper title: *OASIS: Mitigating Harmful Fine-tuning Attacks on LLMs via Orthogonal and Adaptive Safety Alignment Strategy*
Authors: *Jiayu Tang, Guowei Peng, Qiuhao Xie, Yuning Yang, Xiurui Xie, Guisong Liu*

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?
Our work proposes a defense framework (OASIS) specifically designed to mitigate safety risks and protect models against harmful attacks. It does not introduce new potential risks.

☑ **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?
Section 4.1. We utilized the BeaverTails dataset, which inherently contains harmful/offensive content by design to study LLM safety. We strictly used it within a controlled experimental environment to simulate attacks and evaluate our defense mechanisms.

☑ 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). We explicitly state the sample sizes used for both the alignment phase (2,000 safe, 200 harmful) and the user fine-tuning evaluation phase (1,000 samples).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B (Implementation Details) and Appendix C (Additional Hyperparameter Analysis) detail the LoRA configurations, optimizer settings, learning rates, and model-specific selection budgets.

☑ 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 (Experiments). We report comprehensive average scores alongside individual metrics across varying contamination rates (p=0.05 to 0.4) to transparently present the performance trend.

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.?
No human participants were involved in this research.

☒ 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)?
No human participants were involved.

☒ 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)?
No human subjects data was collected.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable as no human subjects were involved.

☑ **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 tools were used exclusively for English language polishing and LaTeX formatting