Responsible NLP Checklist

Paper title: *StealthGraph: Exposing Domain-Specific Risks in LLMs through Knowledge-Graph-Guided Harmful Prompt Generation*
Authors: *Huawei Zheng, Xinqi Jiang, Sen Yang, Shouling Ji, Yingcai Wu, Dazhen Deng*

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?
We discuss potential risks and ethical considerations in a dedicated Ethical Considerations section. Specifically, we note the possible misuse of synthesized harmful prompts and describe mitigation measures such as limiting the release to abstracted prompt templates and examples, with the intended use restricted to LLM safety evaluation and red-teaming purposes.

☑ 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?
We discuss data content and related risks in a dedicated Ethical Considerations section. The study does not involve sensitive personal data, and all domain knowledge is derived from publicly available sources such as Wikidata. While the dataset contains harmful content by design, it is constructed solely for LLM safety evaluation, with mitigation measures such as abstracted prompt templates to minimize misuse.

☑ 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?
We report dataset statistics, including the number of prompts per dataset variant and per domain, in Section 4.1 (Experimental Setup, Datasets).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We describe the experimental setup, including datasets, models, evaluation metrics, and implementation details, in Section 4.1 (Experimental Setup), with full parameter settings provided in Appendix G.

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.

☑ 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?
We report descriptive statistics of experimental results in Section 4 (Experiments), including benchmarking results, safety fine-tuning performance, cross-domain analysis, and ablation studies, with all reported numbers corresponding to single-run evaluations under fixed settings.

☒ **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 study does not involve human participants or 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. No human participants were recruited or compensated in this study.

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. The study does not use data collected from human subjects.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. No human subjects or human data are involved, so ethics board approval was not required.

☑ **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?
In Appendix M (The Use of Large Language Models), we disclose the use of LLMs for limited language polishing, literature identification, and coding assistance. All substantive claims, methodology, experiments, data analysis, and result interpretation were conducted and verified by the authors.