Responsible NLP Checklist

Paper title: *StructBreak: Structural Cognitive Overload-Induced Safety Failures in MLLMs*
Authors: *Yang Luo, Liu Xinran, TianTian Ji, Zhiyi yin, Lingyun Peng, Shuyu Li*

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 1 and Section 7. The paper includes a warning about harmful content in the abstract and Section 1, and discusses the scope of the attacks and need for robust defenses in Section 7.

☑ **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 and Appendix E. The paper uses harmful queries from the SafeBench dataset to evaluate safety. To mitigate risk, we mask sensitive spans in qualitative case studies to avoid amplifying harm.

☑ 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. We detail the dataset statistics, specifically the random sampling of 100 harmful queries from SafeBench, evenly distributed (10 queries each) across ten distinct risk categories.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 and Appendix A detail the target MLLMs, evaluation baselines, Judge model settings, and specific prompt templates used.

☑ 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 4.2, Figure 3, and various tables report Attack Success Rate (ASR), First-Try Success Rate, Average Attempts, and explicit refusal rates across all target models.

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.?
Appendix A.4 and Figure 9 provide the exact tri-label evaluation rubric and definitions given to the human annotators to mirror the automated judge's logic.

☒ 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)?
We state in Appendix A.4 that we recruited three laboratory researchers with relevant domain expertise. We did not report payment details as the annotators were internal lab members conducting this verification as part of their standard research duties.

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)

☑ **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?
We used ChatGPT for grammatical proofreading and phrasing improvements