Responsible NLP Checklist

Paper title: *ECHA: Jailbreaking LVLMs via the Mismatch between Implicit Semantic Reconstruction and Explicit Safety Alignment*
Authors: *Chenxing Xu, junyong jiang, Zehu Zhang, Lu Dong*

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?
In the unnumbered "Limitations" and "Ethics Statement" sections, which are located immediately after Section 5 (Conclusion).

☑ **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 used publicly available safety benchmarks (HADES and SafeBench), which do not contain Personally Identifying Information (PII). Regarding offensive content, we did not take steps to remove or anonymize it because the explicit goal of this red-teaming research is to evaluate model vulnerabilities against harmful queries; removing such content would invalidate the study. Instead of anonymization, we addressed the inherent risks by including an explicit "Content Warning" in the Abstract and detailing strict, controlled data release guidelines in the unnumbered "Ethics Statement" section to mitigate abuse.

☑ 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?
In Section 4.1 (under the "Datasets" paragraph) and Appendix A.2. We clearly reported the total number of evaluation queries and their distribution across different harm categories for both the HADES and SafeBench datasets. Please note that since this is a black-box evaluation study focused on inference, train/test/dev splits are not applicable.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 4.1 (specifically under the "Implementation Details" paragraph). We detailed the target models, baselines, datasets, and the auxiliary model used. Please note that since this is an inference-only black-box jailbreak evaluation rather than a model training task, traditional hyperparameter

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.

search is not applicable. However, we explicitly reported the critical inference hyperparameter used to ensure reproducibility (temperature = 0).

☑ 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?
In Section 4.1 (under the "Implementation Details" paragraph), we explicitly state that we conducted each experiment three times and reported the average Attack Success Rate (ASR). This ensures transparency regarding the summary statistics provided in our results tables.

☑ **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.?
In Appendix B.2 (under the "Human Evaluation of the LLM-based Judge" section), we clearly reported the rigorous assessment criteria and the specific instructions provided to the expert human annotators. We detailed the exact conditions under which an annotator was instructed to classify a response as a successful jailbreak versus an unsuccessful one.

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)?
We selected "N/A" because there was no external recruitment or payment involved in our study. The human evaluation was a small-scale expert meta-evaluation (auditing 80 sampled outputs) conducted internally by expert researchers with domain knowledge, rather than through crowdsourcing platforms or compensated external participants. Therefore, reporting on recruitment and payment adequacy is not applicable.

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)?
We selected "N/A" because our study does not collect, use, or curate personal data from human subjects. The datasets utilized in our experiments (HADES and SafeBench) are existing, publicly available academic safety benchmarks consisting of adversarial text prompts. Since no personal human data is involved, discussing human data consent is completely not applicable to this research.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We selected "N/A" because formal Ethics Review Board (IRB) approval was not required for this study. The research does not involve human subjects experiments or the collection of personal data. The experiments strictly utilize publicly available datasets (HADES and SafeBench) to evaluate the vulnerabilities of pre-trained large vision-language models. The human evaluation component mentioned in our appendix was strictly an internal data-quality audit conducted by expert researchers to verify the automated metrics, not a behavioral study on human participants.

☑ **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?
Section 4.1 details the use of an auxiliary LLM (Qwen3-Next) for generating emoji mappings.