Responsible NLP Checklist

Paper title: *AgenticEval: Toward Agentic and Self-Evolving Safety Evaluation of Large Language Models*
Authors: *Yixu Wang, Xin Wang, Yang Yao, Xinyuan Li, Xibang Yang, Yan Teng, Xingjun Ma, Yingchun Wang*

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?
Ethical Considerations section

☑ **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?
Our evaluation data include adversarial/jailbreaking prompts and may cover sensitive/offensive content categories by design (e.g., hateful/violent or obscene content) for the purpose of safety testing. We do not intentionally collect or include real personally identifying information; test cases are generated/synthesized for controlled, defensive evaluation. We highlight the defensive intent and controlled research setting in Ethical Considerations.

☑ 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 the Experimental Setup section

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In the Experimental Setup section

☑ 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 the Experimental Results section

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.?
The only human involvement was a small-scale internal annotation conducted by the papers authors for validation purposes (no external participants were recruited). Therefore, we did not provide separate participant-facing instruction text beyond the shared internal guideline used by the authors.

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)

☑ **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 AI assistants (e.g., ChatGPT) for language polishing and clarity improvements in the manuscript. All technical claims, experimental results, and final wording were verified and finalized by the authors.