Responsible NLP Checklist

Paper title: *Robust Tool Use via Fission-GRPO: Learning to Recover from Execution Errors*
Authors: *Zhiwei Zhang, Fei zhao, Rui Wang, Zezhong WANG, Bin Liang, Jiakang Wang, Yao Hu, Shaosheng Cao, Kam-Fai Wong*

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 did not include a dedicated discussion of potential risks in the paper. The work focuses on improving robustness in tool-use models and is evaluated on benchmark environments with simulated interactions rather than real-world deployment.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

N/A 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?
(left blank)

☑ 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 report key statistics of the constructed training data, including the number of domains (11), the initial trajectory pool (~2,000), and the final retained training instances (630), along with benchmark settings used for evaluation.

☑ **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 C. We describe the training framework, hardware, optimization settings, rollout group sizes, sequence lengths, sampling parameters, and the error-identification threshold used in our experiments.

☒ 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 task-level benchmark metrics and ablation results, but do not include error bars or multi-seed summary statistics. The reported numbers correspond to the evaluated runs described in the paper.

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.?
We report the purpose and outcome of the human evaluation for non-leakage, but do not include the full annotation instructions in the paper.

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?
Sections 3.3.2 and 4.1. We describe the use of AI systems (including Claude Sonnet 4, Qwen3-235B-A22B-Instruct-2507, and Kimi K2) for data construction, quality filtering, verification, and simulator data preparation.