Responsible NLP Checklist

Paper title: *ExecVerify: White-Box RL with Verifiable Stepwise Rewards for Code Execution Reasoning*
Authors: *Lingxiao Tang, He Ye, Zhaoyang Chu, Muyang Ye, Zhongxin Liu, Xiaoxue Ren, Lingfeng Bao*

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 have dissussed the potential risks at Appendix G.

☑ **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 training data consists only of synthetically generated code snippets and automatically produced inputs. Since it is not collected from people or user logs, the likelihood of containing personal identifiers (PII) or offensive content is low.

☑ 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 2.1.2 reports dataset sizes before/after execution filtering and difficulty filtering. Additional distribution statistics are reported in Appendix A.2A.4.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. Experimental setup and evaluation details are described in Sections 34. Full training hyperparameters are provided in Appendix C, and additional ablations and experimental details are provided in Appendices DE.

☑ 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 specify decoding settings and metrics (e.g., greedy sampling with temperature 0.0 and pass@1) in Section 3/4, and define averaged reporting such as Avg across languages and metrics in Section 4 and Appendix E.

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?**

☒ 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.?
(left blank)

☒ 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?
No human subjects were involved.

☑ **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 use an LLM in the synthesis pipeline to generate constrained code snippets and inputs (Section 2.1), and we use a teacher model to generate CoT and trace translations under unified prompting for controlled comparisons (Appendix E.2).