## Responsible NLP Checklist

Paper title: *MARS$^2$: Scaling Multi-Agent Tree Search via Reinforcement Learning for Code Generation*
Authors: *Pengfei Li, Shijie Wang, Fangyuan Li, Yikun Fu, Kaifeng Liu, Kaiyan Zhang, Dazhi Zhang, Yuqiang Li, Biqing Qi, Bowen Zhou*

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?
*This work focuses on algorithmic and methodological advances in multi-agent reinforcement learning and code generation. It does not involve deployment in real-world systems, human subjects, or sensitive data. As such, we do not identify any immediate or specific potential risks associated with this work.*

☑ **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?
*The datasets used in this work consist of publicly available benchmarks and synthetic or automatically generated data for code generation tasks. They do not contain personal data or information that can identify individuals. Therefore, considerations regarding personally identifying information or offensive content are not applicable.*

☑ 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?
*Dataset statistics and evaluation splits are described in Section 3 (Experimental Setup).*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*The experimental setup and hyperparameter settings are described in Section 3 (Experimental Setup) and Appendix E*

☒ 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?
*Our experiments follow standard evaluation protocols for the considered benchmarks, where results*

*are reported as single-run metrics (e.g., Pass@k) without aggregating over multiple random seeds. As such, additional descriptive statistics such as error bars are not reported.*

☒ **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.?
*This work does not involve human participants or human annotators. Therefore, no instructions were provided, and this question is not applicable.*

☒ 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)?
*No human participants or annotators were involved in this work. As such, there was no recruitment or payment, and this question 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)?
*This work does not involve the collection or curation of data from human participants. All datasets used are publicly available benchmarks or automatically generated data, and therefore issues of participant consent are not applicable.*

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No human subjects or human-generated data collection was conducted in this work. As a result, ethics review board approval was not required, and this question is not applicable.*

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☒ N/A E1. If you used AI assistants, did you include information about their use?
*We did not use any AI assistants in our research, coding, or writing.*