## Responsible NLP Checklist

Paper title: *Closing the Spatial Execution Gap in Digital Whiteboards via Verifiable Reinforcement Learning*
Authors: *Chang Liu, Benjamin Wagley, Zibo Wang, Mehmet E. Belviranli, Bo Wu*

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?
*Our manuscript studies the training of an LLM-based agent to understand and perform operations on a digital whiteboard. We design the benchmark tasks and construct the dataset using an automated generator as part of this work. The entire process does not involve any identifiable risks to individuals, groups, or society.*

☑ **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?
*The dataset used in this work is fully synthetic and generated automatically as part of the benchmark design. It does not contain any real-world personal information, references to individuals, or user-generated content, and therefore does not include personally identifying information or offensive material. As a result, no anonymization, filtering, or content protection steps are required.*

☑ 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; Appendix A.5*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4.2; Section 4.3; Appendix A.5*

☒ 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 evaluation results on a fixed benchmark and do not report descriptive statistics such as variance, standard deviation, or confidence intervals across multiple runs. The benchmark tasks*

*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.*

*are artificial and the evaluation procedure is deterministic; moreover, given the computational cost of running reinforcement learning training over all data points, we therefore report single-run performance.*

☒ **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 any human subjects or annotators. All benchmark tasks and datasets are automatically generated, and no participants were recruited or instructed.*

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)?
*This work does not involve any human participants, annotators, or crowdworkers. Therefore, there was no recruitment or payment process to report.*

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 use or curate any data originating from human participants. All benchmark tasks and datasets are automatically generated and synthetic, so no consent process is applicable.*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*This work does not involve any human subjects, or personal data. Therefore, ethics review board approval or exemption was not required.*

☑ **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?
*Appendix A.8*