## Responsible NLP Checklist

Paper title: *ARCHITECT: Uncertainty-Aware Dynamic Tool Learning via Causal Intervention for Open-World Agents*
Authors: *Zhangyi Wang, Jiexiang Xu, Bingnan Yu, Zongze Li*

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?
*Section 6 (Limitations) discusses potential risks including sandbox security for code execution and the possibility of generated tools being misused. We mitigate these through Docker isolation and validation pipelines.*

☑ **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?
*Our work uses publicly available benchmarks (StableToolBench, MINT, T-Eval, SWE-bench) that do not contain PII or offensive content.*

☑ 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 3.2 reports data statistics: 1,247 tools with 7:1.5:1.5 train/dev/test split (873/187/187). Section 4.1 describes benchmark sizes.*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 3.2 reports hyperparameters: Adam optimizer, lr=5e-4, batch size 32, =0.5, early stopping at epoch 45.*

☑ 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?
*Section 4 reports task success rates across four benchmarks. Each task is evaluated once following official benchmark protocols. Results are reported as single-run values consistent with official evaluation methodology.*

---

*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.?
*Annotation instructions were provided verbally to expert annotators but not included in the paper due to space constraints.*

☒ 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)?
*Annotators were co-authors/lab members; no external recruitment or payment involved.*

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)?
*Not applicable. We use only publicly available benchmark datasets; no personal data was collected or curated.*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No IRB required as annotators were research team members annotating tool failures, not human subjects research.*

☑ **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?
*AI assistants (ChatGPT/Copilot) were used for English grammar polishing and code debugging assistance. All generated content was carefully reviewed, verified, and revised by the authors. The core research ideas, methodology, and experimental design are entirely original work by the authors.*