Responsible NLP Checklist

Paper title: *Action Boundary Blindness: When LLM Agents Cannot Tell Where One Action Ends and Another Begins*
Authors: *Zhangyi Wang, Bingnan Yu, Jiexiang Xu, 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?
Limitations section and Section 1 (Introduction) discuss process fidelity and safety implications for enterprise AI 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?
We use existing public benchmarks (-bench, WebArena, ALFWorld, TheAgentCompany, OSWorld) that do not contain personally identifying information. Our evaluation framework processes agent trajectories without collecting personal data.

☑ 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 4.4 (Experimental Setup) and Table 2 provide dataset statistics including number of tasks (1,655 total), average steps per task, and domain types. Appendix F provides detailed statistics including min/max/median step counts and action space sizes.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.4 (Experimental Setup) and Appendix B (Table 7) report hyperparameters including temperature (0.0), max tokens (4,096), similarity threshold (0.85), alignment parameters (=1, =1, =0.5), and BASR threshold (0.8).

☑ 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?

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.

Section 5 reports statistical significance with p-values (e.g., $p < 0.001$ for key comparisons), correlation coefficients (e.g., Spearman = -0.89, $p < 0.001$), and inter-annotator agreement (Fleiss' = 0.78). Tables 3-6 present results across all models, frameworks, and benchmarks.

☑ **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.?
The paper reports inter-annotator agreement metrics (Fleiss' = 0.78 for primary attribution, = 0.73 for human audit) and sample sizes (N=200, N=150, N=500), but does not include the full annotation guidelines in the paper or appendix. Annotation protocols will be released with the supplementary materials.

☒ 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)?
The paper describes annotators as "expert annotators" and "human experts" but does not provide details on recruitment or compensation. Annotators were research team members with NLP expertise, not crowdworkers.

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)?
Human annotators were research team members who voluntarily participated in the validation study. We used existing public benchmarks (-bench, WebArena, ALFWorld, TheAgentCompany, OSWorld) released under open licenses; no new data was collected from human subjects.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our study involves evaluation of LLM agent trajectories on existing public benchmarks, with annotation by research team members. This does not constitute human subjects research requiring IRB approval. No personal data was collected and annotators were co-authors/collaborators, not external participants.

☑ **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 (Claude, GPT-4) were used for code debugging, literature search assistance, and writing refinement. All scientific contributions, experimental design, data analysis, and core findings were conducted by the authors. AI-generated content was reviewed and verified by the research team.