Responsible NLP Checklist

Paper title: *FineState-Bench: Benchmarking State-Conditioned Grounding for Fine-grained GUI State Setting*
Authors: *Fengxian Ji, Jingpu Yang, Zirui Song, Yuanxi Wang, Zhexuan Cui, Yuke Li, Qian Jiang, Xiuying Chen*

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.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

☑ **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?
We performed manual verification to ensure data quality and filtered out noisy or inappropriate cases, as described in Appendix A (Step 5 of the Construction Pipeline).

☑ 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

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We benchmark off-the-shelf GUI agents without any hyperparameter search/tuning, keeping agent decoding/settings fixed, and we specify the evaluation prompts/protocol in Appendix B.

☒ 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 results are deterministic, dataset-level success rates computed over all 2,209 benchmark instances (not repeated runs), so we report single-run aggregate percentages in 5.25.3.

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.?
We did not conduct a human-participant study; human involvement was limited to dataset annotation/verification as described in Appendix A (Data Collection and Quality Control).

☒ 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)?
while Appendix A describes the human annotation/verification steps used to build FineState-Bench, the paper does not report annotator recruitment, payment, or demographic/adequacy details.

☒ 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)?
FineState-Bench is built from static GUI screenshots curated from OS-Atlas plus supplemental screenshots (with annotation/verification), and we do not collect or use personal subject data requiring individual consent.

☐ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **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, including ChatGPT and Claude, were used during the preparation of this manuscript. Specifically, they were employed to assist with language polishing, grammar correction, and generating template code for data visualization scripts. All AI-generated outputs were critically reviewed, edited, and verified by the authors to ensure accuracy and original contribution.