Responsible NLP Checklist

Paper title: *WindowsWorld: A Process-Centric Benchmark of Autonomous GUI Agents in Professional Cross-Application Environments*
Authors: *Jinchao Li, yunxin li, Chenrui Zhao, Zhenran Xu, Baotian Hu, Min Zhang*

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?
Ethical Considerations and Reproducibility (unnumbered section after Limitations): we describe the human review process and state that the instructions contain no personally identifiable information (PII) or harmful content.

☑ **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?
Ethical Considerations and Reproducibility (unnumbered section after Limitations): the instructions were generated through LLM-assisted pipelines and manually reviewed, and the paper states that they contain no personally identifiable information (PII) or harmful 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 3.4, Table 1, and Figure 4. We report key benchmark statistics, including the number of tasks (181), number of applications (17), the proportion of multi-app tasks (77.9%), average intermediate checkpoints per task (4.97), task difficulty distribution, and application-count distribution.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1, Appendix A.1A.2, and Table 2. We describe the experimental setup, including evaluated models/agents, input modalities, action spaces, and fixed step budgets. No hyperparameter search was performed, since the paper evaluates off-the-shelf models and agents under fixed settings.

☑ 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 4.24.3, Tables 24, and Table 7. We report descriptive statistics including average intermediate and final scores across task levels, latency, step-gap statistics, checkpoint-wise failure distributions, and confidence intervals for the VLM judge validation. The reported quantities are explicitly presented as averages or summary statistics in the paper.

☑ **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.?
No. The paper reports the involvement of human annotators, but it does not include the full text of instructions given to them. The appendix only provides prompts for LLM-based components (e.g., Generator, Refiner, and Environment Generator), not participant instructions.

☑ 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)?
Ethical Considerations and Reproducibility. The paper states that the data annotation and verification process involved four postgraduate researchers, and that annotators were compensated at a rate of 1.5 USD per task, which is described as fair market value for technical evaluation.

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)?
(left blank)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. The paper does not report ethics review board approval or an exemption determination.

☑ **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?
Sections 3.3 and 3.5, plus Appendix D. The paper reports that LLMs were used in the task-generation and refinement pipeline (e.g., DeepSeek-V3.2 for generation) and that Qwen3-VL-Plus was used as the automated judge. The appendix further includes prompts for the Generator, Refiner, and Environment Generator modules.