Responsible NLP Checklist

Paper title: *EVA: Evolving Semantic Adversaries for Red-Teaming GUI Agents Against Environmental Injection Attacks*
Authors: *Yijie Lu, Manman Zhao, Tianjie Ju, Zihe Yan, Xinbei Ma, Yuan Guo, Daizong Ding, Gongshen Liu, Zhuosheng 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?
See the Ethical Considerations section. We explicitly discuss the dual-use risks of automated red-teaming and detail our mitigation strategies. Specifically, all experiments were conducted within the EVA-GUI Benchmark using static, local replicas of web applications (isolated from live platforms) to ensure safety and prevent interaction with real-world users or production servers.

☑ **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?
(left blank)

☑ 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?
See Section 5.1 and Appendix F. We report that the EVA-GUI Benchmark consists of 252 tasks across four domains (Amazon, Gmail, Discord, YouTube), with shopping scenarios (T1-T63) adapted from WebArena and the remainder (T64-T252) synthesized for this work. We also specify the subset size (N=100) used for the visual insensitivity analysis in Appendix A.1.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 5.1 and 5.2. We detail the suite of victim agents, the construction of the EVA-GUI Benchmark, and the baseline configurations. We define the evolutionary search parameters in Algorithm 1 (Section 4.2) and specify the strict budget of 5 iterations in Appendix D. The specific system prompts used to drive the mutation process are provided in Appendix E.

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.

☑ 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?
See Section 5.1 and 5.2. We report the Attack Success Rates (ASR) as percentages averaged over the test set in Table 1. We also report descriptive statistics for the evolutionary process, specifically the mean Mutation Cost (efficiency) and Yield (success per seed) in Table 2. A granular breakdown of these statistics across all scenarios is provided in Appendix D (Table 4).

☒ **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.?
(left blank)

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)?
(left blank)

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)

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?
We selected 'No' regarding the inclusion of a specific section in the paper because the AI usage was strictly limited to code debugging and minor grammatical polishing. The AI tools did not contribute to any scientific ideas, experimental design, data analysis, or the generation of new text content. Therefore, we deemed a dedicated section within the paper unnecessary, but we disclose this usage here for full transparency.