Responsible NLP Checklist

Paper title: *BackdoorAgent: A Unified Framework for Backdoor Attacks on LLM-based Agents*
Authors: *Yunhao Feng, Yige Li, Yutao Wu, Yingshui Tan, Yanming Guo, Yifan Ding, Kun Zhai, Xingjun Ma, Yu-Gang Jiang*

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?
We are a safety assessment framework. Since it is designed to explore safety issues, we are discussing this.

☑ 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?
This work does not involve collecting or releasing datasets containing personally identifying information. We rely on existing public benchmarks and simulated agent environments, and do not include user-identifiable 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?
Yes. We report dataset/task statistics and evaluation settings (e.g., number of tasks/episodes and splits where applicable) in supple.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See in Section 5.

☑ 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?
We report aggregated metrics such as Attack Success Rate (ASR) and clean-task accuracy, and clearly specify whether results are averaged over tasks/episodes or obtained from a single fixed configuration (Section 5). We do not report variance across random seeds.

☒ **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.?
This study did not involve human participants or annotators.

☒ 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)?
We do not use.

☒ 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?
N/A. This work does not use data collected directly from human 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?
We used AI assistants (e.g., ChatGPT/Copilot) for limited support such as language editing/clarity improvements and/or minor coding assistance (e.g., boilerplate code, debugging). The AI tools were not used to generate the core scientific ideas, results, or experimental findings, and all content was reviewed and verified by the authors. We disclose this information here in the Responsible NLP checklist response.