Responsible NLP Checklist

Paper title: *Your LLM Agent Can Leak Your Data: Data Exfiltration via Backdoored Tool Use*
Authors: *Wuyang Zhang, Shichao Pei*

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?
Section: Ethical Considerations (7 equivalent, unnumbered)

☑ **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 section (7 equivalent, unnumbered). No real user data was collected or used. All user personas and session memory contents are fully synthetic, generated by GPT-5 for simulation purposes. Experiments run in isolated sandbox environments with synthetic data only.

☑ 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 (Experiments): 1,000 triggered queries, 5,000 Alpaca queries for FPR, 10,000-document corpus per domain, 3 random seeds. Appendix B: 50 trigger patterns per domain, 2,500 samples per pattern, 125,000 total triggered samples per domain. Appendix C: 5,000 classifier training examples, 500 validation examples.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1: Three domains, trigger configurations, three LLM architectures, seven rerankers, evaluation metrics Appendix B: Full fine-tuning configuration including optimizer, learning rate schedule, hardware, and random seeds Appendix C: Three-phase rewriter training (SFTDPOPPO) with all hyperparameters in Table 7

☑ 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 4.1 (three random seeds, results averaged), Tables 3-5 (mean std notation, e.g., 'std < 1.2 pp'), Figure 3 (error bars with propagated uncertainty), Appendix B (seeds 42, 123, 456 explicitly listed)

☒ **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?
Appendix B: "we use GPT-5 to generate 2,500 triggered training samples" (training data generation) Ethical Considerations: "we use GPT-5 as a simulator with fully synthetic personas" (user behavior simulation for multi-turn examples)