Responsible NLP Checklist

Paper title: *AgentMark: Utility-Preserving Behavioral Watermarking for Agents*
Authors: *Kaibo Huang, Jin Tan, Yukun Wei, Wanling Li, Zipei Zhang, Hui Tian, Zhongliang Yang, Linna Zhou*

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?
The work focuses on watermarking for provenance and IP protection of AI agents, which is itself a safety-enhancing technique. We do not identify significant potential risks beyond standard dual-use considerations.

☑ 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?
(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 (Datasets and Benchmark Details), as well as Appendix G.1. We report the benchmark settings and relevant statistics, including ALFWorld ID/OOD splits, six ToolBench subsets, OASIS platform settings.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Sections 4.3, 5.1, and Appendix G.1. These sections describe the experimental setup and key hyperparameters, including the compared methods, model/temperature settings, RG parameters, ToolBench step cap, RLNC payload length, and robustness-evaluation 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?
See Section 5.2, Figure 2, Figure 3, Table 1, Table 2, and Appendices G.1 and H. We report descriptive statistics throughout, including mean standard deviation over repeated runs, aggregate averages,

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.

Monte Carlo false-positive curves, and summary statistics such as Avg. KL, behavior match rate, and bit recovery rate. The paper also makes clear when results are averaged over three runs, over task subsets, or over repeated trials.

☒ **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?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)