Responsible NLP Checklist

Paper title: *Your LLM Agents are Temporally Blind: The Misalignment Between Tool Use Decisions and Human Time Perception*
Authors: *Yize Cheng, Arshia Soltani Moakhar, Chenrui Fan, Parsa Hosseini, Kazem Faghih, Zahra Sodagar, Wenxiao Wang, Soheil Feizi*

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?
This paper focused on analyzing LLM agent tool-call decisions in dynamic environments with different amount of elapsed time, which, to the best of our knowledge, does not pose any specific risks.

☑ **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?
Our constructed dataset consists of simulated conversation between a human user and LLM agent, and therefore can include human names when completing the task. We conducted human inspection to ensure only fictitious or celebrity names are included. (Sec 2.3) There is no harmful or offensive content in the 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?
Sec 3.3, Sec 3.5.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix D.2

☒ 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?
Our inference results were obtained using temperature=0, which means neglicting numerical nondeterminism in the hardware, the results are deterministic.

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.

☑ **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.?
Appendix A.5.

☑ 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)?
Appendix A.5.

☑ 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)?
As shown in the screenshot in Figure 17, Appendix A.5, our instructions explained the purpose of data collection.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The dataset contains no offensive or harmful content in any way. Recruited annotators were paid with clear rates and participation is voluntary.

☑ **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?
GPT models are used for synthetic data generation. Details are reported in Sec 3.2 and 3.3. AI assistants are otherwise only used in writing for correcting grammar issues and minor refinements.