Responsible NLP Checklist

Paper title: *Mem2ActBench: A Benchmark for Evaluating Long-Term Memory Utilization in Task-Oriented Autonomous Agents*
Authors: *Yiting Shen, Kun Li, Wei Zhou, Songlin Hu*

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?
Yes. Potential risks are discussed in the Limitations section (Section 6) of the paper.

☑ **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?
YES. Ethical Considerations section (Section 6).

☑ 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 statistics in the Methodology section (Section 3) of the paper, including the number of sessions (2,029), the number of generated tasks (400), and the average number of turns per session. Additional details are provided in the experimental setup (Section 4).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
YES. Section 4.1 (Experimental Setup).

☒ 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?
No. We report aggregate evaluation metrics (e.g., F1, BLEU, Tool Accuracy) in Section 4, but do not include variance estimates or error bars across multiple runs. This is because experiments are conducted with deterministic decoding settings (temperature = 0.0), and each configuration is evaluated with 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.

☑ **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.?
YES. Section 3.6 (Human Verification)

☒ 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)?
NO. The annotators are expert researchers with relevant academic backgrounds (e.g., NLP, AI), and were not recruited via crowdsourcing platforms. Payment details are not applicable as this annotation was conducted as part of the research process.

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)?
The dataset is constructed from publicly available sources (ToolACE, BFCL, OASST1), and does not involve direct data collection from human subjects.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This work does not involve intervention with human subjects or collection of sensitive personal data, and the annotation process was limited to expert evaluation of generated data.

☒ **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)