Responsible NLP Checklist

Paper title: *Grounding Agent Memory in Contextual Intent*
Authors: *Ruozhen Yang, Yucheng Jiang, Yueqi Jiang, Priyanka Kargupta, Yunyi Zhang, Jiawei Han*

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?
N/A - This work focuses on general-purpose memory mechanisms for agents (travel planning, debate) and does not involve high-risk applications or generate harmful content.

☑ 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?
Section 3.1, A.1 and A.2. We construct CAME-Bench using a closed-world environment with synthetic entities and LLM-simulated interactions, ensuring no real-world PII is included.

☑ 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 3.1, Table 1, and A.5. We report the number of trajectories (N) for Small/Medium/Large subsets and describe the generation pipeline.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, A.4, and B.2. We specify the backbone models, retrieval parameters, and token budgets.

☑ 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?
Table 1, Section 3.2, and Appendix D. We report answer-set Macro Precision, Recall, and F1 scores across multiple subsets on CAME-Bench and avgerage accuracy on LongMemEval and LoCoMo.

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.6. We provide details on the verification process and instructions given to human annotators for ground-truth 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)?
Appendix A.6

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)?
N/A - The data is synthetic; human subjects were only used for quality verification of the synthetic data, not as data sources themselves.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A - Study involves minimal risk verification of synthetic data.

☑ **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 utilized large language models for polishing text and checking LaTeX formatting. All scientific claims and experimental designs are the authors' own.