Responsible NLP Checklist

Paper title: *RealMem: Benchmarking LLMs in Real-World Memory-Driven Interaction*
Authors: *Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang, Ronghao Chen*

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 6 (Limitations). The benchmark relies on Gemini 2.5 models for data synthesis, which may raise reproducibility concerns. The evaluation scope currently excludes tool-use capabilities, limiting applicability to broader agentic tasks.

☑ 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?
All data in RealMem is synthetically generated via a multi-agent pipeline. User personas (e.g., names, age, demographics) are entirely fictional and do not correspond to any real individuals. No real user data was collected, and no offensive content was introduced during dataset construction. No anonymization steps were necessary.

☑ 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 Appendix A.1 (Table 7). The paper reports comprehensive dataset statistics, including total dialogue turns (14,028), average session number per user (205), average turns per session (6.8), total memories (5,072), total questions (1,415), and their distribution across four query types (Static Retrieval: 1,075; Dynamic Updating: 156; Proactive Alignment: 160; Temporal Reasoning: 24). No train/test/dev split is applied, as RealMem is designed as an evaluation-only benchmark.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 5.1. The paper describes the full experimental setup, including the evaluated memory systems (Mem0, A-mem, MemoryOS, Graph Memory), embedding model choices (e.g., text-embedding-3-small for Mem0), backbone LLMs (GPT-4o-mini and GPT-4o for generation, GPT-4o

as LLM-as-a-Judge), and two context construction settings (memory-only with Top-20 entries vs. session-based with Top-5 sessions). For A-mem, MemoryOS, and Graph Memory, the original recommended hyperparameter settings from their respective papers are followed.

☒ 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 paper reports single-run evaluation results across all methods on the full benchmark test set. Error bars and variance statistics are not reported, as the evaluation relies on deterministic LLM-as-a-Judge scoring (GPT-4o) applied consistently across all systems. However, cross-model judge validation using DeepSeek-V3 (Appendix A.2.2) confirms that rankings are stable across different evaluators.

☑ **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.?
See Appendix A.3. The paper provides the full annotation guidelines given to human evaluators, including four explicit evaluation dimensions (Memory Accuracy, Response Relevance, Information Completeness, and Linguistic Fluency), ranking instructions (14, where 1 = best), a tiebreaker rule (Memory Accuracy as primary tiebreaker), and a blinding procedure to minimize subjective bias. All system identities were strictly hidden from annotators prior to annotation.

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)?
See Appendix A.3. The two annotators are described as having NLP backgrounds, suggesting they are lab members or collaborators rather than paid crowdworkers. No crowdsourcing platform was used, and no payment information is reported in the paper. As internal expert annotators, standard crowdsourcing payment adequacy considerations do not apply.

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)?
All data in RealMem is synthetically generated via a multi-agent pipeline using Gemini 2.5 models. No real user data or third-party personal data was collected or curated. Therefore, data consent procedures are not applicable.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The dataset is entirely synthetically generated and does not involve collection of data from human subjects. The only human involvement is the small-scale internal annotation study (Appendix A.3) conducted by two NLP researchers, which does not constitute human subjects research requiring IRB/ethics board review under standard academic guidelines.

☑ **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?
The paper discloses AI assistant usage in multiple places. Section 5.1 states that GPT-4o-mini was used for memory extraction during dataset construction. Section 6 (Limitations) explicitly acknowledges that the data construction process relies significantly on the Gemini 2.5 series models for data generation, alongside human annotation for label verification. These disclosures cover both the data synthesis and evaluation components of the research.