Responsible NLP Checklist

Paper title: *HiGMem: A Hierarchical and LLM-Guided Memory System for Long-Term Conversational Agents*
Authors: *Shuqi Cao, Jingyi He, Fei Tan*

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?
Our work focuses on the technical mechanisms of memory retrieval. While long-term memory systems for agents have broader societal risks (e.g., privacy, misuse), our proposed method does not introduce new categories of risk beyond those inherent to existing LLM-based conversational agents. A full discussion of these broader risks is beyond the scope of this technical paper.

☑ 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?
The LoCoMo10 dataset used in our evaluation is synthetically generated by AI and does not contain any real user data or Personally Identifiable Information (PII). Therefore, discussions regarding data anonymization or the protection of human subjects are not applicable to this work.

☑ 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 4.1 describes the LoCoMo10 dataset used for our evaluation.

☑ C. Did you run computational experiments?

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The hyperparameters are detailed in Section 4.1. A comprehensive hyperparameter search was not conducted due to the significant computational costs associated with our experimental setup. The chosen parameters were determined through preliminary experiments and were selected to establish a strong performance baseline while maintaining a reasonable computational budget. This configuration effectively demonstrates the core contributions of our method.

☒ 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 reported results are averages over three runs with different random seeds. We provide this information here for transparency regarding how the main results are summarized.

☒ **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?**

☒ E1. If you used AI assistants, did you include information about their use?
AI assistants (e.g., ChatGPT, GitHub Copilot) were used in this research to support coding and text proofreading. For coding, AI was used to generate boilerplate code and assist in debugging. For writing, it was used for grammar and spelling checks. All core algorithms, research ideas, and textual content are original to the authors, who assume full responsibility for the accuracy of all content.