Responsible NLP Checklist

Paper title: *CloneMem: Benchmarking Long-Term Memory for AI Clones*
Authors: *Sen Hu, Zhiyu Zhang, YUXIANG WEI, Xueran Han, Zhenheng Tang, Ronghao Chen, Huacan Wang*

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 9 (Limitations), where we discuss the synthetic nature of the data and potential biases inherited from the LLMs used for generation

☑ **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?
All data in CLONEMEM is synthetically generated through hierarchical generation pipeline using LLMs. No real personal information, user data, or human-authorized content is collected or included. The synthetic personas do not correspond to any real individuals

☑ 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 (Data Statistics)

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 and Appendix D.2 describe all experimental configurations. We do not perform hyperparameter search; instead, we report results across multiple retrieval depths (k {5, 10, 20}) and compare different backbone/embedding combinations as ablations rather than selecting a single best configuration

☒ 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?
Results are reported from a single run per configuration due to the substantial computational cost of running evaluations with multiple backbone models (LLaMA-3.1-8B, GPT-4o-mini), retrieval

configurations (k {5, 10, 20}), and GPT-4o-based LLM-as-a-judge scoring. To reduce variance, each metric is averaged over 1,183 questions across 10 personas covering diverse context lengths and question types.

☑ **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.?
Human review is conducted internally over QA instances. This was not a formal annotation study with external participants, so no participant-facing instructions were distributed.

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?
We used AI assistants for coding assistance, debugging and light editing of paper text for clarity. All research ideas, experimental design, analysis and conclusions are the author's own. Additionally, as described in the paper, LLMs are used as core components of the data construction pipeline itself, which is the central methodological contribution of this work