

Responsible NLP Checklist

Paper title: *EverMemOS: A Self-Organizing Memory Operating System for Structured Long-Horizon Reasoning*

Authors: *Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, Yafeng Deng*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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 Ethical Considerations

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 Ethical Considerations

- 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 Experimental Setup

C. Did you run computational experiments?

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 4.1 Experimental Setup And Section 4.3 Ablation Study

- 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?

Section 4.1 Experimental Setup

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.?

Section A.2 LLM-as-Judge Reliability

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.

- 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)?

Section A.2 LLM-as-Judge Reliability

- 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)?

Section A.2 LLM-as-Judge Reliability

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

The human evaluation study was not submitted to a formal ethics review board, as our institution does not have an IRB for this type of low-risk research. However, the study posed minimal risk to participants: annotators were recruited via Prolific, a platform with established participant protections; they provided informed consent before participating; the task involved only judging the correctness of model answers against reference answers (no sensitive personal data was collected); and participants were compensated at ~\$12/hour, above fair-pay guidelines. These safeguards are described in the Ethical Considerations section and Appendix A.2.

- 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 were used only to assist with code development. All research ideas, experimental design, analyses, and scientific claims are the authors' own.