Responsible NLP Checklist

Paper title: *TIME: Temporally Intelligent Meta-reasoning Engine for Context-Triggered Explicit Reasoning*
Authors: *Susmit Das*

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 paper does not include a dedicated discussion of social or safety risks. It is trained on synthetic or non-sensitive English text and is not targeted at any particular domain, user group, or deployment setting. We explicitly state that the method is not a safety mechanism and do not evaluate bias, fairness, multilingual transfer, or high-stakes use.

☑ **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?
We do not include a dedicated discussion of personally identifying information or offensive content in the paper. All training and evaluation data introduced in this work are synthetic or author-created English dialogues generated from templates, model-assisted pipelines, or manually constructed.

☑ 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?
We report dataset sizes and splits for all training phases and for the evaluation benchmark. Section 4.1 gives counts for each curriculum phase, Section 4.2 provides details on TIMEBench, and Appendix C provides detailed phase-wise statistics, including train/test sizes and sequence length distributions.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 describes the training setup and core hyperparameters for Phases 13. Table 3 and the Deterministic Full-Batch Fine-Tuning subsection report the Phase 4 hyperparameters. Appendix C provides detailed per-phase training configurations.

☑ 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 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.

Section 5 (Results) reports mean TIMEBench scores aggregated over multiple trials per scenario. Section 4.2 (Evaluation Method: TIMEBench) specifies the full protocol. Statistical significance testing with the Wilcoxon signed-rank test is reported in Section 5 and Appendix F.

☒ **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.?
This study did not involve human participants or annotators. All evaluations were conducted using LLM-as-a-Judge, and no instructions were given to human subjects.

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)?
No human participants were recruited or compensated, as the study did not involve human annotation or user studies.

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)?
The study does not use data collected from human subjects. All training and evaluation data are synthetic or author-generated, so consent procedures are not applicable.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics review board approval was not required because the study did not involve human subjects, participant interaction, or human-generated personal data. The work was conducted independently by the author.

☑ **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 explicitly disclose the use of AI systems in this work. Synthetic training data for the early curriculum phases was generated using GPT-4o and Gemini 2.5 Flash, as described in Section 4.1. Evaluation was conducted using LLM-as-a-Judge with GPT-5.2 (2025-12-11 checkpoint), as described in Section 5. Beyond these disclosed roles, AI systems were not used to generate the main scientific claims, experimental results, or conclusions. AI assistance was also used lightly during manuscript preparation for tasks such as LaTeX formatting, wording refinement, and grammar checking. These uses were limited to presentation and clarity and did not generate the core research design, modeling decisions, analyses, results, or interpretations, which were produced independently by the author.