Responsible NLP Checklist

Paper title: *Tree of Agents: Improving Long-Context Capabilities of Large Language Models through Multi-Perspective Reasoning*
Authors: *Song Yu, Xiaofei Xu, KE DENG, Li Li, LIN TIAN*

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?
The work does not involve sensitive data, user interaction, or deployment in high-stakes scenarios, thus posing no identifiable ethical or safety risks.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
In Sections 4.1, 4.3, and 4.4, we cite the datasets, baselines, and base models used.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
All datasets and algorithms are open-source and publicly available.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The datasets and models were used in accordance with their intended purpose for academic benchmarking and evaluation of long-context reasoning capabilities (see Sections 4.1 and 4.3).

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 study uses public benchmark datasets (DetectiveQA, NovelQA, Needle-in-a-Haystack) that do not contain personally identifiable or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 4.1, we describe the datasets used, including their domain coverage (e.g., detective novels), average length, and purpose (long-context QA).

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.

☑ 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 reports average document lengths and number of QA pairs. Table 2 further provides accuracy and none-rate statistics over 100 sampled examples.

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.4 details the model sizes (e.g., LLaMA3.1-8B, DeepSeek-V3), Appendix B details the setup (2 NVIDIA RTX 3090 GPUs), and inference configuration (e.g., FP16, max token length, temperature).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 4.4 and Appendix B, we describe the experimental setup including temperature (0.01), max output length (2048), and chunk size (4096). No hyperparameter search was performed.

☑ 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?
In Table 2 and Table 3, we report mean accuracy and none rate over 3 runs on 100 samples per dataset to reflect statistical reliability.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
In Appendix B, we specify that all baseline methods (e.g., COA, LongRAG, LongLLMLingua) used default settings from their official implementations.

☒ **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.?
No human subjects or annotators were involved in this study.

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 participants were recruited or compensated, as no human subjects or annotators were involved.

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)?
We used publicly available datasets with no personal data, thus no consent was required.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No data collection involving human subjects was conducted; all datasets are publicly available, so IRB approval was not necessary.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No annotators were involved in the study.

☒ **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?
No, we did not use AI assistants in our paper.