Responsible NLP Checklist

Paper title: *Temporal Evidence Chain for Temporal Knowledge Graph Question Answering with Large Language Models*
Authors: *Shihao Liu, Xiaofei Zhou, Bo Wang, Geyuan Zhang*

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 algorithmic improvements for the Temporal Knowledge Graph Question Answering (TKGQA) task using standard public benchmarks (MultiTQ and CronQuestions). The proposed method operates on structured data and does not involve generating open-ended content that could pose safety risks or negative societal impacts.

☑ **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 utilize standard, publicly available academic benchmarks (MultiTQ and CronQuestions) that are derived from public knowledge bases (ICEWS and Wikidata). These datasets consist of facts regarding public figures and historical events, and do not contain private personally identifiable information (PII) of non-public individuals or offensive content.

☑ 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?
See Appendix A.1. We provide detailed statistics for the MultiTQ and CronQuestions datasets, including train/dev/test splits and question category distributions, in Table 7 and Table 8.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 5.1 and Section 5.4. We discuss the experimental setup in Section 5.1 and detail the hyperparameter search (grid search for N and K) and best-found values in Section 5.4 and Figure 4. Implementation details are in Appendix A.3.

☑ 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?
See Section 5.2 and Table 1. We report the Hits@1 metric on the test sets. As stated in Appendix A.3, the LLM inference is deterministic (temperature fixed at 0), so we report the single-run performance.

☒ **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.?
N/A. No human subjects were involved.

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)?
N/A. No human subjects 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)?
N/A. No human subjects were involved.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. No human subjects were involved.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
N/A