Responsible NLP Checklist

Paper title: *Tracking the Limits of Knowledge Propagation: How LLMs Fail at Multi-Step Reasoning with Conflicting Knowledge*
Authors: *Yiyang Feng, Zeming Chen, Haotian Wu, Jiawei Zhou, Antoine Bosselut*

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?
In the sections after conclusion and before reference, we discussed about limitations and ethical considerations.

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

☑ B1. Did you cite the creators of artifacts you used?
In Appendix E.1, we cited the creators of artifacts.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In Appendix E.1, we discussed the license or terms for use.

☑ 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)?
In Appendix E.1, we discussed our use of existing artifacts are consistent with their intended use.

☑ 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?
In Appendix B.1, B.2, B.3, we discussed that we studied multi-hop QA, math, and coding. Multi-hop QA involves Wikidata that may contain information of entities, but such information is public available. The coding and math questions do not contain information that identifies individual people. All data does not include offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Appendix E.1, we reported that all of our data collected is English questions.

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?
In Section 4 and Appendix B, we reported the number of collected examples in our final datasets.

☑ **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?
In Section 5.1 and Appendix C, we reported the number of parameters in all open-source models and the computing infrastructure.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 5.1, 6.1, and Appendix C, we provided detailed experimental setups, including hyperparameters.

☑ 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 3 and 4, we reported 95% of confidence intervals in our experiments.

☑ 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?
The main packages are in Appendix E.1. We do not use existing packages that require setting hyperparameters.

☑ **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.?
In Appendix E.2, we reported the full text of instructions given to participants.

☑ 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)?
In Appendix E.2, we reported that we recruited one PhD student for annotating the quality of our benchmark and compared it with LLM judges.

☑ 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)?
In Appendix E.2, we told the student these results were used to compare with predictions from LLM judges, but the student was not provided with the LLM predictions to avoid cheating.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
In Appendix E.2, we only collected annotation boolean values ("Yes" or "No") from annotators, which doesn't have any ethical issues. Thus, we don't require the approval from ethics review board.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
In Appendix E.2, we reported that the annotator is from the US.

☑ **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?
In Appendix E.2 and E.3, we report that we use AI assistants for data quality check, basic AI code completions, and grammar checking.