Responsible NLP Checklist

Paper title: *Identifying the Achilles' Heel: An Iterative Method for Uncovering Factual Errors in Large Language Models*
Authors: *Wenxuan Wang, Yuk-Kit Chan, Zixuan Ling, SHI Juluan, Youliang Yuan, Jen-tse Huang, Yifei Zhang, Wenxiang Jiao, Zhaopeng Tu, Michael R. Lyu*

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 is a diagnostic evaluation framework designed to uncover factual inaccuracies in LLMs, which is a defensive application aimed at improving model reliability rather than enabling harm. The generated test questions are derived from publicly available factual triplets in Wikidata and contain no harmful or sensitive content, so we did not identify potential risks warranting a dedicated discussion beyond the Limitations section and Appendix A.

☑ **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?
Our data is automatically generated from Wikidata, a publicly available knowledge base, covering neutral academic domains such as history, philosophy, psychology, mathematics, and physics (Appendix F). While some triplets reference public figures whose information is already widely documented in public sources, the data contains no private personal information or offensive content, so no additional anonymization was necessary.

☑ 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 3; Appendix I

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

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

☑ 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,

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.

etc. or just a single run?
Appendix K

☑ **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.?
Appendix L

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)?
(left blank)

☑ 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)?
Appendix M

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☒ **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?
(left blank)