Responsible NLP Checklist

Paper title: *Illusions of Confidence? Diagnosing LLM Truthfulness via Neighborhood Consistency*
Authors: *Haoming Xu, Ningyuan Zhao, Yunzhi Yao, Weihong Xu, Hongru WANG, Xinle Deng, Shumin Deng, Jeff Z. Pan, Huajun Chen, Ningyu 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?
Section "Ethical Statement"

☑ **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?
The dataset was constructed using factual, time-invariant knowledge sourced from established standard QA benchmarks like SimpleQA, HotpotQA, and SciQ (covering STEM, Arts & Culture, Social Sciences, and Sports). Because the data solely pertains to general public facts, checking for PII or offensive content was unnecessary, and thus not discussed.

☑ 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 3.1

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1

☑ 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?
Section 4.2

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.

☑ **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 C.3

☒ 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)?
While we state that "three human experts independently evaluate each candidate," we did not provide details regarding the recruitment platform or the compensation provided to these experts.

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 human participants acted as expert annotators verifying the factual correctness of public knowledge queries (QA pairs). Their personal data was not being curated, rendering data consent protocols typically required for human subjects research inapplicable.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The involvement of humans was strictly limited to the factual annotation and verification of publicly available knowledge datasets (QA benchmarks). Such tasks generally do not constitute human subjects research requiring formal ethics review board approval.

☑ **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?
Appendix A