Responsible NLP Checklist

Paper title: *ContextCheck: Sentence-Level Faithfulness Verification with Context-Aware Disambiguation*
Authors: *Yueqin Yin, Yaxi Li, Xin Liu, Xun Wang, Kaiqiang Song, Simin Ma, Shujian Liu, Sathish Reddy Indurthi, Haoyun Deng, Pengcheng He, Mingyuan Zhou, Song Wang*

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?
We discuss potential risks in the Impact Statement section (Section 6). In particular, incorrect verifier judgments may create a false sense of reliability, especially in high-stakes domains.

☑ **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 datasets used in this work are derived from publicly available sources or generated by language models. We do not collect or use any personally identifying information.

☑ 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?
We report detailed dataset statistics, including the number of samples and dataset composition, in Section 4.1 (Training Data) and Appendix Table 4.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We describe the experimental setup, including model architecture, training procedure, and evaluation settings, in Section 4.1 (Experimental Setup) and Appendix C.

☒ 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?
We report standard evaluation metrics such as Macro F1 and Balanced Accuracy, and include statistical significance analysis in Section D.5. However, we do not report variance estimates (e.g., standard deviation) or results over multiple runs.

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 K.6

☒ 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)?
The human validation study (Appendix C) involved three annotators: two PhD students and one industry engineer who are co-authors or collaborators of this work. No external crowdworkers were recruited and no separate payment was made.

☒ 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)?
All datasets used in this work are publicly available research datasets.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our study does not involve human subjects research requiring IRB 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?
Section J (Appendix). We used LLMs for language polishing and writing refinement.