Responsible NLP Checklist

Paper title: *Fine-Grained Detection of Context-Grounded Hallucinations Using LLMs*
Authors: *Yehonatan Peisakhovsky, Zorik Gekhman, Yosi Mass, Liat Ein-Dor, Roi Reichart*

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?
This research is focused on creating a benchmark for the meta-evaluation of AI models and a methodology for evaluating them. It is designed to help mitigate the existing risk of AI "hallucinations" by studying how LLMs can be used for localizing them. The outputs are tools for researchers, such as the FINAL benchmark and an LLM-based evaluation protocol, and do not present a direct risk of misuse. Our paper is transparent about the method's current limitations.

☑ **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?
Privately revealed to ACL ARR 2026 January Program Chairs, ACL ARR 2026 January Submission637 Area Chairs, ACL ARR 2026 January Submission637 Authors, ACL ARR 2026 January Submission637 Reviewers, ACL ARR 2026 January Submission637 Senior Area Chairs We build on a publicly available dataset and only improve it's annotation

☑ 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 3.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ 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 5.

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 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)?
The authors of the papers were the annotators.

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)?
(left blank)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We build on a publicly available dataset and only improve it's annotation.

☑ **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 sections 3 and 4 (dataset augmentation and then evaluation).