Responsible NLP Checklist

Paper title: *Challenging the Explanation Based on Preceding Tokens: Discovering Transferable Non-Literal Biasing*
Authors: *Yuchen Huang, Junpeng Zhang, Quanshi 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.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

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

N/A 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?
(left blank)

☑ 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?
Relevant data statistics and examples are reported in the paper. We present dataset statistics in Appendix B, along with a subset of test examples for reference.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup, including the models used and the evaluation procedure, is described in Section 2 (Algorithm). As this work is an analytical study on fixed, pre-trained models without training or fine-tuning, no hyperparameter search or tuning is involved.

☑ 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?
Descriptive statistics of the experimental results, including means, standard deviations, and proportions of samples exhibiting specific effects, are reported in Sections 2.1 and 2.2. The paper clearly indicates that results are aggregated over multiple samples.

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?**

☒ N/A 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.?
(left blank)

☒ 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)

☒ 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)

☒ 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?**

☑ E1. If you used AI assistants, did you include information about their use?
AI assistants were used in this work in two limited and explicitly documented ways. First, an LLM was used solely for language polishing to improve clarity and fluency of the manuscript; all technical content, claims, and conclusions were determined by the authors. Second, LLMs were used to generate a dataset following a strictly defined and reproducible generation protocol. The data generation procedure, prompts, constraints, and filtering steps are fully described in Appendix A and Appendix C. No AI system was used to autonomously design experiments, derive conclusions, or make scientific claims.