Responsible NLP Checklist

Paper title: *Reinforcement LearningGuided Adaptive Tuning for Out-of-Distribution Harmful Text Detection*
Authors: *Mengyu Xiang, Tinghao Chen, Boxu Han, Qiudan Li, Shu Wu, Daniel Dajun Zeng*

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?
Sec. Ethical Considerations. This paper proposes an adaptive test-time tuning method for detecting harmful text. It aims to mitigate the negative impact of harmful content, contributing to a more harmonious online social environment, without any inherent risks.

☑ 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?
Sec. Ethical Considerations. All experiments were conducted on publicly available datasets without disclosing any personal information, and the use of the datasets was consistent with the scientific research intent of the original paper.

☑ 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?
Sec. 5.1, Sec. 5.3, Appendix A.2, and Appendix A.3

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sec. 5.3, Sec. 5.7 and Appendix A.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?
Sec. 5.3

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.?
This study did not directly recruit or contact any human annotators or human subjects. All data used were derived from publicly available datasets, annotated in the original work, and anonymized, containing no identifiable personal information.

☒ 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)?
This study did not involve human annotators or human subjects, therefore there were no issues of participant recruitment or compensation.

☑ 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)?
Sec. 5.1. All data used in this study are from publicly available datasets. Informed consent was obtained or the platform's terms of use were followed during the original dataset release process for the collection and use of this data. This study did not directly collect or process any new user data.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Sec. Ethical Considerations.

☒ **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?
The research process and paper writing in this study were completed independently by the authors, without the use of any form of AI-assisted tools in the research, coding, or text writing.