Responsible NLP Checklist

Paper title: *MASH: Evading Black-Box AI-Generated Text Detectors via Style Humanization*
Authors: *Yongtong Gu, Songze Li, Xia Hu*

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?
Yes. In the "Ethics Statement" section, we discussed the potential risks of this technology being misused for academic dishonesty or disinformation dissemination. We also clarified our research motivation as a red-teaming tool to expose the fragility of existing detection paradigms.

☑ 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?
No. In our experiments, we solely utilized publicly available and widely adopted academic benchmark datasets (e.g., MGTBench, MGT-Academic, dmitva, cc_news). These open-source datasets were processed by their original authors prior to release, and our research does not involve any new data collection processes concerning human privacy.

☑ 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?
Yes. In Appendix B (Experimental Details), we detailed the dataset partition ratio (3:1:1:1) , the sample size of the test sets (100 instances per domain) , and the size of the training data we constructed (approximately 6,000 pairs).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. We reported detailed information regarding the model architecture, optimizer (AdamW) , learning rates (210 5 and 510 6) , batch sizes , number of epochs , and beam search parameters in Section 4.1 (Implementation Details) and Appendix B (Implementation and Training Details).

☑ 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,

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.

etc. or just a single run?
Yes. We reported the mean values of various evaluation metrics in the tables within Section 4 (Experiments) , and presented bar charts with error bars indicating standard deviation in Figure 7.

☒ **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.?
N/A. This study does not involve human subjects.

☒ 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)?
N/A. This study does not involve human subjects.

☒ 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)?
N/A. This study does not involve human subjects.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. This study does not involve human subjects.

☑ **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?
Yes. We used AI programming assistants (e.g., Copilot/Cursor) during the coding and debugging phases, and utilized Large Language Models (e.g., ChatGPT/Gemini) for grammar checking and language refinement of the manuscript.