Responsible NLP Checklist

Paper title: *SimMark: A Robust Sentence-Level Similarity-Based Watermarking Algorithm for Large Language Models*
Authors: *Amirhossein Dabiriaghdam, Lele 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?
"Potential Risks" in "Ethical Considerations" section.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1 cites any datasets/models we used.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
"Use of Models and Datasets" in "Ethical Considerations" section.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
"Use of Models and Datasets" in "Ethical Considerations" section.

☒ 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, because of the nature of our work (no fine-tuning, etc.). Also, all datasets used in this work are publicly available and have been widely used in prior research and to the best of our knowledge, do not contain personally identifiable information or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, we have provided the link to their individual Hugging Face model or dataset cards which includes documentation of the artifacts.

☑ 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?
Section 4.1

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.

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and appendices B & D

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and appendices B & D (moreover, there are some ablation studies on hyperparameters available in appendices E, F, G, I, & J.)

☑ 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 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Section 4 and appendices B & D

☑ **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.?
Figure 10 in Appendix M provides the full instructions given to participants, including evaluation criteria, expected time, and privacy assurances.

☒ 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)?
Participants were recruited informally from within our network. No payment or compensation was provided, as participation was voluntary.

☑ 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)?
Consent was obtained via the form in Appendix M (Figure 10), which outlined the task, risks (none), and data usage (research-only). Participants agreed before starting.

☐ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This was a small-scale, low-risk pilot study involving only graduate student volunteers. It did not require formal ethic board review, as no personal or sensitive information was collected, and no risk of harm to participants was identified.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix M

☑ **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?
Footnote in "Ethical Considerations" section.