Responsible NLP Checklist

Paper title: *DSCD: Large Language Model Detoxification with Self-Constrained Decoding*
Authors: *Ming Dong, Jinkui Zhang, Bolong Zheng, Xinhui Tu, Po Hu, Tingting He*

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?
Section 7

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

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section 4

☑ 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)?
Section 4

☑ 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?
Section 4

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The artifacts (code and models) will be released publicly with documentation in the future. Currently, no formal documentation is included in the 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?
We used publicly available datasets, but the paper does not report statistics such as the number of examples or train/dev/test splits. These can be found in the original dataset publications.

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

☑ 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?
The results reported in the paper are averaged over multiple experimental runs. While the paper does not explicitly state average, the reported values reflect the mean performance, providing a stable and representative measure of the method's effectiveness.

☑ 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

☑ **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.?
We used human annotators for a small-scale manual evaluation (spot checks), but the paper does not report full instructions or disclaimers provided to participants.

☒ 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 human evaluation was conducted by lab colleagues and supervising faculty as spot checks. No formal recruitment or payment was involved.

☒ 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)?
All human participants, including lab colleagues and supervising faculty, provided their consent to participate in the spot-check evaluation.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No new data collection was conducted; only publicly available datasets were used. Therefore, ethics review board approval was not applicable.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The data used are publicly available datasets; no new annotators were recruited, so reporting demographic or geographic characteristics of annotators is not applicable.

☑ **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?
While this work studies AI methods, the AI assistants were used only for polishing the writing. No experimental design, data analysis, or research content was generated by AI. All scientific work and results were conducted and verified by the authors.