Responsible NLP Checklist

Paper title: *Data to Defense: The Role of Curation in Aligning Large Language Models Against Safety Compromise*
Authors: *Xiaoqun Liu, Jiacheng Liang, Luoxi Tang, Muchao Ye, Weicheng Ma, Zhaohan Xi*

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?
In appendix D.1, we have include a section named "Potential Risk" to address this discussion.

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

☑ B1. Did you cite the creators of artifacts you used?
In Section 5.1 "Experimental Settings", we have cited the used public datasets, baselines, and attack methods from other papers.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In appendix D.2, we have include a section named "Use of Artifacts," wherein we have highlight "Model Licenses" and "Data and Other Licenses" to discuss the licenses for artifacts we use (model, data, and attack codes).

☑ 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)?
In appendix D.2, we have include a section named "Use of Artifacts," wherein we have highlighted "Artifact Use Consistent With Intended Use" to clearly state that our use of artifacts is consistent to its original intention.

☑ 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?
In appendix D.2, we have include a section named "Use of Artifacts," wherein we have highlighted "Personally Identifiable Information" and "Offensive Content" to clearly state that we have no personally identifiable info and potentially offensive contents are masked to avoid harmful results.

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 5.1 "Experimental Setting" we have elaborated on detailed documentation of used data, baseline, and attack methods.

☑ 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?
(1) In Section 5.1, "Experimental Setting – Dataset and Statistics" we details the test data and its number. (2) In Appendix Table 7, we further detail the training data and its statistics.

☑ **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?
In Appendix B "Experimental Configurations" and Table 7, we have included details of those information.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
(1) In Appendix B "Experimental Configurations" and Table 7, we have included details of those info. (2) In Section 5.1 "Experimental Setting," we further highlight important setup.

☑ 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?
(1) In appendix D.3, we have include a section named "Descriptive Statistics" to address the transparency of reported values. (2) In 5.1 "Experimental Setting," we clearly defined what's the evaluation metrics we use.

☑ 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?
(1) In appendix D.4, we have include a section named "Use of Packages" to introduce used Python packages. (2) In Table 7, we detail the parameter settings.

☒ **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.?
This work is not related to 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)?
This work is not related to 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)?
This work is not related to human subjects.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This work is not related to human subjects.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
This work is not related to 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?
In appendix D.5, we have include a section named "Use of AI" to address this part.