Responsible NLP Checklist

Paper title: *Let Them Down Easy! Contextual Effects of LLM Guardrails on User Perceptions and Preferences*
Authors: *Mingqian Zheng, Wenjia hu, Patrick Zhao, Motahhare Eslami, Jena D. Hwang, Faeze Brahman, Carolyn Rose, Maarten Sap*

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 9 Ethical Considerations

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

☑ B1. Did you cite the creators of artifacts you used?
We cited the dataset CoCoNot in Sections 1 and 3. We added citations of all of the models and tools we used in Sections 3.1, Section 5, Section 6 and Appendix B. We cited annotation tool Potato in Appendix A and annotation platform Prolific in Section 4.

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

☑ 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)?
Appendix B

☑ 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?
We discussed how we verify malicious motivations and generated model responses used in user study in Section 9 Ethical Considerations. The user data doesnt contain any personally identifying information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We discussed the domain coverages and detailed instances of the dataset we created in Section 3 and Appendix A. We also discussed the dataset we used in Section 5.

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.

☑ 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 mentioned the number of instances of the dataset we created in Section 3.

☑ **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?
We listed the model sizes in the Experimental Setup of Sections 5 and 6. We listed the budget details in Appendix B.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We listed the experimental setup and hyperparameters in Appendix B.

☑ 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?
We listed descriptive statistics of user study data in Section 4, and LLM patterns in Section 5 and 6.

N/A 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?
(left blank)

☑ **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 listed the full instructions in Appendix A.3, C.2, and C.3 for annotation and user study, respectively.

☑ 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)?
We reported the recruitment and payment details in Section 4.2 and Appendix C.5.

☑ 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)?
We listed the consent form in Section 4.2, Appendix A.3 and C.2.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our user study protocol was approved by the Institutional Review Board (IRB) of our organization. We mentioned this in Appendix A.3.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We reported the participant information of user study in Appendix A.3 and C.5.

☒ **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?
(left blank)