Responsible NLP Checklist

Paper title: *Understanding and Mitigating Overrefusal in LLMs from an Unveiling Perspective of Safety Decision Boundary*
Authors: *Licheng Pan, Yongqi Tong, Xin Zhang, Xiaolu Zhang, JUN ZHOU, Zhixuan Chu*

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?
Title. Since our work involves constructing an over-refusal dataset, some contents may contain racism, sexuality, or other undesirable content.

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

☑ B1. Did you cite the creators of artifacts you used?
Abstract. We construct a multilingual over-refusal benchmark and a framework for its generation, both of which are publicly released on GitHub.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We use the MIT License in our GitHub repository, intended solely for academic use.

☒ 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)?
We open-sourced our multilingual over-refusal dataset to enable fair and objective evaluation of LLMs refusal capabilities, contributing to both the academic community and the open-source ecosystem.

☒ 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?
Our benchmark do not contain any personal informations.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.

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?
Section 4.1.

☒ **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 4.3 and Appendix C.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C.

☑ 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 4.4.

☒ 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?
We open-sourced our benchmark and generation pipeline.

☒ **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.?
Our study did not involve any risks or sensitive content for participants/annotators, and no disclaimers were necessary. The instructions provided were simple and are included in the paper.

☒ 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 did not recruit or pay any participants for this study.

☒ 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 have only used open-sourced data.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We have only used open-sourced data.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We did not recruit annotators.

☑ **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?
We use ai assistants to generate overrefusal prompts.