Responsible NLP Checklist

Paper title: *Beyond the Safety Tax: Mitigating Unsafe Text-to-Image Generation via External Safety Rectification*
Authors: *Xiangtao Meng, Yingkai Dong, Ning Yu, Li Wang, Zheng Li, Shanqing Guo*

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?
Ethical considerations

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

☑ 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?
Ethical considerations

☑ 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?
3

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☑ 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?
5

☑ **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.?
The manual inspection was conducted by the authors to verify the correctness of the automated labels. Since no external human participants were recruited and the task was limited to objective verification of image content, formal instructional documents and disclaimers were not separately drafted.

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. ACL 2026 used a subset of ARR checklist form.

☒ 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 manual inspection was conducted by the authors to verify the correctness of the automated labels. Since no external human participants were recruited and the task was limited to objective verification of image content, formal instructional documents and disclaimers were not separately drafted.

☒ 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 research primarily utilizes publicly available datasets and automated evaluation frameworks.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This research primarily utilizes publicly available datasets and automated evaluation frameworks. The minimal human involvement was restricted to post-hoc data cleaning and quality assurance of the model's performance, which is exempt from ethics review board oversight according to our institutional guidelines.

☑ **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?
AI assistants were only employed for basic grammar correction and language polishing to improve readability. The core research ideas, experimental design, data analysis, and manuscript drafting were entirely performed by the authors without generating any original academic content via AI.