Responsible NLP Checklist

Paper title: *Reverse Constitutional AI: A Framework for Controllable Toxic Data Generation via Probability-Clamped RLAIF*
Authors: *Yuan Fang, Yiming Luo, Aimin Zhou, Fei Tan*

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?
Yes. Ethics Statement, Limitations and Appendix E. We discuss the potential risks of automated toxic data generation and emphasize that R-CAI is designed as a defensive research tool for red teaming rather than facilitating malicious use (see Introduction). We also analyze the latent capability extraction risk in Ethics Statement, Limitations and Appendix E.

☑ 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?
Yes. Section 4.1 and Appendix C. The synthesized data intentionally contains offensive content for red-teaming research purposes. We used a curated red-teaming dataset of 30,000 prompts as a starting point. No personally identifying information (PII) was collected or generated; any names used in prompts (e.g., in Case Studies) are hypothetical or publicly known figures used to test safety boundaries in a controlled environment.

☑ 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?
Yes. Section 4.1 and Appendix C. We reported the total number of prompts (30,000) and the specific breakdown across four toxic dimensions in Section 4.1 (Experimental Setup).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4. 1 (Experimental Setup); Appendix C (Implementation Details)

☑ 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,

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.

etc. or just a single run?
Yes. We report the mean scores across multiple evaluation runs. For semantic coherence and toxicity scores, we provide comparative results across different alpha values to show stability (see Section 4 and Table 1)

☑ **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.?
Yes. Appendix D. The full text of the instructions and guidelines provided to the expert annotators, including definitions of toxicity levels and safety disclaimers, is provided in Appendix D

☑ 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)?
Yes. Appendix D. We recruited three graduate students specializing in NLP and AI Safety as expert annotators. They were compensated with an hourly rate consistent with local university research assistant standards.

☑ 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)?
Yes. All expert annotators were fully briefed on the nature of the study, including potential exposure to toxic content. Formal consent was obtained from all participants prior to the annotation process.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix

☑ **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?
Appendix E. AI assistants were used for language polishing and formatting guidance. All technical content, experimental design, and results were produced and verified by the authors.