Responsible NLP Checklist

Paper title: *RubricBench: Aligning Model-Generated Rubrics with Human Standards*
Authors: *Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, Chen Ma*

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?
This work focuses on evaluation methodologies and benchmarking. We do not introduce new generative models that could be misused. The data, including safety-related prompts, is curated from existing public benchmarks and is used solely to assess alignment and refusal capabilities, not to generate harmful content.

☑ 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?
The data used in this work is curated from existing publicly available benchmarks (HelpSteer, PPE, RewardBench2) as described in Section 3.2. We did not collect new personally identifying information (PII) from users. The human annotations were provided by a professional team of experts and PhD candidates (described in Appendix A.4), whose individual identities are not disclosed. Regarding offensive content, the "Safety" domain intentionally includes adversarial prompts to evaluate model safety and refusals, which serves the research purpose of alignment benchmarking.

☑ 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?
See Section 3.2, Figure 2, and Appendix A.1.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 4.1 and Appendix A.2.

☑ 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 4.2 and Table 2.

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.

☑ **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 annotation protocol and criteria are described in detail in Section 3.4 and Section 3.5.

☑ 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)?
Appendix A.4 details the recruitment and background of the annotators. External participants were compensated at a rate of $20 USD/hour, which is comparable to the local average hourly wage.

☒ 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)?
The study utilizes data curated from existing public benchmarks (Section 3.2). For the new rubric annotations, the annotators were expert researchers/practitioners (Appendix A.4).

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **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?
(left blank)