Responsible NLP Checklist

Paper title: *Self-Guided Alignment: Adaptive Preference Sensing for Multi-Objective Generation*
Authors: *Ning Wang, Zhanyang Liu, Taotao Zhou, Xinrui Zhang, Zongru Shao, Haojie Zhou*

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.

N/A A2. Did you discuss any potential risks of your work?
This work proposes a training and inference framework for improving alignment in large language models.

☑ **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?
No. We use publicly available, widely used benchmark datasets released by their original authors. These datasets are intended for research use and do not contain personally identifying information beyond what may already exist in public web text. We do not collect new data or perform user-level data aggregation, and no additional personal information is introduced in our work.

☑ 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. See Appendix B.2.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. See 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?
Yes. See Section 4.2, Section 4.3 and Appendix C.2. All reported results are obtained from a single run; we do not report mean or max over multiple runs.

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?**

N/A 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.?
N/A. We do not collect new data involving human participants. All datasets used in this work are publicly available research benchmarks released by prior work, and issues of data consent were handled by the original data creators.

N/A 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)?
N/A. This work does not involve the recruitment of human participants or crowdsourced workers. We do not conduct human studies or collect new human annotations; all experiments are based on existing publicly available datasets.

☒ 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)?
No. We use publicly available, widely used benchmark datasets released by their original authors.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. We use publicly available, widely used benchmark datasets released by their original authors.

☑ **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?
Yes. See Section 4.1 and Appendix C.1. We use ChatGPT as an automatic evaluator to score model outputs for performance comparison.