Responsible NLP Checklist

Paper title: *Gradient-Guided Multi-Judge Prompt Optimization*
Authors: *ChenZhuo Zhao, Xinda Wang, Pu Zhao, Yue Huang, Junting Lu, Ziqian Liu, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang*

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?
Appendix A (Ethical Considerations). We discuss dual-use risks (e.g., misuse to increase jailbreak success and elicit unsafe behavior) and describe mitigation steps including a minimum-necessary-disclosure approach and controlled access/release decisions.

☑ 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?
Appendix A (Ethical Considerations) and Appendix C.1.5 (AdvBench). We note that the safety benchmark contains harmful-behavior instructions and take safety-oriented handling steps, including not disclosing jailbreak prompts and following a minimum necessary disclosure / controlled access release policy to reduce misuse risk

☑ 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 (Implementation Details) and Appendix C.1 (Tasks and Data Details). We report how data is used in optimization/evaluation (e.g., randomly sampling 20% for training capped at 50 instances and evaluating on the remaining examples), and we provide dataset-specific details such as selecting 50 instances for detailed evaluation (and also reporting on the full dataset) and using 50 preference pairs for AlpacaEval prompt optimization.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Privately revealed to ACL ARR 2026 January Program Chairs, ACL ARR 2026 January Submission272 Area Chairs, ACL ARR 2026 January Submission272 Authors, ACL ARR 2026 January Submission272 Reviewers, ACL ARR 2026 January Submission272 Senior Area Chairs Section 4.1

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.

(Implementation Details) and Appendix D.1 (Hyperparameter selection; Table 7). We describe the experimental setup (models used as generator/judges, training split strategy, iterations, and compute) and provide the fixed hyperparameters used across datasets (e.g., iterations, k, window size, max sequence length, and pairwise objective settings).

☑ 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?
Privately revealed to ACL ARR 2026 January Program Chairs, ACL ARR 2026 January Submission272 Area Chairs, ACL ARR 2026 January Submission272 Authors, ACL ARR 2026 January Submission272 Reviewers, ACL ARR 2026 January Submission272 Senior Area Chairs Section 4.2 (Experimental Results and Analysis; Table 2) reports descriptive statistics for open-ended evaluation, including win rate, length-controlled win rate (LC win), and average response length, and explains what each metric represents.

☒ **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.?
(left blank)

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)?
(left blank)

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)?
(left blank)

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?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)