Responsible NLP Checklist

Paper title: *ARES: Adaptive Red-Teaming and End-to-End Repair of Policy-Reward System*
Authors: *Jiacheng Liang, Yao Ma, Tharindu Kumarage, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Aram Galstyan, Charith Peris*

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?
Section "Ethics Considerations" (page 10) discusses potential risks, including the possibility that adversarial prompts or red-teaming data could be misused if deployed outside controlled research settings. The authors note that all models are intended solely for advancing responsible AI alignment research and encourage responsible use by the community to avoid misuse or unintended deployment in unsafe contexts.

☑ 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?
Section "Ethics Considerations" (page 10) states that all adversarial prompt generation and data collection were performed in a contained research environment, with automatic filtering applied to remove explicit, illegal, or personally identifiable content prior to training.

☑ 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 (Evaluation) and Appendix E report relevant dataset statistics. The discovery phase generates a dataset of 4,000 samples (Section 4.3). Baseline comparisons use 3,000 adversarial prompts per method (Section 4.3.1). All evaluation benchmarks are listed with their purpose in Table 9 (Appendix D). Training splits and data composition (HelpSteer2, FalseReject, PKU-SafeRLHF with ~10.8k pairs) are described in Sections 3.2.1 and 4.2.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 ("Experimental Setup") describes model choices (Qwen3-1.7B as Core LLM, SkyworkRM-Qwen3-4B as RM, Qwen3-8B-abliterated as Safety Mentor) and hardware (8 NVIDIA A100 GPUs

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.

on EC2). Appendix E (Tables 1012) provides full hyperparameter tables for (1) adaptive red-teaming discovery, (2) RM fine-tuning, and (3) Core LLM policy optimization (Dr. GRPO), including learning rates, batch sizes, sequence lengths, epochs, and optimizer 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?
Results in Tables 27 report single-run performance metrics across all baselines and ablation conditions. Section 4.5 (Validation of LLM-as-a-Judge) reports inter-annotator agreement with standard deviation (96.0% 1.5%) over n=100 samples. Runtime statistics (e.g., 13 hours total, 9 hours for discovery, 4 hours for repair) are also reported in Section 4.3. The paper notes these are results from controlled single experimental runs; the ablation studies (Tables 45) provide additional systematic comparison.

☑ **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.?
Section 4.5 states that three LLM security researchers served as independent annotators and labeled responses using the same policy rubric as the automated judge. The full judge prompt (including the rubric with 05 scoring criteria) is provided in Appendix E.3 (pages 1617), which constitutes the instructions given to annotators.

☒ 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 three annotators in Section 4.5 are colleagues from the same research team with expertise in LLM security. As internal collaborators rather than paid external participants, no formal recruitment process or monetary compensation was involved. Therefore, reporting on payment adequacy is not applicable to this evaluation setup.

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?**

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