

## Responsible NLP Checklist

Paper title: *AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following*

Authors: *Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, Shengjie Bi, Shishir G Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Kumar Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, Amine Benhalloum, Hany Hassan Awadalla, Manaal Faruqi*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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?

*We observed the issue of potential reward hacking and discussed approaches to mitigate this issue in Section 4.4 and 5.5*

### 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?

*Our collected prompts are generated by human annotators and irrelevant to Personally Identifying Info Or Offensive Content.*

- 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 3*

### C. Did you run computational experiments?

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Section 5.1, the hyperparameters are not mentioned as we use the in-house RLHF infrastructure.*

- 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 5.2, we report the single run result for our large-scale experiment since it is compute expensive.*

*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.?

*We cooperate with the third-party vendor and the information is covered by confidentiality agreements*

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)?

*We cooperate with the third-party vendor and the information is covered by confidentiality agreements*

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)?

*We cooperate with the third-party vendor and the information is covered by confidentiality agreements*

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

*We cooperate with the third-party vendor and the information is covered by confidentiality agreements*

**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)*