Responsible NLP Checklist

Paper title: *What Do LLMs Learn First? Asymmetric Learning Dynamics of Input Complexity and Output Ambiguity in Preference Alignment*
Authors: *Mengyang Li, Jingwen Wang, Pinlong Zhao*

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 I (Appendix I) discusses potential misuse risks of efficient alignment methods and bias from evaluator 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?
The UltraFeedback dataset is a publicly available benchmark. We use existing data without additional collection; its privacy and content policies are documented by the original authors.

☑ 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 5.1 reports dataset statistics including the number of prompts (approximately 60,000) and the discretization scheme.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 and Appendix A detail the experimental setup, hyperparameter values (learning rate, batch size, DPO beta, eta_base, delta), and model configurations.

☑ 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?
All main results (Tables 2-6) report mean and standard deviation over 3 random seeds.

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 human annotation study was conducted informally to validate IC metric correlation; full instruction protocols were not formally documented in the paper. We will add details in the camera-ready.

☒ 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 annotation study involved lab members as annotators rather than paid crowdworkers; no external recruitment or payment was involved.

☐ 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?
See Appendix I (AI Assistance Declaration): AI assistants were used for grammatical correction and text polishing; all content was reviewed by the authors.