Responsible NLP Checklist

Paper title: *When and What to Ask: AskBench and Rubric-Guided RLVR for LLM Clarification*
Authors: *Jiale Zhao, Ke Fang, Lu Cheng*

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?
We did not include a dedicated discussion of potential risks. The paper focuses on evaluating and training clarification behavior in LLMs, and mainly discusses methodological limitations such as the offline benchmark construction, LLM-based user simulation, dependence on a single judge model, and the lack of a comprehensive bias audit.

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

N/A 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?
(left blank)

N/A 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?
Sections 3.2 and 4.1, as well as Appendix D. We report the source datasets, sampling procedure, and the sizes of the constructed evaluation sets (400 AskMind + 400 AskOverconfidence = 800 total multi-turn instances), and also describe the training data sources and quality audit statistics.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 5.1 and 5.2, and Appendix F. We describe the models and evaluation protocol, and report key training hyperparameters including actor learning rate (1e-6), batch size (64), max prompt/response lengths (2048/8192), rollouts per prompt (n=8), total training steps/epochs (500/10), and hardware (8 H200 GPUs).

☒ 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?
We report point estimates for the benchmark metrics in the result tables, but do not provide error

bars, confidence intervals, or multi-seed summary statistics, nor do we explicitly characterize the reported numbers as mean/max over repeated runs.

☒ **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?**

☒ E1. If you used AI assistants, did you include information about their use?
We used AI assistants during research and/or writing, but did not include a dedicated disclosure of their use in the manuscript.