Responsible NLP Checklist

Paper title: *Beyond Self-Report: Bridging the Intention-Behavior Gap in Critical Thinking Assessment via Interpretable Multi-Agent System*
Authors: *Zekun Li, Jifan Yu, Haoxuan Li, Ye He, Daniel Zhang-Li, Shangqing Tu, Joy Jia Yin Lim, Yikun Jiang, Jiaxin Yuan, Yu 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?
See Section [8 Ethics Statement]. We discuss potential biases in AI-based assessment and data privacy concerns regarding student responses.

☑ **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?
See Section [4 & 8] (Data Collection/Ethics Statement). All student identifiers were removed, and we manually audited a subset of the data to ensure no offensive content was present.

☑ 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 and Section 4.2. We detailed the number of participants ($N = 1,161$ for simulation and $N = 70$ for human study), the structure of the scenario-based assessment, and the distribution of samples across different critical thinking dispositions.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 and Appendix (if applicable). We specified the base models used (GPT-4o), the temperature setting (e.g., $T = 0$), and the specific configurations for the multi-agent system (Orchestrator, Inquirer, Curator, and Arbiter). The prompt templates and the AsCoT (Assessment Chain-of-Thought) reasoning steps are also detailed.

☑ 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,

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.

etc. or just a single run?
Section 5.1 and Section 5.2. We reported Pearson correlation coefficients (r), p-values, and mean scores with standard deviations to compare MASA against CCTDI and human expert ratings. We also provided mean cost analysis ($0.41 per participant) and ablation study results to show the stability of our multi-agent framework.

☑ **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 4.2 and Appendix (e.g., Appendix B). We provided the participants with clear instructions on the assessment process, the interaction rules with the multi-agent system, and the specific scenarios they would encounter. Screenshots of the interface and the disclaimer regarding data privacy were also included in the Appendix.

☑ 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)?
Section 4.2. Participants were recruited from a top-tier university via campus mailing lists and social media. We disclosed the compensation details (e.g., a fixed hourly rate or a per-session reward), ensuring the payment was above the local minimum wage and competitive for the students' demographic in China.

☑ 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)?
Section 4.2 and Ethics Statement. Before the experiment, all participants were informed about the purpose of the study and how their dialogue data would be used for research. We obtained explicit consent from each participant, ensuring that all data was anonymized and used strictly for academic analysis.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study was conducted in accordance with the universitys ethical guidelines for human subjects research. Since the data collected was anonymized and the assessment process posed no psychological or physical risk to participants, it was determined to be exempt from a full formal IRB review.

☑ **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?
Section 3 and Acknowledgements. In our research, AI assistants (specifically GPT-4o) were used as the core engine for the multi-agent framework (MASA) to conduct assessments and generate AsCoT reasoning paths. Additionally, AI was utilized for polishing the manuscript's language and assisting in code debugging for the simulation experiments. All final outputs were reviewed and verified by the authors to ensure accuracy and academic integrity.