Responsible NLP Checklist

Paper title: *Can We Predict Before Executing Machine Learning Agents?*
Authors: *Jingsheng Zheng, Jintian Zhang, Yujie Luo, Yuren Mao, Yunjun Gao, Lun Du, Huajun Chen, Ningyu 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?
The paper focuses on accelerating machine learning agents and does not involve direct societal or ethical risks.

☑ **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?
All datasets are sourced from public Kaggle competitions and contain no personally identifying information.

☑ 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 and Appendix B. Table 1 and Table 6 provide detailed statistics for the datasets and the constructed preference corpus.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1, Section 6.3, and Appendix C.8 and C.9 detail the experimental setup, models, and computational 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 4.2 and Table 2 report mean accuracy and standard deviation across multiple independent runs.

☑ **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 experts were the authors themselves performing pass/fail validity checks on execution logs, so explicit participant instructions were not applicable.

☒ 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 experts were internal researchers validating the data, 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)?
Does not apply as no external human subject data was collected.

☐ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Does not apply as the human involvement was strictly limited to code execution validation by the authors.

☑ **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 3.4 explicitly details the use of GPT-5.1 for code generation and verbalization of data analysis reports.