Responsible NLP Checklist

Paper title: *ACE-Router: Generalizing History-Aware Routing from MCP Tools to the Agent Web*
Authors: *Zhiyuan Yao, Zishan Xu, Yifu Guo, Zhiguang Han, Cheng Yang, Shuo Zhang, Weinan Zhang, Xingshan Zeng, Weiwen Liu*

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.

N/A A2. Did you discuss any potential risks of your work?
Our training data is fully synthetic, generated by LLMs, and the evaluation relies on public benchmarks. We do not use real-world user data or personally identifiable information (PII). The only human involvement is a small-scale manual validation of synthesized labels, which involves minimal risk and no sensitive data. Thus, this work does not pose material risks regarding data privacy or ethical compliance.

☑ **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?
The training data used in this work is entirely synthetic, generated by Large Language Models (LLMs). As it contains no real-world user data or personally identifiable information (PII), steps for anonymization or PII protection are not applicable.

☑ 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?
See Section 4.1.1 and Section 4.1.2.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See section 4.1.3.

☑ 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?
See section 4.1.3.

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.?
Yes. The full text of the annotator instructions is provided in Appendix B. In this evaluation, human annotators reviewed 100 randomly sampled synthesized trajectories and labeled the correct tool choice for each example. We then compared these annotations with the synthesized labels and computed Cohens Kappa to assess label reliability.

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)?
See Appendix B. The paper reports a limited manual validation in which human annotators reviewed 100 randomly sampled synthesized trajectories and labeled the correct tool selections. This annotation was conducted solely to verify the reliability of the synthesized labels, rather than as a crowdsourced or paid participant study. Therefore, the recruitment and payment considerations targeted by this question do not apply.

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)?
The data used in this work consist entirely of synthetic data generated by LLMs and publicly available open-source benchmarks.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The only human involvement in this study was a small-scale manual validation, in which annotators reviewed synthetic trajectories and labeled the correct tool choices. We did not collect personal or sensitive data from annotators, and the annotation task involved minimal risk. Therefore, Ethics Review Board approval was not required.

☑ **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 (e.g., ChatGPT) for grammatical error correction and polishing of the text.