Responsible NLP Checklist

Paper title: *SGVEF-LOOP: Coverage-Guided Progressive Topological Exploration and Fact-Grounded Metamorphic Evaluation for MCP Agents*
Authors: *Zenghao Liu, Yansong 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?
This work focuses on the evaluation framework and benchmark construction for MCP Agents. It does not involve human subjects, personal privacy data, harmful content generation, nor does it pose any direct ethical or societal risks.

☑ 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?
This research exclusively utilizes public MCP tools, synthetic metamorphic test cases, and standard evaluation datasets. It does not involve any personally identifying information (PII) or offensive content.

☑ 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?
We report comprehensive statistics of the constructed benchmark and experimental data in Section 5.2, Section 5.3, Appendix A, and Appendix C. The reported statistics include: (1) the scale of the tool library (100 tools over 14 MCP servers and 6 domains); (2) the number of metamorphic test cases (576 valid cases filtered from 600 initial samples); (3) exploration metrics (60 iterative rounds, 1,080 valid tool transitions, 80.54% transition coverage); (4) statistical validation data of the Chao1 estimator across 31 batches; (5) empirical frequency statistics of 5 categories of tool execution failures.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.2

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.

☑ 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 5.4

☒ **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 solely for English grammar proofreading and text polishing.