Responsible NLP Checklist

Paper title: *ETOM: A Five-Level Benchmark for Evaluating Tool Orchestration within the MCP Ecosystem*
Authors: *Jia-Kai Dong, I-Wei Huang, Chun-Tin Wu, YI-TIEN TSAI*

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?
Section 8. We discussed the steps taken to prevent offensive content, protect privacy, and acknowledged potential digital inequalities regarding access to external tools and services.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 5.1 and References. We cited the creators of the Model-Context Protocol, the glama.ai registry, and the various foundation models used in our experiments, including Qwen, GPT-4.1, Llama, and Gemma.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section 8. We stated that all associated licenses for the models and data used permit user access for research purposes and that we have followed all applicable terms of use.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 8. Our use of existing models and the MCP registry is consistent with their intended research and development purposes, and we have adhered to the access conditions provided by the original creators.

☑ 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?
Section 8. We implemented strict measures in our generation prompts to avoid offensive content and conducted manual inspections of a random sample of 200 data entries across all five levels to ensure no personal privacy violations or offensive material.

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 3 and 4, and Appendix B. We provided detailed documentation regarding the MCP tool corpus, task generation logic, data formats, and representative examples for each complexity level.

☑ 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?
Table 2 and Section 3.1. We reported the number of unique servers (491), distinct tools (2375), and total tasks (2075), along with the average plan length and server count for each level of the curriculum.

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sections 5.1, 7, and 8. We reported foundation model sizes (e.g., 4B, 8B, 12B) in the model names. We reported a total computational budget of approximately 500 USD in Section 7 for data generation and evaluation.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 and Appendices C and G. We detailed the four orchestrator architectures, specific configurations for ReAct, ToolShed, and MCP-Zero, and conducted an ablation study on key hyperparameters like retrieval breadth (k) and query expansion.

☑ 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?
Table 3 and Section 5.2. We reported summary statistics including Exact Match (EM) and F1-score for all five levels across different model-architecture combinations.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Section 4 and Table 5. We reported the use of ROUGE-L and METEOR metrics for comparative analysis of retrieval complexity. Implementation details for orchestrators are provided in Appendix C.

☑ **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.?
Sections 3.4, 8, and Appendix I. We described the manual inspection process for quality control and the specific criteria used for human verification of data quality and task validity. Detailed prompt templates used to guide the LLM-assisted annotation are provided in Appendix I.

☑ 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 8. Participants were graduate students and research assistants who were compensated at standard academic rates.

☑ 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 8. We stated that all annotators agreed to participate as their contribution to the research.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our data collection involved internal researchers (students and RAs) performing quality control on non-sensitive, LLM-generated technical tool queries. The process was determined safe and low-risk, so formal IRB approval was not sought.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 8. The annotators were graduate students and research assistants with expertise in AI systems and tool orchestration.

☑ **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 8. We stated that ChatGPT and Gemini were used to assist in proofreading the manuscript and providing code documentation during the writing process.