

Responsible NLP Checklist

Paper title: *OctoTools: A Multi-Agent Framework with Extensible Tools for Complex Reasoning*

Authors: *Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, James Zou*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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 (Broader Impacts and Ethical Considerations).

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?

(left blank)

- 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 3.1 and Appendix B.1 detail the dataset statistics, including the sampling process for validation (100 examples) and test sets (200 examples) across all 16 benchmarks.

C. Did you run computational experiments?

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

Section 3.1, Section 3.3, and Appendix B detail the experimental setup, baseline configurations, prompt templates, and reasoning/time budgets (e.g., 10 steps, 300 seconds).

- 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 3.1 states that results are reported as the average accuracy with standard deviation across three trials. Table 1 explicitly reports these mean and standard deviation () values.

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.?

(left blank)

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.

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?
Section 8 (Broader Impacts and Ethical Considerations, final paragraph) explicitly details the use of LLMs strictly for writing assistance (phrasing, clarity, and grammar), noting they were not involved in generating core scientific arguments.