Responsible NLP Checklist

Paper title: *MemTR: Enhancing Tool-Calling Reliability via Uncertainty-Triggered FFN-Space Retracing*
Authors: *hongtao duan, Lu Jiang, Minying Zhang, Xiaobing zhu, Tianpeng Bu, Hao Jiang, Xinyu Wei, lulu hu*

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 9: Ethics Statement

☑ **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?
We used a small calibration set to select the entropy threshold and mixing weight. This calibration set was constructed from publicly available tool-calling prompts/tools (or synthetic prompts) and does not contain personally identifying information to the best of our knowledge. We did not collect user data, and therefore did not apply additional anonymization beyond what is provided in the original sources.

☑ 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 4: Experiments and Appendix D.0.1: Calibrated uncertainty-triggered MemTR decoding

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2, Appendix D, Section 5 (Figure 6)

☒ 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?
We report point estimates from single deterministic evaluations on fixed benchmarks (no repeated runs with multiple random seeds), so we do not provide error bars or confidence intervals. All reported numbers are single-run metrics computed over the full test sets.

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. Annotation guidelines and labeling criteria for the calibration set (e.g., format-validity, correct tool name, and required argument keys/values) are described in Appendix D.0.1.

☒ 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)?
Not applicable: annotations were performed by the authors; no recruitment or payment.

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)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No IRB/ethics board review was sought because we did not collect data from human subjects, we only used publicly available (or synthetic) benchmark data and created an internally annotated calibration set that contains no personally identifying information.

☑ **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 an AI assistant (e.g., ChatGPT) for minor English proofreading and formatting suggestions, and for routine coding convenience (e.g., boilerplate/debugging). It was not used to design the method, create/label evaluation data, choose experimental settings, or generate/select reported results. All experiments and analyses were conducted and verified by the authors.