Responsible NLP Checklist

Paper title: *CRITICTOOL: Evaluating Self-Critique Capabilities of Large Language Models in Tool-Calling Error Scenarios*
Authors: *Shiting Huang, Zhen Fang, Zehui Chen, Siyu Yuan, Junjie Ye, Yu Zeng, Lin Chen, Qi Mao, Feng Zhao*

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?
The LLM benchmark proposed in this paper has been designed with potential risks in mind. The dataset and evaluation methods do not contain any potential for malicious or unintended harmful effects and uses, and there are no privacy concerns involved. Furthermore, we do not use a lot of resources to train the model, thus minimizing environmental impact.

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

☑ B1. Did you cite the creators of artifacts you used?
In Sec.2, Sec.A, Sec.B and Reference of the paper, we list in detail the authors of all artifacts used and provide corresponding citations.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The license information of the artifacts we used is publicly available in their respective repositories, and we follow common academic citation practices.

☑ 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)?
In Sec. 2, we discuss whether the use of existing artifacts is consistent with their intended purpose and explain the intended use of the artifacts we create.

☒ 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?
Our dataset has been processed to ensure that it does not contain any personally identifiable information or offensive content.

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.?
Sec. A provides documentation of the artifacts, including coverage of domains, number of data, etc.

☑ 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?
Sec.2 and Sec. B reports statistics about the data we use and create.

☑ **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?
Privately revealed to ACL ARR 2025 May Program Chairs, ACL ARR 2025 May Submission1383 Area Chairs, ACL ARR 2025 May Submission1383 Authors, ACL ARR 2025 May Submission1383 Reviewers, ACL ARR 2025 May Submission1383 Senior Area Chairs In Sec.3 we provide the number of parameters in the model, but other information is not the focus of our study.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Our study does not involve hyperparameter search.

☒ 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?
The variability of the results is not a focus of the study.

☒ 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?
The use of software packages is standardized and does not require special instructions.

☒ **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)

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(left blank)

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)