Responsible NLP Checklist

Paper title: *BioTool: A Comprehensive Tool-Calling Dataset for Enhancing Biomedical Capabilities of Large Language Models*
Authors: *Xin Gao, Ruiyi Zhang, Meixi Du, Peijia Qin, Pengtao Xie*

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?
No. The paper discusses the technical limitations of the current framework in the "Limitations" section, such as its restriction to one-hop tool calling and the lack of an independent agent architecture. However, it does not include a dedicated discussion on broader potential societal risks or dual-use concerns associated with releasing this dataset.

☑ **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?
No. The dataset is constructed by querying authoritative public biological databases (NCBI, Ensembl, UniProt) for generic biomedical entities such as genes, proteins, and taxonomy. It relies exclusively on open scientific data and does not contain human subjects' personally identifying information 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?
Yes. Section 3.2 and Figure 3 detail the distribution of the dataset across data sources, tool types, and biological domains. The paper explicitly reports the total number of query-API pairs (7,040) and the test set size (1,408 samples). Appendix B also discusses data subsets ranging from 10% to 100%.

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No. While Section 4.1 outlines the experimental setup, baseline models, and evaluation metrics, the paper does not explicitly report the detailed hyperparameter search space or the best-found hyperparameter values (e.g., learning rate, batch size) used during the fine-tuning of the open-source LLMs.

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?
Yes. Section 4.1 explicitly states that the experimental results reported in the paper represent the average performance across three independent runs.

☑ **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.?
Section 4.3 and Appendix D describe the general criteria used by the human annotators (e.g., judging informativeness, task fulfillment, and scientific correctness)

☒ 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)?
No. The paper mentions that evaluations were conducted by "two annotators with college-level bioinformatics backgrounds", but it does not detail the recruitment process, platform used, or specific compensation provided to these annotators.

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. The paper does not mention ethics review board (IRB) approval, as the human involvement was strictly limited to evaluating the scientific accuracy of model outputs, rather than conducting formal human subjects research.

☑ **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?
I use AI Assistants to proofread the manuscript for grammatical corrections and phrasing improvements.