Responsible NLP Checklist

Paper title: *LLM4Cell: Taxonomy and Evaluation of LLM and Agentic Models for Single-Cell Biology*
Authors: *Sajib Acharjee Dip, Adrika Zafor, Bikash Kumar Paul, Uddip Acharjee Shuvo, Muhit Islam Emon, Xuan Wang, Liqing Zhang*

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?
Yes. Potential risks are discussed in the main paper and appendix, including data bias, privacy and consent concerns in single-cell datasets, overinterpretation of language-grounded outputs, and the reliability of agentic reasoning systems. These issues are addressed in the evaluation rubric and discussion of ethical, fairness, and privacy dimensions (Section X.X), as well as in the Limitations and Open Problems sections.

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

N/A 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?
No. As a survey, this work does not introduce new datasets, perform model training, or define train/test/dev splits. Dataset statistics such as sample counts and resolutions are reported at a high level in the supplementary registry where relevant, but detailed experimental splits are outside the scope of this study.

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

N/A C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
(left blank)

N/A 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?
(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.

☒ **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)

☑ **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 AI-assisted tools (e.g., ChatGPT) for language refinement, formatting suggestions, and drafting non-technical text such as summaries, responses, and figure captions. All scientific content, analysis, and conclusions were developed and verified by the authors.