

Responsible NLP Checklist

Paper title: *UrbanGeoEval: A City-Scale Benchmark for Evaluating Large Language Models in Geospatial Reasoning*

Authors: *Mutian Bao, Qiuyi Qi, Tian Liang, Jinjian Zhang, Wei Zhou, Ming Kong, Linjian Mo, Qiang Zhu*

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?

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?

Ethical Considerations

- 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?

3

C. Did you run computational experiments?

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

5

- 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?

5

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.?

We recruited annotators through a third-party professional data annotation platform, which operates under its own labor and confidentiality agreements; neither those agreements nor the platform's

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.

internal instruction package can be redistributed by us. A summary of the task design, real-world scenario sourcing, visual verification, dual-layer validation, annotator demographics and compensation is provided in Section 3.2.3 and Appendices A.2, A.3, and the Ethical Considerations section.

- 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)?

Ethical Considerations

- 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)?

Ethical Considerations

- 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?

During the preparation of this manuscript, we utilized AI assistants (e.g., large language models) for the following purposes: 1) to assist with literature searches; and 2) to improve the grammar, phrasing, and overall clarity of our writing. All AI-generated text and suggestions were critically reviewed, edited, and verified by the authors. The authors retain full responsibility for the final content of the manuscript, including its accuracy, originality, and any potential errors.