Responsible NLP Checklist

Paper title: *EVGeoQA: Benchmarking LLMs on Dynamic, Multi-Objective Geo-Spatial Exploration*
Authors: *Jianfei Wu, Zhichun Wang, Zhensheng Wang, Zhiyu He*

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?
All data used in this study are publicly available and do not involve any ethical concerns, personal information, or protected data. Therefore, the research poses no risk to individuals or groups and does not require ethical approval.

☑ **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?
Our data do not contain any personally identifiable 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?
Section 3 and Table 1.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and 5 detail the GeoRover framework, tool definitions, and prompting techniques (Few-Shot, CoT).

☑ 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 the primary metrics (Hits@K) as averages over the test set in Table 2. Additionally, we report the average tool invocation frequencies in Table 3 and the distribution of error causes in Figure 4.

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.?
The manual verification and error analysis were conducted exclusively by the authors.

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)?
The manual verification and error analysis were conducted exclusively by the authors.

☒ 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)?
All data used in this study were obtained from publicly accessible sources under their standard terms of use and the manual verification and error analysis were conducted exclusively by the authors. Therefore, no additional consent was required.

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 explicitly state in Section 3.4 that we employed Large Language Models (specifically Qwen2.5-72B) for data synthesis and query paraphrasing during the dataset construction process. we used AI assistants to check for grammatical errors and polish the writing.