Responsible NLP Checklist

Paper title: *RSMeM: Knowledge-Enhanced Memory Evolution for Remote Sensing Agents with Systematic Evaluation*
Authors: *Bingxian Wu, Yu Zhang, Zonghao Guo, Tang Liu, Chen Qian, Yuxiang Lu, Xingbo Du, Yanghao Li, Yidan Zhang, Chi Chen, Ling Yao, Chenghu Zhou, Maosong Sun*

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?
Our work focuses on improving tool-use accuracy for geoscientific analysis tasks. The system operates within a controlled benchmark environment and does not involve generation of harmful content, personal data, or dual-use capabilities.

☑ 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?
EarthBench consists of remote sensing imagery and geoscientific analysis tasks. It does not contain 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?
Section 4.1 reports the dataset size (248 instances), modality distribution (RGB, Spectrum, Earth Products), and domain coverage (3 domains, 12 subdomains).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.2 reports the scoring function hyperparameters (=0.2, =0.08, =0.1, =0.2). Section 4.1 describes the experimental setup including backbone models, tool configuration, and evaluation protocol.

☑ 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?
Appendix (DeepSeek-V3.2 Results with Variance) reports mean std over n=3 independent runs with error bars for all five metrics.

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?**

☑ 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.?
Appendix A.2 describes the expert-driven three-tier curation process, including expert selection criteria and the systematic top-down decomposition protocol.

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 knowledge base was curated by co-authors with advanced degrees in Remote Sensing and Geospatial Information Science, not by external crowdworkers.

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)?
The knowledge base was curated by co-authors as part of this research. EarthBench is a publicly available benchmark.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects were involved beyond co-author domain experts contributing to knowledge base construction. No ethics review was required.

☑ **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?
AI writing assistants were used for English polishing and LaTeX formatting of the manuscript. All scientific content, experimental design, and analysis were conducted by the authors.