PEACE: Empowering Geologic Map Holistic Understanding with MLLMs

Supplementary Material

A. Geologic Map

Geologic map is a specialized type of map that depicts the distribution, characteristics, and chronological relationships of rock units as well as the occurrence of structural features such as faults and folds. These maps are essential tools for geologists and earth scientists as they provide a visual representation of the geological characteristics of a specific area. Typically, as shown in Figure A1, a geologic map comprises several key elements, including the title, scale, legend, main map, index map, cross section, stratigraphic column, and other components. These elements collectively contribute to the coherence and utility of a geologic map. Specifically, please refer to the following content.

Title indicates the physical region, map type, author, and other pertinent information.

Scale demonstrates the relationship between distances on the map and physical distances on the ground.

Legend explains the symbols and colors used to represent different rock types, ages, and geological features. For detailed information on the legend units, refer to the legend component in Figure A1.

Main Map depicts the geological characteristics of the mapped area, including distributions of rock types, ages, folds, and faults.

Index Map illustrates the spatial relationship with the neighboring regions.

Cross Section provides a vertical slice through the Earth, showing the arrangement of rock units below the surface.

Stratigraphic Column displays the sequence, thickness, and types of rock layers present in a particular area.

Other Components Besides the above 7 key components frequently found in geologic maps, there are additional supplementary components that provide further geological explanation for the region.

B. Evaluation Metrics

The metrics are designed to measure the quality of answers generated by AI-based methods for each question in GeoMap-Bench.

B.1. Overall Score

S_{all} is the overall score of an AI-based method on GeoMap-Bench, where M denotes the number of abilities to be measured in it, including extracting, grounding, referring, reasoning, and analyzing.

$$S_{all} = \frac{1}{M} \sum_{i=1}^{M} S_i(T, Q, A, L) \tag{A1}$$

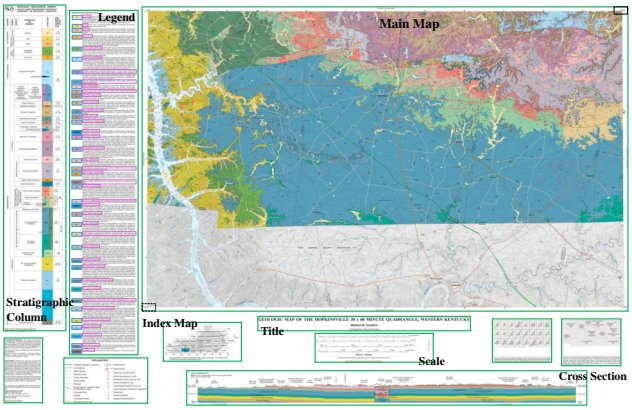

Figure A1. Example of a geologic map and its components. All components and legend units are enclosed within bounding boxes for interpretive understanding.

B.2. Ability Score

S_i is the ability score of an AI-based method measured for i-th ability in GeoMap-Bench, where N represents the number of questions pertaining to that ability. T, Q, A, and L indicate the sets of question types, questions, AI-responded answers, and expert-labeled answers respectively. The j-th instance of these sets are denoted as t_j, q_j, a_j, and l_j.

$$S_i(T, Q, A, L) = \frac{1}{N} \sum_{j=1}^{N} S_{i,t_j}(q_j, a_j, l_j) \tag{A2}$$

B.3. Type Score

S_{i,t_j} is the type score for the j-th question of type t_j within the i-th ability. This score can correspond to one of the following types: S_{mcq} for multiple-choice questions, S_{fitb} for fill-in-the-blank questions, and S_{eq} for essay questions.

Multiple-choice Question. S_{mcq} is the type score of a multiple-choice question, where q, a, l are a element of sets Q, A, L respectively.

$$S_{mcq}(q, a, l) = \begin{cases} 1.0, & a = l \\ 0.0, & \text{otherwise} \end{cases} \tag{A3}$$

Fill-in-the-blank Question. S_{fitb} is the type score of a fill-in-the-blank question.

$$S_{fitb}(q,a,l) = \begin{cases} IoU_{det}(a,l), & \text{all grounding tasks} \\ IoU_{set}(a,l), & \text{set extracting tasks} \\ S_{mcq}(q,a,l), & \text{otherwise} \end{cases} \quad \text{(A4)}$$

where all grounding tasks encompass tasks of both grounding by name and grounding by intention, and set extracting tasks include tasks of index map extracting and longitude-latitude extracting.

IoU_{det} is the intersection over union metric to evaluate the accuracy of a predicted bounding box against the ground-truth bounding box.

$$IoU_{det}(b_1, b_2) = \frac{I(b_1, b_2)}{U(b_1, b_2)} \quad \text{(A5)}$$

where b_1 and b_2 are two bounding boxes. I and U are functions to calculate the intersection area and union area of two bounding boxes respectively.

IoU_{set} is the intersection over union metric to evaluate the overlap of two sets.

$$IoU_{set}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{(A6)}$$

where A and B are two sets, which could be either discrete, such as neighboring regions, or continuous, like longitude and latitude range.

Essay Question. S_{eq} is the type score of an essay question. To avoid any order-related bias of answer-judging agent, it is measured twice per question by keeping and switching the order of two answers.

$$S_{eq}(q,a,l) = \frac{1}{2}(1 - J(q,l,a) + J(q,a,l)) \quad \text{(A7)}$$

J is a answer-judging agent powered by GPT-4o [3], with its prompt detailed in Section C.2. For the given essay question, it is designed to determine which of the two answers is better based on principles of diversity, specificity, and professionalism.

$$J(q,a_1,a_2) = \begin{cases} 1.0, & a_1 \text{ is better than } a_2 \\ 0.0, & a_1 \text{ is worse than } a_2 \\ 0.5, & a_1 \text{ and } a_2 \text{ are comparable} \end{cases} \quad \text{(A8)}$$

where a_1 and a_2 are two input answers of judging agent.

C. Evaluation Prompt

There are two types of prompts used in the evaluation process, the question answering (QA) prompt and the answer judging (AJ) prompt of essay question. We introduce them in the following subsections, where variables are represented in the format ${var_name}.

C.1. Question Answering Prompt

- QA Prompt

Image prompt:
${selected sub-images in geologic map}

Instruction prompt:
Extracted information: ${information}
Injected knowledge: ${knowledge}
This is a ${question type} question.
Based on the provided text and image, reason and answer the question in JSON format only, for example: {"reason": "XXX", "answer": "XXX"}

Question:
${question}
Answer:

C.2. Answer Judging Prompt

- AJ Prompt

Image prompt:
${entire image of geologic map}

Instruction prompt:
Please evaluate which of the two answers below is better for the essay question ${question}, consider the following criteria:
1. Diversity: The answer should address various aspects of the question, providing a well-rounded perspective.
2. Specificity: The answer should be detailed and precise, avoiding vague or general statements.
3. Professionalism: The answer should be articulated in a professional manner, demonstrating expertise and credibility.

Answer1:
${answer1}
Answer2:
${answer2}

Question: which answer is better?
A. Answer1 is better than Answer2
B. Answer1 is worse than Answer2
C. Answer1 and Answer2 are comparable

Only respond answer with A, B or C in JSON format, for example: {"answer": "C"}
Answer:

D. Evaluation Setting

Base Model. We set all the random seeds to 42, the temperature to 0, and the maximum tokens to 2048 for base models. Among them, we enable structured mode to enforce responses in JSON format for GPT-4o and GPT-4o-

mini. This functionality is not applied to other open-source MLLMs as they do not support it. The system prompt is set to "You are an expert in geology and cartography with a focus on geologic map.".

Detection Model. We use YOLOV10 [7] as the detection framework to train the map component detector and legend unit detector models. The training settings are as follows: input images are resized to 640×640 for both detectors, SGD is employed as optimizer with initial learning rate of 0.01 and finnally linear decay to 0.0001. The weight decay is set to 0.0005, and the total number of epochs is 500. The models are trained on single GPU (80GB NVIDIA Ampere A100), where the batch size is 32. We select and annotate approximately 1k original geologic maps as training dataset, ensuring no overlap with the GeoMap-Bench dataset. During the inference stage, the Intersection over Union (IoU) threshold for Non-Maximum Suppression (NMS) is set to 0.8.

GEE APIs. In Google Earth Engine (GEE) [2], we use API of "WorldPop/GP/100m/pop" [1, 4] image collection to retrieve population density data and API of "ESA/WorldCover/v200" [8] image collection for land cover data. The scale for both collections is set to 100.

Scientific DBs. We use USGS earthquake database [6] to retrieve records of historical earthquake data with magnitudes greater than 2.5 occurring since the 1970s. For active faults database, we use GEM DB [5], which currently encompasses most of the deforming continental regions on Earth, with the exceptions of the Malay Archipelago, Madagascar, Canada, and a few other areas.

E. Additional Experiment

E.1. Overall Performance Comparison

We compare the performance of different methods evaluated on the entire GeoMap-Bench dataset at both the ability and the task levels. To visually present these results, we use radar charts, as shown in Figure A2 and Figure A3. The results demonstrate that (1) Currently, GPT-4o is the best publicly available MLLMs on GeoMap-Bench across various abilities and tasks. (2) Our method, GeoMap-Agent, significantly outperforms all the public MLLMs using GPT-4o as the base model.

E.2. Improvement from Prompt Enhancement

In the last PEQA module, GeoMap-Agent is further improved by enhancing its prompt from 4 aspects. Aside from the first context enhancement, which relies on the global metadata and external knowledge from the previous two modules, the other three can be applied independently. To evaluate the effectiveness, experiments are conducted with and without the last 3 enhancements in the PEQA module, as demonstrated in Table A1.

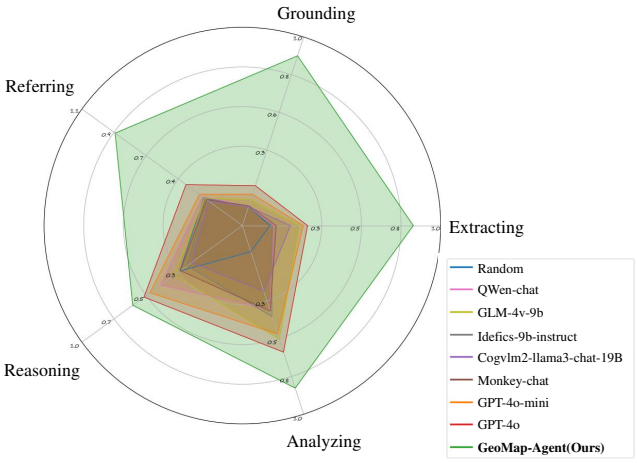

Figure A2. Overall performance comparison on different abilities.

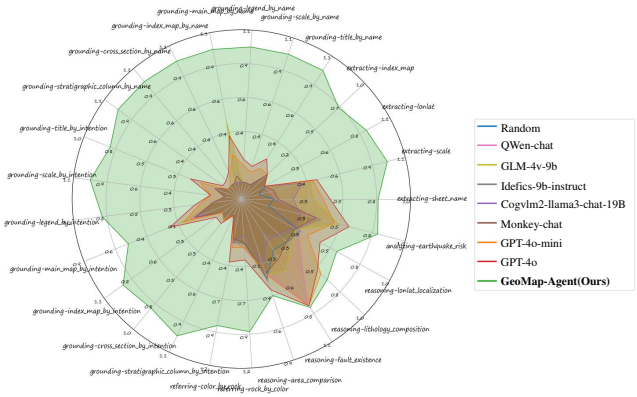

Figure A3. Overall performance comparison on different tasks.

E.3. Time Cost Analysis

The proposed GeoMap-Agent effectively addresses efficiency issues in several ways. Firstly, the HIE module digitalizes each map only once, and all the questions of each map share the digitalized metadata, avoiding repeated digitalization. Secondly, the DKI module injects only necessary knowledge from the expert group, the first round of conversation would filter out most of the knowledge types, rather than leveraging all available knowledge. Lastly, the PEQA module inputs only the focused regions of the map as a prompt to the base model, significantly reducing tokens compared to the baseline method, which takes the whole map as input. As shown in Table A2, the average time cost of GeoMap-Agent and GPT-4o on different subsets are comparable. However, the accuracy of GeoMap-Agent is significantly improved.

Dataset	Ability	enhance prompt w/	enhance prompt w/o
USGS Set	Ext.	**0.379**	0.208
	Gro.	**0.123**	0.100
	Ref.	**0.415**	0.398
	Rea.	0.491	**0.494**
	Ana.	**0.733**	0.683
	Ove.	**0.428**	0.376
CGS Set	Ext.	**0.326**	0.230
	Gro.	**0.258**	0.157
	Ref.	0.331	**0.359**
	Rea.	**0.547**	0.521
	Ana.	**0.584**	0.542
	Ove.	**0.409**	0.361

Table A1. Performance comparison of GeoMap-Agent on GeoMap-Bench with and without prompt enhancement. All other settings remain the same, including the use of GPT-4o as base model and **excluding the HIE and DKI modules**.

Dataset	GPT-4o (s/question)	GeoMap-Agent (s/question)
USGS Set	6.26	7.43
CGS Set	8.08	12.58

Table A2. Comparison of time cost between GPT-4o and GeoMap-Agent on a machine with an AMD EPYC 7763 64-Core Processor and no GPU.

F. Other Tools

F.1. Lithological Mapping Table

To incorporate lithological knowledge into GeoMap-Agent, our professional geologists compile a 3-level lithological table (rock type, rock category, and lithology), containing 335 items in English and 256 items in Chinese, which is scalable as well. A sample of the English lithological table is presented in Table A3.

F.2. Legend Unit Extractor

The legend unit is a standardized component across different geologic map sources. We develop a tool for information extraction within each legend unit, encompassing both text and color extraction. This process is based on the bounding box pairs of text unit and color unit detected by legend unit detector described in Section D. For text extraction, we employ the base model to process each cropped legend text unit, using the prompt "Only output the OCR result of the given image.". For color extraction, we calculate the median color in each cropped legend color unit.

Class	Subclass	Lithology
Sedimentary	Clastic	conglomerate
		tillite
		breccia
	Carbonate	limestone
		marl
	...	...
Volcanic	Acid volcanic	trachydacite
		keratophyre
		quartz keratophyre
	Alkali volcanic	analcimite
		leucitite
	...	...
Intrusive	Acid intrusive	tonalite
		plagiogranite
	Alkaline intrusive	foid diorite
		foid gabbro
	...	...
Metamorphic	Slate	siliceous slate
		charcoal slate
		sandy slate
	Schist	graphitic schist
		actionlite schist
		amphibole schist
	...	...

Table A3. Sampled lithological mapping table. There are a total of three levels, compiled by geological experts.

References

[1] Andrea E Gaughan, Forrest R Stevens, Catherine Linard, Peng Jia, and Andrew J Tatem. High resolution population distribution maps for southeast asia in 2010 and 2015. *PloS one*, 8 (2):e55882, 2013. 3

[2] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. 3

[3] OpenAI. hello-gpt-4o. (2024). 2

[4] Alessandro Sorichetta, Graeme M Hornby, Forrest R Stevens, Andrea E Gaughan, Catherine Linard, and Andrew J Tatem. High-resolution gridded population datasets for latin america and the caribbean in 2010, 2015, and 2020. *Scientific data*, 2 (1):1–12, 2015. 3

[5] Richard Styron and Marco Pagani. The gem global active faults database. *Earthquake Spectra*, 36(1_suppl):160–180, 2020. 3

[6] USGS. Earthquake hazards program. 1977. 3

[7] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024. 3

[8] Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, et al. Esa worldcover 10 m 2021 v200. 2022. 3