GR1: Reinforcement-Enhanced LLM for Geoscience Reasoning

Yule Xie[1], Jiaxin Ding[1,*], Cheng Deng[2], Shiqing Gao[1], Junran Zhang[1], Sibo Zhang[1], Zeyuan Wang[1], Ke Wu[1], Xin Ding[1], Luoyi Fu[1], Meng Jin[1], Xinbing Wang[1]
[1]Shanghai Jiao Tong University, [2]University of Edinburgh
jiaxinding@sjtu.edu.cn

Abstract

Reinforcement learning (RL) has recently shown remarkable ability to enhance reasoning in large language models (LLMs), yet its potential in scientific domains beyond mathematics remains largely unexplored. Geoscience questions couple broad factual knowledge with multi-step inference and often rely on visual evidence such as maps, cross-sections, and diagrams, making them a challenging but verifiable testbed for RL-based reasoning. To enable this study, we introduce **GeoMC-10K**, a dataset of 10,000 geoscience multiple-choice questions spanning physical to human geography and high-school to professional levels; over 30% of the questions are image dependent. To support text-only RL on these multimodal questions, we design **GeoM2T**, a multi-agent framework that converts multimodal questions into descriptive text while preserving answerability and difficulty. Fine-tuning LLaMA-3.1-8B and Qwen-3-8B with Group Relative Policy Optimization (GRPO), incorporating a factual reward mechanism, yields **GR1**, which achieves absolute accuracy improvements of 5.9% and 13.3%, respectively, and it generalizes to out-of-distribution geoscience benchmarks. Together, GeoMC-10K, GeoM2T, and GR1 establish a scalable benchmark and baseline for RL-enhanced geoscience reasoning. Code is available at https://github.com/xyl-alter/GR1.

1 Introduction

Recent advances in large language models (LLMs) have underscored the transformative impact of embedding complex reasoning capabilities into general-purpose systems. Notable breakthroughs, such as OpenAI's O1 and DeepSeek's R1, have demonstrated reinforcement learning's (RL) substantial capacity to enhance multi-step reasoning beyond what supervised instruction tuning achieves. However, most existing studies focus primarily on mathematical reasoning (Qwen, 2024; Shinn et al., 2023; Lewkowycz et al., 2022), where challenges are primarily symbolic and answers are easily verifiable. Whether similar RL-driven gains extend to scientific domains, where questions intertwine factual knowledge, causal inference, and heterogeneous evidence, remains less understood.

Geoscience is a natural and consequential testbed for scientific reasoning. Unlike mathematics, geoscience reasoning requires a dual capability: the recall of extensive factual knowledge (e.g., geology, climatology, and human geography) and the execution of multi-step inferential reasoning grounded in natural and human systems. This dual requirement aligns well with the strengths of RL: its reward-driven sampling mechanism has already proven effective at boosting multi-step reasoning in mathematical tasks (Christiano et al., 2017; Ouyang et al., 2022); and by up-sampling high-reward factual outputs (e.g., correctly selecting Washington over New York as the capital of the United States) while down-sampling incorrect ones, it can also consolidate and even discover domain-specific knowledge (Levine, 2018). This makes RL a highly feasible strategy for training geoscience models. Existing domain-specific geoscience LLMs, such as K2 (Deng et al., 2024) and GeoGalactica (Lin et al., 2023), have been trained primarily through supervised fine-tuning (SFT) and instruction tuning, leaving RL-based training of geoscience models an appealing and largely unexplored pathway.

Progress has been impeded by two domain-specific practical barriers. First, applying RL requires large-scale, high-quality, and verifiable benchmarks, but existing geoscience resources are often small, narrow, or lack standardized evaluation protocols. Second, a substantial fraction (more than 30%) of real-world geoscience questions are inherently multimodal, involving maps, cross-sections, and diagrams, while RL for multi-

*Jiaxin Ding is the corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 22730–22743

modal LLMs remains immature (Zhang et al., 2024; Ma et al., 2025; Chen et al., 2025a). Naïvely converting figures to text can either omit critical cues (hurting answerability) or introduce unintended hints (collapsing difficulty), making it difficult to build a faithful training signal for RL.

To address these barriers, we present a unified benchmark-pipeline-model contribution for RL-enhanced geoscience reasoning. We introduce **GeoMC-10K**, with 10,000 multiple-choice questions (MCQs) from NYSED Earth Science exams, AP tests, China's Gaokao, Engineering Geology MCQs, and ASBOG practice problems, covering a wide spectrum from physical to human geography and from high school to professional levels. To enable text-based RL on the image-dependent items, we propose **GeoM2T**, a multi-agent "Extractor-Critic-Advisor" pipeline that leverages LLM self-refinement to iteratively produce text-only equivalents that remain both information-complete and difficulty-matched to the original multimodal questions. Finally, we fine-tune LLaMA-3.1-8B (Dubey et al., 2024) and Qwen-3-8B (Qwen, 2025) backbones using Group Relative Policy Optimization (GRPO), incorporating a **factual reward** mechanism that grants partial credit for semantically accurate scientific statements, guiding the model to ground its reasoning in verified domain knowledge rather than surface patterns.

The resulting model, **GR1**, achieves absolute accuracy improvements of 5.9% and 13.3% over the base models respectively, with the factual reward accounting for over 17% of the overall performance gains. To the best of our knowledge, GR1 is the first LLM explicitly optimized for geoscience reasoning via reinforcement learning. These results show that RL can effectively enhance domain-specific reasoning in LLMs, and validate GR1 as a strong baseline for future geoscience research.

Our contribution can be summarized as follows:

- **GeoMC-10K**, a large-scale corpus of 10,000 real-world MCQs from diverse sources covering physical to human geography and high-school to professional levels, providing a standardized evaluation benchmark for geoscience reasoning tasks.
- **GeoM2T**, a multi-agent pipeline leveraging the self-refinement capabilities of LLMs to extract, critique, and iteratively refine image-derived information, thereby converting each original multimodal item into text-only equivalents while preserving answerability and difficulty, enabling text-based RL training.
- **GR1**, a RL-trained geoscience reasoning model fine-tuned from LLaMA-3.1-8B and Qwen-3-8B using GRPO with lightweight factual supervision on our dataset, yielding accuracy gains of 5.9% and 13.3% respectively.

2 Related Works

2.1 RL for Reasoning in LLMs

Reinforcement learning (RL) has recently emerged as a promising framework for enhancing the reasoning capabilities of large language models (LLMs), particularly in multi-step problem-solving tasks. Methods such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), Direct Preference Optimization (DPO) (Rafailov et al., 2023), and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) guide model updates based on reward signals derived from reasoning quality or final correctness. In the mathematical domain, numerous studies have explored RL-based fine-tuning to enhance multi-step reasoning in LLMs (Lewkowycz et al., 2022; Shinn et al., 2023). In interdisciplinary domains, mainly medicine, recent efforts such as HuatuoGPT-O1 (Chen et al., 2024) and Med-U1 (Zhang et al., 2025) further confirming RL's value in specialized, interdisciplinary settings.

2.2 Domain-Specific LLMs for Geoscience

While reinforcement learning has demonstrated effectiveness across a wide range of domains, its potential for enhancing reasoning in geoscience tasks remains underexplored. Prior efforts to build geoscience-specialized models, including K2 (Deng et al., 2024), GeoGalactica (Lin et al., 2023), and GeoGPT (Science Custom Publishing, 2024), rely primarily on domain-specific pre-training and supervised instruction tuning. We will provide a detailed introduction to these models in the experimental section.

2.3 Verifiable Geoscience Problems

Outcome-level reinforcement learning relies on tasks whose correctness can be automatically and unambiguously verified. In the domain of mathematics, several large-scale datasets have been developed to support this paradigm, including Big-Math (Albalak et al., 2025), GSM8K (Cobbe et al., 2021), MathQA (Amini et al., 2019), etc.

In the geoscience domain, however, verifiable datasets remain scarce. Representative efforts include GeoSignal and GeoBench from the K2 project, as well as the geoscience subset of ScienceQA (Lu et al., 2022). GeoSignal contains open-ended alignment-style questions and lacks ground-truth verifiability. GeoBench comprises approximately 1.5K multiple-choice questions derived solely from AP test sources and excludes original imagery references, limiting both its scale and multimodal fidelity. ScienceQA includes 4,108 geoscience-related multiple-choice questions, but the majority are shallow in complexity and low in topical diversity—for example, 1,475 questions about U.S. state capitals and 473 on climate-vs-weather distinctions.

To date, no existing geoscience dataset satisfies the combination of large scale, automatic verifiability, and high reasoning complexity required for outcome-level RL fine-tuning. This gap motivates our construction of a new benchmark, introduced in the next section.

3 Data Collection and Curation

To underpin our reinforcement learning experiments, we built GeoMC-10K, a unified corpus of over 10,000 geoscience multiple-choice questions drawn from authoritative educational and professional sources. In the following, we first provide a brief overview of each data source, then introduce GeoM2T, a pipeline that converts multimodal questions into text-only format.

3.1 Data Collection

Our GeoMC-10K dataset aggregates questions from six widely used geoscience assessments, providing both scale and domain breadth. These include high-school level exams—New York State Earth Science Regents, China's Gaokao, and Advanced Placement (AP) Earth Science tests-professional certification samples (ASBOG exam questions and Engineering Geology MCQs), and the Geoscience Concept Inventory (GCI), a research-validated tool in geoscience education. Together, they span physical and human geography across varying difficulty levels and modalities. Detailed descriptions of data sources are provided in Appendix A.1, and data leakage checks are reported in Appendix A.2.

Aggregating questions from the six sources described above, we assembled GeoMC-10K, a corpus of over 10,000 geoscience multiple-choice items. Table 1 summarizes its distribution. As shown in Table 1, over 30% of GeoMC-10K items include embedded figures, and preliminary LLaMA-8B evaluations reveal a notable drop in accuracy on these multimodal questions compared to purely textual ones. This rich visual content, while central to authentic geoscience reasoning, complicates reinforcement learning workflows that require text-only inputs, thus underscoring the necessity of our multimodal-to-text conversion pipeline.

Sources	Num. of Questions (LLaMA-8B Acc.)		
	Total	Plain-text	Multimodal
NYSED	2617 (74.8%)	1515 (80.7%)	1102 (66.7%)
Gaokao	2904 (53.2%)	806 (54.3%)	2098 (52.8%)
APTest	1395 (73.7%)	1214 (74.4%)	181 (68.5%)
Eng.	2915 (54.3%)	2915 (54.3%)	0 (–)
ASBOG	113 (57.3%)	113 (57.3%)	0 (–)
GCI	56 (56.1%)	56 (56.1%)	0 (–)
Total	10000 (62.1%)	6619 (64.1%)	3381 (58.2%)

Table 1: Statistics of the GeoMC-10K dataset. Values in parentheses denote corresponding Llama-8B accuracy.

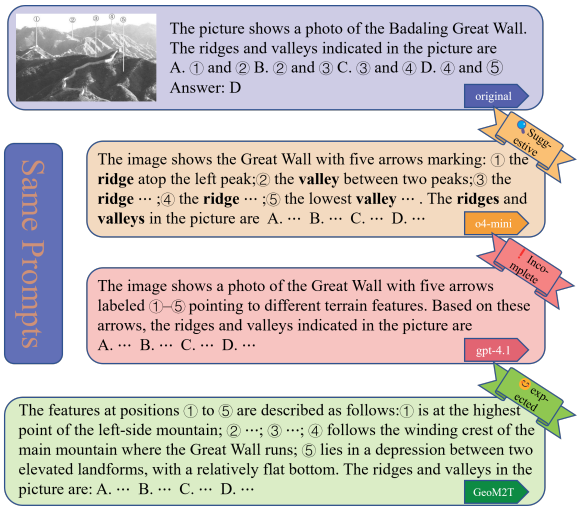

Figure 1: We evaluated the zero-shot performance of commercial LLMs in converting multimodal geoscience questions to text. Both GPT-4.1 and o4-mini-high failed under the same prompt: the former omitted key information, while the latter leaked the answer.

3.2 Multimodal-to-Text for Geoscience MCQs

Transforming multimodal geoscience questions into purely textual formats requires a delicate balance between preserving essential information for accurate problem solving and maintaining the original level of difficulty. While more exhaustive extraction of key information enhances answerability,

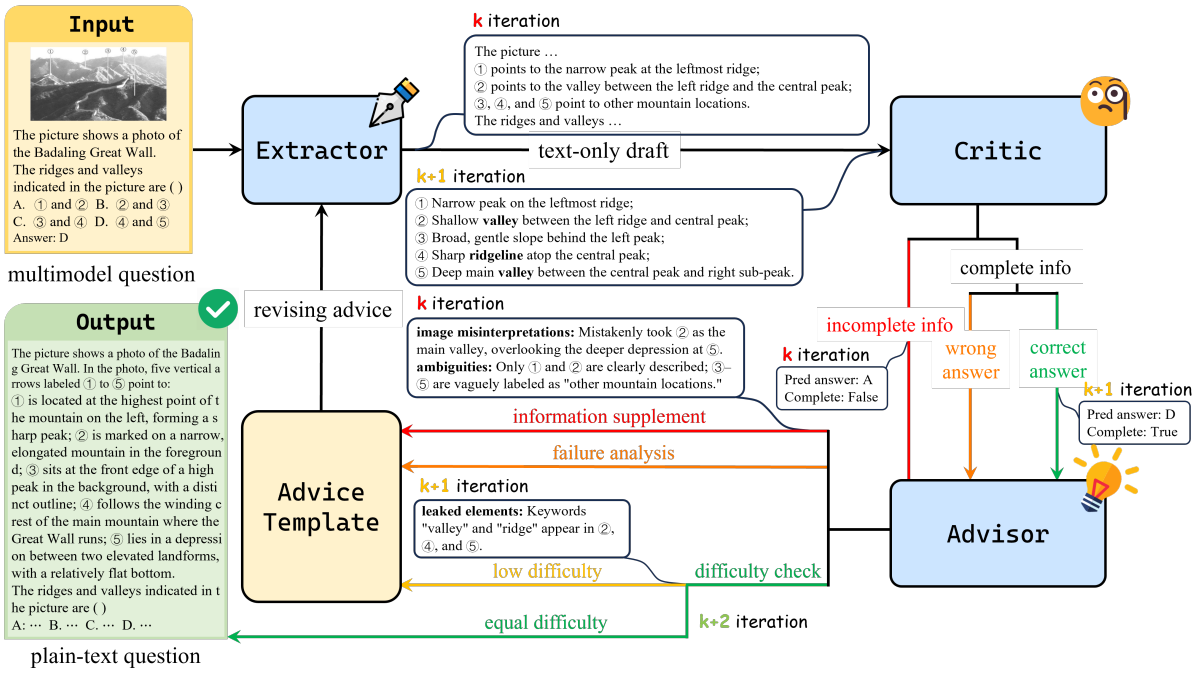

Figure 2: Overview of the GeoM2T pipeline. It involves three agents: the Extractor proposes drafts, the Critic tests their answerability, and the Advisor suggests targeted revisions for refinement.

it can also unintentionally reduce the complexity of the question. Although commercial LLMs, such as GPT-4o, exhibit strong capabilities in understanding and solving multimodal tasks, our experiments show that they often either omit crucial visual cues or contain suggestive descriptions, as is shown in Figure 1. This leads to question drafts that are either insufficiently informative to support valid answers or overly simplified to the point where the challenge is diminished.

Recent research highlights the impressive self-refinement capabilities of LLMs, and our findings corroborate this: iterative revision of an initial draft often yields better outcomes than attempting to generate a complete solution in a single pass through prompt engineering. To harness this potential, we design **GeoM2T**, a **Geo**science **M**ultimodal-**to**-**T**ext pipeline composed of three cooperative agents—Extractor, Critic, and Advisor—that iteratively refine question drafts until both completeness and difficulty criteria are satisfactorily met. Figure 2 visualizes our GeoM2T pipeline.

Specifically, given a question with one or more images, the Extractor first produces a draft that verbalizes the key visual elements in the stem and options. The Critic then attempts to solve the question based on this draft and assigns one of three labels: incomplete, if crucial visual cues are missing; complete-but-incorrect, if the draft appears self-contained but leads to a wrong answer; and adequate, if the draft supports the correct answer at a comparable difficulty level.

Based on this label, the Advisor generates targeted feedback: for incomplete drafts it enumerates missing cues that must be added, for complete-but-incorrect drafts it highlights misinterpreted or ambiguous descriptions that mislead the solver, and for adequate drafts it checks for unintended answer hints or missing distractors that may alter the original difficulty. The revised draft, together with the Advisor's feedback, is fed back to the Extractor, and this loop is repeated for a small number of iterations (five in our experiments) until the Critic judges the question to be both answerable and difficulty-matched. Implementation details are provided in Appendix C.1.

4 Factuality-aware Geoscience Reasoning

Given that the base model has already achieved a moderate level of accuracy on our dataset, we skip the supervised fine-tuning (SFT) stage and directly employ Group Relative Policy Optimization (GRPO) to make full use of the dataset to stimulate the model's reasoning potential.

Reward Design. Rewards play a crucial role in guiding the RL training target. During generation, we prompt the model to output an intermediate reasoning segment, factual knowledge segments

used to solve the problem, and a final answer using the structured format <think> ... </think>, <knowledge> ... </knowledge>, and <answer> ... </answer>.

Our motivation stems from the dual nature of geoscience, which combines STEM-like reasoning (e.g., quantitative analysis) with humanities-like factual recall (e.g., conceptual knowledge). Intuitively, we draw inspiration from a familiar classroom experience: in STEM subjects, answers are primarily judged by correctness, whereas in humanities, teachers often award partial credit for key points. Following this analogy, our reward design aims to capture both absolute correctness and partial factual reasoning.

Moreover, from a reinforcement learning perspective, it also mitigates a zero-gradient issue in GRPO, where purely outcome-based rewards collapse to identical values when all sampled trajectories are incorrect, yielding zero groupwise advantage and no policy update. By assigning partial credit to factual knowledge, the factual reward differentiates such trajectories and enables non-zero updates even when final answers are wrong.

The final reward function consists of three parts:
Accuracy Reward: A reward of 1.0 is assigned if the predicted answer matches the ground-truth answer, and 0.0 otherwise.
Format Reward: A reward of 1.0 is granted if the model's output conforms to the required <think>, <knowledge>, and <answer> format, and 0.0 otherwise.
Factual Reward: When the final answer is incorrect, we compute a soft factual similarity between the model-generated knowledge claims and the ground-truth claims extracted by `gpt-3.5-turbo` using the *VeriScore* (Song et al., 2024) approach. Specifically, let p_i denote a factual claim produced by the model within the <knowledge> block, and g_j denote a required gold factual claim automatically extracted by *VeriScore*, we first encode all claim pairs (p_i, g_j) with a `RoBERTa` cross-encoder to obtain semantic similarities $S(p_i, g_j)$. We then apply the Hungarian algorithm to obtain optimal one-to-one matchings:

$$M^* = \arg\max_{M} \sum_{(p_i,g_j)\in M} S(p_i, g_j),$$

and compute the soft Jaccard similarity as:

$$R_{\text{softJ}} = \frac{\sum_{(p_i,g_j)\in M^*} S(p_i, g_j)}{|P| + |G| - \sum_{(p_i,g_j)\in M^*} S(p_i, g_j)}.$$

This value serves as a continuous factual reward, providing partial credit for semantically relevant knowledge. And when the final answer is correct, we set factual reward to 0 since the accuracy reward is already set to 1. This avoids double-counting rewards and prevents the model from over-optimizing verbose factual descriptions when correctness is already achieved.

The overall reward is the average of the three components, encouraging the model to produce well-formed and factually grounded answers.

A recent study on reasoning in scientific domains (Chen et al., 2025b), has also introduced a factuality-oriented reinforcement learning framework. However, their approach relies on commercial LLM-based factuality judges to evaluate each generated sample during training. Such online factuality scoring incurs substantial API token costs and significantly increases training time due to network latency. In contrast, our method empowers the model itself to summarize the factual knowledge it relies on within the <knowledge> segment during generation. This allows us to perform factual claim extraction only once during dataset preprocessing, while preserving the low-cost and efficient computation characteristics of GRPO. We discussed the consistency between extracted claims and human annotations in Appendix B.2.

Reinforcement Fine-Tuning. GRPO is a RL algorithm that updates the policy based on groupwise relative performance, without relying on well-trained reward models such as PPO. We employ GRPO with a KL-divergence regularization term for RL fine-tuning. Further algorithmic details can be found in the Appendix B.1.

5 Experiments

5.1 GeoM2T Pipeline Evaluation

To evaluate the effectiveness of the proposed GeoM2T pipeline, we conduct experiments on 1,000 multimodal geoscience questions randomly sampled from the Gaokao examination corpus. Evaluation metrics are as follows:

- **Answerability:** We use GPT o4-mini-high to check whether the transformed questions retain essential information. As this model performs well on high school-level text question, failing to answer correctly suggests that key visual cues were lost or misrepresented.
- **Difficulty:** For questions GPT answers cor-

	Model	Size	NYSED	gaokao	APTest	Eng.	ASBOG	GCI	Reason	Knowledge	All
	GeoGalactica	30B	38.62	20.56	23.08	36.88	26.07	10.00	24.74±0.32	33.14±0.95	30.34±0.74
LLaMA	K2	7B	25.51	18.63	21.24	27.46	21.74	16.67	16.72±3.68	26.78±3.41	23.43±3.50
	LLaMA3.1-8B	8B	75.64	53.29	75.72	53.89	75.00	70.54	51.64±1.37	68.12±0.67	62.64±0.68
	SFT	8B	76.02	55.99	77.95	53.22	80.43	71.43	53.80±0.93	68.61±0.85	63.69±0.52
	GR1-LLaMA†	8B	80.90	60.90	81.31	58.75	**83.70**	**81.25**	59.53±1.32	73.07±0.51	68.57±0.35
	LLaMA3.1-70B	70B	**87.01**	**69.72**	**88.68**	**59.02**	**83.70**	70.54	**70.45**±1.34	**75.23**±0.42	**73.64**±0.55
	GeoGPT-LLaMA	70B	84.36	65.28	86.95	54.93	80.43	76.79	64.62±0.27	73.03±0.85	70.24±0.54
Qwen	Qwen3-8B	8B	78.83	57.44	81.48	53.85	78.26	87.50	55.04±0.75	70.84±0.32	65.59±0.21
	SFT	8B	78.78	61.06	80.72	52.65	78.26	87.50	56.92±0.83	70.69±0.28	66.11±0.32
	SFT+RL	8B	80.10	68.65	80.70	64.03	81.16	85.71	63.48±0.60	76.37±0.51	72.08±0.49
	GR1-Qwen†	8B	**92.55**	**74.42**	86.00	**68.43**	81.99	86.73	**75.57**±0.83	**80.57**±0.36	**78.91**±0.37
	Qwen3-32B	32B	87.40	60.81	85.06	63.50	**89.13**	**93.75**	59.66±0.77	78.59±0.43	72.29±0.42
	GeoGPT-Qwen	72B	89.91	73.89	**90.07**	60.21	75.36	71.43	74.31±0.91	76.85±0.52	76.01±0.32
	number	–	489	563	297	614	23	14	665	1335	2000

Table 2: Accuracy of all models on the GeoMC-10K test set as well as the Reasoning and Knowledge oriented subsets. All results are reported in percentages, with mean ± standard deviation.

rectly, we evaluate their difficulty using LLaMA-8B, which is one of our fine-tuning target, by measuring (1) accuracy and (2) average reasoning length. Lower accuracy and longer responses suggest higher complexity and better retention of the original challenge.

We compare GeoM2T against two families of baselines: prompt-engineering methods (zero-shot, few-shot, and answer-aware zero-shot) and image-captioning methods (PromptCap and IC3). Details of each baseline are provided in Appendix C.2. Experimental results are shown in Table 3.

With zero-shot prompting as the baseline, few-shot prompting yields slightly lower accuracy and longer outputs, as the provided examples (e.g., "a depression between two elevated landforms" instead of "valley") obscure direct clues, increasing difficulty but making it harder for the Extractor to accurately describe key elements. Providing the correct answer improves answerability by guiding the model's focus towards ground truth, but also lowers difficulty by introducing preferences for correct answers.

IC3 combines outputs from multiple models and temperatures for broader coverage, yet lacks mechanisms to preserve challenge, resulting in performance similar to the answer-aware setting. PromptCap underperforms due to its limited capacity for extracting fine-grained, task-relevant features, reflecting the constraints of lightweight vision-language models.

In contrast, GeoM2T achieves clear gains in both answerability and reasoning difficulty. Moreover, when we replace the o4-mini-high backbone with the much weaker Gemini 2.5-Flash model, GeoM2T* retains over 90% of GeoM2T†'s performance, suggesting that our pipeline is robust to the choice of backbone model.

Method	**Accuracy**		**Generated Words↑**
	GPT↑	**LLaMA↓**	
Zero-shot	63.64%	55.47%	171.44
Few-shot	59.47%	54.48%	175.48
Zero-shot + Ans	66.63%	56.08%	170.21
PromptCap	32.49%	58.89%	150.73
IC3	66.01%	56.60%	175.50
GeoM2T† (Ours)	**90.69%**	**52.75%**	**181.49**
GeoM2T* (Ours)	83.75%	53.73%	180.75

Table 3: Performance comparison of different methods. Higher GPT accuracy means better answerability; lower LLaMA accuracy and longer generated words by LLaMA means higher difficulty. **GeoM2T**† (and other LLM based methods) uses GPT (o4-mini-high) as its backbone, while **GeoM2T*** uses the much weaker Gemini (2.5-Flash) backbone.

5.2 Effects of Fine-Tuning.

5.2.1 Experimental Setup

Data Splitting. Keeping the proportion of each data source unchanged, we split the GeoMC-10K dataset into training, validation, and test sets with a 7:1:2 ratio. We further divide the test set into two subsets based on few-shot prompting results from the GPT-4.1 model: Knowledge, comprising questions that primarily assess factual understanding; and Reasoning, comprising questions that require multi-step logical deduction. This division reflects the dual emphasis of geoscience LLMs on factual

knowledge retrieval and multi-step inferential reasoning, and facilitates the analysis of reinforcement learning's impact on these two question types.

Model Training. We fine-tune two widely used 8B-parameter base models: LLaMA-3.1-8B-Instruct and Qwen-3-8B, using our training set of 7,000 questions. Training is conducted for 3 epochs on six NVIDIA H20 GPUs with a constant learning rate of 3×10^{-6}. To mitigate policy collapse during reinforcement learning, we apply a KL-divergence penalty in the GRPO objective, setting the weight coefficient β to 0.07 for the LLaMA based model and 0.04 for the Qwen based model. To reduce noise from weakly related pairs, we mask similarity scores below 0.3 when computing factual rewards.

We evaluate our fine-tuned models against three categories of baselines:

General LLMs. This includes the original base model (LLaMA-3.1-8B, Qwen-3-8B) and larger variants in the same families (LLaMA-3.1-70B, Qwen-3-32B).

Geoscience-Specific LLMs. We include three specialized model series: K2, GeoGalactica, and GeoGPT. K2, the first geoscience LLM, is built on LLaMA-7B and further pre-trained on 5.5B tokens of geoscience texts, followed by instruction tuning on the GeoSignal dataset. GeoGalactica is a 30B-parameter model obtained by pre-training Meta's Galactica on 65B tokens of geoscience data and fine-tuning on 1M expert-level instructions. GeoGPT comprises a series of models: LLaMA3.1-70B GeoGPT and Qwen2.5-72B GeoGPT, both trained via SFT and instruction tuning.

Training Variants. We also include SFT as a baseline using the question stem, GPT-generated reasoning, and the ground-truth answer. We further consider a hybrid SFT-RL baseline where the training data are evenly split between SFT and reinforcement learning. While performing SFT before RL is a common training strategy, in our case the base model already possesses a certain level of reasoning and instruction-following ability, making direct RL training feasible.

5.2.2 Evaluation on Geoscience Tasks

We evaluate GR1 and various baselines on the GeoMC-10K test set; results are reported in Table 2. Unless otherwise noted, results in this section are averaged over four runs with different random seeds, and we report mean accuracy and standard deviation. Key observations are summarized below.

	Model	Reason	Knowledge	All (ΔAll)
LLaMA	LLaMA-8B	51.64	68.12	62.64 (base)
	GR1 (w/o F)	57.35	71.78	66.98 (↑ +4.34)
	GR1 (w/ F)	**59.53**	**73.07**	**68.57** (↑ +5.93)
Qwen	Qwen-8B	55.04	70.84	65.59 (base)
	GR1 (w/o F)	71.00	80.26	77.18 (↑ +11.59)
	GR1 (w/ F)	**75.57**	**80.57**	**78.91** (↑ +13.32)

Table 4: Ablation study on the factual reward on the GeoMC-10K test set. "w/o F" and "w/ F" indicate training without and with the factual reward, respectively. All results are reported in percentages.

RL Fine-Tuning Boosts Accuracy. Applying GRPO fine-tuning to the LLaMA-8B and Qwen-8B models produces consistent gains on both the LLaMA and Qwen model families. For LLaMA-3.1-8B, the overall accuracy improves from 62.64% to 68.57%, yielding a gain of 5.93 percentage points (ppt). Similarly, Qwen-8B benefits substantially from GRPO fine-tuning, with accuracy increasing from 65.59% to 78.91% (+13.32 ppt). These results confirm that RL-based fine-tuning guided by factual rewards significantly enhances performance on geoscience fields.

Small Model, Strong Capability. After GRPO fine-tuning, GR1-qwen (8B) achieves 78.91% accuracy, achieving best performance over all models. Within the LLaMA family, RL fine-tuning also narrows the performance gap between smaller and larger models, indicating that well-designed RL objectives can compensate for limited model capacity to a large extent.

Greater RL Gains for Reasoning-Intensive Questions. We observe a markedly larger improvement in accuracy on reasoning oriented questions compared to knowledge oriented questions. For Qwen-8B, reasoning accuracy increases from 55.04% to 75.57% (+20.53 ppt), while knowledge accuracy improves from 70.84% to 80.57% (+9.73 ppt). Similarly, for LLaMA-3.1-8B, the reasoning subset improves by 7.69 ppt, compared to a 4.95 ppt gain on the knowledge subset. These shows that RL-based fine-tuning amplifies both the model's factual recall and its multi-step reasoning capabilities, with the latter exhibiting more improvements.

5.3 Ablation Study on Factual Reward

To evaluate the effectiveness of the proposed factual reward, we conduct an ablation study by comparing models trained with and without this reward under the same GRPO framework. As shown in Ta-

ble 4, incorporating the factual reward consistently improves both reasoning and knowledge performance across model families. Overall, the factual reward provides an additional 1.6–1.8 ppt improvement in total accuracy over standard RL, accounting for roughly 17.2% of the overall performance gain.

Notably, the improvements brought by the factual reward are more evident on reasoning-oriented questions than on the knowledge subset. For example, on Qwen-8B, the reasoning accuracy increases by 4.47 ppt, compared to a 1.77 ppt gain on the knowledge subset. This observation suggests that the factual reward does not simply encourage memorization of factual content. Instead, it supports the model in recalling relevant factual information from its parametric knowledge during the reasoning process, thereby facilitating multi-step reasoning that integrates evidence and prior knowledge.

These results represent a meaningful step toward factuality-aware reasoning, demonstrating that factual knowledge can be effectively utilized during the reasoning process with an efficient training strategy that avoids reliance on external LLMs.

5.4 Out-of-Distribution (OOD) Evaluations

To evaluate the generalization ability of our models under distribution shifts, we propose two OOD benchmarks:

- **IESO** (International Earth Science Olympiad): We collect MCQs from official IESO competitions held between 2013 and 2019. It evaluates knowledge-level OOD generalization and represents a more challenging setting that requires richer domain-specific knowledge and stronger causal reasoning capabilities.

- **GeOpen**: GeOpen is an open-ended question answering dataset constructed from Chinese postgraduate entrance examination problems. It introduces a format-level distribution shift by requiring free-form answer generation rather than multiple-choice selection, thereby testing the model's robustness to task changes.

For the GeOpen benchmark, We use GPT-4.1-mini as a judge. We randomly verify over 200 samples with two geoscience experts with Ph.D. degrees, where the average score difference between expert judgments and model-based ones is 0.12. More details can be found in Appendix D.

	Model	Size	IESO	GeOpen
Llama	LLaMA3.1-8B	8B	53.50±2.26	55.60±0.43
	GR1-LLaMA†	8B	54.76±3.53	62.42±0.13
	GR1 (w/o F)†	8B	54.24 (↓0.52)	56.94 (↓5.48)
	LLaMA3.1-70B	70B	**70.68**±2.30	70.38±0.48
	GeoGPT-LLaMA	70B	66.74±2.30	**78.09**±1.59
Qwen	Qwen3-8B	8B	53.27±2.29	58.50±0.42
	GR1-Qwen†	8B	**74.85**±2.24	**94.83**±0.22
	GR1 (w/o F)†	8B	73.14 (↓1.71)	76.02 (↓18.8)
	Qwen3-32B	32B	67.19±2.71	88.22±0.32
	GeoGPT-Qwen	72B	73.51±2.23	82.55±1.22
	Number	–	84	1035
	Difficulty	–	competition	postgraduate

Table 5: Results on OOD datasets and tasks, including the multiple-choice IESO benchmark and the open-ended GeOpen benchmark. Accuracy is reported in percentage (mean ± standard deviation). "w/o F" indicates ablation settings without the factual reward.

As shown in Table 5, GR1 consistently improves over its corresponding base models on both OOD benchmarks. Among all evaluated models, GR1-Qwen achieves the best performance on both datasets. Moreover, when the factual reward is removed, performance drops are consistently observed. These observations suggest that our training approach is robust to distribution shifts and generalizes well to different geoscience tasks.

6 Conclusion

In this work, we introduced GeoMC-10K, a 10,000 item geoscience MCQ dataset, developed a multi-agent GeoM2T pipeline to convert multimodal questions into both answerability- and difficulty-preserving text-only ones, and demonstrated that GRPO-based RL fine-tuning of LLaMA-8B and Qwen-8B yields substantial accuracy gains.

To further enhance factual reliability, we proposed a factual reward that encourages models to explicitly summarize and utilize the factual knowledge involved in reasoning, thereby advancing efficient factuality reasoning based on the traditional GRPO framework. The GR1-Qwen model achieves the best performance on the test set, surpassing many larger-scale general-purpose and geoscience-specific models. The experiments show that in subjects like geoscience where factual knowledge and multi-step reasoning are both essential, reinforcement learning can still produce strong performance gains. Together, these contributions establish a new benchmark and baseline for geoscience-specific reasoning LLMs.

7 Acknowledgement

This work was supported by NSF China under Grant No. 92579104, T2421002, 62525209, T2542021.

8 Limitations

First, although GeoMC-10K substantially expands the scale of verifiable geoscience questions, it primarily contains multiple-choice items, which may not capture the full diversity of open-ended or quantitative reasoning tasks in geoscience. Extending the dataset to include free-response and multimodal problems would allow broader evaluation.

Second, the GeoM2T pipeline transforms image-dependent questions into text-only versions; while our iterative refinement procedure helps preserve difficulty and fidelity, the resulting descriptions may still lose certain spatial or visual nuances that are critical in geoscientific interpretation. Future multimodal RL frameworks could directly integrate visual signals instead of relying on text conversion.

Third, our factual-reward mechanism is based on semantic similarity between generated knowledge statements and reference facts, which provides partial credit but may not fully capture deeper causal correctness or conceptual soundness.

9 Ethical considerations

All data used in this study were collected from publicly accessible educational resources and examination archives available on the internet. The data include geoscience multiple-choice questions from open educational websites, governmental exam repositories, and publicly released test materials. No personal or sensitive information was involved, and all sources were used strictly for research purposes in accordance with fair use and open-data principles.

To ensure compliance and academic integrity, we carefully verified that all materials were publicly available without access restrictions or copyright notices prohibiting research use. We acknowledge that automated data collection may raise ethical concerns regarding data ownership and consent. Therefore, we have restricted the use of the resulting dataset to non-commercial, academic research only, and we will release it under a research license after publication. The models trained on this dataset are intended solely for research purposes and should not be deployed in real-world decision-making contexts without human oversight.

References

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and et al. 2025. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and et al. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI conference on artificial intelligence*, 34(05):7432–7439.

David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David Ross, and John Canny. 2023. Ic3: Image captioning by committee consensus. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8975–9003.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025a. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. 2025b. Learning to reason for factuality. *arXiv preprint arXiv:2508.05618*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2963–2975.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Sergey Levine. 2018. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and et al. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, and et al. 2023. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Yan Ma, Steffi Chern, Xuyang Shen, Yiran Zhong, and Pengfei Liu. 2025. Rethinking rl scaling for vision language models: A transparent, from-scratch framework and comprehensive evaluation scheme. *arXiv preprint arXiv:2504.02587*.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. 2025. Training on the benchmark is not all you need. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

D Paperno, G Kruszewski, A Lazaridou, QN Pham, Raffaella Bernardi, S Pezzelle, M Baroni, G Boleda, and R Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016-Long Papers*, 3:1525–1534.

Team Qwen. 2024. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*.

Team Qwen. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Science Custom Publishing. 2024. Beyond boundaries: Geogpt's evolution in geosciences. *Science*. Sponsored letter.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. 2025. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. *arXiv preprint arXiv:2506.12307*.

A GeoMC-10K

A.1 Data Sources

GeoMC-10K draws mainly on six publicly available sources detailed below:

- **NYSED Earth Science Exams.** We obtain the PDF archives of New York State Education Department (NYSED) Earth Science Regents exams, which is administered three times annually from June 1998 to January 2025. Each exam contains 50-75 multiple-choice questions covering physical and environmental geology. We parse these PDFs using the Mineru project, which extracts text, images, and metadata with high fidelity.

- **China's Gaokao Geography Papers.** We collect the geography sections of China's National College Entrance Examination (Gaokao) from publicly available repositories. Some geography papers are embedded within the comprehensive liberal arts exam, alongside history and politics, and we employ GPT to extract the geography portion. These tests are renowned for their rigor and breadth, placing particular focus on evaluating comprehension of both natural and human geography components. Since the source data is presented in Chinese, we use GPT for translation after processing.

- **AP Earth Science Exams.** We also collet AP (Advanced Placement) examinations, which are exams offered in the US by the College Board and are taken each May by students. We collect and clean 1,395 multiple choice questions about geology, geography, and environmental science.

- **Engineering Geology MCQs.** We collect about 3,000 professional-level engineering geology questions from Sanfoundry. These items span subjects including soil mechanics, rock classification, groundwater flow, and metamorphic processes, making it a challenge for LLMs. Both this dataset and the previous one were obtained by parsing web pages.

- **ASBOG Exam Outline Questions.** We extract the sample multiple-choice questions provided at the end of the official ASBOG exam specifications (Association of State Boards of Geology). Although only a few dozen items are included, they cover key professional topics and are representative under relevant topics.

- **GCI.** We also curate the Geoscience Concept Inventory (GCI), a research-validated assessment tool designed to probe fundamental misconceptions in geoscience education. This set comprises 56 multiple-choice questions focusing on core concepts in physical and environmental geology. We manually extract both this dataset and the previous one since the number of items is small.

We provide 2,000 test cases in the supplementary material. The training and validation sets will be released upon acceptance of the paper.

A.2 Data Leakage

Since all data sources used in our benchmark are publicly available, a natural concern arises: could parts of the dataset have already been seen during the pre-training of foundation models? To address this, we follow the data contamination detection procedure proposed in (Ni et al., 2025) and measure the potential overlap between our benchmark and the pre-training corpora of LLaMA-8B and Qwen-8B. The resulting contamination scores are 0.22 and 0.24, respectively. Compared with the datasets examined in the reference study (Figure 3), these values indicate a relatively low level of potential data leakage, suggesting that GeoMC-10K serves as a fair and qualified benchmark.

It should also be noted that this detection method only evaluates potential overlaps in the question stems, without considering whether the answers appear in the training corpora. Given that humans—and by extension, large models—cannot fully solve a question by merely memorizing its

stem, the actual impact of such partial contamination is likely to be even smaller.

B Factuality-aware Geoscience Reasoning

B.1 GRPO Implementation Details

GRPO operates by sampling a batch of prompts $q \sim P(Q)$, and for each prompt, generates a group of candidate completions $\{o_i\}_{i=1}^{G} \sim \pi_{\text{old}}(O|q)$. Each output token sequence o_i receives a scalar reward r_i, and the groupwise-normalized advantage for the i-th sample is computed as:

$$\hat{A}_i = \frac{r_i - \mu_r}{\sigma_r},$$

where μ_r and σ_r denote the mean and standard deviation of rewards within the group.

The optimization objective is defined as:

$$\mathcal{L} = -\sum_{i=1}^{n} \hat{A}_i \cdot \log \pi_\theta(y_i \mid x) + \beta \cdot \text{KL}(\pi_\theta \parallel \pi_{\text{ref}})$$

This objective encourages the model to assign a higher probability to relatively better-performing samples within each group while penalizing divergence from the reference policy π_{ref} via a KL term.

B.2 Human Consistency of Extracted Facts

Since the factual reward is constructed from automatically extracted gold facts, a natural concern is whether potential noise in the extracted claims could bias the reinforcement learning process. To assess the reliability of the extracted gold facts, we conduct a targeted human consistency analysis.

We randomly sample 100 records from the question set and perform a manual inspection. For each sample, the we check whether the extracted claims contain

1. redundant claims (i.e., claims that are factually correct but unnecessary or duplicated)
2. missing claims (i.e., essential factual statements that should have been extracted but are absent)

We quantify the quality of extracted claims using a Jaccard-style consistency score:

$$\text{Consistency} = \frac{|\mathcal{C}| - |\mathcal{C}_{\text{red}}|}{|\mathcal{C}| + |\mathcal{C}_{\text{miss}}|},$$

where $\mathcal{C}$ denotes the set of extracted claims, $\mathcal{C}_{\text{red}}$ denotes redundant claims, and $\mathcal{C}_{\text{miss}}$ denotes missing claims identified by human inspection. Across the 100 sampled questions, the average consistency score reaches **0.898**. Notably, we do not observe any missing-claim cases in the inspected samples, indicating that the extraction process has high recall with respect to core factual content.

These results suggest that the automatically extracted gold facts exhibit strong consistency with human judgment. Combined with the substantial performance gains brought by the factual reward in our main experiments, this analysis supports that the proposed factual reward is effective and robust, even when derived from automatically constructed supervision signals.

B.3 Evaluation on Standard Benchmarks

Although our focus is geoscience reasoning, domain-specific fine-tuning may cause catastrophic forgetting and harm generalization (Kirkpatrick et al., 2017; Luo et al., 2023). Thus, we evaluate whether our RL-based adaptation maintains general language capabilities.

To this end, we evaluate both the base and fine-tuned LLaMA and Qwen models on a diverse set of standard benchmarks: MMLU (Hendrycks et al., 2020), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), BoolQ (Clark et al., 2019), GSM8K (Cobbe et al., 2021) and LAMBADA (Paperno et al., 2016). As shown in Table 6, the fine-tuned variants maintain accuracy on par with or slightly above their respective baselines, and perplexity on LAMBADA remains largely unchanged. These results confirm that our RL-based geoscience adaptation preserves the model's broad capabilities, ensuring robust performance across both specialized and general tasks.

	Acc↑					Ppl↓
Model	MMLU	HS	PIQA	BQ	G8	LMD
llama-8B	.683	.797	.822	.872	.782	4.486
GR1-llama	.681	.796	.820	.871	.770	4.410
qwen-8B	.749	.760	.793	.875	.873	5.371
GR1-qwen	.750	.778	.804	.877	.870	5.879

Table 6: General-domain evaluation results across standard benchmarks. Acc↑: accuracy (higher is better); Ppl↓: perplexity (lower is better). HS = HellaSwag, BQ = BoolQ, G8 = GSM8K, LMD = LAMBADA.

C GeoM2T

C.1 GeoM2T Implementation Details

GeoM2T consists of three cooperative agents—an Extractor, a Critic, and an Advisor—that work together to convert multimodal questions into text-only form. **Initial Extraction.** Given a multimodal problem with one or more images, the Extractor invokes a pretrained multimodal LLM (e.g., GPT-4V) with a prompt template designed to elicit all visible elements. The output is a first-pass plain-text question draft, including stem and choices, in which visual details are rendered as natural language.

Critique Assessment. The Critic then evaluates this draft along two dimensions: whether it includes all information needed to solve the question, and whether the Critic can answer it correctly. These judgments fall into three cases:

- **Incomplete Draft.** If key visual elements are missing, the draft is flagged as "incomplete".
- **Complete but Incorrect.** If the draft is judged to be complete yet the Critic's predicted answer diverges from the ground truth, the draft is flagged as "incorrect".
- **Complete and Correct.** If all key details are present and the Critic predicts the correct answer, the draft is flagged as "adequate".

Advisor-Driven Correction. Depending on the Critic's verdict, the Advisor Agent applies one of three targeted feedback routines:

- **Information Supplement Routine** (for Incomplete Drafts): the Advisor compares the original multimodal question with the draft and enumerates omitted facts or relationships that must be integrated.
- **Failure Analysis Routine** (for Complete but Incorrect Drafts): the Advisor inspects the original multimodal question, the proposed draft alongside the Critic's incorrect prediction and its rationale, and the correct answer, identifying any (a) misinterpretation of visual signals, (b) ambiguous or imprecise expressions, and (c) critical facts that should be emphasized to guide towards the correct answer.
- **Difficulty Maintenance Routine** (for Adequate Drafts): the Advisor checks for (a) unintentional answer hints that lower difficulty, (b) missing distractive information that originally in the images, and (c) judge whether the overall difficulty of the draft is equivalent to the source.

Each routine produces structured natural-language suggestions, which are injected back into the Extractor's prompt using predefined templates. The Extractor ingests the Advisor's suggestions and generates a refined draft. This new draft is re-evaluated by the Critic, and if it still fails any checks, the loop repeats. Termination occurs when the Advisor confirms that the draft is fully difficulty matched, yielding the final pure-text question.

C.2 GeoM2T Evaluation Baselines

We compare our GeoM2T pipeline against five baseline methods for generating text-only question conversions. Unless otherwise specified, all API calls use the GPT o4-mini-high model with a temperature parameter of 0.0. For GeoM2T, we perform 5 rounds of iterative optimization.

1. **Zero-shot prompting.** The model is prompted to convert a multimodal question into text, with explicit instructions to extract visual cues while preserving both answerability and difficulty. This prompt is also used as the initialization step in GeoM2T.
2. **Few-shot prompting.** On top of the zero-shot setup, five curated examples of high-quality conversions are provided. These examples are manually selected to preserve visual information and maintain question difficulty.
3. **Zero-shot with answer.** The zero-shot prompt is augmented with the ground-truth answer, offering an upper-bound setting where answer-aware extraction can better retain relevant visual content.

Among conventional tasks, image captioning is the most closely related to ours, as it aims to generate textual descriptions of visual content. Accordingly, we include two image captioning methods as comparative baselines:

1. **PromptCap.** PromptCap (Hu et al., 2023) is a fine-tuned image captioning model that generates descriptive captions from images. Unlike API-based methods, PromptCap is lightweight and locally deployable, representing small vision-language models (VLMs).

2. **IC3.** IC3 (Chan et al., 2023) generates candidate captions with different small VLMs and temperatures, and then merge them with a LLM. We adapt it to our task by generating four candidates with GPT-4.1 and o4-mini-high under zero-shot prompts at temperatures 0.0 and 1.0, and then consolidate them into a single, text-only problem statement.

D LLM as a Judge for the GeOpen Benchmark

For GeOpen, we evaluate models at the level of the final answer span. Concretely, we first extract the text inside the `<answer> ... </answer>` tag from each model output and treat it as the predicted answer. This answer is then paired with the original question and passed to GPT-4.1-mini, which acts as an automatic grader. The grader assigns one of three scores: 1.0 for fully correct answers that cover all key points without major factual errors, 0.5 for partially correct answers that capture some key ideas but are incomplete or contain minor inaccuracies, and 0.0 for incorrect or irrelevant answers.

The final GeOpen score for a model is computed as the average of these per-question scores. For outputs that do not strictly follow the required format, we heuristically recover the final answer span and multiply the resulting score by a discount factor of 0.75 to penalize format violations.