Geo-Expert: Towards Expert-Level Geological Reasoning via Parameter-Efficient Fine-Tuning

Chenyou Guo [*1] **Zongqi Liu** [*1] **Yizhou Zhang** [*1] **Zhaorui Jiang** [*2 3] **Ze Liu** [1]

Abstract

While general-purpose Large Language Models (LLMs) applied to Geology often hallucinate when reasoning about subsurface structures and deep-time evolution, current AI in Earth sciences predominantly targets surface remote sensing and GIS. To bridge this gap, we introduce Geo-Expert, a family of parameter-efficient geological LLMs fine-tuned on a custom-curated, high-quality instruction dataset processed with a reproducible synthesis workflow and geology-specific prompts and filters. We investigate the impact of model scaling and architecture by fine-tuning three base models: Qwen3-8B, Qwen3-32B, and Gemma-3-27B, with Low-Rank Adaptation (LoRA) method. Our extensive evaluation on a novel domain-specific benchmark, Geo-Eval, reveals that a domain-aligned 8B model can outperform open-weight 70B generalists and proprietary GPT-4o on specialized geological reasoning, while a 32B variant approaches frontier reasoning models. The optimized 8B model further offers a favorable parameter-efficiency profile for deployment. This work provides a reproducible recipe for democratizing scientific LLMs and establishes a baseline for geological artificial intelligence.

1. Introduction

Geology plays a vital role in underground resource exploration, geologic hazard mitigation, and understanding planetary evolution. Unlike general knowledge domains, geological reasoning requires interpreting complex spatio-temporal relationships and a large amount of professional data. However, LLMs often struggle with the specialized logic required in geology. General-purpose models frequently exhibit severe hallucinations when asked to perform a rigorous deduction on stratigraphy or structural geology, limiting their utility in geological research and industrial applications.

Current large models in the vertical field of geography predominantly focus on Geography Knowledge Agent and Earth Observation, demonstrating proficiency in general geospatial intelligence and utilizing computer vision to analyze imagery for surface monitoring. They effectively handle tasks related to surface geography, environmental monitoring, and human-geography interactions. However, a critical gap remains between general *Geography* and specialized *Geology*. While existing models can retrieve facts about Earth's surface, they often lack the deep-domain reasoning required to understand the solid Earth such as interpreting subsurface stratigraphic sequences, reconstructing deep-time tectonic evolution, or analyzing complex petrogenetic processes.

To bridge this gap, we introduce Geo-Expert, a suite of parameter-efficient geological LLMs. We construct a high-quality, domain-specific instruction dataset derived from five canonical geology textbooks. Using a document-to-instruction synthesis workflow, we synthesize structured instruction-response pairs with geology-specific prompts and filters to instill rigorous domain knowledge. We investigate model scaling and architecture effects by fine-tuning three base models: Qwen3-8B, Qwen3-32B, and Gemma-3-27B, employing Low-Rank Adaptation (LoRA) to improve training efficiency.

In sum, our main contributions are as follows:

- **Dataset Construction:** We present a reproducible pipeline for converting unstructured geological textbooks into high-quality instruction-tuning datasets.
- **Model Scaling Analysis:** We provide a comparative analysis of model architectures (Qwen vs. Gemma) and parameter scales (8B vs. 27B/32B), showing that smaller models can be efficient deployment candidates while larger models remain helpful for complex structural reasoning.
- **Geo-Expert Benchmark:** We establish Geo-Eval, a benchmark for geological reasoning, demonstrating

[1]Ocean University of China [2]Peking University [3]Monash University. Correspondence to: Zhaorui Jiang <zrjiang25@stu.pku.edu.cn>, Ze Liu <liuze@ouc.edu.cn>.

Accepted at the AI for Science workshop (ICML 2026). Copyright 2026 by the author(s).

that our fine-tuned models substantially outperform general-purpose baselines in specialized tasks.

2. Model Framework and Benchmark Construction

2.1. Data Sources

Geological reasoning is fundamentally grounded in established physical principles and deep-time evolutionary models, necessitating data sources of the highest scientific rigor. To construct a high-fidelity instruction dataset, we identified five canonical textbooks as the data source. These sources were strategically selected to cover the full spectrum of solid earth science, balancing theoretical depth, cognitive difficulty, and practical application:

- *Principles of Sequence Stratigraphy I & II* (Catuneanu) (Catuneanu, 2006) — This text serves as the standard for understanding temporal frameworks in sedimentation and basin analysis. It is renowned for its highly abstract conceptual models, challenging the model's deep-time evolutionary reasoning.
- *Structural Geology* & *Advanced Structural Geology* (Fossen, 2016) — These volumes provide the theoretical foundation for rock mechanics, stress analysis, and rheology. They feature a steep learning curve with complex mathematical derivations, establishing a rigorous physical baseline.
- *Tectonics and Sedimentation: Implications for Petroleum Systems* (Gao, 2012) — A critical resource bridging the gap between plate tectonics, basin formation, and hydrocarbon exploration. It provides complex, interdisciplinary macro-scale scenarios.
- *Structural Analysis and Synthesis* (Rowland et al., 2020) — It focuses on practical application, including 3D geometric interpretation and structural mapping techniques. This volume strictly evaluates the model's spatial imagination and geometric deduction capabilities.

2.2. Dataset Construction

Transforming static textbook content into a dynamic instruction-tuning dataset involves a pipeline designed to maximize data fidelity and reduce noise. We implemented a multi-stage process: digitization, heuristic cleaning, and instruction synthesis.

Text Extraction and Sanitization. We first used MinerU (OpenDataLab, 2024), a high-precision PDF-to-Markdown tool, to extract textual content while preserving the structure of the original textbooks. Following the extraction, we developed a custom Python-based cleaning module to process the raw Markdown. This pipeline first segments the text into discrete blocks based on paragraph lines containing metadata keywords like page numbers and copy notices. Furthermore, to prevent redundancy caused by layout overlaps, we implemented a global deduplication mechanism using a hash-based method to filter out identical text blocks across the entire corpus.

Instruction Synthesis. To transform the cleaned Markdown segments into a high-quality, geoscientific instruction-tuning corpus, we implement a lightweight workflow for document chunking, question generation, answer generation, and dataset export. Our domain-specific contribution is the geology source curation, prompt design, filtering rules, and downstream training/evaluation protocol. The synthesis process consists of three key stages:

1. **Chapter-Aware Recursive Chunking:** Unlike simple fixed-length splitting, the framework utilizes a semantic segmentation algorithm based on the Markdown header structure. This ensures that each text chunk retains complete semantic context within its hierarchical scope, preventing the fragmentation of complex geological descriptions.

2. **Domain-Structured Question Generation:** An LLM first constructs a hierarchical domain tree (extracting outlines and concepts) to bind specific tags to text segments. Questions are then dynamically generated based on these domain tags and character density, minimizing redundancy and ensuring broad coverage of geological concepts.

3. **Chain-of-Thought (CoT) Answer Construction:** To capture the rigorous logic required for geological reasoning, the system employs reasoning-oriented teacher models, primarily DeepSeek-R1 with source-grounded prompts, to generate candidate answers. Each generated item is constrained to cite or paraphrase the provided chunk rather than external memory, and malformed or source-inconsistent outputs are filtered before training. This process produces not only the final answer but also intermediate reasoning traces, improving supervision for multi-step deduction while keeping the answer aligned with the source material.

This automated pipeline enabled us to efficiently scale the dataset to 11,518 high-quality instruction pairs while preserving the authoritative accuracy of the original canonical sources.

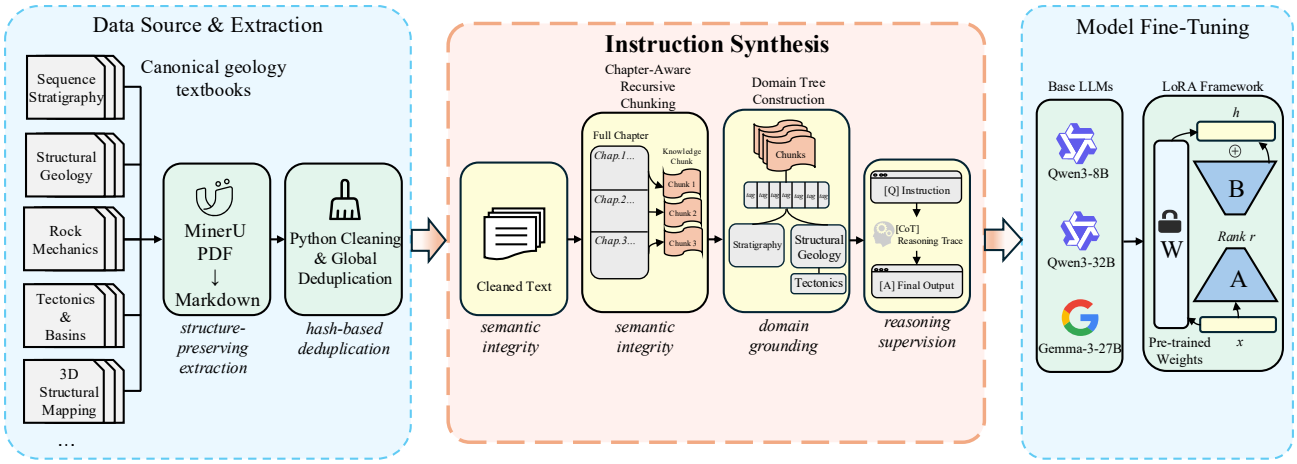

Figure 1. **Overview of the Geo-Expert framework.** The pipeline consists of three main stages: (1) extracting and sanitizing text from canonical geology textbooks; (2) synthesizing high-quality, CoT-enhanced instruction pairs via a domain-structured generation pipeline; (3) applying parameter-efficient fine-tuning on base LLMs.

2.3. Model Fine-Tuning

We performed instruction tuning using the ms-swift framework (ModelScope, 2024), applying distinct hyperparameter strategies tailored to the scale of each model. We targeted all linear layers (all-linear) for Low-Rank Adaptation (LoRA) (Hu et al., 2021).

For the compact Qwen3-8B model (Yang et al., 2024), we adopted a standard efficiency configuration with LoRA rank r=32, a=32, and a learning rate of 2e-5 using FP16 precision. This lightweight setup, with gradient checkpointing enabled, allowed for rapid iteration on a single NVIDIA RTX 5090.

In contrast, to effectively adapt the deeper architectures of Gemma-3-27B (Team, 2024) and Qwen3-32B, we scaled up the adaptation capacity, increasing the LoRA rank to r=64 and alpha to a=128. To mitigate the high memory footprint and ensure numerical stability during training, we employed BF16 (bfloat16) precision and activated gradient checkpointing with a gradient accumulation step of 4. This optimized configuration enabled the fine-tuning of these 27B+ parameter models on a 4×RTX 5090 setup, demonstrating a feasible recipe for fine-tuning 27B+ models on consumer-grade GPUs.

2.4. Benchmark Construction: Geo-Eval

Standard benchmarks in Earth sciences predominantly focus on remote sensing tasks or general geography trivia, failing to capture the rigorous deductive logic required for solid earth geology. To bridge this evaluation gap and comprehensively evaluate the true reasoning capabilities of LLMs in the vertical domain of geology, we constructed Geo-Eval, a novel domain-specific benchmark consisting of high-fidelity, open-ended reasoning questions. Unlike traditional static datasets manually extracted from exams, Geo-Eval was constructed through an automated adversarial mining pipeline coupled with expert human-in-the-loop verification. The construction process proceeded in four distinct phases:

1. **Large-Scale Generative Extraction:** We initially deployed DeepSeek-R1 (DeepSeek-AI et al., 2025) to perform a comprehensive semantic analysis of the source geology textbooks. Through this process, we extracted a foundational pool of 2,591 complex geological questions and synthesized their corresponding preliminary reference answers.

2. **Comparative Inference:** To evaluate the baseline domain capability, both our fine-tuned Qwen3-8B-Geo and DeepSeek-R1 were tasked with generating independent responses to the entire pool of 2,591 questions.

3. **Difficulty-Aware Evaluation Paradigm (LLM-as-a-Judge):** To move beyond static, undifferentiated testing, we introduce a *difficulty-aware evaluation paradigm* designed to identify the true cognitive frontier of geological reasoning. We utilized GLM-4.5 (Zeng et al., 2024) as an impartial evaluator to blindly score both models' responses on a 10-point scale. Rather than evaluating on a random distribution, we systematically isolated a subset of 387 "hard boundary" questions where the score differential between our Qwen3-8B-geo and the frontier generalist DeepSeek-R1 was marginal (score difference ≤ 4). This dynamic calibration of difficulty yields boundary samples where a strong generalist no longer has a decisive margin over the domain-aligned model. Consequently, this design

shifts the focus from basic factual recall to a stricter, boundary-aware stress test of expert-level geological reasoning.

4. **Expert Vetting and Ground Truth Finalization:** To improve scientific validity, the isolated 387 candidate questions were submitted to geology professors for manual screening, refinement, and calibration. The final test set uses expert-written reference answers rather than model-generated answers.

The resulting finalized dataset serves as a highly targeted, expert-verified ground truth. By deliberately focusing on these high-complexity boundary problems, Geo-Eval provides a robust metric for assessing whether a model has truly internalized deep geological logic, moving far beyond superficial pattern matching.

2.5. Selection Bias and Benchmark Scope

We acknowledge that the boundary mining process uses Qwen3-8B-Geo as one comparator, which can introduce selection bias because the initial candidate pool is partly shaped by this model's strengths and weaknesses. Our mitigations reduce, but do not eliminate, this coupling: (i) the 387 candidate questions were independently re-vetted, refined, and, where necessary, rewritten by geology professors, and the reference answers used for scoring are expert-authored ground truth, not model outputs; (ii) the final evaluation employs GPT-4o as a reference-guided judge against this expert-verified ground truth, decoupling scoring from the mining process; (iii) two of our three reported models (Qwen3-32B-Geo and Gemma-3-27B-Geo) were not used in the boundary-mining loop, yet exhibit consistent gains. We therefore interpret Geo-Eval as a hard, boundary-aware stress test rather than a fully independent estimate of performance over the entire geology question distribution. Constructing a larger benchmark from an entirely model-independent sampling process remains important future work.

3. Benchmark and Evaluation: Geo-Eval

3.1. Benchmark Content

To avoid one-dimensional evaluation, we rigorously partition the Geo-Eval test dataset into three cognitive levels:

- **Concept:** Focuses on the identification and definition of core terminology, such as mineralogy, petrology, and stratigraphy.
- **Process:** Focuses on the logical deduction of dynamic geological mechanisms, including tectonic movements and diagenesis.
- **Engineering:** Focuses on the comprehensive application analysis of geological conditions in practical scenarios, such as engineering site selection and geological hazard prevention.

Because Geo-Eval contains 387 expert-vetted questions, it is intended for targeted stress testing rather than exhaustive subdiscipline coverage. The three-way partition provides a coarse view of reasoning type, but the current size limits finer-grained claims within each category.

3.2. Benchmark Baselines

After applying Geo-Eval to different models, we adopt the widely recognized LLM-as-a-Judge paradigm, utilizing GPT-4o as a reference-guided LLM evaluator comparing against expert-verified ground truth. We impose strict domain-fact-checking constraints within the prompt to generate a quantitative score ranging from 0 to 10. We select 11 highly representative models for a comprehensive comparison, categorized into three groups: proprietary state-of-the-art models (e.g., GPT-5.4, DeepSeek-V3.2), open-weight generalist models of varying parameter scales (e.g., Llama-3.1-70B, Qwen3-32B), and our domain-specific models fine-tuned via the proposed Geo-Expert framework. We acknowledge that LLM-as-a-Judge evaluation can introduce evaluator-specific bias. The reference-guided prompt is designed to reduce free-form preference effects by anchoring scores to expert answers, but it is not a substitute for a full human reliability study. A multi-judge or human-adjudicated subset is an important next step for strengthening Geo-Eval.

3.3. Main Results and Insights

The experimental results demonstrate the advantages of the Geo-Expert framework on the boundary-aware Geo-Eval setting. We analyze the results from three core dimensions:

1. **Remarkable Performance Breakthrough** As shown in Table 1, general open-weight base models struggle with high-depth geological questions. However, after fine-tuning with the Geo-Expert framework, all models achieve a substantial improvement in performance. For instance, the average score of Qwen3-32B rises from 5.00 to 6.82, and Gemma-3-27B rises from 5.16 to 6.59. Notably, in the "Engineering" dimension, which most directly tests applied deductive reasoning, the score of Qwen3-8B rises from 4.41 to 6.44 after fine-tuning. This suggests that, in specialized STEM branches, relying solely on general knowledge injected during pre-training can be insufficient. Furthermore, a paired t-test on the 387 samples shows that the performance gains of the Geo-Expert models are statistically significant within this benchmark ($p = 3.7121 \times 10^{-106} < 0.001$), although the non-

Table 1. Performance comparison of various models on the Geo-Eval Benchmark. The best results in each column are highlighted in bold, and the second-best are underlined. The Δ column indicates the absolute performance gain achieved by our Geo-Expert fine-tuning framework compared to the respective base model. *** indicates statistical significance ($p = 3.7121 \times 10^{-106} < 0.001$) compared to the corresponding base models via paired t-test; this statistic should be interpreted within the boundary-mined benchmark rather than as an i.i.d. estimate over all geology questions.

Model	Size	Concept	Process	Engineering	Average	Δ
Proprietary SOTA						
GPT-5.4 (Frontier)	-	**7.35**	**7.10**	**7.00**	**7.15**	-
DeepSeek-V3.2-Instruct	-	<u>6.80</u>	6.75	6.67	6.74	-
GPT-4o	-	6.10	5.90	5.80	5.93	-
Open-weight Generalists						
Gemma-3-27B-IT	27B	5.30	5.10	5.08	5.16	-
Qwen3-32B	32B	5.20	4.90	4.90	5.00	-
Qwen3-8B	8B	4.80	4.68	4.41	4.63	-
GLM-4-9B	9B	4.70	4.50	4.51	4.57	-
Llama-3.1-70B-Instruct	70B	4.30	4.10	3.96	4.12	-
Geo-Expert (Ours)						
Qwen3-32B-geo	32B	6.78	<u>6.79</u>	<u>6.90</u>	<u>6.82</u>***	+1.82
Gemma-3-27B-geo	27B	6.70	6.60	6.47	6.59***	+1.43
Qwen3-8B-geo	8B	6.10	6.27	6.44	6.27***	+1.64

random boundary-mined construction means that the statistic should be interpreted as evidence of consistent per-question gains on Geo-Eval rather than as a population-level estimate.

2. **Extreme Parameter Efficiency** One of the most striking findings of this study is the parameter efficiency exhibited by the fine-tuned models. The Qwen3-8B-geo model, with 8B parameters, outperforms Llama-3.1-70B-Instruct across all three dimensions (overall score 6.27 vs. 4.12), and also exceeds the proprietary GPT-4o baseline (5.93) on this benchmark. This result suggests that domain alignment can be more valuable than parameter count in narrow expert settings. Conversely, a compact model with refined domain supervision may be a practical alternative to larger generalist models in resource-constrained geological scenarios.

3. **Competitive Frontier Comparison** Our best-performing model, Qwen3-32B-geo, achieves a score of 6.82, exhibiting balance across the Concept (6.78), Process (6.79), and Engineering (6.90) dimensions. This performance exceeds GPT-4o in our evaluation and remains below the strongest proprietary baseline, GPT-5.4 (7.15), by 0.33 points on average. Considering its smaller parameter scale and local deployability, Geo-Expert models provide a practical cost-performance tradeoff for academic and engineering use.

Table 2. Training and hardware-efficiency summary for Geo-Expert models. Runtime and memory are taken from the ms-swift training logs. Score/B is the Geo-Eval average divided by total model parameters in billions.

Model	Train.	GB	Hours	Score/B
Qwen3-8B-geo	87M	21.8	2.23	0.757
Gemma-3-27B-geo	454M	87.5	4.97	0.236
Qwen3-32B-geo	537M	106.9	6.03	0.205

3.4. Hardware Efficiency

To make the deployment claim more concrete, Table 2 summarizes training efficiency from our logs and a simple score-per-parameter proxy. The 8B model obtains the highest Geo-Eval score per billion parameters and can be trained on a single consumer GPU. We did not conduct a controlled inference-latency benchmark across all proprietary and open models because several baselines were accessed through APIs, but the local hardware profile indicates why the 8B model is the most practical deployment target.

3.5. Qualitative Analysis

To intuitively demonstrate the effectiveness of the Geo-Expert framework in mitigating catastrophic domain-shifting errors, we present a qualitative comparison between GPT-4o and our fine-tuned model (Qwen3-8B-geo) on a highly specialized structural geology question, as shown in Table 3.

Table 3. Qualitative comparison of model outputs. The base model suffers from cross-domain hallucination, whereas Geo-Expert accurately anchors to geology.

Prompt:
What phenomena does the local thickening of a wedge cause, and how is this deformation accommodated?

GPT-4o: **[Score: 0/10]**
Cross-domain Hallucination:
...Local thickening may alter the structural stress distribution, leading to **concrete cracking** or gear fatigue. To adjust this deformation, it is recommended to use **carbon fiber cloth for reinforcement**...

Geo-Expert: Qwen3-8B-geo **[Score: 9/10]**
Accurate Geological Reasoning:
...Local thickening of a wedge leads to **internal stress concentration** and an increase in thrust... This deformation is primarily accommodated through **thrust fault sliding, extensional collapse, and lateral material flow**...

As illustrated in Table 3, the base model, lacking deep geological corpus alignment, exhibits a severe cross-domain hallucination. It incorrectly interprets the polysemous term "wedge" as a mechanical or civil engineering component, generating completely irrelevant advice such as using carbon fiber cloth for concrete reinforcement. This literal interpretation leads to a complete failure in answering the geological prompt.

In stark contrast, the Geo-Expert fine-tuned model successfully anchors the semantic context to structural geology. It accurately explains the mechanical responses (stress concentration, fault reactivation) and correctly identifies the complex tectonic accommodation mechanisms (thrust sliding, lateral material flow). This striking contrast (0 vs. 9) highlights a critical vulnerability of generalist LLMs: when encountering polysemous terms across disciplines, they tend to default to high-frequency general concepts. The Geo-Expert framework effectively recalibrates the model's prior distribution, ensuring precise semantic anchoring and rigorous factual reasoning in the target scientific domain.

4. Related Work

4.1. Large Language Models in Earth Sciences

The application of AI in Earth sciences has grown rapidly, spanning medium-range global weather forecasting (Lam et al., 2023) and scientific language modeling (Taylor et al., 2022). Within geoscience specifically, two complementary lines of work have emerged. The first builds *foundation models* via continued pre-training on broad geoscience corpora: K2 (Deng et al., 2024) adapts LLaMA on geoscience literature and pairs it with a curated knowledge graph, while GeoGalactica (Lin et al., 2024) extends Galactica with over 65B geoscience-specific tokens. These models improve factual recall across general geoscience topics but do not explicitly target multi-step deductive reasoning. The second line develops *geospatial agents and benchmarks*: GeoGPT (Zhang et al., 2023) demonstrates autonomous spatial query execution and facility siting from natural language, and UnivEARTH (Kao et al., 2025) evaluates LLM agents on Earth observation tasks using NASA satellite data.

Despite this progress, a critical gap persists in *solid-earth* geology. Foundation models like K2 and GeoGalactica are trained predominantly on surface-oriented literature (remote sensing, hydrology, atmospheric science), while agent-based systems operate on 2D geospatial workflows. Neither paradigm addresses the core challenges of structural geology and stratigraphy—cross-cutting relationship inference, polyphase tectonic restoration, and deep-time evolutionary reasoning grounded in abstract physical principles. GEO-EXPERT targets this neglected subdomain, shifting the adaptation objective from factual retrieval over surface phenomena to structured reasoning about subsurface geological processes.

4.2. Domain-Specific Adaptation and PEFT

General-purpose LLMs remain prone to hallucination in specialized STEM domains, motivating a growing body of domain-adapted models in medicine (Li et al., 2023), finance (Wu et al., 2023), and law (Cui et al., 2023). A common limitation of these efforts is their emphasis on *factual recall*—e.g., symptom lookup or terminology matching—rather than multi-step deductive reasoning. Recent work on reasoning-oriented distillation (Mukherjee et al., 2023; Yu et al., 2023) suggests that explicit chain-of-thought supervision is essential for eliciting logical inference, yet this insight has rarely been applied to domain-specific scientific adaptation. Parameter-Efficient Fine-Tuning (PEFT), particularly Low-Rank Adaptation (LoRA) (Hu et al., 2021), offers a practical path to internalize domain knowledge while preserving base-model capabilities. However, its effectiveness is fundamentally bounded by the quality of the supervision signal: tuning on raw domain text teaches the model *what to say* but not *how to reason*. In the geological setting this gap is especially acute, as correct answers often require chaining stratigraphic principles, cross-cutting relationships, and tectonic context across multiple inference steps. GEO-EXPERT addresses this by coupling LoRA with a CoT-enhanced instruction synthesis pipeline, shifting the adaptation objective from surface-form memorization to structured geological reasoning.

4.3. Scientific Benchmarks and LLM-as-a-Judge

Evaluating LLMs in specialized scientific domains poses a significant challenge. Traditional benchmarks often rely on static, multiple-choice questions extracted from stan-

dardized exams (Hendrycks et al., 2021). While useful for assessing factual recall, these static datasets fail to capture the multi-step deductive logic required in real-world scientific research. To overcome the limitations of human evaluation scalability, the "LLM-as-a-Judge" paradigm has been increasingly adopted to assess open-ended, complex reasoning tasks (Zheng et al., 2024). Geo-Eval advances this paradigm by introducing an adversarial mining pipeline. By isolating "hard boundary" samples where general-purpose models fail but domain-aligned models succeed, Geo-Eval provides a much more rigorous and discriminative metric for assessing true expert-level geological reasoning.

5. Conclusion and Future Work

In this work, we present Geo-Expert, a framework that addresses the critical reasoning bottlenecks of generalist LLMs in highly specialized scientific domains. At the heart of our contribution is a robust, scalable dataset construction and instruction synthesis pipeline. Rather than passively fine-tuning, this pipeline achieves deep domain alignment by systematically recalibrating the model's semantic representations. By doing so, Geo-Expert successfully transforms text generation into well-rounded scientific reasoning.

Beyond achieving strong domain performance, this research provides evidence that parameter count alone is not sufficient in narrow expert domains. Our empirical results show that a domain-aligned 8B model can outperform a 70B open-weight generalist on Geo-Eval. This supports the practical value of high-quality data curation and targeted reasoning supervision for vertical scientific adaptation. We do not claim a new fine-tuning algorithm; rather, the contribution is an applied, reproducible recipe and benchmark for solid-earth geological reasoning.

Our work has several limitations. First, the training corpus is built from five canonical textbooks, biased toward structural geology, stratigraphy, and tectonics; mineralogy, geochemistry, and geophysics are under-represented. Second, the Geo-Eval benchmark contains 387 expert-vetted questions, sufficient for boundary-aware stress testing but smaller than general-purpose benchmarks and not large enough for fine-grained subcategory claims. Third, the benchmark construction and LLM-as-a-Judge scoring protocol may still contain selection and evaluator bias despite expert vetting. Fourth, the current comparison does not include direct runs of K2 or GeoGalactica on Geo-Eval, and we have not yet tested whether the same pipeline transfers to a second STEM discipline. Finally, the current framework is text-only and does not yet handle the inherently multimodal nature of geological data (cross-sections, well logs, field photos).

While Geo-Expert establishes a formidable foundation for text-based geological AI, we acknowledge that Earth sciences are inherently multimodal. Future work will extend this parameter-efficient framework by integrating Vision-Language Models (VLMs) to parse stratigraphic profiles and remote sensing imagery. Coupled with Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to dynamically track evolving literature, we aim to evolve Geo-Expert from a static reasoning engine into a dynamic, autonomous agent for geoscience discovery.

Availability and Responsible Release

Code availability: the reproducibility code, including Markdown cleaning, instruction-synthesis prompts, fine-tuning command templates, evaluation scripts, and small self-authored examples, is available at `https://github.com/Kunling-AI/Geo-Expert`. Because the instruction-tuning corpus is derived from copyrighted textbooks, we do not redistribute raw textbook text, the full textbook-derived training corpus, or long generated passages that would reconstruct the sources. Model checkpoints and the full benchmark set are not included in the public code repository at this stage. The released scripts and examples are intended to support independent reimplementation with licensed source material and local model paths.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

We are grateful to the Editors and anonymous reviewers for their constructive reviews and insightful suggestions, which significantly improve the manuscript. This research was financially funded by National Natural Science Foundation of China (No. 42202231, 42121005, 92141005). This work was supported by a grant from the Postdoctoral Applied Research Foundation of Qingdao City. Furthermore, we appreciate the editors and the reviewers for their critical, careful, and constructive reviews. Figures in this paper were prepared by the authors based on the Geo-Expert data-processing, instruction-synthesis, and fine-tuning workflow.

References

Catuneanu, O. *Principles of sequence stratigraphy*. Elsevier, 2006.

Cui, J., Li, Z., Yan, Y., Chen, B., and Yuan, L. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023.

DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Deng, C., Zhang, T., He, Z., Xu, Y., Chen, Q., Shi, Y., Fu, L., Zhang, W., Wang, X., Zhou, C., Zhuang, L., et al. K2: A foundation language model for geoscience knowledge understanding and utilization. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 161–170, 2024.

Fossen, H. *Structural geology*. Cambridge university press, 2016.

Gao, D. (ed.). *Tectonics and Sedimentation: Implications for Petroleum Systems*. AAPG Memoir 100. American Association of Petroleum Geologists, 2012.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Kao, C. H., Zhao, W., Revankar, S., Speas, S., Bhagat, S., Datta, R., Phoo, C. P., Mall, U., Vondrick, C., Bala, K., and Hariharan, B. Towards llm agents for earth observation. *arXiv preprint arXiv:2504.12110*, 2025. Accepted at ICML 2025 Workshop TerraBytes.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wiratunga, P., Ewalds, T., Ferraro, A., Pritzel, A., et al. Learning skillful medium-range global weather forecasting. *Science*, 382 (6677):1416–1421, 2023.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. Chatdoctor: A medical chat model fine-tuned on a large language model meta-llama. *arXiv preprint arXiv:2303.14070*, 2023.

Lin, Z., Ding, C., Yao, H., Hao, Z., Zheng, J., et al. Geogalactica: A scientific large language model in geoscience. *arXiv preprint arXiv:2401.00434*, 2024.

ModelScope. ms-swift: Scalable lightweight infrastructure for fine-tuning. https://github.com/modelscope/swift, 2024.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

OpenDataLab. Mineru: A high-precision pdf to markdown extraction tool. https://github.com/opendatalab/MinerU, 2024.

Rowland, S. M., Duebendorfer, E. M., and Schiefelbein, I. M. *Structural analysis and synthesis: a laboratory course in structural geology*. John Wiley & Sons, 2020.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Elvis, E., Pumperla, M., Toyama, Y., Guillemotte, T., and Jones, L. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Team, G. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Yang, A., Yang, B., Zhang, B., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2024.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Zeng, A., Liu, H., Du, Z., et al. Glm-4: An open language model family. *arXiv preprint arXiv:2406.12793*, 2024.

Zhang, Y., Wei, C., Wu, S., He, Z., and Yu, W. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Hao, S., Wu, Z., Ba, S., Zhuang, Y., Lin, Z., Li, Z., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2024.

A. Implementation Details

In this section, we provide the detailed hyperparameter configurations and hardware setups used for fine-tuning the Geo-Expert model family. All experiments were conducted using the `ms-swift` framework to ensure standardized PEFT across different model architectures.

To balance training efficiency and memory footprint, we customized the LoRA settings based on the base models' parameter scales. For the compact Qwen3-8B model, we utilized a single NVIDIA RTX 5090 GPU with FP16 precision. For the larger architectures (Gemma-3-27B and Qwen3-32B), we scaled the hardware to a 4×RTX 5090 and employed BF16 mixed-precision training. Gradient checkpointing was enabled in all runs. The detailed hyperparameter settings are summarized in Table 4.

Table 4. Hyperparameter configurations for fine-tuning the Geo-Expert model family.

HYPERPARAMETER	8B MODEL	27B/32B MODELS
LORA TARGET MODULES	ALL LINEAR	ALL LINEAR
LORA RANK (r)	32	64
LORA ALPHA (α)	32	128
LORA DROPOUT	0.05	0.05
LEARNING RATE	2E-5	1E-4
OPTIMIZER	ADAMW	ADAMW
PRECISION	FP16	BF16
GRADIENT CHECKPOINTING	TRUE	TRUE
BATCH SIZE	1 (PER DEVICE)	4 (PER DEVICE)
GRADIENT ACCUMULATION	1	4
HARDWARE	1×RTX 5090	4×RTX 5090

B. Prompt Templates

B.1. LLM-as-a-Judge Prompt for Geo-Eval

To conduct a rigorous and impartial evaluation during the hard sample mining phase of Geo-Eval, we utilized GLM-4.5 as the evaluator. We designed a zero-shot prompt template that constrains the judge to focus purely on geological factual accuracy and logical deduction, penalizing superficial or hallucinatory responses. The specific prompt used is as follows:

> *"You are an expert professor in solid earth geology. Your task is to evaluate the responses of two AI models to a complex geological question.*
> ***Question:*** *[Insert Question]*
> ***Model A Response:*** *[Insert Model A Answer]*
> ***Model B Response:*** *[Insert Model B Answer]*
> ***Evaluation Criteria:*** *1. Factual Accuracy (40%): Does the response align strictly with established geological principles (e.g., Andersonian faulting, sequence stratigraphy)? 2. Deductive Logic (40%): Is the chain of thought coherent? 3. Professional Terminology (20%): Are the geological terms used accurately without cross-domain confusion?*
> ***Task:*** *Score each model from 0 to 10. You must provide a brief justification for your score, explicitly pointing out any hallucinations or logical flaws. Output your final scores in the format: Score A: [X], Score B: [Y]."*

B.2. Reference-Based Evaluator Prompt for Final Scoring

To conduct the final quantitative evaluation, we utilized GPT-4o as a reference-guided evaluator. Unlike zero-shot blind testing, this phase employs a strict Reference-Based Evaluation paradigm, where the model compares candidate responses against expert-verified ground truth. The prompt explicitly enforces terminology flexibility to accurately accommodate the nuances of geological nomenclature.

```
You are an authoritative Professor of Geology and an expert academic evaluator.
Your task is to grade a Model's Answer to a complex geological question against
```

```
a provided professional Reference Answer.

[Grading Criteria]
1. Score from 0 to 10.
2. Focus on SEMANTIC CORRECTNESS and GEOLOGICAL ACCURACY, not verbatim matching.
3. Terminology Flexibility: Be lenient with accepted geological synonyms
   (e.g., "igneous" vs "magmatic", "thrust fault" vs "reverse fault"
   depending on context).
4. Completeness: A 10/10 answer must contain all core geological facts mentioned
   in the reference.
5. Penalty: Deduct heavily for hallucinations, factually incorrect statements,
   or mixing up crucial concepts (e.g., misidentifying rock types or eras).

[Output Format]
Output ONLY a JSON object with this exact structure:
{"score": 8, "reason": "brief professional explanation"}

Question: {question}
Reference Answer: {reference}
Model Answer: {model_ans}
```

B.3. Instruction Synthesis Details

The instruction synthesis stage uses DeepSeek-R1 as the primary reasoning-oriented teacher model. For reproducibility, we release a command-line implementation that performs Markdown chunking, source-grounded question generation, CoT answer construction, filtering, and ShareGPT-style export. For each cleaned chunk, the prompt contains the local Markdown content, chapter path, domain tags from the domain tree, and an explicit requirement that the answer remain grounded in the provided source chunk. We use temperature 0.6 and a maximum generation length of 2048 tokens for candidate generation. Post-processing removes invalid chat schemas, duplicate questions, empty answers, and responses that fail simple source-grounding checks such as introducing unrelated disciplines or unsupported entities. The resulting records are stored in a chat format with `user` and `assistant` messages. When a reasoning trace is generated, it is wrapped in `<think>` tags before the final answer.

B.4. Instruction-Tuning Data Example

As described in Section 2.2, our instruction synthesis pipeline generates high-quality, reasoning-enhanced QA pairs. Below is a translated JSON snippet extracted from our training corpus, demonstrating the required `<think>` trajectory before final generation.

```
[
  {
    "messages": [
      {
        "role": "user",
        "content": "How is the method of measuring vorticity using rotated
                    porphyroblasts applied to the study of rocks near the
                    Alpine Fault?"
      },
      {
        "role": "assistant",
        "content": "<think>\n1. Identify core concepts: vorticity measurement,
                    rotated porphyroblasts, Alpine Fault.\n2. Recall geological
                    context: The Alpine Fault is a major plate boundary fault
```

```
                    with intense ductile deformation.\n3. Formulate explanation:
                    Porphyroblasts record kinematic history during shear
                    deformation...\n</think>\n\nThe method of measuring
                    vorticity using rotated porphyroblasts is a sensitive
                    kinemetric technique used to study the behavior of rocks
                    during shear deformation. By analyzing the rotational
                    dynamics of these crystals within the ductile shear zones..."
      }
    ]
  }
]
```


B.5. Geo-Eval Hard Boundary Sample Analysis

In Section 2.4, we detailed the automated mining of "hard boundary" samples where general-purpose models fail but domain-expert models succeed. The following real evaluation cases illustrate severe cross-domain hallucinations by the base model and precise domain anchoring by Geo-Expert.

Case 1: Semantic Disambiguation of Polyphase Deformation
Question: What is polyphase deformation?
Ground Truth: In structural geology, Polyphase Deformation (or Superimposed Deformation) refers to the complex process where the same geological body undergoes two or more distinct regional tectonic stress fields over a long period of geological history...

Base Model Response (Score: 0.0 / 10):
[Exposed Thinking Process]: *Hmm, the user is asking about "polyphase deformation". First, I need to determine the context of this term. It could appear in materials science, engineering, or computer graphics. For instance, in computer graphics, it might involve multiple steps of deformation algorithms...*
[Final Output]: *"Polyphase deformation" is an interdisciplinary concept. In materials science, it refers to different stages of deformation under varying temperatures or stress... In computer graphics, it refers to the combination of multiple animation steps...*

Geo-Expert Response (Score: 9.0 / 10):
[Final Output]: *In structural geology, Polyphase Deformation refers to the complex superimposed deformation process generated by two or more distinct tectonic stress fields over geological time. Each independent deformation event leaves specific structural traces in the rock, such as folds, faults, or foliations. Studying polyphase deformation is of decisive significance for reconstructing paleostress fields and understanding tectonic evolution...*

Case 2: Disciplinary Anchoring of "Deformation" Terminology
Question: What is the academic definition of deformation terminology?
Ground Truth: The academic definition of deformation refers to the process and result of relative positional changes among internal particles of an object under external factors (e.g., stress, temperature), leading to changes in shape, size, orientation, or volume. It is a core concept in structural geology and solid mechanics...

Base Model Response (Score: 0.0 / 10):
[Exposed Thinking Process]: *Hmm, the user is asking about the academic definition of "deformation terminology". This could have different interpretations in linguistics, philosophy, or sociology. For instance, in linguistics, it might refer to inflection or word morphing. Or it could mean a term that is morphed or transformed across contexts...*
[Final Output]: *"Deformation terminology" does not have a universally accepted definition. Its meaning varies by discipline: In linguistics, it may refer to word inflection or semantic shifts. In philosophy, it may refer to conceptual evolution. In sociology, it implies the contextualization of a term across different social groups...*

Geo-Expert Response (Score: 9.0 / 10):
[Final Output]: *The academic definition of deformation terminology refers to the scientific concepts used to describe the*

shape changes of rocks under stress, encompassing major types such as elastic, plastic, and brittle deformation. Elastic deformation refers to reversible shape changes... Furthermore, it includes key concepts like stress, strain, and strain rate, as well as deformation mechanisms (e.g., dislocation creep). These terms collectively form a systematic description of rock deformation processes, providing a theoretical foundation for studying geological structures.