Large Language Models Encode Geoscience Knowledge

Cheng Deng [1] **Zhouhan Lin** [2] **Yi Xu** [2] **Le Zhou** [2] **Jin Zhuo** [2] **Tianhang Zhang** [2] **Shu Wang** [3] **Jiaxin Ding** [2] **Luoyi Fu** [2] **Weinan Zhang** [2] **Junxian He** [4] **Xinbing Wang** [2] **Chenghu Zhou** [3]

[1]*University of Edinburgh* [2]*Shanghai Jiao Tong University* [3]*Institute of Geographical Science and Natural Resources Research, CAS* [4]*The Hong Kong University of Science and Technology*. Correspondence to: Zhouhan Lin lin.zhouhan@gmail.com.

1. Introduction

Large Language Models (LLMs) have transformed natural language processing and are increasingly driving AI for Science (AI4S), demonstrating strong reasoning, question answering, and code generation abilities through large-scale data-driven learning. Extending LLMs to domain-specific sciences offers significant potential for accelerating discovery.

Geoscience, spanning geophysics, meteorology, and environmental science, produces extensive textual resources. Knowledge extraction traditionally relies on domain knowledge systems [1], knowledge graphs [2], and semantic models [3]. Although NLP has been applied to classification [4], topic modeling [5], entity recognition [6, 7], knowledge graph construction [8], question answering [9], and summarization [10], these approaches remain fragmented and task-specific.

General-domain LLMs such as Galactica [11], LLaMA [12], and GLM [13] achieve strong overall performance but lack authoritative geoscience training data. Domain adaptation efforts such as K2 [14] show promising results but are limited by model scale and corpus size. Scaling domain LLMs introduces challenges in data curation, training stability, and systematic multi-source integration.

To address these gaps, we propose a 30B-parameter geoscience LLM optimized for professional academic interaction, demonstrating improved domain reasoning and reduced hallucination compared to its base model. Our contributions are fourfold: (1) we develop GeoGalactica, a domain-specific geoscience foundation model with enhanced factual grounding; (2) we design an academic data preprocessing toolchain for constructing high-quality corpora from large-scale PDF collections; (3) we propose a data-centric training recipe covering continued pre-training, supervised fine-tuning, and deployment; and (4) we release model checkpoints (including 8-bit quantized versions) together with new geoscience benchmark datasets to support open research.

2. Methods

2.1 Data Collection

To address the scarcity of large-scale geoscience corpora, GeoCorpus. For training the model to follow instruction, we curate approximately 6M expert-filtered documents and expand the GeoSignal dataset from K2. Our pipeline emphasizes large-scale aggregation, copyright-compliant parsing, deduplication, and structured normalization.

GeoCorpus contains 5.98M geoscience-related documents collected from MAG, OpenAlex, CommonCrawl, The Pile, arXiv, and related sources. After cleaning and deduplication, the final training corpus contains 78B tokens, balanced across geoscience and auxiliary domains. Special tokens (e.g., `[START_FIGURE]`, `[START_TABLE]`, `[START_REF]`, `[START_FORMULA]`) are introduced to standardize heterogeneous document structures.

We design **GeoSignal** v2 to support knowledge-intensive geoscience tasks, including knowledge graph construction (NER, RE), academic applications (summarization, keyword extraction), QA, education dialogue, and geospatial tasks. The dataset integrates three sources: (1) literature-structured signals (via Grobid citation parsing and reference linking), (2) knowledge-intensive structured websites converted into instruction-response pairs using key-value matching, (3) self-instruct data generated from curated geoscience problem seeds and expert-verified ChatGPT outputs. After filtering and manual verification, we construct 100K high-quality instruction samples for supervised fine-tuning.

2.2 Model Training

Further Pre-training. Following Meta AI's initial pre-training, we further adapt Galactica on *GeoCorpus* to enhance domain-specific geoscientific capabilities. Training is conducted on a ROCm-based accelerator cluster comprising 512 nodes (2048 accelerators in total), using the Megatron-LM [15] framework with 3D parallelism (model, pipeline, and data parallelism). We set model parallel size to 4 and pipeline parallel size to 16 for optimal memory utilization.

All documents are tokenized, concatenated with *EOS* markers, and cropped to a fixed sequence length of 2048 tokens, resulting in 30 million samples and 7,324 training steps. Training proceeds at approximately 3 minutes per step. Although nominal completion required 16 days, node instability extended the overall duration to nearly one month. Checkpoints were saved every 100 steps and later converted back to HuggingFace format for downstream usage. Figure 1 shows stable loss convergence and gradient norm behavior during domain adaptation.

Supervised Fine-Tuning. To improve geoscientific reasoning while preserving general language ability, we perform two-stage supervised fine-tuning (SFT) following the K2 recipe. Training is implemented with DeepSpeed (ZeRO-3 + gradient checkpointing) on 128 nodes (512 accelerators). We maintain the

Dataset	#blockNum	#tokenNum	#itemNum	#tokenSize	#batchRatio
GeoCorpus	25,743,070	52,721,798,004	5,548,479	98.21G	80%
ArXiv	6,691,886	13,704,981,558	742,835	25.53G	10%
Codedata	6,066,725	12,424,652,670	3,456,887	23.14G	10%
Total	38,501,681	78,851,432,232	9,748,201	146.88G	100%

Table 1: Data distribution of the corpus used for training GeoGalactica

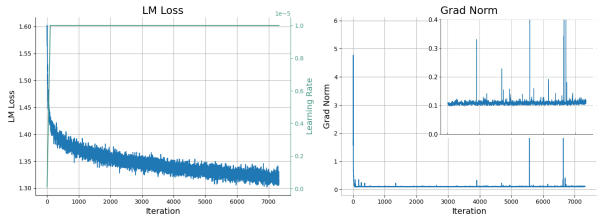

Fig. 1: Training curve during the further pre-training.

pre-training learning rate schedule (max LR 1×10^{-5}, 100-step warm-up) with Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.05). The sequence length is limited to 512 with a global batch size of 512.

Stage 1 aligns the model using Alpaca instruction data. Stage 2 further enhances domain reasoning using GeoSignal v2. And we train the GeoGalactica with geoscientific tools calling data. The full SFT process completes within one day.

Deployment and Quantization. To improve accessibility in resource-constrained environments, we apply GPTQ post-training quantization to compress the model to INT8 precision. This reduces memory requirements from approximately 130GB to 30GB with minimal accuracy degradation, enabling single-GPU deployment. Although GPTQ is designed for zero-shot quantization, we construct a 1,000-document geoscience calibration dataset to better match the model's domain distribution.

3. Results

We evaluate GeoGalactica through objective benchmarks, hallucination analysis, and quantization validation, aiming to measure both scientific reasoning and practical usability in geoscience.

3.1 Performance on Geoscience Benchmarks

We conduct the evaluation **GeoBench (NPEE + APTest)** and the professional **ASBOG** geology examination (113 multiple-choice questions). Results are summarized in Table 2. GeoGalactica achieves the best performance on ASBOG (53.0), substantially outperforming prior geoscience models. It is competitive on NPEE and shows balanced performance across research-oriented tasks. Larger academic models tend to perform well on advanced knowledge but underperform on APTest, indicating bias toward research-level corpora.

3.2 Hallucination Analysis

We evaluate factual consistency using the Focus metric [16] on 18 Earth-science entities derived from Wikipedia. Results (Table 3) show GeoGalactica achieves the highest Focus score (0.7685), substantially outperforming both Galactica-30B and K2. This suggests improved factual grounding after domain-specific continued pre-training and SFT.

Model	NPEE	APTest	ASBOG
Random	27.1	20.0	25.0
ChatGPT	**48.8**	20.0	25.6
K2-7B	39.9	29.3	27.1
Galactica-30B	41.2	38.5	22.9
Galactica-Alpaca-30B	42.6	**44.1**	23.8
GeoGalactica	46.6	36.9	**53.0**

Table 2: Objective benchmark results.

Model	Focus Score
K2-7B	0.6121
Galactica-30B	0.3478
GeoGalactica	**0.7685**

Table 3: Hallucination evaluation (Focus metric).

Model	Perplexity	ASBOG
GeoGalactica (FP32)	3.71	53.0
GeoGalactica (FP16)	3.75	52.5
GeoGalactica (INT8)	3.88	51.2

Table 4: Quantization evaluation.

3.3 Quantization Validation

To support deployment in resource-constrained environments, we evaluate FP32, FP16, and 8-bit GPTQ versions of GeoGalactica. As shown in Table 4, quantization causes only minor performance degradation while maintaining competitive results.

4. Related Works

LLM have enabled domain-specific foundation models across science, biomedicine, finance, and geoscience [17, 18, 19, 11, 20, 21, 14], underscoring the critical role of large-scale, high-quality domain data and data-centric training strategies. Domain-adaptive fine-tuning [22, 23, 24] has been widely adopted to specialize general foundation models for downstream tasks at lower cost, while prompt engineering [25, 26, 27] further enhances task adaptability without additional large-scale retraining.

References

[1] Shu Wang, Yunqiang Zhu, Yanmin Qi, Zhiwei Hou, Kai Sun, Weirong Li, Lei Hu, Jie Yang, and Hairong Lv. A unified framework of temporal information expression in geosciences knowledge system. *Geoscience Frontiers*, 2022.

[2] Cheng Deng, Yuting Jia, Hui Xu, Chong Zhang, Jingyao Tang, Luoyi Fu, Weinan Zhang, Haisong Zhang, Xinbing Wang, and Cheng Zhou. Gakg: A multimodal geoscience academic knowledge graph. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.

[3] Rahul Ramachandran, Muthukumaran Ramasubramanian, Pravesh Koirala, Iksha Gurung, and Manil Maskey. Language model for earth science: Exploring potential downstream applications as well as current challenges. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 4015–4018, 2022.

[4] Qinjun Qiu, Zhong Xie, Liang Wu, and Wenjia Li. Geoscience keyphrase extraction algorithm using enhanced word embedding. *Expert Systems with Applications*, 125:157–169, 2019.

[5] Christopher JM Lawley, Michael G Gadd, Mohammad Parsa, Graham W Lederer, Garth E Graham, and Arianne Ford. Applications of natural language processing to geoscience text data and prospectivity modeling. *Natural Resources Research*, pages 1–25, 2023.

[6] Qinjun Qiu, Zhong Xie, Liang Wu, and Liufeng Tao. Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques. *Earth Science Informatics*, 13:1393–1410, 2020.

[7] Qinjun Qiu, Zhong Xie, Liang Wu, and Wenjia Li. Dgeosegmenter: A dictionary-based chinese word segmenter for the geoscience domain. *Computers & geosciences*, 121:1–11, 2018.

[8] Chengbin Wang, Xiaogang Ma, Jianguo Chen, and Jingwen Chen. Information extraction and knowledge graph construction from geoscience literature. *Computers & geosciences*, 112:112–120, 2018.

[9] Cheng Deng, Bo Tong, Luoyi Fu, Jiaxin Ding, Dexing Cao, Xinbing Wang, and Chenghu Zhou. Pk-chat: Pointer network guided knowledge driven generative dialogue model. *arXiv preprint arXiv:2304.00592*, 2023.

[10] Kai Ma, Miao Tian, Yongjian Tan, Xuejing Xie, and Qinjun Qiu. What is this article about? generative summarization with the bert model in the geosciences domain. *Earth Science Informatics*, pages 1–16, 2022.

[11] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *ArXiv*, abs/2211.09085, 2022.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[13] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, P. Zhang, Yuxiao Dong, and Jie Tang. Glm-130b: An open bilingual pre-trained model. *ArXiv*, abs/2210.02414, 2022.

[14] Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Le Zhou, Luoyi Fu, Weinan Zhang, Xinbing Wang, Cheng Zhou, Zhouhan Lin, and Junxian He. Learning a foundation language model for geoscience knowledge understanding and utilization. *ArXiv*, abs/2306.05064, 2023.

[15] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[16] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*, 2023.

[17] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

[18] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[19] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim DabrXie2023DARWINSDavolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *ArXiv*, abs/2303.17564, 2023.

[20] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 2022.

[21] Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*, 2023.

[22] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2019.

[23] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342, 2019.

[24] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *ArXiv*, abs/2010.02559, 2020.

[25] Yifan Zhang, Cheng Wei, Shangyou Wu, Zhengting He, and Wenhao Yu. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*, 2023.

[26] Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, et al. Impressiongpt: an iterative optimizing framework for radiology report summarization with chatgpt. *arXiv preprint arXiv:2304.08448*, 2023.

[27] Zhiyuan Peng, Xuyang Wu, and Yi Fang. Soft prompt tuning for augmenting dense retrieval with large language models. *arXiv preprint arXiv:2307.08303*, 2023.