Large Language Models Encode Geoscience Knowledge

Anonymous ACL submission

Abstract

Large language models (LLMs) have shed light on potential inter-discipline applications to foster scientific discoveries of a specific domain by using artificial intelligence (AI for science, AI4S). In this study, we introduce A Data-centric Recipe for advancing the application of Large Language Models (LLMs) in the realm of geoscience. Leveraging the versatility of LLMs and their potential for interdisciplinary applications, particularly in Artificial Intelligence for Science (AI4S), we propose a methodology to tailor an open-source LLM to the geoscience domain, with potential for broader interdisciplinary use. This involves further pre-training the model with a comprehensive geoscience text corpus and fine-tuning it using a custom instruction tuning dataset. Our efforts culminate in multiple size of LLM specialized for geoscience tasks. Through rigorous evaluation on geoscience examinations and open-domain questions, our model exhibits state-of-the-art performance across a diverse array of Natural Language Processing tasks within the geoscience domain.

1 Introduction

The advent of Large Language Models (LLMs) marks a seminal point in the evolution of natural language processing (NLP), heralding an era where the amalgamation of artificial intelligence (AI) with diverse scientific domains promises to redefine the frontiers of research and application. LLMs, through their unparalleled proficiency in a vast array of tasks such as reading comprehension, open-ended question answering, and code generation, have showcased the profound impact of harnessing extensive datasets to drive innovation and problem-solving in areas previously constrained by traditional methodologies. This synergy between AI and science, particularly under the umbrella of AI for Science (AI4S), is poised to catalyze significant advancements and discoveries.

Figure 1: Goods helped GeoGalactica construction.

In the landscape of AI4S, the incorporation of NLP within geoscience emerges as a compelling exploration, bridging computational intelligence with the intricate study of Earth's phenomena. Geoscience, encompassing disciplines like geophysics, meteorology, and environmental science, traditionally leans on empirical and theoretical methods to decipher Earth's complex systems. Nonetheless, the exponential growth of data within this field necessitates a paradigm shift towards integrating AI and computer science techniques, promising to accelerate research breakthroughs and effectively tackle global challenges such as climate change and natural disaster resilience.

In the field of geoscience, domain-specific geoscientific knowledge is usually presented in various forms of text data, such as scientific literature, textbooks, patents, industry standards, etc., which traditionally require the utilization of knowledge systems (Wang et al., 2022), knowledge graphs(Deng et al., 2021), or semantic models (Ramachandran et al., 2022) to extract a structured form of these knowledge. More broadly, applying NLP techniques for geoscience use cases has been widely accepted (Zhang and Xu, 2023), ranging from less complex tasks such as document classification (Qiu et al., 2019), topic modeling (Lawley et al., 2023), and entity recognition(Qiu et al., 2020, 2018), to more complex tasks such as knowledge graph construction (Wang et al., 2018), question answering (Deng et al., 2023a) and summarization (Ma et al., 2022).

While general domain LLMs like Galactica (Taylor et al., 2022), LLaMA (Touvron et al., 2023), and GLM (Zeng et al., 2022) have achieved impressive performance across various NLP tasks, they lack the domain-specific knowledge required for geoscience applications. These models have been trained on general datasets that lack authoritative geoscience-related data, limiting their adequacy in addressing the unique challenges posed by the geoscience domain. Although some recent attempts to adapt the LLaMA-7B model for geoscience using geoscience-specific data, such as the K2 (Deng et al., 2023b) model, has shown promising results, this primitive attempt is constrained by its model size and data scale, which consequently may not fully capture the complexity of geoscientific terminology and concepts. However, training a larger LLM comes with new technical challenges, since many aspects of the process become fundamentally different as the model scales up. For example, the stability of training will become more vulnerable, and the training data needs to be scaled up accordingly, resulting in a more systematic way of managing different data sources, etc.

Addressing these challenges necessitates the development of a geoscience-specific LLM, leveraging a comprehensive and meticulously curated dataset to transcend the constraints of current models. This initiative aims to not only tailor a model for the geoscience domain but also refine the dataset and training pipeline to enhance model performance and applicability.

In this paper, we introduce a robust framework for assembling a vast geoscience dataset. This endeavor has led to the creation of GeoSignal-v2, a comprehensive dataset facilitating supervised fine-tuning, alongside the development of tools for the efficient processing of diverse data forms into a coherent training corpus.

The culmination of these efforts is the GeoGalactica (Lin et al., 2023), a LLM with 30 billion parameters, fine-tuned for geoscience applications. This model stands as a testament to the potential of tailored LLMs in revolutionizing geoscientific research, outperforming general-domain models in both benchmark tests and human evaluations across a variety of geoscience-related tasks.

In addition to establishing a roadmap for encoding geoscience knowledge, the main contribution of the paper can be listed as follows:

1. **A Domain-specific LLM:** Our construction of GeoGalactica represents a geoscience LLM that focuses on interacting with humans and generating contents on highly professional academic topics. And showing lower hallucination compared to original Galactica.

2. **A Toolchain for Data Cleaning:** A high-quality training dataset is crucial for successfully training large language models. Therefore, our contribution to the community includes developing an efficient academic data preprocessing toolchain to construct a clean training corpus from PDF documents [1].

3. **A recipe for training domain-specific LLM:** This work provides a comprehensive recipe for training and inferencing domain-specific Large Language Models (LLMs), using geoscience cases as the example, showcasing a step-by-step approach tailored to encode deep geoscience knowledge efficiently.

4. **Full model parameters and benchmarks:** Our work has made all model parameters open source, including both the original and the 8-bit quantized models, along with new benchmark data in the geoscience domain. This allows the open community to observe and iterate on the model's capabilities in geoscience.

2 Related works

With the advent of large-scale language models, numerous disciplines, including geoscience, have witnessed the evolution of domain-specific pre-trained models, trained on specialized corpora (Beltagy et al., 2019; Gu et al., 2021; Wu et al., 2023; Taylor et al., 2022; Luo et al., 2022; Bi et al., 2023). These models undergo large-scale pre-training on domain-specific texts, resulting in foundational models imbued with domain knowledge. It should be highlighted that these models point out the importance of data, and the data-centric training realm is gradually emerging. Meanwhile, (Lee et al., 2019; Huang et al., 2019; Chalkidis et al., 2020) have fine-tuned these base models using domain-specific data, creating models that are custom-tailored to specific downstream tasks at a reduced cost. These

[1]The toolchain is open-sourced on Github repos: `example_url` and `example_url`

efforts have significantly advanced the development of domain-specific Large Language Models (LLMs) through dedicated data integration and model training.

Recently, (Zhang et al., 2023b; Ma et al., 2023; Peng et al., 2023) have delved into prompt engineering to unlock the potential of models without additional training. This approach offers the possibility of unifying various geoscience tasks and further decreasing the cost of deploying large models in domain applications. In geoscience, the exploration of large models is still in its nascent stages. (Deng et al., 2023b) have amassed a considerable amount of high-quality data from geoscience Wikipedia and research literatures, and further fine-tuned the base model, leading to remarkable scientific proficiency and knowledge in geoscience. For the first time, our work employs a large corpus of geoscience documents and textbooks, which were meticulously cleaned using a dedicated toolchain to construct large-scale geoscience models, ensuring data quality. Moreover, our work encompasses the entire process of "further pre-training, supervised fine-tuning, augmented learning" for large foundational models for geoscience, bringing the largest scale and highest quality proprietary language models to the geoscience field from a data-centric perspective. This will open up immense possibilities for future research conducted by geoscience researchers.

3 A Data-centric Recipe for Geoscience LLM Construction

3.1 Data collection and cleaning

To address the lack of geoscience knowledge, we gathered approximately six million geoscience related documents curated by experts. Additionally, we expanded the GeoSignal dataset from K2 to enhance support for NLP tasks in geoscience. We elaborate on our dataset construction process below.

3.2 Customized Pre-training dataset: GeoCorpus

We've developed a comprehensive geoscience document collection, amassing over 5.98 million documents across disciplines like geology and geography, sourced primarily through Microsoft Research (MAG) and supplemented with data from Openalex, CommonCrawl, The Pile, and arXiv, and so on. Our methodology includes sophisticated data collection and deduplication techniques, leveraging diverse sources and copyright-compliant methods to parse and anonymize PDFs. The resulting dataset, optimized for storage efficiency and structured for advanced parsing, forms the basis of a 78B token training corpus, strategically balanced across geoscience and supplementary domains, to support cutting-edge AI research in geoscience.

We also employed tokenization method to cope with special tokens, such as [START_FIGURE], [START_TABLE], [START_REF], and [START_FORMULA], to unify text extracted from various sources into a standardized protocol.

3.3 The Customized SFT dataset: GeoSignal Version 2

Through extensive research, we've explored NLP tasks tailored to geoscience, identifying various tasks. However, we've noticed untapped unsupervised signals within. Tasks include Geoscience Knowledge Graph (NER, RE, text-to-graph transformation), Academic Applications (keyword extraction, summarization, information retrieval), General Applications (Q&A, geoscience education conversations, text classification), and Geographical Applications (POI queries, multimodal Q&A).

These signals can be reconstructed using professional geoscience data websites. We've categorized data into literature-related, geoscience-related, and self-instruction-related, the latter distilled from ChatGPT and annotated by geoscience experts for constructing high-quality question-answering datasets.

Domain General Natural Language Instruction: We integrated four platforms to restructure signals from various geoscience-related platforms. Deep Literature and DataExpo serve as datasets for referential relationships. Using Grobid, we convert documents into XML, identifying in-text citations and corresponding references. GSO provides valuable supervised signals by extracting synonyms and definitions. GAKG's rich graphical information generates binary pairs for sequence-to-sequence supervised data.

Restructured Knowledge-intensive Instruction: To construct restructured knowledge-intensive instruction data, we first search for authoritative websites covering paleontology, dinosaurs, fossils, rocks, and other geoscience fields. We then filter these sites, focusing on those with structured data available for extraction. For structured websites, we implement processing similar to K2,

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	-

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

matching structured data using Key-Value pairs to create natural Instruction and Response pairs.

Self-Instruct: Following methods outlined in Alpaca and Baize, we generate instructional tuning data by utilizing problem seeds to generate answers from ChatGPT. For geoscience, we generate 1000 questions per subject and make these problem seeds public available.

For overall data collection, we compile the following totals and select a proportion for supervised fine-tuning. After manual verification and cleaning, we finalize a dataset of 100K samples as GeoSignal Version 2 for instructional data during supervised fine-tuning.

Finally, the detailed statistic of the instruction tuning data is shown in Table 7.

4 Training

Building upon the insights gleaned from GLM-130B (Zeng et al., 2022), we outline the frameworks and strategies for our training phase.

4.1 Further Pre-training

Following initial pre-training by Meta AI, we further pre-train the Galactica using GeoCorpus. This process aims to refine the model's understanding and generation capabilities within specific domains or styles.

We leverage a accelerators cluster with *ROCm* software stack, coupled with the Megatron-LM framework, to conduct further pre-training. The computing cluster comprises *512 nodes*, each equipped with a 32-core CPU, 128GB of memory, and 4 pieces of 16G memory accelerators, totaling *2048 accelerators*. The Megatron-LM framework employs 3D parallelism strategies, including pipeline-parallel, model-parallel, and data-parallel approaches, to maximize GPU performance while minimizing communication overhead. With four acceleration cards per node, we set the model parallel size to *4* for optimal efficiency. Additionally, with a mini-batch size of *1*, we configure the pipeline-parallel size to *16* to fully utilize memory resources.

Before the training, we referred to and modified the code available on Hugging Face for converting Hugging Face's GPT-2 to Megatron's GPT-2. The conversion parameters can be adjusted based on the actual scale of pipeline parallelism (PP), model parallelism (MP), and data parallelism (DP) during runtime.

All the training samples are preprocessed through tokenization. The tokenized results of each document are concatenated using an end-of-sentence (eos) marker. Subsequently, we crop the concatenated sequences into fixed lengths of *2048*, resulting in *30 million* training samples, corresponding to *7,324* training steps. Prior to formal training, we conduct preliminary experimental analyses of node failures and save checkpoints at 100-step intervals. After transforming the initial checkpoint format into the required Megatron-LM format, the pre-training process commences. Over a span of *16 days*, the computing cluster completes the further pre-training at a speed of *3 minutes per step*. However, due to frequent node failures, the actual training duration extends to nearly a month. Following pre-training, we convert the checkpoints into the Hugging Face format for subsequent applications.

4.2 Supervised Fine-Tuning (SFT)

LLMs undergo SFT post-pre-training on a more focused dataset under human supervision, adapting the model to specific tasks or enhancing performance in certain areas.

We employed SFT to boost the geoscientific reasoning of large-scale models on specific tasks, ensuring effective transfer of language capabilities while maintaining pre-training generalization.

We utilized DeepSpeed frameworks, primarily utilizing the accelerators cluster. SFT truncated to 128 nodes and 512 accelerators, maintaining pre-training learning rate schedule (max LR: 1e5) with linear warmup (100 steps) and Adam optimizer

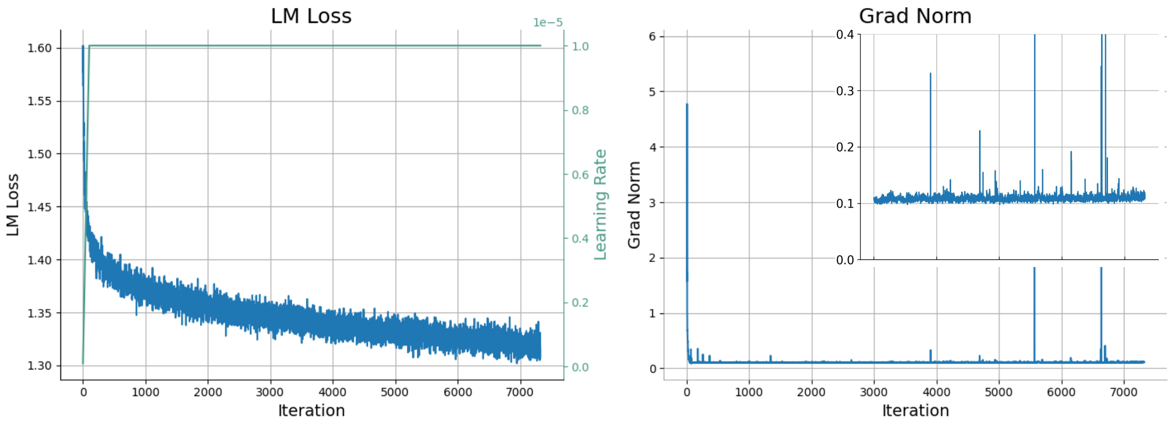

Figure 2: Training curve during the further pre-training.

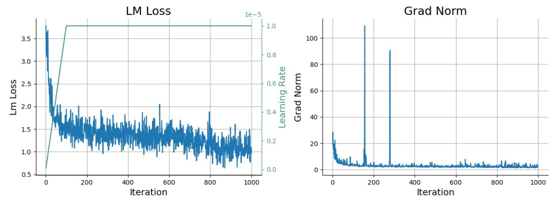

Figure 3: Training curve during the SFT on Geosignal.

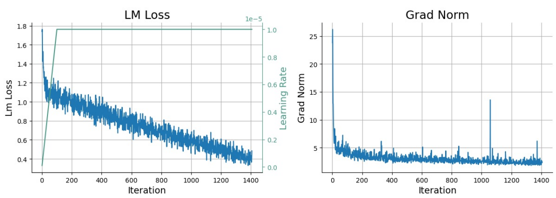

Figure 4: Training curve during the tools SFT.

($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay: 0.05, $\epsilon : 1e - 8$).

We also utilized DeepSpeed ZeRO3 and gradient checkpoint for memory optimization, limiting input sequence length to 512. Global batch size is set to 512 due to DeepSpeed limitations. Default Huggingface trainer framework settings is used, training conducted on Alpaca dataset for three epochs, completed within one day with Megatron-LM support.

We implemented SFT in two stages refer to K2's recipe, aligning model with humans via Alpaca instruction tuning data in the first stage and utilizing GeoSignal v2 in the second stage, the learning curve is shown as Figure 3.

Moreover, we also do geoscience data-centric tool learning, the learning curve is shown as Figure 4.

4.3 Deploy

In various specialized fields, researchers often aim to utilize models with lower resources and costs, and geoscience field is no exception. Using GEOGALACTICA requires a minimum of 140GB of GPU memory, a significant expenditure for many independent research institutions. Therefore, we also offer a post-training quantization method for GEOGALACTICA, reducing its memory consumption from 130GB to 30GB while maintaining considerable capabilities, we will illustrate in experiment session.

We do quantization via GPTQ method(Frantar et al., 2022), which does post-training quantization of large language models. We adopt this method to enable the compression of GEOGALACTICA to 8 bits per weight with minimal accuracy loss, significantly reducing computational and storage requirements. GPTQ allows the execution of model GEOGALACTICA on a single GPU, offering considerable speedups and making geoscience generative AI more accessible and efficient.

To better serve the communities in low-resource vertical fields, we have selected *1,000* documents from various geoscience fields to form the GPTQ dataset used for quantisation.

We believe that even though the GPTQ aims to work on any kind of data, remaining actually zero-shot, using a dataset more appropriate to the GEOGALACTICA training can improve quantization accuracy.

5 Experiments

Once model training is complete, we proceed to evaluate its scientific and geoscientific knowledge.

Evaluation is divided into two parts, including automated evaluation over new version of GeoBench to exam the geoscience knowledge comprehension of the models, and functional evaluation to test the model over selecting geoscience tasks

5.1 Benchmarks

We evaluate the model abilities of geoscience knowledge comprehension over three benchmarks, including general knowledge evaluation benchmark **MMLU**, geoscience knowledge evaluation benchmark **GeoBench**, and the **ASBOG** test, proposed by this paper.

MMLU. The MMLU divided into math and non-math sections by Galactica, indicates improvement in specific model skills (algebra, biology, chemistry, mathematics) after processing 6 million geoscience-related literature documents. Enhancement in mathematics, machine learning attributed to mathematical geology, biological geoscience, chemical geology papers, showcasing geoscience's interdisciplinary nature. However, physics performance favors original Galactica over our model, while unrelated disciplines (medical genetics, medicine, electrical engineering) show decline. Furthermore, Our model and original Galactica demonstrate similar average performance in math-related MMLU sections.

GeoBench. GeoBench is proposed by K2 (Deng et al., 2023b) for assessing geoscientific task performance, consisting tasks **NPEE** and **APTest**. There are 183 multiple-choice questions in NPEE and 1,395 in total in the AP Test, constituting the objective task set. Meanwhile, K2 gathers all 939 subjective questions in NPEE to be the subjective tasks set and use 50 to measure the baselines with human evaluation.

ASBOG. The ASBOG Fundamentals of Geology Examination is a requirement for a person to become a Licensed Professional Geologist and to offer geologic services to the public in States that register geologists by examination. We collect 113 pieces if the textual multiple choices questions.

Through these evaluations, we aim to comprehensively assess the model's abilities and compare its performance against automated benchmarks and human assessments, ensuring competence in scientific and geoscientific domains.

5.2 Automatic Evaluation

Our tests on GeoBench reveals larger academic models outperforming NPEE but underperforming in AP Study, indicating a bias towards advanced knowledge due to training on academic research achievements like literatures. This highlights the need to address basic knowledge deficiencies for future improvements.

Baselines	NPEE	APTest	ASBOG
Random	27.1	20.0	25.0
ChatGPT	**48.8**	20.0	25.6
Gal-6.7B	25.7	29.9	23.9
LLaMA-7B	21.6	27.6	22.1
K2-7B	39.9	29.3	<u>27.1</u>
Gal-30B	41.2	<u>38.5</u>	22.9
GalAlp-30B	42.6	**44.1**	23.8
GEOGALACTICA	<u>46.6</u>	36.9	**53.0**

Table 2: Comparison among baselines on Objective tasks.

Surprisingly, machine learning has experienced significant enhancement, likely due to the inclusion of GitHub code in our corpus. In summary, subjects closely related to geoscience, including those logically connected to geology and its subfields, have shown notable progress. However, disciplines like physics indicate that the original Galactica outperforms our GEOGALACTICA and subjects unrelated to geosciences, such as medical genetics, medicine, and electrical engineering, have shown a decline in performance. It is noteworthy that GEOGALACTICA and the original Galactica are generally at a similar stage regarding average performance in math-related subjects within the MMLU. The results are in subsection B.1.

After assessing mathematical subjects, we analyzed excluded subjects. Overall, our model slightly outperforms original Galactica in average non-math-related MMLU subjects. Notably, global facts, US History, and World History show significant improvement, likely due to history's intertwining with geoscience. This underscores geoscience's profound impact on global progress. Moreover, in conceptual physics, learning from geoscience documents improves model understanding, indicating misalignment with traditional education. However, Models struggle to apply geoscience-related knowledge to college and high school-level problems. The results are in subsection B.2.

5.3 Functional Evaluation

We invited ten geoscience researchers participating in voting and scoring. Model performance is compared with five other large-scale platforms in open testing. In this section, we evaluated five open models alongside our model. including MOSS, Qwen, ChatGPT, Yiyan (Ernie Bot), ChatGLM.

We adopted K2's Human Evaluation framework to define evaluation metrics for open-ended questions, comprising scientificity, correctness, and coherence, scored between 1 to 3:

Scientificity: Assesses if the generated content aligns with geoscience professional discourse, with scores indicating the quality from not good (1) to very good (3).

Correctness: Judges if the information provided is convincing and accurate from a geoscience expert's perspective, with scores ranging from incorrect (1) to correct (3).

Coherence: Evaluates the consistency and smoothness of the text in discussing a specific topic, graded from not good (1) to very good (3).

These metrics enable the calculation of cumulative scores. For functional questions of the large model, the evaluation metric is relative ranking. Participants receive responses from all six models for the same input, and expert judges rank these models in order from 1 to 6. The overall ranking of each model is then determined. Ten geoscience practitioners, including six students and four teachers, were invited for this evaluation process.

The tasks and corresponding scores are presented as follows:

In the evaluation of functional tasks, we have chosen to utilize our model specifically for analyzing scientific research literature, aiming to enhance comprehension and interpretation. When external information input is unnecessary, we rely on the consistent output provided by the ChatALL interface. Since the overall evaluation involves ranking, lower scores are preferred. The tasks include:

- **Knowledge-based Associative Judgment Question**: Questions are formulated based on the knowledge trees in GSO to determine the presence or absence of knowledge system relationships.

- **Research Paper Titling Task**: Abstracts from 20 geoscience research papers are randomly selected and inputted into the model to generate titles, demonstrating the model's grasp of knowledge points and familiarity with the field.

- **Geoscience Research Functionality**: To ensure fairness in incorporating external research papers, we use our own PDF parsing solution for interpretation and rely on consistent output from the ChatALL interface. For GEOGALACTICA, interactions are conducted through our UI interface, producing outputs accordingly.

  In interpreting scientific literature, we often inquire about speech writing based on the article's content, summarization assistance, and recommendation of prerequisite knowledge points. We assessed five papers covering various domains of Earth sciences and written in different styles.

The tasks and corresponding scores are presented as follows:

We can envision that functional characters represent the services currently available from scientific large language models. Despite the persistent illusions created by large language models, it's challenging to directly influence these disciplines from an educational and instructional standpoint. However, we can offer simple aids such as question generation, summarization, rapid reading, and information extraction. Our goal is to facilitate research in the geosciences, thereby enhancing the efficiency of scholarly research in this field.

Fortunately, we came across Galpaca-30B on Hugging Face, which significantly reduced the carbon emissions from our finetuning experiments. This model utilized Alpaca's instructions to learn from the dataset and was applied to SFT on Galactica-30B. Upon horizontal comparison as an ablation experiments, Galpaca-30B performed notably worse than the original Galactica and GEOGALACTICA in the majority of disciplines. This indicates that instruction learning in the general domain can significantly impact the performance of specialized domain models during practical evaluations.

5.4 Hallucination Detection

As a large language model designed to support academic research, we must address the issue of illusions. Although our current hallucination tests in academic verticals are limited, we can examine the model's performance from a factual knowledge perspective. We compared the previous geoscience model K2, our base model Galactica-30B, and our GEOGALACTICA, using Wikipedia knowledge of 18 keywords across 18 fields of Earth Science as a

		MOSS	Qwen	ChatGPT	Yiyan	ChatGLM	GeoGalactica
Noun Definition	Scientificity	291	419	337	236	278	339
	Correctness	302	435	351	276	291	361
	Coherence	351	435	357	305	347	393
Beginner LevelQ&A	Scientificity	116	191	219	176	160	176
	Correctness	120	177	214	174	156	173
	Coherence	147	207	225	187	184	202
Intermediate Level Q&A	Scientificity	143	178	210	180	161	162
	Correctness	154	180	206	186	163	169
	Coherence	178	193	207	189	179	171
Advanced Level Q&A	Scientificity	166	202	137	190	172	185
	Correctness	173	199	133	192	171	187
	Coherence	194	209	181	200	194	206

Table 3: We report the results of the selected baselines on Q&A tasks.

		MOSS	Qwen	ChatGPT	Yiyan	ChatGLM	GeoGalactica
Knowledge-based Associative Judgment	Sum of Rank	579	557	600	570	752	725
Research Paper Titling Task	Sum of Rank	805	426	326	561	440	451
Geoscience Research Functionality	Writing	114	135	62	178	106	135
	Summary	164	185	86	139	168	100
	Extraction	115	232	51	160	169	212

Table 4: We report the results of the selected baselines on functional tasks.

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

Table 5: Focus score over geoscience entities explanation.

	Perplexity	ASBOG
GeoGalactica (FP32)	**3.71**	**53.0**
GeoGalactica (FP16)	**3.75**	**52.5**
GeoGalactica-8bit-GPTQ	**3.88**	**51.2**

Table 6: Quantization Accuracy Evaluation over Perplexity and ASBOG Test.

reference point for evaluation using Focus (Zhang et al., 2023a).

5.5 Quantization Accuracy

In numerous geoscience contexts, researchers require large language models to perform geoscientific reasoning, such as summarizing documents and offering concise insights into interdisciplinary materials. However, smaller models like K2 (Deng et al., 2023b) and OceanGPT (Bi et al., 2023) struggle with complex challenges due to limitations in scalability. To address this, we employ quantization to reduce the model size. To guarantee the accuracy of quantization, we utilize 1,000 geoscience documents for post-training the GeoGalactica. Additionally, we initially pre-train and fine-tune the GeoGalactica with FP32 precision and then convert it to FP16 using PyTorch's quantization techniques. There is a slight decrease in the model's performance, which we attribute to the computational differences of the accelerators. Table 6 Shows the results of the quantized GeoGalactica.

6 Conclusion

In conclusion, our study underscores the transformative potential of domain-specific Large Language Models (LLMs) in geoscience, achieving notable advancements in understanding Earth's dynamics. The development of the GeoGalactica model exemplifies how targeted AI can address critical environmental challenges, marking a pivotal step towards harnessing AI for scientific discovery. This endeavor not only sets a new benchmark for AI applications in the sciences but also reinforces the importance of open science, inviting collaboration and further innovation in the AI-driven exploration of our natural world.

Limitations

Discipline

The scope of our research is confined to the field of geoscience. The generalization of the recipe of our data-centric road-map across different domains remains an open question. Meanwhile, it may still face challenges with very niche or cutting-edge topics within the field that are not well-represented in its training data. The researchers should adopt their own data to fine-tune the model to their own needs.

Computational Resources

The GEOGALACTICA model demands substantial computational resources for both training and inference processes. This high demand can restrict its accessibility, particularly for institutions and researchers with limited resources, and may also impede its use in real-time applications where rapid response is necessary. Even the quantized version of the GEOGALACTICA, which is designed to reduce the computational footprint, still necessitates the use of multiple consumer-grade accelerators to effectively deploy and run the model. This requirement can be a barrier to entry for smaller organizations or individual researchers who may not have access to such hardware.

Ethics Statement

The dataset utilized for training the GEOGALACTICA model is comprised of publicly accessible documents. We have meticulously ensured that all data was collected and processed with utmost respect for the privacy and intellectual property rights of the original authors. Our approach strictly avoids the use of any personal data, and we have diligently attributed all information to its respective sources. It is important to acknowledge that, like all large language models (LLMs), GEOGALACTICA might inadvertently inherit biases from its training data. These biases could potentially impact the fairness and accuracy of the model's outputs. As a result, the model may sometimes generate content that deviates from factual accuracy, a phenomenon commonly referred to as "hallucinations." Therefore, we strongly advise users and readers to exercise discretion when interpreting the outputs generated by GEOGALACTICA.

References

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

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

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

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

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

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. 2023b. Learning a foundation language model for geoscience knowledge understanding and utilization. *ArXiv*, abs/2306.05064.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *ArXiv*, abs/2210.17323.

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

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

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

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

Zhouhan Lin, Cheng Deng, Le Zhou, Tianhang Zhang, Yi Xu, Yutong Xu, Zhongmou He, Yuanyuan Shi, Beiya Dai, Yunchong Song, Boyi Zeng, Qiyuan Chen, Tao Shi, Tianyu Huang, Yiwei Xu, Shu Wang, Luoyi Fu, Weinan Zhang, Junxian He, Chao Ma, Yunqiang Zhu, Xinbing Wang, and Cheng Zhou. 2023. Geogalactica: A scientific large language model in geoscience. *ArXiv*, abs/2401.00434.

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

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

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

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

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

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

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

Rahul Ramachandran, Muthukumaran Ramasubramanian, Pravesh Koirala, Iksha Gurung, and Manil Maskey. 2022. 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.

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

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. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

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

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

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

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. 2022. Glm-130b: An open bilingual pre-trained model. *ArXiv*, abs/2210.02414.

Hao Zhang and Jin-Jian Xu. 2023. When geoscience meets foundation models: Towards general geoscience artificial intelligence system. *arXiv preprint arXiv:2309.06799*.

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

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

A SFT data in GeoSignal

The data used to do the supervised fine-tuning is in Table 7:

B MMLU Evaluation Results

In this section, we append the materials of the MMLU test result for GeoGalactica, Galactica, and GalAlpaca.

B.1 Math-related tasks in MMLU

The Table 8 shows the results of math-related tasks in MMLU.

B.2 Non-Math tasks in MMLU

The Table 9 shows the results of none-math-related tasks in MMLU.

		Signals	tuples	#NumofSamples
DDE Scholar		Title (with Abstract)	(abstract; title)	2,690,569
		Abstract (with Publications Fulltext)	(fulltext; abstract)	2,601,879
		Category (with abstract)	(abstract; category)	12,321,212
		Related Paper (with abstract)	(source abstract; target abstract; reference sentence)	40,047,777
		One Sentence Summary (with abstract)	(abstract; question; answer)	2,690,569
		Reference resolution	(sentence; pronoun.; reference item) [including citation]	2,329,820
DDE DataExpo		Title	(abstract; title)	216,036
		Summary & Abstract	(fulltext; abstract)	216,036
GAKG	**GAKG**	Principal Concepts	(sentence; entity; types)	3,892,102
		Relations	(abstract; sentence; head entity; relation; tail entity)	30,123
		Paper table caption	(table caption; refering sentence)	2,772,166
		Paper illustration caption	(illustration caption; refering sentence)	9,128,604
		Paper table content	(table caption; table content)	2,772,166
		Paper illustration content	(illustration caption; illustration content)	9,128,604
	GSO	Factual knowledge	(sentence; facts; improper statement)	114,392
		Taxonomy	(upper term; term)	112,298
		Synonyms	(term; synonym term)	23,018
		Word description	(word; description; source)	110,209
	GA-Dialogue	Future content and Previous content	(corrupted text; corrupted positions; target spans)	5,434
GeoOpenData	**dinosaur**	Factual knowledge	(property; property value)	11,348
	fossilcalibrations	Factual knowledge	(property; property value)	1,749
	fossilontology	Factual knowledge	(property; property value)	3,210
	mindat	Factual knowledge	(property; property value)	51,291
	ngdb	Factual knowledge	(property; property value)	148,212
	opendatasoft	Factual knowledge	(property; property value)	37,823
	rruff	Factual knowledge	(property; property value)	32,778
	usgsearthquake	Factual knowledge	(property; property value)	37,284
WordNet		Synonyms	(term; synonym term)	6,408
		Word description	(word; description; source)	27,123
Wikipedia		Title	(term; abstract)	3,033,595
		Summary & Abstract	(fulltext; abstract)	753,920
		Entity mentions	(paragraph; entities)	3,688,926
		Relation	(text; subject; property; object)	630,210
IODP		Title	(abstract; title)	2,839
		Summary & Abstract	(fulltext; abstract)	2,638

Table 7: GeoSignal Statistics Table.

Subject	Our model	GAL 30B	GalAlp 30B
Abstract Algebra	**0.300**	0.250	0.320
Astronomy	0.461	**0.500**	0.474
College Biology	**0.576**	**0.576**	0.514
College Chemistry	**0.370**	0.320	0.350
College Computer Science	0.400	**0.410**	0.370
College Mathematics	0.320	**0.350**	**0.350**
College Medicine	0.480	**0.520**	0.445
College Physics	0.284	**0.333**	0.294
Econometrics	**0.377**	0.368	0.368
Electrical Engineering	0.538	**0.579**	0.503
Elementary Mathematics	**0.328**	0.310	0.288
Formal Logic	**0.302**	0.270	0.278
High School Biology	**0.565**	0.561	0.535
High School Chemistry	0.360	**0.399**	0.355
High School Computer Science	0.500	0.480	**0.510**
High School Mathematics	**0.311**	0.256	0.304
High School Physics	0.298	**0.364**	0.325
High School Statistics	0.333	**0.352**	0.319
Machine Learning	**0.411**	0.339	0.366
Medical Genetics	0.550	**0.580**	0.520
Average	**0.4032**	**0.40585**	**0.3894**

Table 8: We report the results of the three models in math.

Subject	**GeoGal 30B**	**Gal 30B**	**GalAlp 30B**
Anatomy	0.496	**0.541**	0.533
Business Ethics	**0.430**	0.420	0.470
Clinical Knowledge	0.532	**0.555**	0.491
Computer Security	0.600	**0.650**	0.620
Conceptual Physics	**0.481**	0.434	0.417
Global Facts	**0.390**	0.300	0.340
High School European History	0.533	**0.606**	0.491
High School Geography	**0.581**	0.540	0.515
High: School Gov & Politis	0.534	**0.565**	0.461
High School Macroeconomics	**0.408**	0.405	0.367
High School Microeconomics	0.424	**0.458**	0.424
High School Psychology	0.613	**0.628**	0.556
High School US History	**0.436**	0.352	0.319
High School World History	**0.620**	0.456	0.446
Human Aging	**0.552**	**0.552**	0.511
Human Sexuality	0.511	**0.565**	0.481
International Law	0.612	**0.644**	0.554
Jurisprudence	**0.491**	0.472	0.444
Logical Fallacies	0.423	**0.472**	0.442
Management	0.573	**0.602**	0.515
Marketing	0.641	**0.705**	0.607
Miscellaneous	**0.522**	0.501	0.470
Moral Disputes	**0.480**	0.462	0.468
Moral Scenarios	0.238	0.244	**0.245**
Nutrition	**0.536**	0.520	0.448
Philosophy	0.444	**0.492**	0.431
Prehistory	0.503	**0.522**	0.435
Professional Accounting	**0.344**	0.312	0.319
Professional Iaw	0.326	0.326	**0.327**
Professional Medicine	0.438	**0.449**	0.379
Professional Psychology	0.472	**0.505**	0.449
Public Relations	**0.473**	0.445	0.455
Security Studies	**0.424**	0.408	0.322
Sociology	0.537	**0.547**	0.483
US Foreign Policy	**0.550**	0.510	0.540
Virology	**0.434**	0.422	0.410
World Religion	0.421	**0.427**	0.380
Average	**0.487**	**0.486**	**0.448**

Table 9: We report the results of the three models in social sciences.