Learning Foundation Language Models for Geoscience Knowledge Understanding and Utilization

Cheng Deng[1], Tianhang Zhang[1], Zhongmou He[1], Qiyuan Chen[2], Yuanyuan Shi[1], Le Zhou[1]
Luoyi Fu[1], Weinan Zhang[1], Xinbing Wang[1], Chenghu Zhou[3], Zhouhan Lin[1], Junxian He[1]
[1]Shanghai Jiao Tong University, [2]University of Waterloo
[3]Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences
davendw@sjtu.edu.cn,{lin.zhouhan,junxianh2}@gmail.com
Corresponding authors: Zhouhan Lin and Junxian He

arXiv:2306.05064v1 [cs.CL] 8 Jun 2023

ABSTRACT

Large language models (LLMs) have achieved great success in general domains of natural language processing. In this paper, we bring LLMs to the realm of geoscience with the objective of advancing research and applications in this field. To this end, we present the first-ever LLM in geoscience, **K2**, alongside a suite of resources developed to further promote LLM research within geoscience. For instance, we have curated the first geoscience instruction tuning dataset, **GeoSignal**, which aims to align LLM responses to geoscience-related user queries. Additionally, we have established the first geoscience benchmark, **GeoBenchmark**, to evaluate LLMs in the context of geoscience. In this work, we experiment with a complete recipe to adapt a pretrained general-domain LLM to the geoscience domain. Specifically, we further train the LLaMA-7B model on over 2 million pieces of geoscience literature (3.9B Tokens) and utilize GeoSignal's supervised data to fine-tune the model. Moreover, we share a protocol that can efficiently gather domain-specific data and construct domain-supervised data, even in situations where manpower is scarce. Experiments conducted on the GeoBenchmark demonstrate the effectiveness of our approach and datasets. [1]

KEYWORDS

Geoscience Language Model, Domain Adaptation

1 INTRODUCTION

Geoscientists have long faced challenges in integrating data from various sources and disciplines due to differences in terminologies, formats, and data structures, which subsequently leads to numbers of natural language tasks in geoscience such as geological and geographical named entity recognition [10], spatial and temporal relation extraction [27] to build geoscience knowledge graph [7], geology reports and literatures summarization [26], and representation learning via geoscience language models [33]. However, language models in geoscience are sparse and remain limited in scale [8]. This situation stands in stark contrast with the prosperity of large language models (LLMs), such as ChatGPT [31] and GPT-4 [32], in general natural language processing (NLP), where notable successes have been achieved.

Despite their effectiveness in general domains, current LLMs often fall short in catering to the needs of geoscientists. This shortfall is largely attributed to the lack of reliable knowledge concerning geoscience problems, given that the related geoscience data seldom exist in the commonly used pretraining text corpora such as C4 [35] and the Pile [12]. Moreover, top-performing LLMs like ChatGPT only offer services via APIs, which presents roadblocks for external domain research and advancement. To mitigate these issues and foster research and application within the geoscience domain, we introduce the first-ever open-source LLM for geoscience, referred to as **K2** (*The second highest mountain in the world, which we believe in the future larger and more powerful geoscience language models will be created*). K2, a GPT-like language model comprising 7 billion parameters, is based on the pre-trained LLaMA [42] model but specializes on the geoscience domain. Along with the introduction of K2, this paper also explores a roadway to collect geoscience text corpus, constructs geoscience instruction supervised data, and builds geoscience NLP tasks benchmarks, in alignment with the Deep-time Digital Earth (DDE, [44])[2] big science plan.

The training of K2 consists of two stages, the pretraining stage and the instruction tuning stage, as depicted in Figure 1. During pretraining, we continue pretraining the LLaMA-7B model on a geoscience text corpus that we preprocessed from geoscience papers. Then we perform instruction tuning [4, 23, 36], where we further train the model to follow human instructions. To this end, we have curated **GeoSignal**, an instruction tuning dataset created by unifying 8 diverse geoscience NLP task data with prompts, such as relation extraction, entity recognition, classification, and summarization. We also construct **GeoBenchmark**, an evaluation dataset comprising more than 1500 objective questions and 939 subjective questions collected from National Postgraduate Entrance Examination (NPEE) on Geoscience and AP Test Geology, Geography and Environmental science. GeoBenchmark serves to track the progress and drive the development of geoscience language models. Through our concerted efforts in data collection and training, the resulted K2 model is a foundation language model that can be used to design multiple geoscience applications, making it benefit geoscience researchers and practitioners [28].

Our contributions can be listed as follows:

- We introduce K2, a foundation langugage model in geoscience field. K2 can answer geoscience questions and follow geoscientists' instructions via suitable prompts with its professionalism in geoscience.
- We construct GeoSignal, the first-ever geoscience-supervised instruction data. To evaluate K2 on geoscience tasks and the following language models in geoscience, we build GeoBenchmark, the first NLP task benchmarks in geoscience.

[1]We release the full version of training data and models after the final draft.

[2]https://www.iugs.org/dde

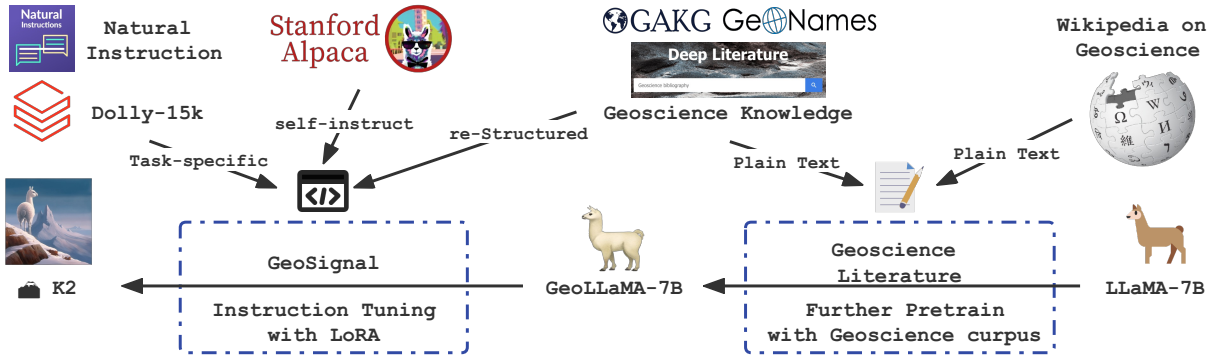

Figure 1: Pipeline of training K2, including two steps, one is further pretrain for absorption of geoscience knowledge, another one is instruction tuning, deploying to make the model align to human, instructed by human, and response like a human.

- Taking geoscience as an example, we build up a paradigm to construct the domain text corpus and domain-supervised instruction data and explore a recipe to train a domain-specific language model.
- Compared with similar-size baseline models, K2 outperforms both subjective and objective geoscience tasks. At last, we release the code, K2 weights, GeoSignal, and GeoBenchmark at Github.

The rest of the paper is arranged as follows: Section 2 will introduce the related work of K2. In section 3, the detail of data collection and supervised instruction data construction will be illustrated. Further, we will share our further pretrain detail and parameter-efficient instruction tuning processes in section 4, while in section 5, we will evaluate the K2 and do an ablation study. Finally, we will discuss the topics raised, lessons learned, potential applications, and future work related to the K2.

2 RELATED WORK

Foundation Language Models. Since the appearance of ChatGPT [31], there has been a large number of large language models for use as a foundation model to solve real-life problems. Since the models that provide only online demos and APIs, like ChatGPT, GPT-4[32] and Yiyan (https://yiyan.baidu.com/) are not suitable and convenient for further pretraining and developing. The open-source models like CodeGen [30], LLaMA[42], GLM [48], becomes the foundation models for many other instruction-tuned LLMs like Alpaca [39], Baize [46], Vicuna [3], Koala [13], and Dolly [6].

Domain Language Models. Large language models become the foundation model to address the issues in many other domains. In life science field, Med-PaLm [38], MedGPT [19], BioGPT [25], and Bio-Megatron [37] The large language model is useful and reliable in the biomedicine field [43]. In natural science field, Geographic-BERT [22], MGeo [9], ERNIE-GeoL [17] and GeoBERT [8] are typical cases in geography and geology, while MatSciBERT [14] is the one in material science. In academic scenario, SciBERT [2] and Galactica [40] are two examples.

Parameter-Efficient Tuning on LLMs. Conventional fine-tuning needs to update all the parameters in LLMs, leading to inefficient and leaving a large carbon footprint as the models grow along with the scaling law [18]. Soft Prompt tuning [20] frozen language models to perform specific downstream tasks. Prefix-tuning [21] draws inspiration from prompting for language models, allowing subsequent tokens to attend to this prefix as if it were "virtual tokens". In addition, Adapter [15] make the parameters of the original network remain fixed, yielding a high degree of parameter sharing, and LoRA [16] views the update of the weights as the result of two tunable low-rank matrices multiplication.

3 DATA COLLECTION AND CURATION

To train K2, we collect geoscience text corpus and geoscience-oriented data from various resources. Then we re-structure the data into signals and build up the instruction tuning dataset GeoSignal. This valuable information can serve for learning knowledge for geoscience tasks and instruct models for aligning with humans and experts. Moreover, we develop GeoBenchmark to compare language models focusing on geoscience. We will publicly make our datasets available after the final draft on Github.

3.1 Pretraining Data

Our text corpus for further pretraining on LLaMA-7B consists of *3.9 billion tokens* from geoscience papers published in selected high-quality journals in earth science and mainly collected by GAKG [7].

3.1.1 *Geoscience Text Corpus Collection.*

Geoscience Open Access Literatures. With the support from DDE (Deep-Time Digital Earth Big Science Program), we can have the resources and chances to access materials and data strongly related to geoscience, including *531* journals, and *4,274,716* papers' metadata. We use *1,122,094* open-access papers' PDFs organized by GAKG to build the text corpus.

Wikipedia pages about Earth science. Wikipedia is an import resource we take into account for text corpus collection, and the root node of the Wikipedia category of geoscience we take into consideration is *"https://en.wikipedia.org/wiki/Earth_science"*. We mine all the child and related topics connected to it and finally gain *767,341* Wikipedia pages.

In brief, The statistic of the collection of geoscience text corpus is shown in Table 1.

3.1.2 *Text Corpus Preprocessing.*

[START_FIGURE]

Fig. 2. SEBLUP estimates of perceived disorder in Manchester (division in 6 quantiles).

[END_FIGURE]

A. Figure Processed Text.

Table 5
Summary of covariates and coefficients of correlation of each variable with direct estimates of perceived disorder.

	Min	First quartile	Median	Mean	Third quartile	Max	Spearman coeff
Proportion BME	0.03	0.15	0.25	0.31	0.42	0.89	0.31*
Crime per capita	0.04	0.11	0.16	0.17	0.23	0.84	0.29*
Proportion unemployed	0.06	0.28	0.36	0.35	0.43	0.69	0.29*
Income deprivation	0.01	0.12	0.24	0.24	0.34	0.52	0.35*
Population churn	0.21	0.33	0.39	0.39	0.44	0.68	0.13*
Mixed land-uses	0	0	0	0.19	0	1	0.10*

* p-value < 0.01, *p-value < 0.05.

[START_TABLE]
Table 5

Summary of covariates and coefficients of correlation of each variable with direct estimates of perceived disorder.

None	Min	First quartile	Median	Mean	Third quartile	Max	Spearman coeff
Proportion BME	0.03	0.15	0.25	0.31	0.42	0.89	0.31 a
Crime per capita	0.04	0.11	0.16	0.17	0.23	6.84	0.29 a
Proportion unemployed	0.06	0.28	0.36	0.35	0.43	0.69	0.29 a
Income deprivation	0.01	0.12	0.24	0.24	0.34	0.52	0.35 a
Population churn	0.21	0.33	0.39	0.39	0.44	0.68	0.13*
Mixed land-uses	0	0	0	0.19	0	1	0.10*
a p-value < 0.01, *p-value < 0.05.	None	None	None	None	None	None	None

[END_TABLE]

B. Table Processed Text.

Simulation is an established practice in the imaging radar community that serves many purposes, including inSAR accuracy assessment [1], [2], analysis of new algorithms [3], [4] and new sensor designs [5]–[7], and improvement of image interpretation algorithms [8]–[10].

Simulation is an established practice in the imaging radar community that serves many purposes, including inSAR accuracy assessment [START_REF]Terrain height measurement accuracy of interferometric synthetic aperture radars [END_REF], [START_REF]A time-domain raw signal simulator for interferometric SAR[END_REF], analysis of new algorithms [START_REF]Application of angular correlation function of clutter scattering and correlation imaging in target detection[END_REF], [START_REF]Clutter effects on ground moving target velocity estimation with SAR along-track interferometry[END_REF] and new sensor designs [START_REF]Multi-frequency and multipolarization SAR system analysis with simulation software developed at CSA[END_REF]– [START_REF] Spatial considerations in SAR speckle simulation[END_REF], and improvement of image interpretation algorithms [START_REF]Radar image simulation[END_REF] – [START_REF]Enhanced simulation of radar backscatter from forests using LiDAR and optical data[END_REF].

C. Citation Processed Text.

$$I = sn \tag{2}$$

with a speckle pdf given by

$$P(n) = \exp\{-n\}. \tag{3}$$

[START_FORMULA]I = sn (2)[END_FORMULA]

with a speckle pdf given by

[START_FORMULA]P (n) = exp{−n}.[END_FORMULA]

D. Formula Processed Text.

Figure 2: Tokenization processed text. A. shows an example of a figure marker, we only choose to preserve the captions; B. shows an example of a table marker, we transfer the tables into the form of Markdown; C. shows the tokenization of the citations, we replace the reference numbers into reference papers' title to preserve the readability of the text corpus.

Data source	Document	Tokens
Geoscience papers	1,122,094	3.9B
Geoscience papers Metadata	4,274,716	0.1B
Wikipedia page	767,341	1.5B
Total	**6,164,151**	**5.5B**

Table 1: The details of the text corpus used to train K2.

PDF Parsing. We build an automatic PDF parsing toolkit based on the GROBID library [1]. We use Markdown as the format for all papers in the corpus to preserve readability and consistency. Finally, we use regular expressions and rule-based scripts to clean the data, removing the text obstructing reading, garbled, and impurity data. The script will be released shortly.

Tokenization. Tokenization is an essential part of text corpus design. To make the language model understand the academic papers, we utilize specialized tokens for different modalities as follows, and the examples are shown in Figure 2.

- **Illustrations:** we use special tokens [START_FIGURE] and [END_FIGURE] to annotate the captions of the illustrations in the papers.
- **Tables:** Two special tokens [START_TABLE] and [END_TABLE] are used to locate the position of the table in the passage. In this process, we transform the tables in the PDFs into the format of Markdown.
- **Citations:** We use special tokens [START_REF] and [END_REF] to annotate the citations.
- **Formulas:** For mathematical content or formulas, we filter and clean the irregular formulas parsed from PDFs through regular expressions and rule-based methods. Further we use special tokens [START_FORMULA] and [END_FORMULA] to capture them.

3.2 Instruction Tuning Data: GeoSignal

Further, we collect well-organized instruction tuning data, such as natural instruction [29], AI2 Reasoning Challenge, stanford-alpaca [39], and Dolly-15k [6] for human-alignment and further build up expert-alignment data with a semi-manual pipeline called **GeoSignal**, the statistics of these instruction tuning data are shown in Table 2.

Instruction Tuning Data	Prompts	Data Type
GPT4-Alpaca	52,002	Self-instruct
Dolly-15K	15,011	Task-specific
Natural Instruction	2,446	Task-specific
AI2 Reasoning Challenge	7,787	Task-specific
GeoSignal	82,202	Knowledge Intensive

Table 2: Datasets used to train K2 during the instruction tuning process.

3.2.1 Align-to-Human. Expert is a human who specializes in a given domain; therefore, we collect several well-construct supervised datasets, including self-instruct and human-annotated.

- **Alpaca-GPT4:** Alpaca-GPT4[3] is instruction-following data generated by the techniques named Self-Instruct [45], and all the samples are in the form of *<instruction, input, output>*, which we choose to follow.
- **Dolly-15k:** databricks-dolly-15k [6] is an open-source dataset of instruction-following records generated by thousands of Databricks employees, including brainstorming, classification, closed QA, generation, information extraction, open QA, and summarization. We organize them all into *<instruction, input, output>* format.
- **Natural Instruction:** Natural Instruction [29] maintains many tasks and their natural language definitions/instructions. Its v1.x dataset consists of 61 tasks. The v2.x dataset contains over 1.5k tasks. We select *objective tasks* elaborately from the v2.x dataset and organize them into *<instruction, input, output>* format.
- **AI2 Reasoning Challenge:** AI2 Reasoning Challenge (ARC) [5] is a dataset of *7,787* genuine grade-school level, multiple-choice science questions. As it is well-formed, we sample randomly and organize it into *<instruction, input, output>* format.

3.2.2 Align-to-Expert. Referring to reStructured pre-training [4, 47], signals are the data we can use to train models and usually exist in databases and websites. Many data sources and materials have different types of geoscience signals in geoscience, as illustrated in Figure 3. These signals could be restructured into input-output pairs as instruction tuning samples. For example, with a paper's abstract and title information, we can restructure such signals into

[3] https://github.com/tloen/alpaca-lora

a title generation task given the abstract. In addition, and most importantly, with the support of several applications and products of DDE, we collect a large quantity of geoscience expertise data and re-structure it with prompts into a unified sequence-to-sequence format, namely GeoSignal. The databases and websites we use are as follows:

- **GAKG:** GAKG [7] is a multimodal Geoscience Academic Knowledge Graph organizing geoscience papers' illustrations, text, and bibliometric data.
- **DDE Scholar:** DDE Scholar (https://ddescholar.acemap.info/), a geoscience academic literature search engine, contains more than 3 million papers and 4 million scholars' information in the field of earth sciences.
- **DataExpo:** DataExpo [24] is a one-stop dataset service and has indexed over 960,000 datasets from more than 27,000 repositories in the context of Deep-time Digital Earth Program.
- **GSO:** GSO (https://gso.acemap.info/) is a large-scale ontology of research areas that was automatically generated using the hierarchical topic modeling, which consists of more than 120 thousand research interests in the field of geoscience.
- **Geoscience QA:** We crawler *4* question and answer platform, and *7* geoscience-related databases, using OpenAI [31] for template generation and with the help of the human expert, we finally have a clean and correct geoscience Q&A dataset. The distribution of each part is shown in the Github Repo for the limitation of this manuscript.

For a better understanding of geoscience signals, we list the main signals we take into consideration in bellowing:

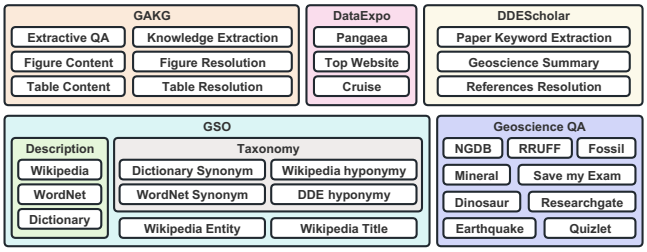

Figure 3: The components of GeoSignal.

- **G1: Paper content:** The title, abstract, full-text of geoscience literature. This signal naturally exists on DDE Scholar, GAKG, and DataExpo, and can be used in summarization tasks.
- **G2: Category:** The category of a geoscience paper or term. This signal typically exists on DDE Scholar, GAKG, and Wikipedia. It can be used for the text classification task.
- **G3: Reference Paper:** This signal exists in the reference lists and introduction of papers and is useful for text comprehension and summarization.
- **G4: Paper table and illustration:** Tables and figures in geoscience papers provide captions and content mentioned in the passage, which can be used for question-answering tasks.
- **G5: Entity mentions:** The entities within a given text. This signal can be found in GAKG and Wikipedia and can be useful for named entity recognition tasks.
- **G6: Relations:** The relationships between different geoscience entities. This information exists in human-annotated datasets such as GAKG and GSO. This signal is useful for finding synonyms and hyponymy terms in geoscience.
- **G7: Word description:** The definition of a word. Various geoscience resources contain this signal, such as Geoscience Dictionary, WordNet, Wikipedia, and GSO. This signal is useful for the task of explanation.
- **G8: Synonyms & Taxonomy:** The Synonyms and hyponymy relation between terms in geoscience. Geoscience Dictionary and GSO contain this signal, which is useful for finding synonyms and hyponymy terms in geoscience.
- **G9: Text Comprehension:** This signal typically exists in geoscience academic platforms and other text material containing question and answer pairs and is useful for question answering.
- **G10: Factual knowledge:** Geoscience facts, e.g., Dolomite is a carbonate rock. This signal typically exists in some geoscience-related QA platforms, useful for question-answering and fact verification.

Based on these signals, we restructure the data for tuning on tasks of *Question Answering, Named Entity Recognition, Relation Extraction, Fact Verification, Summarization, Text Classification, Word Semantics, and Explanation*, and we sample and clean the data to build the instruction tuning data GeoSignal. The statistics of GeoSignal is listed as Table 3.

Tasks	Samples	Total
Question Answering	11,360,163	15,349
Named Entity Recognition	6,252,268	2,400
Relation Extraction	1,200	600
Fact Verification	168,424	8,000
Summarization	3,279,336	800
Text Classification	8,313	2,000
Word Semantics	826,194	6,400
Explanation	731,374	4,200
Entire GeoSignal	**22,627,272**	**39,749**

Table 3: The statistics of GeoSignal categorized by tasks.

3.3 Evaluation on Expertise in Geoscience: GeoBenchmark

Lastly, in order to evaluate the language models for solving geoscience questions and the capacity to understand and utilize the geoscience knowledge, we extract the data from various Question-answer websites, crawl several open-source test websites, and finally construct a benchmark, named **GeoBenchmark**.

NPEE. First, we collected National Postgraduate Entrance Exam questions on geology and geography in the past five years. We choose the text-only questions and translate them into English since the base model is LLaMA. After verifying the translated questions and corresponding answers, we got *182* multiple-choice questions, *150* fill-in-the-blank questions, *454* word-explanation tasks, and *335* essay questions. Since the fill-in-the-blank questions, word-explanation tasks, and essay questions are hard to evaluate, we make them subjective tasks, while the multiple-choice questions are objective tasks.

APTest. we also collect AP (Advanced Placement) examinations are exams offered in the US by the College Board and are taken each May by students. We collect and clean *1395* multiple-choice questions about geology, geography, and environmental science.

To sum up, There are *183* multiple-choice questions in NPEE, and *1,395* in total in AP Test, constituting the objective task set. Meanwhile, we gather all *939* subjective questions in NPEE to be the subjective tasks set and use *50* to measure the baselines with human evaluation. In the experiments sessions, we further discuss the evaluation metrics on these tasks.

4 TRAINING THE K2

In this section, we establish a recipe for tuning a large language model on a specific domain and share the settings we adopt to train the K2.

4.1 Geoscience Domain Adaptation Recipe

Since geoscience is a relatively secondary or arcane field of study, there are few language models for such scenarios. However, advanced natural language models and tools can help geoscientists with data mining and knowledge discovery in their research fields. Therefore, learning a language model for knowledge understanding, summary, and QA is necessary. Meanwhile, geoscience has a rich knowledge accumulation, such as academic papers and scientific reports, which has established a data foundation for training large-scale language models. Consequently, Based on the data in the field of geosciences, we explored a recipe for scientific domain adaptation and finally obtained K2.

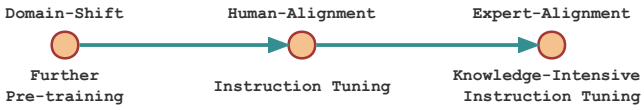

Figure 4: Training recipe for domain language models.

As shown in Figure 4, scientific domain adaptation has three main steps. First, we use domain-specific text corpus to further pretrain the base model. In this paper, we use LLaMA as the base model. Second, since instruction tuning can make the language models generate content following human instructions, we can first do instruction tuning with general instruction-tuning data, such as Alpaca, and natural instruction. Lastly, after learning the paradigm to follow the instructions, the model can learn more information from the restructured domain knowledge, which we call expertise-instruction tuning. In the ablation experiments, we will further verify the correctness of this recipe.

4.2 Further Pretrain

During the stage of further pretrain on geoscience text corpus, We initialize the LLaMA-7B [42] checkpoints with the 8-bit integer format (int8) parameters, using bf16 and tf32 as the floating point formats, and train it on *3.9B* tokens from *2,455,040* samples consisting more than *2 million* well-preprocessing geoscience literature.

The entire parameters of LLaMA-7B (6.7B trainable parameters) are further pretrain for one epoch on 4 NVIDIA A100-SXM-40GB GPUs and the training for *214 hours*. In this stage, we set a learning rate of 1e-5, with a global batch size of 128 and a micro-batch size of 2. The incremental steps of the train are *30,140 steps* (1,000 for warm-up). Finally, we will call the model obtained after the further pretrain *GeoLLaMA* for better distinction.

4.3 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) helps achieve the mission of training in a low-resource setting. As mentioned in [16], the weight updates during the fine-tuning process also have a low "intrinsic rank" during adaptation. Therefore, according to LoRA, a hidden layer $h = W_0x, W_0 \in R^{d*k}$, the modified forward pass yields: $h = W_0x + \Delta Wx = W_0x + BAx$, where $B \in R^{d*r}, A \in R^{r*k}$ and $r << min(d, k)$ are two low "intrinsic rank" matrix containing trainable parameters. Moreover, after further pretraining the LLaMA, the adaptation to the field of geoscience is more comprehensive. During the instruction tuning stage, the target is to train the model to align with humans and experts. We use Low-Rank Adaption to tune the model.

In instruction tuning, we set a learning rate of 1e-4 with a global batch size of 128. As for the LoRA setup, we set lora_r as eight while lora_alpha as 16. We set the lora_target_modules as k_proj, q_proj, and v_proj, based on our experimental observation. The instruction tuning via LoRA only trains *6M* parameters on one single NVIDIA GeForce RTX 3090 for 23 hours. In order to make the model perform better and inject part of the geoscience knowledge in the instruction tuning stage, we first use alpaca instruction tuning data to train GeoLLaMA, which we recognize as Human-alignment. Then we resume from the checkpoint obtained and continue fine-tuning the model using GeoSignal for further training. Based on our experimental observation, the performance does not perform better if we mix these training data.

5 EVALUATION

This section illustrates the evaluation methods and results of K2 and related baselines. **GeoBenchmark** consists of two kinds of tasks, one is subjective, and one is objective. We will release our evaluation pipeline on Github Repo. In this part, we choose four baselines models: Galactica [40], MPT-7B [41], Vicuna-7B [3], LLaMA-7B [42] and Alpaca-7b [39].

5.1 Objective tasks in GeoBenchmark

For objective tasks like multiple-choice tasks (GeoBenchmark-AP and multiple-choice and true-false questions in GeoBenchmark-GNPEE), we prompt appropriately, ending with the phrase "*The answer is*", calculate the $Softmax$ of the probability of next token among the choice label (*e.g., A, B, C, D, sometimes E*), and finally gain the score of the **Accuracy** based on these test ground truth.

First, we evaluate all the saved checkpoints, as shown in Figure 5. We can find that as the tokens seen by the model gradually scale up, the model's performance on our benchmark is improving. This result indicates that the model learns geoscience knowledge in further pretrain.

Moreover, compared with the baselines, we can see that K2 outperforms the NPEE dataset. However, in the AP Test, K2 is similar to the Galactica model since geoscience learned in high school is human geography and environmental science, including in the training corpus of Galactica.

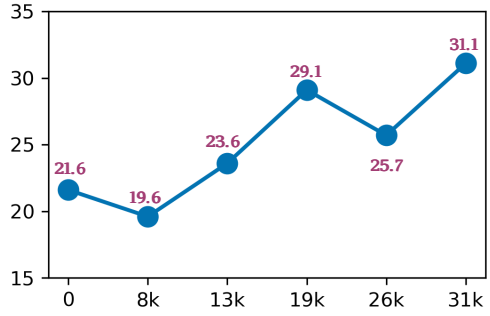

Figure 5: Each score at selected training steps of K2 on the Objective tasks in GeoBenchmark.

Baselines	NPEE	APTest
Gal-6.7B	25.7	**29.9**
LLaMA-7B	21.6	27.6
MPT-7B	28.4	26.0
Vicuna-7B	26.4	16.8
Alpaca-7B	31.1	29.1
K2-7B (Ours)	**39.9**	29.3

Table 4: Comparison among baselines on Objective tasks in GeoBenchmark. The best number is bolded, while the second best is underlined.

5.2 Subjective tasks in GeoBenchmark

For subjective tasks (mainly in GeoBenchmark-NPEE), we use automatic methods, **GPTScore** [11] and **Perplexity** to evaluate the quality of the output. GPTScore utilizes generative pre-trained models' emergent abilities (e.g., zero-shot instruction) to score generated texts. In addition, perplexity is computed with GPT-2 [34] on the generated text and measures the fluency of the generations. Furthermore, referring to geoscientists, we collect 50 open geoscience questions and gather ten geoscience research practitioners to evaluate the output of baseline models. We evaluate the models on three metrics, 1) rationality: whether the generated content of the model is technical rationality or not; 2) correctness: whether the content generated by the model is reliable or not; 3) consistency: whether the generated content always stays in the topic. All the scores scale from 1 (poor) to 3 (good), with 2 indicating acceptable content. The complete results of the subjective tasks are in Table 5.

Baselines	Automatic Evaluation		Human Evaluation		
	perplexity	GPTScore	rationality	correctness	consistency
Gal-6.7B	34.57	-2.3598	1.96	1.74	1.79
LLaMA-7B	40.07	-1.9531	2.24	2.04	2.01
GeoLLaMA-7B	**32.32**	**-1.9457**	2.15	1.89	2.03
Alpaca-7B	40.07	-1.9536	2.09	1.93	**2.34**
K2-7B (Ours)	**32.32**	-1.9487	**2.38**	**2.13**	2.14

Table 5: Comparsion on subjective tasks in GeoBenchmarks. The best number is bolded, while the second best is underlined.

As we can see, K2 performs better on rationality and correctness. At the same time, consistency stays competitive. The results indicate that our model better understands geoscience and can utilize scientific knowledge.

5.3 Ablation on Expert-Alignment

To better understand the recipe for aligning the model with humans and experts, we deploy the ablation experiments to explore the detail. We treat the data constructed by self-instruct or human-annotated in the general domain or for dialogue generation as human-alignment data. At the same time, view the data annotated by experts in specific domains as expert-alignment data. As shown in Table 6, using task-special data, such as dolly-15k, fails to achieve a good performance, while using self-instruct data, such as Alpaca-GPT4, is still not as effective as using knowledge-intensive data. Surprisingly, we have discovered that the results are unsatisfactory if we mix the knowledge-intensive data GeoSignal with human-alignment data Alpaca. It is better to use Alpaca to align the model to follow human instruction and then use the GeoSignal to align with the experts. Here LoRA is deployed only on attention layers.

Model	NPEE	APTest
GeoLLaMA → Dolly	27.0	26.3
GeoLLaMA → Alpaca-GPT4	34.4	26.5
GeoLLaMA → GeoSignal	37.2	27.4
GeoLLaMA → GeoSignal mix Alpaca-GPT4	33.8	23.4
GeoLLaMA → Alpaca-GPT4 → GeoSignal (K2)	**39.9**	**29.8**

Table 6: Results when using different instruction tuning data. The best number is bolded, while the second best is underlined.

6 CONCLUSION AND DISCUSSION

In this paper, we introduce K2, the first-ever large language model in the geoscience field. K2 can answer geoscience questions and follow geoscientists' instructions with its geoscience professionalism. We construct the first geoscience-supervised instruction data, GeoSignal. Meanwhile, we build GeoBenchmark, the first NLP benchmark in geoscience to evaluate the capability on geoscience knowledge understanding and utilization. On the geoscience benchmarks collected, K2 shows its professionalism and effectiveness compared with other similar-size language models.

Limitations. Like other language models, hallucination, toxicity, and stereotypes exist in K2. Since LLaMA's pretraining data are from before 2020 and K2 only use open-access geoscience papers, the information, and knowledge may not be up-to-date. Moreover, LLaMA only supports 20 languages and has limited support for non-English languages, which K2 inherits as well.

Potential Applications. K2 shows the potential of adapting language models to a scientific field with domain barriers. As a language model, K2 can understand geoscience materials and modify the statement about geoscience with suitable prompts. Since K2 is a generative language model, it can generate paragraphs and statements on word description and answer generation based on the given questions. In this way, K2 acts like a knowledge base and gives the geoscientist a professional assistant.

In the future, we will maximize the advantages and minimize the limitation of using K2 and provide better services to data mining communities and geoscientists for further research.

REFERENCES

[1] 2008–2023. GROBID: A machine learning software for extracting information from scholarly documents. https://github.com/kermitt2/grobid.

[2] 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*.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://vicuna.lmsys.org

[4] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *ArXiv* abs/2210.11416 (2022).

[5] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *ArXiv* abs/1803.05457 (2018).

[6] Databricks. 2023. Hello Dolly: Democratizing the magic of ChatGPT with open models. https://www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html

[7] 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* (2021).

[8] Huseyin Denli, HassanJaved Chughtai, Brian Hughes, Robert Gistri, and Peng Xu. 2021. Geoscience Language Processing for Exploration. *Day 3 Wed, November 17, 2021* (2021).

[9] Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang-Wei Zhang, and Yao Xu. 2023. A Multi-Modal Geographic Pre-Training Method. *ArXiv* abs/2301.04283 (2023).

[10] Majigsuren Enkhsaikhan, Wei Liu, Eun-Jung Holden, and Paul Duuring. 2021. Auto-labelling entities in low-resource text: a geological case study. *Knowledge and Information Systems* 63 (2021), 695 – 715.

[11] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. *ArXiv* abs/2302.04166 (2023).

[12] Leo Gao, Stella Rose Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv* abs/2101.00027 (2020).

[13] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A Dialogue Model for Academic Research. Blog post. https://bair.berkeley.edu/blog/2023/04/03/koala/

[14] Tanishq Gupta, Mohd Zaki, N. Krishnan, and Mausam. 2021. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials* 8 (2021), 1–11.

[15] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *International Conference on Machine Learning*.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv* abs/2106.09685 (2021).

[17] Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. 2022. ERNIE-GeoL: A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022).

[18] Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *ArXiv* abs/2001.08361 (2020).

[19] Zeljko Kraljevic, Anthony Shek, Daniel M Bean, Rebecca Bendayan, James T. H. Teo, and Richard J. B. Dobson. 2021. MedGPT: Medical Concept Prediction from Clinical Narratives. *ArXiv* abs/2107.03134 (2021).

[20] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. *ArXiv* abs/2104.08691 (2021).

[21] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* abs/2101.00190 (2021).

[22] Xiao Liu, Juan Hu, Qi Shen, and Huan Chen. 2021. Geo-BERT Pre-training Model for Query Rewriting in POI Search. In *Conference on Empirical Methods in Natural Language Processing*.

[23] S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. *ArXiv* abs/2301.13688 (2023).

[24] Bin Lu, Lyuwen Wu, Lina Yang, Chenxing Sun, Wei Liu, Xiaoying Gan, Shiyu Liang, Luoyi Fu, Xinbing Wang, and Cheng Zhou. 2023. DataExpo: A One-Stop Dataset Service for Open Science Research. *Companion Proceedings of the ACM Web Conference 2023* (2023).

[25] 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* (2022).

[26] Kai Ma, Miao Tian, Yongjian Tan, Xuejing Xie, and Qinjun Qiu. 2021. What is this article about? Generative summarization with the BERT model in the geosciences domain. *Earth Science Informatics* 15 (2021), 21 – 36.

[27] Xiaogang Ma, Chao Ma, and Chengbin Wang. 2020. A new structure for representing and tracking version information in a deep time knowledge graph. *Comput. Geosci.* 145 (2020), 104620.

[28] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, G. Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2023. On the Opportunities and Challenges of Foundation Models for Geospatial Artificial Intelligence. *ArXiv* abs/2304.06798 (2023).

[29] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Natural Instructions: Benchmarking Generalization to New Tasks from Natural Language Instructions. *arXiv preprint arXiv:2104.08773* (2021).

[30] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Haiquan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis.

[31] OpenAI. 2022. Introducing ChatGPT. (2022). https://openai.com/blog/chatgpt

[32] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).

[33] José Padarian and Ignacio Fuentes. 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. *SOIL* (2019).

[34] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[35] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv* abs/1910.10683 (2019).

[36] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Rose Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. *ArXiv* abs/2110.08207 (2021).

[37] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Bio-Megatron: Larger Biomedical Domain Language Model. *ArXiv* abs/2010.06060 (2020).

[38] K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Lee Kai Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Y. Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomavsev, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. *ArXiv* abs/2212.13138 (2022).

[39] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

[40] 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 (2022).

[41] MosaicML NLP Team. 2023. Introducing MPT-7B: A New Standard for Open-Source, ly Usable LLMs. (2023). www.mosaicml.com/blog/mpt-7b

[42] 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 (2023).

[43] Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, and Jie Fu. 2021. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *ArXiv* abs/2110.05006 (2021).

[44] Chengshan Wang, Robert M. Hazen, Qiuming Cheng, Michael H. Stephenson, Chenghu Zhou, Peter A. Fox, Shu-zhong Shen, Roland Oberhänsli, Zeng-qian Hou, Xiaogang Ma, Zhiqiang Feng, Junxuan Fan, Chao Ma, Xiumian Hu, Bin Luo, Juanle Wang, and Craig M. Schiffries. 2021. The Deep-Time Digital Earth

program: data-driven discovery in geosciences. *National Science Review* 8 (2021).
[45] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. *ArXiv* abs/2212.10560 (2022).
[46] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. *ArXiv* abs/2304.01196 (2023).
[47] Weizhe Yuan and Pengfei Liu. 2022. reStructured Pre-training. *ArXiv* abs/2206.11147 (2022).
[48] 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 (2022).