SciDFM: A Large Language Model with Mixture-of-Experts for Science

Liangtai Sun[1] **Danyu Luo**[1] **Da Ma**[1] **Zihan Zhao**[1] **Baocai Chen**[1]

Zhennan Shen[1] **Su Zhu**[3] **Lu Chen**[1,2]* **Xin Chen**[2] **Kai Yu**[1,2]*

[1]X-LANCE Lab, Department of Computer Science and Engineering
MoE Key Lab of Artificial Intelligence, SJTU AI Institute
Shanghai Jiao Tong University, Shanghai, China
[2]Suzhou Laboratory, Suzhou, China
[3]AI Speech Co, .Ltd., Suzhou, China
{slt19990817, chenlusz, kai.yu}@sjtu.edu.cn

Abstract

Recently, there has been a significant upsurge of interest in leveraging large language models (LLMs) to assist scientific discovery. However, most LLMs only focus on general science, while they lack domain-specific knowledge, such as chemical molecules and amino acid sequences. To bridge these gaps, we introduce SciDFM, a mixture-of-experts LLM, which is trained from scratch and is able to conduct college-level scientific reasoning and understand molecules and amino acid sequences. We collect a large-scale training corpus containing numerous scientific papers and books from different disciplines as well as data from domain-specific databases. We further fine-tune the pre-trained model on lots of instruction data to improve performances on downstream benchmarks. From experiment results, we show that SciDFM achieves strong performance on general scientific benchmarks such as SciEval and SciQ, and it reaches a SOTA performance on domain-specific benchmarks among models of similar size. We further analyze the expert layers and show that the results of expert selection vary with data from different disciplines. To benefit the broader research community, we open-source SciDFM at `https://huggingface.co/OpenDFM/SciDFM-MoE-A5.6B-v1.0`.

1 Introduction

The advent of Large Language Models (LLMs) [1, 2, 3] has ignited a revolution in the realm of artificial intelligence and has pushed the field of AI for Science (AI4S) to an unprecedented new height. LLMs have demonstrated promising performances in assisting and accelerating scientific discovery [4, 5, 6, 7], such as protein design [8], weather forecasting [9], and geoscience [10]. Despite remarkable achievements in science, LLMs primarily focus on general scientific knowledge represented in text form, ignoring domain-specific contents such as molecules in chemistry and proteins in biology, which are fundamental to advances in these fields.

To overcome this limitation and fully exploit the potential of LLMs for scientific discovery, we introduce SciDFM, a mixture-of-experts Large Language Model trained from scratch with 18.2

*The corresponding authors are Lu Chen and Kai Yu.

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

billion parameters in total and 5.6 billion parameters activated. SciDFM integrates a mixture-of-experts (MoE) [11, 12, 13] architecture into a transformer-based [14] framework, aiming at enhancing its sophisticated scientific reasoning and understanding capabilities and better modeling similarities and differences across different disciplines and modalities, i.e. text, molecule, protein. In this paper, we detail the pretraining and instruction tuning process of SciDFM, including training corpus and settings. SciDFM leverages a carefully curated corpus of scientific literature and domain-specific databases for pretraining to capture vast scientific knowledge, and is also trained on a large general corpus to retain general knowledge, consuming about 1.1T tokens in total. We meticulously fine-tune SciDFM using a set of instruction-following data containing about 9.3M samples, including interpreting molecular structures and amino acid sequences, thereby improving the performance on downstream benchmarks and bridging the gap in domain-specific knowledge.

To illustrate the prowess of SciDFM, we conduct extensive experiments on several scientific benchmarks. Empirical evaluations affirm the efficacy of SciDFM, demonstrating its superiority on both general scientific benchmarks like SciEval [15] and SciQ [16], as well as achieving state-of-the-art (SOTA) performance in domain-specific tasks among models of similar size, such as Mol-Instructions [17]. Our analysis further delves into the results of the expert layer selection, revealing their adaptability to different scientific disciplines, and demonstrating the effectiveness of the MoE model in multidisciplinary scenarios. To benefit the broader research community, we will make SciDFM openly accessible.

2 SciDFM

In this section, we introduce the pretraining and instruction tuning details of SciDFM, including the training data construction, model architecture and infrastructure.

2.1 Pretraining

2.1.1 Model Architecture

SciDFM is based on a transformer architecture [14], and follows modifications of Llama [2], i.e. RMSNorm, RoPE and SwiGLU. SciDFM uses the same hyper-parameters of OpenLLaMa-3B [18], the details are shown in Table 1. And in order to better model knowledge of different disciplines, we replace the feed-forward block with Mixture-of-Expert (MoE) layers [11, 12].

We also use the tokenizer of OpenLLaMa-3B, which is trained from scratch using the Bype-Pair Encoding (BPE) method. To better encode the molecules and amino acid sequences and distinguish them from normal characters for better modeling, we treat each chemical atom and amino acid character as a single token and add them into the vocabulary, with special identifiers wrapped [4]. For example, molecules C(C(=O)O)N will be encoded as C,(,C,(,=,O,),O,),N, and amino acid sequences MIRLGAPQTL will be encoded as M,I,R,L,G,A,P,Q,T,L, where the special identifiers are omitted.

Parameter	Value
dim	3200
n_layers	26
head_dim	100
ffn_hidden_dim	8640
n_heads	32
n_kv_heads	32
context_len	8192
vocab_size	32192
num_experts	8
topk_experts	2

Table 1: Hyper-parameters of SciDFM

2.1.2 Data Construction

To enhance the understanding and reasoning abilities of SciDFM on science domain, we collect a large-scale training corpus containing a large number of open-access scientific papers and books of different disciplines. And to acquire domain-specific knowledge, we also include data from some databases. Furthermore, in order to maintain the generic capabilities of SciDFM, we use data from some open-source generic corpora. The details of our pretraining corpus are shown in Table 2. Our pretraining corpus contains about 300B science-domain tokens and 270B general-domain tokens, with 570B tokens in total. We train SciDFM for two epochs, and for the second epoch, we re-sample data of C4, CC, and Github subsets of SlimPajama [24, 25]. And for the-stack dataset, we only

Data Source	Domain	# Tokens (B)
AMPS [19]	Math	1.07
OEIS	Math	0.08
Proof-Pile-2 [20]	Math	25.32
MathGLM-dataset [21]	Math	9.66
MathPile [22]	Math	0.21
PubChem Compound	Chemistry	18.34
USPTO	Chemistry	0.13
Uniprot	Biology	0.52
BioRxiv	Biology	2.77
MedRxiv	Biology	0.31
RefSeq Genome	Biology	1.38
Uniref	Biology	16.85
GeoNames	Geography	0.82
the-stack [23]	Science Code	9.65
Scientific Papers	General Science	182.65
Scientific Books	General Science	7.00
SlimPajama-Arxiv [24]	General Science	29.78
LibreText	General Science	0.05
Total Science	-	306.59
WikiBooks	General	0.14
En-Zh-Trans	General	3.10
Wikipedia-Zh	General	0.56
Baike	General	23.16
SlimPajama(w/o Arxiv) [24]	General	241.64|241.75
Total General	-	268.60|268.71

Table 2: The training corpus details of SciDFM.

use data from programming languages that are relevant to scientific computing, such as Matlab and Scilab.

2.1.3 Training Details

Following Llama [2], SciDFM is trained from scratch using the AdamW optimizer [44] with $\beta_1 = 0.9, \beta_2 = 0.95$. We use a cosine learning rate schedule, such that the final learning rate is equal to 10% of the initial learning rate. We use a weight decay of 0.1 and gradient clipping of 1.0, and we set the macro batch size to 4M tokens and sequence length to 8192 tokens. For MoE layers, we set the auxiliary loss factor to 0.02 and set the expert capacity factor to 1.0. In total, we train SciDFM for two epochs, resulting in about 1.1T tokens fed. We use a learning rate of $3e-4$ for the first epoch and $3e-5$ for the second epoch, while the other settings remain the same as the above. We train SciDFM on a cluster of 16 nodes, with 8 A800 GPUs on each node for about two months.

2.2 Instruction Tuning

To improve the performance of SciDFM on downstream benchmarks, we collect a number of instruction-tuning data from open-source datasets. The details are shown in Table 3. We fine-tune the pre-trained SciDFM for 5 epochs in total. During fine-tuning, we use a learning rate of 2e-5, and we set the sequence length to 2048 and macro batch size to 32. The other settings are the same as the pretraining stage.

3 Evaluation

In this section, we show the performance of SciDFM on some general and domain-specific benchmarks, and analyze the results of expert selection in different domains.

Data	Domain	# Samples
Arxivphy	Physics	30231
MATH [19]	Math	7500
MetaMathQA [26]	Math	395000
MathInstruct [27]	Math	262040
OrcaMath [28]	Math	200035
DeepLoc2 [29]	Biology	22642
BioASQ [30]	Biology	8021
MedMCQA [31]	Biology	182822
MedQA [32]	Biology	10178
PubMedQA [33]	Biology	500
Mol-Instructions [17]	Chemistry & Biology	1863630
ChemDFM-sft [5]	Chemistry	1818154
SciQ [16]	General Science	11679
Camel [34]	General Science	110000
SciInstruct [7]	General Science	82640
WebInstructSub [35]	General Science	2327291
MMLU [36]	General	99842
LIMA [37]	General	1000
ROPES [38]	General	10924
QASC [39]	General	8134
OpenBookQA [40]	General	4957
Dolly [41]	General	15011
SlimOrca-Dedup [42]	General	363491
GPT4All [43]	General	806199
Other	General	994896
Total(after dedup)	-	9325310

Table 3: Details of instruction tuning dataset.

Model	①	②	③	④	⑤	⑥	⑦	⑧	Avg.
Llama2-7B	27.06	57.00	36.43/46.59	3.94	3.96	26.32	29.84	66.80	32.95
Galactica-6.7B	46.28	74.20	44.28/61.83	2.80	6.32	30.48	36.46	48.80	38.91
Llama2-13B	33.88	78.10	56.66/72.35	22.82	3.90	32.68	34.28	77.80	45.45
ChatGLM2-6B	54.25	75.80	57.08/73.57	25.09	7.18	27.42	34.21	60.40	45.94
Galactica-30B	54.24	83.10	57.85/75.04	13.65	8.66	37.71	48.43	58.80	48.35
Llama3-8B	59.70	90.00	71.16/84.05	5.91	7.00	48.78	52.74	26.60	49.59
ChatGLM3-6B	51.13	77.60	60.84/75.97	60.27	23.52	24.59	31.39	51.80	50.53
SciGLM-6B	61.22	88.70	77.47/86.57	42.23	16.40	42.81	44.94	73.60	59.12
ChatGLM3-6Bb	60.34	89.00	**78.58/87.37**	59.82	22.64	42.73	45.14	74.40	61.96
Llama3-8B-it	**64.91**	**91.60**	76.45/87.33	**76.57**	26.26	**56.48**	**59.31**	72.00	67.44
SciDFM(ours)	62.48	88.00	64.76/81.48	59.14	**27.28**	44.54	53.10	**78.00**	61.56

Table 4: Main Results on 8 general scientific language understanding and reasoning tasks. ChatGLM3-6Bb stands for `ChatGLM3-6B-base`, Llama3-8B-it stands for `Llama3-8B-Instruct`. The columns from ① to ⑧ stand for SciEval, SciQ, ARC, GSM8K, MATH, MedQA, MedMCQA. PubMedQA, respectively, where the results of ARC are shown in the form of ARC_C/ARC_E.

.

3.1 Evaluation Setup

Evaluation Tasks Summarization of evaluation tasks are shown in Table 5. These evaluation datasets cover a wide range of subjects, including math, chemistry, physics, biology, protein and molecule.

- SciEval [15] is a comprehensive and multidisciplinary benchmark designed to assess the scientific research capabilities of Large Language Models.

Dataset	# Samples	Subject
SciEval	15,901	♢, △, □, ♡
SciQ	1,000	♢, △, □, ♡
ARC_C	1,172	♢, △, □, ♡
ARC_E	2,376	♢, △, □, ♡
GSM8K	1,319	□
MATH	5,000	□
MedQA	1,273	△
MedMCQA	4,183	△
PubMedQA	500	△
Mol-Instructions	24,324	♢, △
MoleculeNet	11,767	♢

Table 5: Overview of evaluation benchmarks. (♢: Chemistry, △: Biology, □: Math, ♡: Physics)

Model	Avg.G	Avg.M	Avg.B
LLaMa2-7B	41.77	3.95	40.99
Galactica-6.7B	56.65	4.56	38.58
LLaMa2-13B	60.25	13.36	48.25
ChatGLM2-6B	65.18	16.14	40.68
Galactica-30B	67.56	11.16	48.31
LLaMa3-8B	76.23	6.46	42.71
ChatGLM3-6B	66.39	41.90	35.93
SciGLM-6B	78.49	29.32	53.78
ChatGLM3-6B-base	78.82	41.23	54.09
Llama3-8B-Instruct	80.07	51.42	62.60
SciDFM(ours)	74.18	43.21	58.55

Table 6: Average Results of General Science, Math and Biology tasks on scientific language understanding and reasoning tasks. Avg.G, Avg.M and Avg.B stand for average performance on general science, math and biology tasks respectively.

- SciQ [16] comprises high-quality, domain-targeted multiple-choice science exam questions, created through a novel method that combines crowd-sourcing with AI-assisted document and answer option selection.
- ARC [45] presents a sophisticated question set, designed to advance AI research in complex question answering within the context of grade-school science, composed of an easy subset and a challenging subset.
- GSM8K [46] is composed of diverse and linguistically rich grade school math word problems, designed to benchmark and improve the multi-step mathematical reasoning abilities of LLMs, revealing their limitations in handling complex reasoning tasks.
- MATH [19] contains challenging competition mathematics problems, each with detailed solutions, designed to assess and enhance mathematical problem-solving and reasoning capabilities of LLMs, accompanied by an auxiliary pretraining dataset to bolster fundamental math understanding.
- MedQA [32] is a pioneering multiple-choice dataset for open-domain question answering in the medical field, encompassing a number of questions sourced from professional medical board exams.
- MedMCQA [31] is a large-scale medical multiple-choice question-answering dataset, spanning 2,400 healthcare topics and 21 subjects, designed to challenge models with diverse reasoning skills across various medical domains.
- PubMedQA [33] is a biomedical question-answering dataset based on PubMed abstracts, requiring quantitative reasoning over research texts.
- Mol-Instructions [17] is a specialized, meticulously curated dataset containing diverse biomolecular instructions, designed to enhance LLMs' understanding and prediction capabilities within the realms of molecular, protein, and broader biomolecular texts.

- MoleculeNet [47] is a comprehensive benchmark dataset for molecular machine learning, featuring curated public datasets, standardized evaluation metrics, and open-source implementation of various molecular featurization and learning methods.

In our experiments, we take Mol-Instructions and MoleculeNet as domain-specific tasks, and take the remaining benchmarks as general scientific language understanding and reasoning tasks.

Model	bace	bbbp	ClinTox	HIV	Tox21	Avg
LLaMa2-13B-chat	26.0	60.3	45.7	29.0	51.7	42.54
GPT4o(0513)	62.5	61.5	51.6	65.9	55.2	59.34
Galactica-30B	72.7	59.6	82.2	75.9	68.5	71.78
ChemDFM-13B	**78.4**	**66.7**	89.9	**73.6**	**79.8**	77.68
SciDFM(ours)	76.4	64.8	**98.5**	71.5	75.6	77.36

Table 7: The Results of molecular property prediction tasks on MoleculeNet in AUC-ROC scores.

Evaluation Methods Since SciDFM is an instruction-following model by default, we conduct all experiments using zero-shot settings. And most of the models we select for comparison are able to follow instructions:

- Galactica [4] is a large language model specifically designed to store, combine, and reason about vast amounts of scientific knowledge, outperforming existing models on various scientific tasks and aiming to serve as a new, advanced interface for scientific research. We select `Galactica-30B` and `Galactica-6.7B` for comparison.
- Llama [2] is a series of open-source powerful language models, ranging from 7 billion to 70 billion parameters, trained on massive public datasets, and outperforms many of the available open-source models on common benchmarks. We select `Llama3-8B`, `Llama3-8B-Instruct`, `Llama2-7B` and `Llama2-13B` for comparison.
- ChatGLM [48, 49] is a series of advanced language models, excel in various metrics and tasks, rivaling or surpassing counterparts like GPT-4, thanks to their extensive training on multilingual data, specialized alignment techniques, and the ability to integrate diverse tools dynamically. We select `ChatGLM2-6B`, `ChatGLM3-6B` and `ChatGLM3-6B-base` for comparison.
- SciGLM [7] is a suite of scientific language models that enhance college-level scientific reasoning through a self-reflective instruction annotation framework, addressing data scarcity in the science domain, and improving upon ChatGLM in handling complex scientific and mathematical problems without compromising language understanding. We select `SciGLM-6B` for comparison.
- ChemDFM [5] is specifically trained for Chemistry, combining knowledge from chemical literature and general domains to excel in understanding, reasoning, and applying chemical information, outperforming generic LLMs and even GPT-4 on chemical tasks. We select `ChemDFM-13B` for comparison.

3.2 Main Results

General Scientific Benchmark Table 4 presents the evaluation results on eight general scientific language understanding and reasoning tasks. The results show that SciDFM reaches a better performance on average than Galactica-series models, Llama-series models except `Llama3-8B-Instruct` and ChatGLM-series models except `ChatGLM3-6B-base`. In Table 6, we also present the average performance of general science, math and biology tasks on the above eight benchmarks, in which SciEval, SciQ and ARC belong to general science task, GSM8K and MATH belong to math task, MedQA, MedMCQA and PubMedQA belong to biology task. We find that SciDFM outperforms all models except `Llama3-8B-Instruct` on math and biology domain, while it is weak in general science tasks. In conclusion, SciDFM can reach a similar performance to top-tier models of similar amount of compute, while it is weaker compared to models that are larger and trained using more data.

Domain-specific Scientific Benchmark Table 7 presents the performance of molecular property prediction tasks on MoleculeNet. From the results shown in AUC-ROC scores, we find that SciDFM

Model	Exact↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Description Guided Molecule Design						
ChatGLM3-6B	0.001	199.47	0.103	0.174	0.059	0.236
Galactica-6.7B	0.000	44.152	0.134	0.248	0.088	0.992
Text+Chem T5	**0.097**	41.819	0.352	0.474	0.353	0.721
Mol-Instructions	0.002	**41.367**	0.231	0.412	0.147	**1.000**
SciDFM(ours)	0.084	43.414	**0.586**	**0.704**	**0.443**	0.994
Reagent Prediction						
ChatGLM3-6B	0.000	154.73	0.044	0.144	0.046	0.275
Galactica-6.7B	0.000	35.021	0.156	0.257	0.097	0.946
Text+Chem T5	**0.239**	20.413	**0.705**	**0.789**	**0.652**	0.762
Mol-Instructions	0.045	27.262	0.313	0.509	0.262	**1.000**
SciDFM(ours)	0.192	**17.527**	0.476	0.576	0.452	0.999
Retrosynthesis						
ChatGLM3-6B	0.000	59.062	0.636	0.695	0.570	0.492
Galactica-6.7B	0.000	30.760	0.036	0.127	0.051	0.995
Text+Chem T5	0.000	49.323	0.039	0.186	0.052	0.313
Mol-Instructions	0.044	23.167	0.237	0.364	0.213	**1.000**
SciDFM(ours)	**0.665**	**6.45**	**0.916**	**0.937**	**0.888**	0.998

Table 8: The Results of molecular generation tasks on Mol-Instructions. Exact stands for exact matches, and validity stands for valid molecules. RDK, MACCS and Morgan are three kinds of molecular fingerprints.

Model	Protein Function	Functional Description	Catalytic Activity	Domain/Motif
GPT4o(0513)	0.06	0.05	0.07	0.06
ChatGLM	0.15	0.14	0.13	0.10
Mol-Instruction	0.43	0.44	0.52	0.46
SciDFM(ours)	**0.60**	**0.72**	**0.76**	**0.55**

Table 9: The Results of protein understanding tasks on Mol-Instructions. All tasks are evaluated in ROUGE-L score.

outperforms most LLMs except `ChemDFM-13B`. To further evaluate model's ability on domain-specific benchmarks, we present the performance of molecule and protein understanding tasks on Mol-Instructions in Table 8 and Table 9. We find that SciDFM can reach a SOTA performance on molecule and protein understanding tasks.

3.3 Expert Choices Analysis

In this subsection, we conduct analysis on expert choice results on data from different domains. Formally, we denote the output of the ith attention layer as $h_i \in \mathbb{R}^{l \times d}$, where l is the sequence length and d is the hidden dimension, and we denote the weight of the ith gate network in the corresponding MoE layer as $W_g \in \mathbb{R}^{d \times e}$, where e represents the number of experts. Then, we have $g_i = h_i \cdot W_g \in \mathbb{R}^{l \times e}$, representing the probability of each token being assigned to each expert. Suppose the number of hidden layers is N, for a given text T, we define the expert choice results as:

$$e_i = \text{Softmax}(\sum_{j=1}^{l} g_i[j,:]) \in \mathbb{R}^e, \tag{1}$$

$$E_T = \text{Concat}([e_1, e_2, \ldots, e_N]) \in \mathbb{R}^{Ne}. \tag{2}$$

We randomly select 100 research papers each in the fields of math, chemistry, biology, and physics, and also select 100 chemical molecules and 100 amino acid sequences as analysis data. For each text

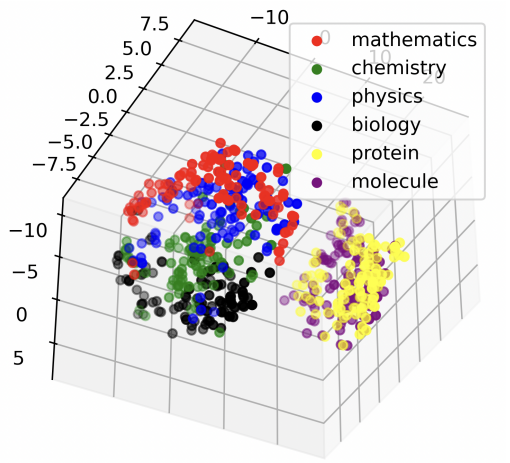

Figure 1: Visualization of t-SNE results using data from different domains.

T, we calculate E_T using the above formulas. Then, we use the t-SNE [50] algorithm to reduce them to three dimensions and visualize them.

The visualization result is shown in Figure 1. It can be found that research papers in mathematics, chemistry, physics, and biology demonstrate a clear clustering pattern, indicative of discipline-specific language characteristics, while chemical molecules and amino acid sequences do not exhibit such a clustering phenomenon. In addition, the separation of molecular and protein data from other categories is stark, likely due to the unique vocabularies inherent to these domains, which are not shared with the remainder of the dataset.

Furthermore, the spatial proximity observed in the visualization adds another layer of insight: the clusters of mathematics and physics are in close proximity, as are those of chemistry and biology, with chemistry also showing affinity towards physics. This aligns well with the interrelationships and overlapping nature of knowledge between these scientific disciplines, as reflected in their linguistic characteristics.

4 Related Works

The success of pretraining language models like BERT [51] and GPT [52] makes researchers wonder whether the language model can bring about improved performance in the field of Science.

Domain-Specific Language Model for Science BioGPT [53] is a domain-specific generative language model pre-trained on large-scale biomedical literature, which outperforms previous models on six biomedical-related tasks. Based on case studies, the researchers further demonstrated the advantages of BioGPT in generating fluent descriptions for biomedical terms in biomedical literature. ProGen2 [54] is a protein language model pre-trained on a corpus of more than one billion protein sequences including genome, metagenome, and immune library databases. ProGen2 shows optimal performance in capturing observed evolutionary sequence distributions, generating new protein sequences, and predicting protein fitness without additional fine-tuning. Med-PaLM [55] is a large language model (LLM) designed to provide high-quality answers to medical questions, which is an instruction prompt-tuned version of Flan-PaLM [56] specialized for the medical domain. They reveal limitations of Flan-PaLM in scientific grounding, harm, and bias through evaluation, while Med-PaLM significantly reduces the gap (or even compares favorably) to clinicians on several of these axes, according to both clinicians and lay users. MTL-BERT [57] proposes to use large-scale pre-training, multi-task learning, and SMILES enumeration to alleviate the data sparsity problem. It mines the rich contextual information in SMILES strings through self-supervised pre-training, and then fine-tunes the pre-trained model simultaneously using multiple downstream tasks. At the same time, it combines SMILES enumeration as a data augmentation strategy to increase data diversity. Experimental results show that MTL-BERT can achieve optimal performance on molecular datasets. ChemDFM [5] is the first LLM towards Chemical General Intelligence (CGI), which is trained on

34B tokens from chemical literature, textbooks, and instructions as well as various data from the general domain. It can store, understand, and reason over chemical knowledge and languages while still possessing advanced free-form language comprehension capabilities. Extensive quantitative evaluation shows that ChemDFM can significantly outperform the representative open-sourced LLMs.

General-domain Language Model for Science SciBERT [58] is a pre-trained language model based on the BERT model architecture, which aims to address the lack of high-quality, large-scale labeled scientific data. SciBERT uses a large multi-domain scientific publication corpus for pre-training to improve the performance of downstream scientific benchmarks and has achieved state-of-the-art performance on multiple tasks. Galactica [4] is a large language model that can store, combine and reason about scientific knowledge, which is trained on a large scientific corpus of papers, reference material, knowledge bases and many other sources. Galactica outperforms previous models on a range of scientific tasks and sets a new state-of-the-art on downstream tasks such as PubMedQA and MedMCQA. SciGLM [7] is a suite of scientific language models designed to enhance college-level scientific reasoning. It utilizes a self-reflective instruction annotation framework to address data scarcity in the science domain. SciGLM significantly improves upon ChatGLM by effectively handling complex scientific and mathematical problems, all while maintaining strong language understanding capabilities.

Compared to prior works, SciDFM either can achieve a better performance, or is more generalized. With the utilization of Mixture-of-Experts architecture, SciDFM can better model similarities and differences across different disciplines and modalities and have stronger sophisticated scientific reasoning and understanding capabilities.

5 Conclusion

In this paper, we introduce SciDFM, a mixture-of-experts LLM able to conduct college-level scientific reasoning and understand molecules and amino acid sequences. We show the pretraining and instruction-tuning process of SciDFM in detail, including data, architecture and hyper-parameters. We conduct evaluation on eight general scientific language understanding and reasoning tasks and two domain-specific tasks. From the results, we show that SciDFM achieves strong performance on general scientific benchmarks such as SciEval and SciQ, and it reaches a SOTA performance on domain-specific benchmarks among models of similar size. We further analyze the expert choices of MoE layers and show that the results of expert selection vary with data from different disciplines and exhibit clustering phenomena related to their relationships.

Acknowledgments and Disclosure of Funding

This work was supported by the National Science and Technology Major Project 2023ZD0120703, the China NSFC Projects (U23B2057, 62106142 and 62120106006), and the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[3] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

[5] Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*, 2024.

[6] Zhao, Zihan and Chen, Bo and Li, Jingpiao and Chen, Lu and Wen, Liyang and Wang, Pengyu and Zhu, Zichen and Zhang, Danyang and Wan, Ziping and Li, Yansi and others. ChemDFM-X: Towards Large Multimodal Model for Chemistry. *arXiv preprint arXiv:2409.13194*, 2024.

[7] Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning, 2024.

[8] Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, Yue Wang, Zun Wang, Tao Qin, and Rui Yan. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*, 2024.

[9] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

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

[11] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[12] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[13] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pages 18332–18346. PMLR, 2022.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061, 2024.

[16] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

[17] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net, 2024.

[18] Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023.

[19] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[20] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics, 2023.

[21] Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, Yuyi Guo, Jinfeng Bai, and Jie Tang. Gpt can solve mathematical problems without a calculator. *arXiv preprint arXiv:2309.03241*, 2023.

[22] Zengzhi Wang, Rui Xia, and Pengfei Liu. Generative ai for math: Part i – mathpile: A billion-token-scale pretraining corpus for math. *arXiv preprint arXiv:2312.17120*, 2023.

[23] Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. The stack: 3 tb of permissively licensed source code. *Preprint*, 2022.

[24] Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. Slimpajama: A 627b token cleaned and deduplicated version of redpajama, 2023.

[25] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023.

[26] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

[27] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

[28] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math, 2024.

[29] Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(W1):W228–W234, 2022.

[30] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.

[31] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[32] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[33] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[34] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large scale language model society, 2023.

[35] Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *arXiv preprint arXiv:2405.03548*, 2024.

[36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[37] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. Reasoning over paragraph effects in situations. In *MRQA@EMNLP*, 2019.

[39] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*, 2020.

[40] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

[41] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*, 2023.

[42] Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, "Teknium", and Nathan Hoos. Slimorca dedup: A deduplicated subset of slimorca, 2023.

[43] Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. *GitHub*, 2023.

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[45] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[46] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[47] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[48] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[49] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[51] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[52] Alec Radford. Improving language understanding by generative pre-training. 2018.

[53] 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*, 23(6):bbac409, 2022.

[54] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

[55] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[56] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[57] Xiao-Chen Zhang, Cheng-Kun Wu, Jia-Cai Yi, Xiang-Xiang Zeng, Can-Qun Yang, Ai-Ping Lu, Ting-Jun Hou, and Dong-Sheng Cao. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning bert enhanced by smiles enumeration. *Research*, 2022:0004, 2022.

[58] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.