Towards Efficient Large Language Models for Science: A Review

Anonymous ACL submission

Abstract

Large language models (LLMs) have ushered in a new era for processing complex information in various fields, including science. The increasing amount of scientific literature allows these models to acquire and understand scientific knowledge effectively, thus improving their performance in a wide range of tasks. Due to the power of LLMs, they require extremely expensive computational resources, intense amounts of data, and training time. Therefore, in recent years, researchers have proposed various methodologies to make scientific LLMs more affordable. The most well-known approaches align in two directions. It can be either focusing on the size of the models or enhancing the quality of data. To date, a comprehensive review of these two families of methods has not yet been undertaken. In this paper, we (I) summarize the current advances in the emerging abilities of LLMs into more accessible AI solutions for science, and (II) investigate the challenges and opportunities of developing affordable solutions for scientific domains using LLMs.

1 Introduction

Recently, the advancement of large language models (LLMs) has equipped us with the capability to address complex tasks that demand an understanding of both structure and language. The key factors that make LLMs so rapid are the huge amount of generated data and the advancement in computational architectures. With regard to scientific data itself, this domain has witnessed a constantly and rapidly increase in number of publications. For example, there were more than 2.4 million scholarly papers on ArXiv[1] (up to 2024) and 36 million publications on PubMeb[2] (up to 2022). The exponential growth enables us to leverage the success of language models to effectively learn scientific knowledge. Recently, (Ho et al., 2024) reported that there are about 117 language models constructed for the scientific domain. Tasks such as Text Classification, Summarization, or Named-Entity Recognition are effectively handled by most of these models, which have shown impressive performance on various benchmarks.

In order to perform sophisticated problem-solving tasks, the scientific language models are designed to have complex structures with vast scale. In particular, recent LLMs for science such as Galactica (Taylor et al., 2022) are equipped with groundbreaking architectures. They surpass most of the evaluations on reasoning, problem solving, and knowledge understanding. However, these LLMs face inevitable drawbacks, as they require a substantial amount of resources, for example, a large-scale high-quality dataset and a high training or inference cost (OpenAI et al., 2024). Whereas, these resources are not available in many cases, such as low-resource languages or small organizations with limited computational access. Therefore, limitations related to accessibility, cost, and adaptability pose substantial challenges to fully utilize the capabilities of scientific LLMs. In this review, we present two main contributions:

- We provide a comprehensive overview of the latest developments of the application of Large Language Models (LLMs) in scientific fields. This includes discussing how LLMs have been tailored to solve complex scientific problems, and their integration into existing studies.

- We delve into examining the technical and economic barriers to deploying LLMs for science, exploring cost-effective strategies and innovations, and identifying opportunities for

[1] https://arxiv.org/stats/monthly_submissions

[2] https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

reducing expenses without compromising performance.

2 Related surveys

There are few surveys on pre-trained language models (PLM) for science (Ho et al., 2024; Kalyan et al., 2022; Wang et al., 2023) and to make LLMs more accessible (Wan et al., 2024b; Xu et al., 2024). Regarding scientific language models, (Ho et al., 2024) presented the first comprehensive review of scientific language models (SciLM), describing more than 110 models, evaluating their performance across various domains and tasks, and addressing future research challenges. The survey examined six main aspects: time scope, target language models, domains, scientific texts, languages, and modalities, and provided a unique evolutionary overview of SciLMs in recent years. Specifically, in the biomedical sector, (Wang et al., 2023) reviewed the latest advancements of PLMs in the biomedical field and their applications in downstream biomedical tasks. The authors explored the motivations for PLMs in the biomedical sector, outlined key concepts, and proposed a taxonomy that classifies existing biomedical PLMs from multiple perspectives. In another related survey by (Kalyan et al., 2022), the authors examined the fundamental concepts of transformer-based PLMs, including pre-training methods, pre-training tasks, fine-tuning methods, and embedding types specific to the biomedical field. The survey introduced a taxonomy for transformer-based BPLMs, reviewed all the models, investigated various challenges, and suggested potential solutions.

According to (Wan et al., 2024b), while LLMs are at the forefront of the AI revolution, their impressive abilities require significant resources. As model sizes increase, the GPU hours needed for training increase exponentially, enhancing performance but also increasing costs. Furthermore, inference operations significantly add to the financial burden of running LLMs. Although enlarging the size of the model improves performance, it reduces inference throughput (increases inference latency), which poses obstacles in extending their adoption to a wider range of customers and applications affordably. The substantial resource requirements of LLMs underscore the critical necessity of devising methods that improve their efficiency. In the survey of (Wan et al., 2024b), a fairly detailed number of approaches based on three aspects is listed: model-centric, data-centric, and frameworks. However, their survey lacks investigation on the application of listed methods in different domains. Furthermore, in the survey by (Xu et al., 2024), they focused on making use of the power of proprietary LLM (such as models from the GPT family) by using knowledge distillation. Knowledge distillation for LLMs is a technique in which the hidden 'knowledge' from proprietary models is "injected" into open-source language models. These approaches seek to reduce the performance gap between cutting-edge proprietary and open-source LLMs. Knowledge distillation uses the advanced capabilities of leading proprietary models such as GPT-4 (OpenAI et al., 2024), employing them as benchmarks to improve open-source LLMs. This method resembles an experienced instructor transferring expertise to a student, with the student models adopting the performance traits of the teacher LLMs.

Despite the existing surveys on making LLMs more accessible, these works presented methods and techniques primarily in a broader domain. Meanwhile, in previous reviews on scientific language models, the authors encouraged finding efficient and low-cost solutions for scientific adaptation and leveraging LLMs for science. Therefore, our review focuses on investigating recent efficient approaches for scientific LLMs and potential research directions.

3 Advancement in efficient LLMs for Science

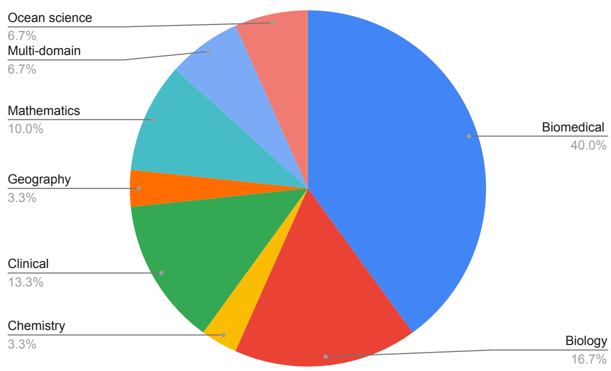

Figure 1: Distribution of efficient LLMs for Science.

This section discusses the latest developments in the application of Large Language Models (LLMs) within the scientific field. The purpose of this study is to investigate the capabilities of LLMs in scientific research. In this review, we attempt to encompass a broad range of science-related topics, including biology, biomedicine, mathematics, geoscience, ocean science, and other natural sciences. Figure 1 shows the distribution of efficient methods leveraging LLMs for each scientific domain in our review.

3.1 Biology

In the field of biology, there has been a trend towards studying increasingly large language models. Yet, the substantial computational and memory requirements for fine-tuning these models pose significant challenges for many academic laboratories and small biotechnology firms. (Sledzieski et al., 2023) implemented parameter-efficient fine-tuning (PEFT) on ESM2 model (Lin et al., 2023) to predict protein-protein interactions. Employing the PEFT technique LoRA, the model surpassed the performance of fully fine-tuned model while consuming less memory, illustrating that effective deployment of large protein language models is feasible even for groups with constrained computational resources. The study further highlighted that the efficacy of this method could be enhanced by utilizing more informative embeddings produced by LLMs. Research on the PEFT LoRA method adapted for the ESM2 model was also conducted in (Zeng et al., 2023b) focusing on signal peptides (SP) prediction. Other PEFT techniques such as Adapter Tuning and Prompt Tuning were also explored. It was noted that Prompt Tuning underperformed compared to other previous models, likely due to the size of the model. While Adapter Tuning improved performance, it required a considerably larger number of training parameters relative to LoRA. Future enhancements were suggested, including combining PEFT techniques to improve interpretability for identifying SP-related motifs and integrating structure-aware language models to include protein structural data. Another PEFT approach, adaptive LoRA (AdaLoRA), was utilized in the study by (Zhan and Zhang, 2023). This study introduced AdaLoRa with random sampling (AdaLoRA-RS) on OPT-350M to enhance the understanding of genomic language complexities. When compared to other models, DNABERT and Nucleotide Transformer, AdaLoRA+RS demonstrated performance on par with fully fine-tuned models across 13 genomic datasets while using less than 2% of the training parameters. The experimental findings further showed that pre-trained language models such as OPT-125M outperformed the specialized DNA model HR-500M, utilizing only 25% of the parameters.

3.2 Biomedical domain

The advancement of LLMs has also greatly impacted biomedical research. Over the past decade, vast unlabelled datasets such as PubMed, PMC, MIMIC, and ScienceDirect have become available in biomedicine. Models like GPT-4 and Med-PaLM 2 have shown exceptional performance in various biomedical NLP tasks. However, these models, with their hundreds of billions of parameters, are expensive in terms of computational resources, require data transmission over the Internet, and are trained on proprietary data sources.

In 2022, (Li et al., 2022) leveraged this unlabelled information to introduce BioKnowPrompt, a prompt-tuning PLM framework tailored for extracting relationships from biomedical texts. Additionally, prompting can be challenging for certain phenomena and may struggle with highly imbalanced training data. Follow up, with the introduction of ChatGPT and GPT-4 (OpenAI et al., 2024), many researches have leveraged its power for data augmentation. (Zhang et al., 2023a) created HuatuoGPT based on LlaMa model, employing both refined data from ChatGPT and data from doctors for health consultations. This model is superior at producing patient-friendly and doctor-like responses and outperformed existing medical open-source LLMs. DoctorGLM presented by (Xiong et al., 2023) also used the data generated by ChatGPT for medical dialogues in Chinese. The training process of DoctorGLM can handle a considerable number of question-answer pairs per hour per GPU, with a relatively low cost per training session. Furthermore, the inference operations of DoctorGLM demand minimal GPU memory, enabling execution on standard consumer hardware, thus making it accessible for numerous research facilities and healthcare centers. They also mentioned that the model can be deployed on even more affordable GPU when applying PEFT method such as LoRA. The superiority of GPT-4 also demonstrated by (Hsueh et al., 2023). The authors succeeded in using prompt engineering for ChatGPT(GPT-4) to generate answers for biomedical questions. Although their method outperformed the fine-tuned BioBERT model, they discussed that there were rooms for improvements such as determining key information before prompting. (Bolton et al., 2024) introduced BioMedLM, a 2.7 billion parameter GPT-style autoregressive model trained exclusively on PubMed abstracts and full articles. Their result highlights that smaller models have the potential to serve as transparent, privacy-preserving, cost-effective, and environmentally friendly solutions for biomedicine. Another approach using GPT-3.5 for biomedical purposes was presented by (Bao et al., 2023). The researchers employed GPT-3.5 to extract medical knowledge triples from a knowledge graph through a department-focused method based on patient query patterns from real-world consultations, producing 50,000 samples. Additionally, (Liu et al., 2023) addressed the issue of securely managing medical data in the modern digital age, where confidentiality is a major concern. Utilizing advancements in large language models such as ChatGPT and GPT-4, the researchers introduced DeID-GPT, an innovative framework designed to automatically identify and mask personal information in medical texts. Their method not only achieved high accuracy in maintaining text integrity but also set a new standard for the application of LLMs in healthcare settings focused on privacy protection.

While proprietary LLMs are usually huge, untrainable and their architecture are unclear, researchers adapt instruction-tuning technique to open-source smaller LLMs for solving biomedical problems. (Wu et al., 2023) systematically adapted the open-source general LLM, LLaMA, for biomedical tasks by injecting domain-specific data and instruction-tuning tailored to medical contexts. The PCM-LLaMA model, an open-source language model designed for medical purposes, demonstrates superior results on various medical benchmarks, surpassing both ChatGPT and LLaMA-2 while utilizing considerably fewer parameters. Additionally, (Luo et al., 2023b) presented BioMedGPT, a multi-modal generative pre-trained model tailored for biomedical applications. The design of BioMedGPT highlights the critical importance of knowledge distillation in bridging complex biological information with natural language, enabling substantial advancements in the discovery of drugs and therapeutic targets. (Peng et al., 2024) conducted comparison on GatorTron using soft-prompting in various configurations. The study revealed that soft prompting surpassed hard prompting, unfrozen Large Language Models (LLMs) display robust few-shot learning abilities and adaptability across different institutions, using frozen LLMs reduces computational costs to between 2.5% and 6% relative to earlier methods that utilized unfrozen LLMs, while still attaining optimal outcomes with large-scale unfrozen LLMs. To enhance performance and generalizability beyond traditional benchmarks, (Zhang et al., 2024b) introduced MedInstruct-52k, a diverse dataset generated with GPT-4 and ChatGPT. Fine-tuning LLaMA-series models on this dataset resulted in AlpaCare, which outperformed previous medical LLMs by up to 38.1% in medical instruction-following tasks and showed consistent improvements in general domain benchmarks based on human evaluations. Parameter-efficient fine-tuning is also an effective approach to reduce the training time and cost when performance domain adaptation. (Han et al., 2023) utilized the LLaMA foundation models with 7 billion and 13 billion parameters, fine-tuning them over five epochs with learning rates specifically adjusted for each model variant. They applied Low-Rank Adaptation (LoRA) to improve efficiency by lowering GPU memory usage and reducing training time. Additionally, the author incorporated 8-bit matrix multiplication to further decrease computational requirements, making it more feasible to deploy these models in medical applications with strict resource limitations.

3.3 Clinical domain

The clinical domain has also experienced a transition from traditional pre-trained language models to the effective use of LLMs. (Gema et al., 2024) introduced a two-step PEFT framework based on the LLaMA model, which was evaluated within the clinical domain. This framework integrates a specialized PEFT adapter layer for clinical domain adaptation with another adapter for downstream tasks. It was tested on various datasets for clinical outcome prediction and compared to language models trained specifically for clinical purposes. This research is the first to propose a comprehensive empirical analysis of the interaction between PEFT techniques and domain adaptation in the crucial real-world setting of clinical applications. (Goswami et al., 2024) examined the effectiveness of prompt engineering and parameter-efficient fine-tuning to summarize hospital discharge summary (HDS) articles. The objective was to ensure that these models accurately interpret medical terminology and contexts, generate concise summaries, and extract key themes. The study used LLaMA-2 as the base model and fine-tuned it with QLoRA

(Quantized Low-Rank Adapters) to minimize memory usage without sacrificing data quality. Chinese patent medicine (CPM), a vital component of traditional Chinese medicine (TCM) that utilizes Chinese herbs, was explored by (Liu et al., 2024) using LLMs. The researchers introduced the first CPM instructions (CPMI) dataset and fine-tuned the ChatGLM-6B base model, resulting in CPMI-ChatGLM. They employed parameter-efficient fine-tuning with consumer-grade graphics cards and investigated LoRA, P-Tuning v2, along with various data scales and configurations. Comparative experiments with similar-size LLMs demonstrated the leading performance of CPMI-ChatGLM in recommending CPM, highlighting its potential for clinical support and data analysis in TCM research.

3.4 Mathematics

Large language models like GPT-4 have demonstrated exceptional performance in complex mathematical reasoning, yet open-source models are typically pre-trained on large-scale internet data without specific optimization for mathematical tasks. Addressing this limitation, (Luo et al., 2023a) introduced WizardMath, enhancing mathematical reasoning in LLaMa-2 through Reinforcement Learning from Evol-Instruct Feedback (RLEIF). WizardMath outperformed ChatGPT-3.5, Claude Instant-1, PaLM-2, and Minerva on GSM8k, as well as Text-davinci-002, PaLM-1, and GPT-3 on MATH, highlighting RLEIF's efficacy. Derived from this foundation, (Yue et al., 2023) introduced MAmmoTH, a series of open-source LLMs specialized for mathematics. MAmmoTH-7B achieved a 33% accuracy rate on MATH, surpassing WizardMath-7B by 23%, underscoring the importance of diverse problem coverage and hybrid rationales in developing advanced math models. Additionally, (Gou et al., 2024) presented TORA, integrating natural language reasoning with external computational tools like computation libraries and symbolic solvers to tackle challenging mathematical problems. TORA models significantly outperformed existing open-source models on ten mathematical reasoning datasets, achieving average improvements of 13%-19%. TORA-7B achieved 44.6% accuracy on the competition-level MATH dataset, outperforming WizardMath-70B by 22% absolute, demonstrating the effectiveness of integrating computational tools with language models for mathematical problem-solving.

3.5 Geoscience

In the field of geoscience, (Deng et al., 2023) introduced K2, the first ever LLM tailored for geoscience applications. The authors developed critical resources to improve LLM research within geoscience, including GeoSignal, the first geoscience instruction tuning dataset, and GeoBench, the inaugural geoscience benchmark for evaluating LLMs. In their study, they detailed the process of adapting a pre-trained general-domain LLM, specifically the LLaMA-7B model, to the geoscience domain by further training it in a 5.5 billion token corpus of geoscience texts and fine-tuning it with GeoSignal's supervised data. The authors also provided a protocol for efficiently gathering and constructing domain-specific supervised data, even with limited manpower. The experimental results on GeoBench confirmed the effectiveness of their approach and datasets in improving understanding and application of geoscience knowledge, marking a significant advancement in the integration of LLMs within geoscientific research and practice.

3.6 Chemistry

In the quest to enhance crystal property prediction, recent studies have turned their attention to utilizing textual descriptions of crystal structures. Conventional techniques mainly employ graph neural networks (GNNs) to model these structures (Huang et al., 2024b; Ruff et al., 2023; Yan et al., 2024), but they often face challenges with the complex interactions between atoms and molecules. A novel approach presented by (Rubungo et al., 2024) includes the development of a benchmark dataset called TextEdge, which offers detailed text descriptions of crystal structures along with their properties. Moreover, the authors introduce LLM-Prop, an innovative method using large language models (LLMs) to predict the physical and electronic properties of crystals based on their textual descriptions. Additionally, it surpasses a domain-specific fine-tuned BERT model, MatBERT, despite having significantly fewer parameters.

3.7 Ocean Science

Ocean science, crucial for understanding the vast reservoirs of life and biodiversity covering over 70% of our planet, has yet to fully benefit from advancements in large language models (LLMs). Despite their success in various fields, LLMs often fall short in meeting the specialized needs of oceanographers due to the complexity and richness of ocean data. To address this gap, (Bi et al., 2024) introduced OCEANGPT, the first LLM specifically tailored for ocean science. Comprehensive experiments demonstrated that OCEANGPT not only possessed a high level of knowledge expertise in ocean science but also showed preliminary capabilities in embodied intelligence for ocean technology. Furthermore, (Zheng et al., 2023) introduced MarineGPT, the first vision-language model specifically designed for the marine domain. MarineGPT, developed using the Marine-5M dataset of over 5 million marine image-text pairs, aimed to make ocean knowledge more accessible and improve marine vision and language alignment, addressing the inadequacies of existing general-purpose MLLMs in understanding and responding to domain-specific intents.

3.8 Multi-scientific domains

In their respective studies, (Xie et al., 2023) introduced DARWIN, a series of tailored LLMs optimized specifically for scientific disciplines such as material science, chemistry, and physics. Built upon the foundational LLaMA-7B model, DARWIN achieved significant advances in automating the generation of scientific text instruction, thus improving its performance in various scientific tasks and reducing the dependency on closed-source LLMs. Similarly, (Zhang et al., 2024a) presented SciGLM, a suite of scientific language models designed for college-level scientific reasoning. Using a self-reflective instruction annotation framework, SciGLM addressed data scarcity challenges in the science domain by improving both base models like ChatGLM3-6B-Base by 4.87% and larger-scale models by 2.67%. This approach enhances the model's ability to conduct diverse scientific discovery tasks while preserving its language understanding capabilities.

4 Challenges and future directions

Current studies on the application of LLMs in science have made significant progress. We summarize the existing methods and scientific LLMs in Table 1. Most of these studies have initially harnessed the power of LLMs to address problems in scientific fields such as biology and biomedicine. However, many issues remain unresolved. This section will present some research gaps with potential for further exploration.

4.1 Data Collection

Challenges The lack of labeled data is a common issue faced by researchers when training language models in various scientific fields. Despite the abundance of unlabeled scientific data, it is not utilized efficiently to train language models. (Ho et al., 2024) summarized that among 117 language models for scientific fields, most previous work focused on the biomedical domain, with more than 87% pre-trained language models in this area. The author also noted that these language models typically have fewer than 1 billion parameters (e.g., BERT-based models) and do not leverage open-source LLMs. **This creates a problem where unlabeled data in other scientific domains are underutilized.** Collecting high-quality labeled data for model training is notoriously time-consuming and labor-intensive.

Potential directions Existing solutions such as active learning for small language models (SLMs) and in-context learning for large language models (LLMs) have somewhat mitigated the lack of labeled data, but still rely heavily on human intervention. (Xiao et al., 2023) addressed this issue by introducing FreeAL, a collaborative learning framework where an LLM acts as an active annotator and an SLM filters high-quality in-context samples for label refinement. Extensive experiments on eight benchmark datasets showed that FreeAL significantly improved zero-shot performance for both SLMs and LLMs without human supervision. (Zhang et al., 2023c) introduced LLMaAA, which uses LLMs as annotators in an active learning loop to efficiently select data for annotation, demonstrating superior performance in named entity recognition and relation extraction tasks with fewer annotated examples. (Huang et al., 2024a) tackled the challenge of high quality annotations under limited budgets with SANT, a selective annotation framework utilizing error-aware triage and bi-weighting mechanisms, setting a new benchmark for triage-based annotation studies.

4.2 Data Selection

Challenges Determining the optimal data volume crucial for maximizing the effectiveness of Large Language Models (LLMs) remains a persistent challenge, necessitating further research to establish clear guidelines. Additionally, developing robust methodologies to filter out low-quality data continues to be an ongoing concern in leveraging

Methods		Models
Efficient Fine-tuning		ESM2-LoRA (Sledzieski et al., 2023), PEFT-SP(Zeng et al., 2023b), AdaLoRA+RS (Zhan and Zhang, 2023), BioMedGPT (Luo et al., 2023b), MedAlpaca (Han et al., 2023), Clinical LLaMA-LoRA (Gema et al., 2024), LLaMa-QLoRA (Goswami et al., 2024), CPMI-ChatGLM (Liu et al., 2024)
Instruction Tuning		BioKnowPrompt (Li et al., 2022), NCU-IISR (Hsueh et al., 2023), Alpacare (Zhang et al., 2024b), GatorTron (Peng et al., 2024), K2 (Deng et al., 2023), WizardMath (Luo et al., 2023a), MAmmoTH (Yue et al., 2023), TORA (Gou et al., 2024), OCEANGPT (Bi et al., 2024), MarineGPT (Zheng et al., 2023), SciGLM (Zhang et al., 2024a)
Knowledge distillation	Black box	HuatouGPT (Zhang et al., 2023a), DoctorGLM (Xiong et al., 2023), DISC-MedLLM (Bao et al., 2023), DeID-GPT (Liu et al., 2023)
	White box	BioMedLM (Bolton et al., 2024), PCM-LLaMA (Wu et al., 2023), LLM-Prop (Rubungo et al., 2024), DARWIN (Xie et al., 2023)

Table 1: Summary of previous work on efficient LLMs for science.

LLMs effectively.

Potential Directions In general domain, (Zhou et al., 2023) proposed that a minimum of 1000 well-curated, high-quality data samples could be sufficient to align LLMs, as pre-training already provides essential knowledge. (Chen et al., 2024b) introduced a new data selection method using a robust LLM such as ChatGPT to independently filter out low-quality data. They developed AlpaGasus, a model refined with just 9,000 high-quality samples from the initial dataset. More recently, (Li et al., 2024) presented Superfiltering, which used smaller models such as GPT-2 to extract a high-quality subset from a dataset. Despite these advancements, the challenges of selecting optimal data for refining LLMs and determining the necessary data volume persist because of the abundance of unlabeled scientific data.

4.3 Utilizing multiple LLMs

Challenges The majority of current models originate from a single LLM, yet it is commonly recognized that models trained with diverse data sources possess distinct advantages. Consequently, the question arises: **Can knowledge from multiple LLMs be integrated into a single smaller model?**

Potential directions In an effort to create a "BabyLM," (Timiryasov and Tastet, 2023) trained an ensemble of GPT-2 and small LLaMA models on the 10M-word BabyLM dataset, then distilled this ensemble into a small, 58M-parameter LLaMA model. The distilled model outperformed both its teachers and a similar model trained without distillation, suggesting that distillation can retain and even exceed the performance of teacher models, particularly on small datasets. (Wan et al., 2024a) subsequently developed 'knowledge fusion' to combine the strengths of multiple LLMs, validating their approach with Llama-2, MPT, and OpenLLaMA across various benchmarks. This method improved the target model's performance in reasoning, common sense, and code generation. Additionally, (Chen et al., 2024a) introduced MAGDI to enhance reasoning in small models by distilling interactions between multiple large LLMs using Multi-Agent Interaction Graphs (MAGs). MAGDI outperformed traditional distillation methods and improved reasoning and efficiency in smaller models. Despite these advancements, the scientific community still lacks research on leveraging knowledge from multiple LLMs.

4.4 Addressing Catastrophic Forgetting

Challenges Prior studies have investigated optimizing LLMs to enhance their directive-following and knowledge transfer abilities, leveraging advancements in LLM technology. However, **persistent optimization with specific datasets can lead to catastrophic forgetting**.

Potential directions In the scientific domain, (Yue et al., 2023) introduced MAmmoTh, an ensemble of open-source LLMs designed to tackle mathematical challenges using the MathInstruct dataset, overcoming catastrophic forgetting seen in prior models like WizardMath (Luo et al., 2023a). Meanwhile, continual learning (CL) research focuses on dynamically enhancing models while preserving prior knowledge. Methods such as Lifelong-MoE (Chen et al., 2023), CITB (Zhang et al., 2023d), and DCL (Zeng et al., 2023a) utilize strategies like expert addition, regularization, task distribution modeling, and knowledge distillation to address catastrophic forgetting. Despite these efforts, maintaining original model capabilities and transferring knowledge across domains remain challenging.

4.5 Multimodality

Challenges In the scientific domain, there is a growing interest in multi-modal models (Ho et al., 2024), developed by further training on mono-modal or multi-modal models from general domains, leveraging their strong performance. However, several challenges persist. The scientific domain often lacks sufficient data compared to general domains, making it difficult to adequately train or fine-tune multi-modal language models. **Incorporating this multi-modal information into scientific language models is crucial for advancing research.**

Potential directions Numerous studies focus on developing adapters that convert non-language data to be processed within the same embedding space as language (Dai et al., 2023; Zhu et al., 2023). These architectures aim to handle non-language information while preserving the robust problem-solving capabilities of LLMs. Although proprietary LLMs like GPT-4 can process multiple scientific data types, prompting these models requires significant resources. Therefore, it is recommended to find efficient methods to make LLMs more accessible and introduce multimodality in scientific fields, enabling the full potential of multi-modal models in the scientific domain to be harnessed.

4.6 Further reduce the cost

Challenges Despite the impressive capabilities of modern LLMs, their substantial resource demands highlight **the critical need for effective solutions to address these challenges**. Based on Table 1, in previous work within the scientific domain, common ways to reduce costs have included Instruction Tuning and Efficient Fine-Tuning. Continued research and development in other methodologies are crucial to making LLMs more accessible and sustainable.

Potential directions In other domains, various efficient approaches have been studied, such as Quantization (Frantar et al., 2023; Kim et al., 2023; Tao et al., 2022), Parameter Pruning (Ma et al., 2023; Zhang et al., 2023b), and Memory Efficient Fine-Tuning (Dettmers et al., 2023; Malladi et al., 2023). The question of how to further decrease the cost of LLMs remains unsolved. For instance, Memory Efficient Fine-Tuning techniques, such as QLoRA (Goswami et al., 2024), which optimizes memory usage during fine-tuning, also offer potential solutions.

5 Conclusion

The rapid advancement of large language models (LLMs) has significantly enhanced our ability to address complex tasks requiring deep linguistic and structural understanding. The growth of scientific data has enabled effective learning of scientific knowledge through LLMs. However, despite their impressive performance in tasks like reasoning and problem-solving, these models remain resource-intensive and often inaccessible to smaller organizations and low-resource languages. Our review highlighted various cost-effective techniques for utilizing LLMs in scientific domains. We address the challenges in fully harness the potential of LLMs for science and ensure their broader accessibility and applicability in scientific research.

6 Limitations

Our work is based on results and suggestions of as many papers as possible we can find. We also mostly emphasize text-based scientific information, setting aside other forms such as images, videos, audio, and structured knowledge like knowledge graphs (KGs) and databases for future consideration. Our review primarily highlights the most recent advancements in the last three years, specifically from 2023 and 2024. However, our review may hold a potential of missed out some the most recent studies. We leave this as future improvements. Moreover, due to space limitations, we provide only concise summaries of the reviewed methods.

References

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *Preprint*, arXiv:2308.14346.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2024. Oceangpt: A large language model for ocean science tasks. *Preprint*, arXiv:2310.02031.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. *Preprint*, arXiv:2403.18421.

Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2024a. Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. *Preprint*, arXiv:2402.01620.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024b. Alpagasus: Training a better alpaca with fewer data. *Preprint*, arXiv:2307.08701.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cu. 2023. Lifelong language pretraining with distribution-specialized experts. *Preprint*, arXiv:2305.12281.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuanyuan Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 2023. K2: A foundation language model for geoscience knowledge understanding and utilization. *Preprint*, arXiv:2306.05064.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. 2023. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Preprint*, arXiv:2208.11580.

Aryo Pradipta Gema, Pasquale Minervini, Luke Daines, Tom Hope, and Beatrice Alex. 2024. Parameter-efficient fine-tuning of llama for the clinical domain. *Preprint*, arXiv:2307.03042.

Joyeeta Goswami, Kaushal Kumar Prajapati, Ashim Saha, and Apu Kumar Saha. 2024. Parameter-efficient fine-tuning large language model approach for hospital discharge paper summarization. *Applied Soft Computing*, 157:111531.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. *Preprint*, arXiv:2309.17452.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Xanh Ho, Anh Khoa Duong Nguyen, An Tuan Dao, Junfeng Jiang, Yuki Chida, Kaito Sugimoto, Huy Quoc To, Florian Boudin, and Akiko Aizawa. 2024. A survey of pre-trained language models for processing scientific text. *Preprint*, arXiv:2401.17824.

Chun Yu Hsueh, Yu Zhang, Yu Wei Lu, Jen Chieh Han, Wilailack Meesawad, and Richard Tzong Han Tsai. 2023. Ncu-iisr: Prompt engineering on gpt-4 to stove biological problems in bioasq 11b phase b. *CEUR Workshop Proceedings*, 3497:114–121. Publisher Copyright: © 2023 Copyright for this paper by its authors.; 24th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF-WN 2023 ; Conference date: 18-09-2023 Through 21-09-2023.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Ido Dagan. 2024a. Selective annotation via data allocation: These data should be triaged to experts for annotation rather than the model. *Preprint*, arXiv:2405.12081.

Jiao Huang, Qianli Xing, Jinglong Ji, and Bo Yang. 2024b. Ada-gnn: Atom-distance-angle graph neural network for crystal material property prediction. *Preprint*, arXiv:2401.11768.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, 126:103982.

Young Jin Kim, Rawn Henry, Raffy Fahim, and Hany Hassan Awadalla. 2023. Finequant: Unlocking efficiency with fine-grained weight-only quantization for llms. *Preprint*, arXiv:2308.09723.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *Preprint*, arXiv:2402.00530.

Qing Li, Yichen Wang, Tao You, and Yantao Lu. 2022. Bioknowprompt: Incorporating imprecise knowledge into prompt-tuning verbalizer with biomedical text for relation extraction. *Information Sciences*, 617:346–358.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Can Liu, Kaijie Sun, Qingqing Zhou, Yuchen Duan, Jianhua Shu, Hongxing Kan, Zongyun Gu, and Jili Hu. 2024. Cpmi-chatglm: parameter-efficient fine-tuning chatglm with chinese patent medicine instructions. *Scientific Reports*, 14.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *Preprint*, arXiv:2303.11032.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *Preprint*, arXiv:2308.09583.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023b. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *Preprint*, arXiv:2308.09442.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. LLM-pruner: On the structural pruning of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. 2023. Fine-tuning language models with just forward passes. In *Thirty-seventh Conference on Neural Information Processing Systems*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Cheng Peng, Xi Yang, Kaleb E Smith, Zehao Yu, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics*, 153:104630.

Andre Niyongabo Rubungo, Craig Arnold, Barry Rand, and Adji Bousso Dieng. 2024. LLM-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions.

Robin Ruff, Patrick Reiser, Jan Stühmer, and Pascal Friederich. 2023. Connectivity optimized nested graph networks for crystal structures. *Preprint*, arXiv:2302.14102.

Samuel Sledzieski, Meghana Kshirsagar, Minkyung Baek, Bonnie Berger, Rahul Dodhia, and Juan Lavista Ferres. 2023. Democratizing protein language models with parameter-efficient fine-tuning. *bioRxiv*.

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4821–4836, Dublin, Ireland. Association for Computational Linguistics.

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

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 279–289, Singapore. Association for Computational Linguistics.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024a. Knowledge fusion of large language models. In *The Twelfth International Conference on Learning Representations*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024b. Efficient large language models: A survey. *Preprint*, arXiv:2312.03863.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie fu. 2023. Pre-trained language models in biomedical domain: A systematic survey. *Preprint*, arXiv:2110.05006.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, Imran Razzak, and Bram Hoex. 2023. Darwin series: Domain specific large language models for natural science. *Preprint*, arXiv:2308.13565.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *Preprint*, arXiv:2304.01097.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *Preprint*, arXiv:2402.13116.

Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. 2024. Complete and efficient graph transformers for crystal material property prediction. *Preprint*, arXiv:2403.11857.

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

Min Zeng, Wei Xue, Qifeng Liu, and Yike Guo. 2023a. Continual learning with dirichlet generative-based rehearsal. *Preprint*, arXiv:2309.06917.

Shuai Zeng, Duolin Wang, and Dong Xu. 2023b. Peft-sp: Parameter-efficient fine-tuning on large protein language models improves signal peptide prediction. *bioRxiv*.

Huixin Zhan and Zijun Frank Zhang. 2023. Parameter-efficient fine-tune on open pre-trained transformers for genomic sequence. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning. *Preprint*, arXiv:2401.07950.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023a. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Mingyang Zhang, Hao Chen, Chunhua Shen, Zhen Yang, Linlin Ou, Xinyi Yu, and Bohan Zhuang. 2023b. Loraprune: Pruning meets low-rank parameter-efficient fine-tuning. *Preprint*, arXiv:2305.18403.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023c. Llmaaa: Making large language models as active annotators. *Preprint*, arXiv:2310.19596.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2024b. Alpacare:instruction-tuned large language models for medical application. *Preprint*, arXiv:2310.14558.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. 2023d. CITB: A benchmark for continual instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9443–9455, Singapore. Association for Computational Linguistics.

Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. 2023. Marinegpt: Unlocking secrets of ocean to the public. *Preprint*, arXiv:2310.13596.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.