

# XIWU: A BASIS FLEXIBLE AND LEARNABLE LLM FOR HIGH ENERGY PHYSICS

Zhengde Zhang<sup>1</sup>, Yiyu Zhang<sup>1</sup>, Haodong Yao<sup>1</sup>, Jianwen Luo<sup>2</sup>, Rui Zhao<sup>1</sup>, Bo Huang<sup>1</sup>, Jiameng Zhao<sup>3,1</sup>, Yipu Liao<sup>1</sup>, Ke Li<sup>1</sup>, Lina Zhao<sup>1</sup>, Jun Cao<sup>1</sup>, Fazhi Qi<sup>1,\*</sup>, and Changzheng Yuan<sup>1,†</sup>

<sup>1</sup>Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

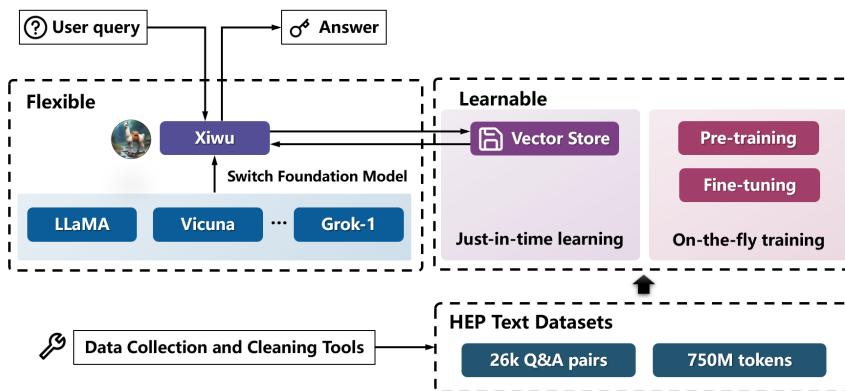
<sup>2</sup>School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

<sup>3</sup>School of Computer and Artificial Intelligence, ZhengZhou University, Henan 450066, China

## ABSTRACT

Large Language Models (LLMs) are undergoing a period of rapid updates and changes, with state-of-the-art (SOTA) model frequently being replaced. When applying LLMs to a specific scientific field, it's challenging to acquire unique domain knowledge while keeping the model itself advanced. To address this challenge, a sophisticated large language model system named as Xiwu has been developed, allowing you switch between the most advanced foundation models and quickly teach the model domain knowledge. In this work, we will report on the best practices for applying LLMs in the field of high-energy physics (HEP), including: a seed fission technology is proposed and some data collection and cleaning tools are developed to quickly obtain domain AI-Ready dataset; a just-in-time learning system is implemented based on the vector store technology; an on-the-fly fine-tuning system has been developed to facilitate rapid training under a specified foundation model.

The results show that Xiwu can smoothly switch between foundation models such as LLaMA, Vicuna, ChatGLM and Grok-1. The trained Xiwu model is significantly outperformed the benchmark model on the HEP knowledge Q&A and code generation. This strategy significantly enhances the potential for growth of our model's performance, with the hope of surpassing GPT-4 as it evolves with the development of open-source models. This work provides a customized LLM for the field of HEP, while also offering references for applying LLM to other fields, the corresponding codes are available on Github <https://github.comzhang/zhengde0225/Xiwu>.



**Keywords** LLM · Deep Learning · Artificial Intelligence · Particle Physics · HEP

\*Corresponding Author: [qfz@ihep.ac.cn](mailto:qfz@ihep.ac.cn)

†Corresponding Author: [yuancz@ihep.ac.cn](mailto:yuancz@ihep.ac.cn)

---

<div style="background-color: #f0f0f0; padding: 5px; border-radius: 5px;"> <span style="font-size: 2em; color: #ccc;">?</span> What is HEPS?         </div> <div style="background-color: #fff; padding: 10px; border-radius: 5px;"> <span style="font-size: 1.5em; color: #ccc;">💡</span> GPT-4 :         </div> <div style="background-color: #f0f0f0; padding: 10px; border-radius: 5px;">           HEPS stands for <b>Harmonised European Prevalence of Infection Study</b>. It is ...         </div>	<div style="background-color: #e0e0e0; padding: 5px; border-radius: 5px;"> <span style="font-size: 2em; color: #ccc;">?</span> Xiwu :         </div> <div style="background-color: #fff; padding: 10px; border-radius: 5px;"> <span style="font-size: 1.5em; color: #ccc;">💡</span> Xiwu :         </div> <div style="background-color: #e0e0e0; padding: 10px; border-radius: 5px;">           HEPS stands for High Energy Photon Source. It is ...         </div>
<div style="font-size: 1.2em; color: #ccc;">🎓 Comment:</div> GPT-4 generated incorrect answers without self-awareness, while Xiwu generated correct answers.	
<div style="background-color: #f0f0f0; padding: 5px; border-radius: 5px;"> <span style="font-size: 2em; color: #ccc;">?</span> Please provide an example of simulate particle decay on BOSS         </div> <div style="background-color: #fff; padding: 10px; border-radius: 5px;"> <span style="font-size: 1.5em; color: #ccc;">💡</span> GPT-4 :         </div> <div style="background-color: #f0f0f0; padding: 10px; border-radius: 5px;"> <b>BOSS (BesIII Object-oriented Simulation Software)</b> is a software package used for event simulation at the BESIII experiment. Here is a simplified example of how a particle decay simulation might be conducted using BOSS:  <b>Initialization:</b> Before the simulation can begin, the software must be initialized. This includes setting the parameters for the experiment, such as the type of particle to be simulated, the energy levels, etc.  <b>Generation of Initial Particles:</b> The software generates a set of initial particles based on the parameters set in the initialization step. These particles are then "fired" into the simulated detector.  <b>Tracking:</b> Each particle's path through the detector is tracked and recorded. This includes any interactions the particle has with the detector material, such as scattering or decay.  <b>Decay Simulation:</b> If the particle is unstable (i.e., it can decay into other particles), the software simulates this decay process. This involves randomly determining when the particle will decay (based on its known lifetime) and what particles it will decay into (based on known decay modes).  <b>Data Recording:</b> All data from the simulation, including the initial and final states of each particle, are recorded for later analysis.  <b>Analysis:</b> After the simulation is complete, the recorded data can be analyzed to learn more about the particle's properties and behavior. This might include things like determining the particle's lifetime, measuring its decay modes, etc.            Please note that this is a simplified description. Actual simulations would be much more complex and require a deep understanding of particle physics.       </div> <div style="background-color: #e0e0e0; padding: 10px; border-radius: 5px;"> <span style="font-size: 2em; color: #ccc;">?</span> Xiwu :         </div> <div style="background-color: #fff; padding: 10px; border-radius: 5px;"> <span style="font-size: 1.5em; color: #ccc;">💡</span> Xiwu :         </div> <div style="background-color: #e0e0e0; padding: 10px; border-radius: 5px;">           In BOSS (BESIII Offline Software System), a common particle simulation is:  <math display="block">e^+ e^- \rightarrow \psi(2S) \rightarrow J/\psi \pi^+ \pi^-</math>           The decay card for this example is as follows:  <pre>Decay psi(2S) 1.0 J/psi pi+ pi- JPIPI; Enddecay  Decay J/psi 0.5 e+ e- PHOTOS VLL; 0.5 mu+ mu- PHOTOS VLL; Enddecay  End</pre>           In addition, it is necessary to modify the control file of the KKMC generator, such as:  <pre>*****job options for generator (KKMC)***** #include "\$KKMCROOT/share/jobOptions_KKMC.txt"  KKMC CMSEnergy = 3.686; KKMC BeamEnergySpread=0.0009; KKMC NumberOfEventPrinted=1; KKMC GeneratePsiPrime=true;</pre>           In this example, the entire simulation process is conducted at a collision energy of 3.686 GeV and uses the KKMC generator.            This simulated particle decay process is commonly used in high-energy physics experiments to study the properties and behavior of particles. I hope this example is helpful to you !         </div>	
<div style="font-size: 1.2em; color: #ccc;">🎓 Comment:</div> GPT-4 generated an incorrect concept of BOSS, offering a solution that seems correct but is useless, while Xiwu correctly explained the concept of BOSS, providing the correct decay card and example code for generator.	

Figure 1: The hallucination of GPT-4 when answering domain questions.

## 1 Introduction

**Large Language Models** The Large Language Models (LLMs) such as GPT-4 [1] and LLaMA[2, 3] have exhibited capabilities beyond expectations in terms of general intent understanding, robust continuous dialogue, intelligent interaction correction, and adequate logical reasoning, actively propelling the shifts towards an intelligent paradigm in scientific research.

**High Energy Physics** High Energy Physics (HEP) is a crucial branch of physics that delves into the characteristics and interactions of the most basic elements of matter. It serves as an essential field for unraveling the universe's underlying principles and laws. Creating Large Language Models (LLMs) for High Energy Physics (HEP) can play a pivotal role in streamlining research activities, including literature review, coding, data management, analysis of physics, interpretation of results, and manuscript composition. This advancement can free researchers from mundane tasks that demand less creativity, thereby amplifying the efficiency and productivity of scientific investigations.

**Hallucinations** However, the LLMs can suffer from **hallucinations**, i.e. "produce content that is nonsensical or untruthful in relation to certain sources" [4, 5], which is particularly harmful in scientific fields. As shown in Figure 1, even the most powerful LLM (GPT-4) can struggle with domain-specific questions or generating scientific code of HEP. The capability of general LLMs to precisely tackle specialized topics or develop code for focused areas may be constrained.

**Way to Reduce Hallucinations** There are two methods can enhance the model's ability to handle domain-specific tasks: one is further training the model, including both secondary pre-training and fine-tuning, and the other is Retrieval-Augmented Generation (RAG). The foundation models are often referred to as L0 (Level 0) models, such as LLaMA and GPT-3. Models fine-tuned for specific tasks based on the foundation model are called L1 models, such as LLaMA-Chat, Vicuna [6] and GPT-3.5 [7], which is further adjusted for conversational tasks. Models that are further fine-tuned with domain-specific knowledge can be termed as L2 models.

**HEP-Xiwu LLM** Here, we introduce the **Xiwu** (/ʃər'wu:/), a L2 large language model that exhibits fewer hallucinations in HEP field. The most significant features of this model are its basis flexibility and learnability. Basis flexible means that the foundational model can be changed, becoming increasingly powerful with the upgrade of open-source models. Learnable implies that once the foundational model is updated, it can quickly be taught domain knowledge. Specifically, a just-in-time learning system based on a vector store has been implemented, allowing for instant knowledge updates to the Xiwu at a low cost.

**Why Consider Flexible?** Training a foundation LLM is extremely costly. For a specific scientific field, on one hand, training LLM from scratch is unaffordable; on the other hand, even if trained LLM from scratch, it might not keep up with the pace of development of open-source models. "Using the strengths of others to improve oneself," we believe that having a flexible foundational model is the best strategy to maintain the most advanced performance of the model at the lowest cost.

**Contributions** The contributions are as follows:

1. The seed fission technology is proposed and realized, which has proven its effectiveness in acquiring training Q&A data related to the field with only one seed. It can be widely applied in various fields.
2. A basis flexible LLM system has been strategized and implemented, evolving from the initial LLaMA to now support upgrades to LLaMA2 and Vicuna, and beyond.
3. An just-in-time learning system based on RAG is realized, capable of acquiring knowledge instantly.
4. Xiwu, the first LLM specilized for high energy physics outperforms the foundation model in accuracy for domain-specific knowledge question answering, and exceeds GPT-4 in BOSS (BESIII Offline Software System) code generation.
5. The WebUI application has been deployed into HepAI platform, the related codes is open-sourced on GitHub.

## 2 Related Work

**Language Models** The introduction of deep learning has greatly propelled the development of language models. Early applications of neural networks to language models, such as Feedforward Neural Networks[8], laid the groundwork. The emergence of RNN[9, 10, 11] and LSTM[12] models advanced language model by effectively capturing temporal dependencies. However, their inability to process sequences in parallel limited their efficiency. In 2017, the introduction of the Transformer model[13], with its self-attention mechanism, overcame these limitations, significantly boosting the handling of long-distance dependencies and computational efficiency, propelling the development of language models into a new era. Building on this architecture, BERT[14] and OpenAI's GPT series[5, 4] were developed for substantial advancements in NLP tasks through pre-training and fine-tuning. Subsequent works [15, 16, 17, 18] further refined and enhanced these methodologies, leading to continuous performance improvements. In 2020, The release of GPT-3[19], a significantly larger laguage model, underscored the effectiveness of few-shot and zero-shot learning, prompting a

significant trend toward model scaling. By 2022, the launch of ChatGPT[7] marked milestones in NLP, showcasing the potential of Large Language Models (LLMs) as efficient AI communicators. Moreover, the open-source model Vicuna[6], which is based on fine-tuning LLAMA[2], made a substantial contribution to the open-source community by achieving up to 90% of GPT-4's[1] performance, sparking widespread research interest.[20, 21, 22, 23].

**LLMs in Specific-Domain** Recent research has shown that BERT models, when fine-tuned for specific domains, can achieve significant performance improvements. This trend is exemplified by models such as BioBERT[24](targeted at the biomedical domain), SciBERT[25] (for scientific literature), and ClinicalBERT[26](for clinical data). However, the high costs associated with pre-training large-scale, domain-specific models like Galactica[27] and BloombergGPT[28] underscore the value of domain-specific fine-tuning as a more practical and cost-effective strategy. For instance, Flan-PaLM[29] constructs a specialized large model for the clinical medicine domain by combining few-shot learning, chain-of-thought (CoT) prompting strategies, self-consistency prompting, and instruction fine-tuning. WizardMath[30] fine-tunes LLAMA2[3] using the Reinforcement Learning from Evol-Instruct Feedback method to create a large model for mathematical reasoning. Meanwhile, SciGLM[31] develops a specialized large model for the scientific domain by fine-tuning with high-quality datasets built using a self-reflexive annotation framework. These studies showcase the potential and value of fine-tuning large language models for specific domains.

**LLMs in HEP** In the field of high-energy physics, ChATLAS has trained a domain-customized model using documents such as Twiki, ATLAS docs, and the E-group Archive, providing AI services to the collaboration group. AccGPT, developed based on LLAMA, is used for accelerator auxiliary control at CERN. Based on ChatGPT, the ATLAS Open Data Higgs Analysis Guide can assist individuals interested in physics analysis who are not specialized in high-energy physics. LLMs are being experimented with in various types of task of HEP.

### 3 Methodology

#### 3.1 Overall Architecture

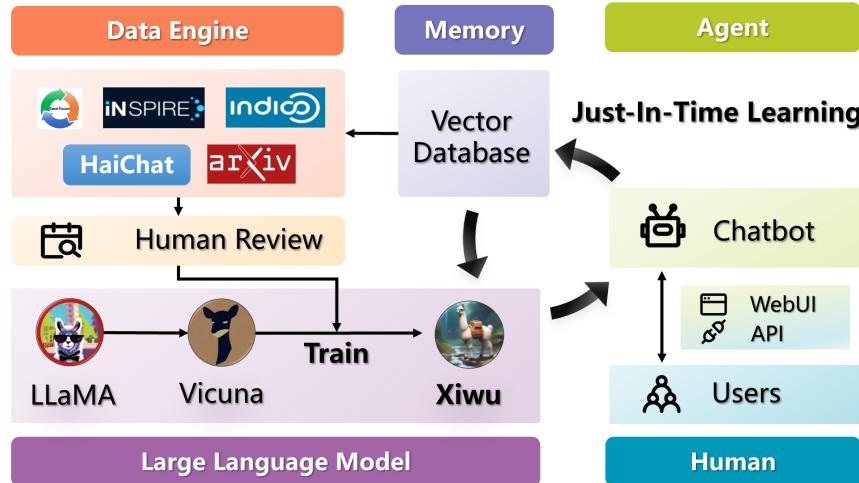


Figure 2: The architecture of Xiwu large language model system.

The architecture of Xiwu LLM is shown in FIG. 2. It consists of four parts: data engine, large language model, memory based on external knowledge lib and intelligent agent with its interface.

**Data Engine** The data engine is primarily used to collect text data from eight fields related to HEP. The specific methods of data collection include seed fission technology, real concerns from ChatBot, knowledge from highly cited full-text literature, and a large number of literature abstracts. The data is cleaned and organized using LLM, and reviewed by professional staffs.

**Xiwu LLM** Based on the Level 0 Model LLaMA open-sourced by Meta, the Lmsys team has implemented a Level 1 Vicuna through secondary training. Currently, Xiwu is a Level 2 model trained on high-energy physics data based on Vicuna. With the upgrades of LLaMA and Vicuna, Xiwu's current base model is Vicuna-1.5 (corresponding to LLAMA2), with parameter sizes of 7B and 13B. Plans are in place to adapt to even more advanced models.

**Memory Module** RAG has been widely used to mitigate hallucinations in LLMs, essentially enhancing the model's capabilities through external knowledge library. Similarly, we vectorize HEP data using an Embedding model and store it in a vector database as the long-term memory for Xiwu. Additionally, the process of adding or modifying information in the vector database represents a low-cost, rapid, and just-in-time learning process.

**Intelligent Agent** The intelligent agent is structured around the Xiwu model, supplemented by an onboard knowledge base. Utilizing our proposed HepAI Distributed Deployment Framework (HepAI-DDP), we've successfully implemented both an Application Programming Interface (API) and a ChatBot Web User Interface (WebUI). This advancement significantly simplifies the interaction process between the intelligent agent and its human counterparts, be they scientists or developers.

**Learning Loops** There are two learning loops in the system. The first loop is a just-in-time learning system, wherein users can directly embed accurate information into the knowledge base via the WebUI or API. Xiwu then retrieves and synthesizes this information to provide answers to queries. The second loop involves stable, unchanging information being integrated into the HEP text datasets, which are subsequently used to fine-tune the weights of Xiwu model. The former loop facilitates the rapid assimilation of evolving information, while the latter progressively enables the model to "understand" knowledge.

### 3.2 Datasets

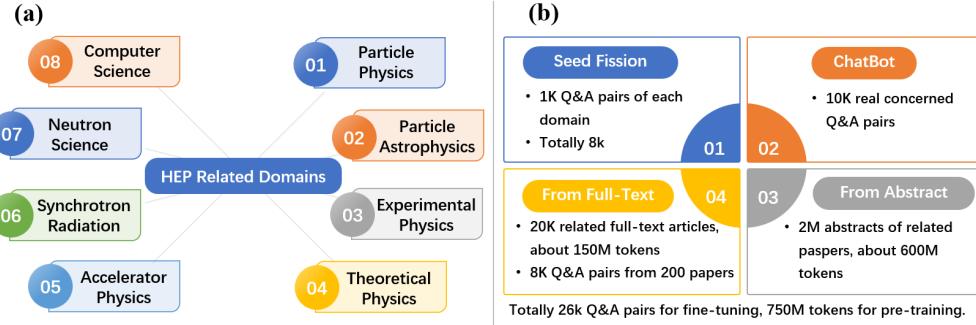


Figure 3: The data resources and acquisition methods. (a) Eight domains related to High Energy Physics that are of our concern; (b) Four methods employed to gather the dataset

The FIG. 3 shows HEP text data from eight sub-domains and the data collection methods. As shown in FIG. 3 (a), in order to ensure the diversity and breadth of the model, we extensively gathered Q&A pairs and literatures from eight HEP related domains, including particle physics, particle astrophysics, experimental physics, theoretical physics, accelerator physics, synchrotron radiation, neutron science and computer science. Data of each field is collected using the methods shown in FIG. 3 (b) on the right. There are four methods we used to collect data, including seed fission technology, chat robot backend, full-text of highly-cited papers and abundant abstracts of the papers. Through those methods, totally 26k Q&A pairs for fine-tuning and 750M tokens for pre-training are collected and cleared up. The details will be described below.

#### 3.2.1 Data from Seed Fission Technology

**Seed Fission** The seed fission technology was proposed by us to quickly obtain related Q&A pairs in a specific domain. This technology allows a single seed such as *Particle Physics*, can fission into a multitude of diverse and in-depth Q&A pairs. By employing this technique, we successfully generated more than 1K Q&A pairs for each HEP-related domain as shown in FIG. 3 (a). After a thorough human review, we obtained about 8K high-quality data entries.

**Fission Process** As shown in FIG. 4, our approach involved designing three chat robots through prompt engineering: Newbie, Expert and Checker. The Newbie asking questions with curiosity based on information. The Expert providing answers based on information and reliable sources as much as possible. The Checker is a topic supervisor that responsible for selecting topics and determining when to transition away from the current topic. The fission process is as follows:

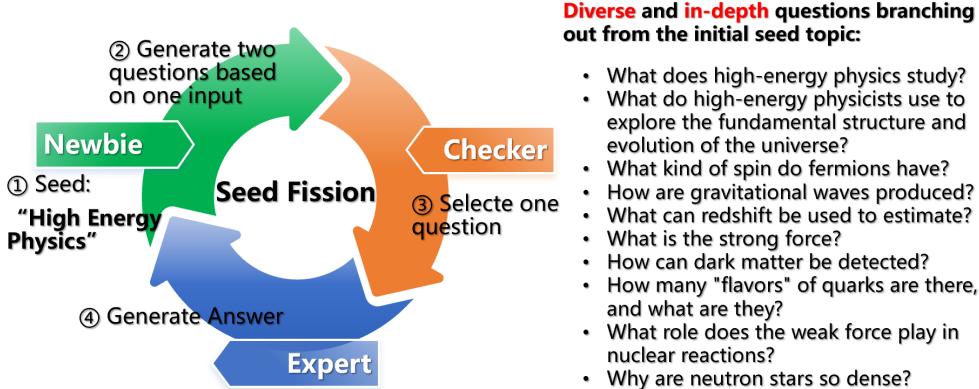


Figure 4: The seed fission technology for getting diverse and in-depth data

1. The initial topic is considered a seed, such as "Particle Physics," is given to the Newbie.
2. The Newbie generates 2 to 10 random questions based on the input. A single input leading to multiple outputs is thus referred to as fission.
3. These questions are then reduplicated and filtered by the Checker, selecting the most interesting ones, or deciding to exit the current topic to choose another candidate topic. The selected question is directed to the next robot, while the remaining questions go into a candidate question list.
4. The Expert answers questions based on the knowledge base and search engine it is equipped with, and provides the source of the answers.
5. The answer serves as input information for the Newbie, who then poses more questions.

**Data Quality** By using "Particle Physics" as the "seed," 50 question-answer pairs and 2,822 candidate questions can be generated. Some samples are shown in FIG. 4, from just this seed alone, concepts related to high-energy physics such as fermions, spin, gravitational waves, and redshift can be derived, thus the data generated is diverse and in-depth.

**Advantage** The significant advantage of this technique is that it allows us to generate a large volume of relevant and diverse question-answer datasets with depth using just one topic as a guide.

**Limitations** Initially, a notable limitation is the tendency for both Newbie and Expert robots to become ensnared within similar topics. Subsequently, we introduced the Checker to oversee this process, effectively addressing the issue. Another limitation is that the quality of the generated answers depends on the robot itself, and there is a possibility of generating incorrect or even harmful information, equipping the Expert with a trustworthy knowledge base, knowledge retrieval capabilities, and incorporating human verification can to some extent solve this problem.

### 3.2.2 Data from Chat Robot

**HaiChat Sever** We developed HaiChat [32] to provide generative model services for HEP researchers. The backend includes API for close models, local deployment of open-source models, and Xiwu model. High Energy Physics (HEP) researchers utilize the HaiChat service in their daily work, generating real questions of interest and corresponding answers. This serves as an important data source for further improving the performance of the Xiwu model.

**Data and Quality** As of now, HaiChat has generated about 600,000 Q&A pairs, of which 8%, or 48,000, are related to high-energy physics. To ensure the quality and accuracy of the data, we employed a multi-step filtering and cleaning process. First, we utilized the ChatGPT model to automatically filter out irrelevant or inaccurate dialogues from the collected question-answer pairs. Then, we performed manual review and editing to further enhance the quality and usability of the data.

### 3.2.3 Data from Highly Cited Papers

**Data Sources** For the eight sub-domains mentioned above, We analyzed the citation data of papers over the past five years and used tools to download a carefully selected set of 20,000 papers.

**HaiNougat PDF Parser** Based on the Nougat [33] model, we trained a PDF parsing algorithm, HaiNougat [34], using a domain-specific dataset, which achieves more accurate parsing of formulas and tables of HEP.

**Q&A Pairs Generation and Human Verification** Based on the parsed data, we extracted the core contributions, viewpoints, and results of the papers through prompt engineering, obtaining relevant question-and-answer pairs. To ensure quality, these were reviewed by humans.

**Data Collected** From those highly cited papers, about 8,000 Q&A pairs and 150M tokens were collected.

### 3.2.4 Abstracts Data from arXiv

Kaggle offers comprehensive coverage of research papers across various domains, but it does not provide the full text. We downloaded approximately 2 million paper abstracts from the aforementioned eight related fields on Kaggle. The quality of these abstract data is very high, they can be directly used for the secondary pre-training of the Xiwu model.

### 3.2.5 Final Dataset

After data collection and cleaning, a AI-Ready dataset of HEP texts was ultimately formed that can be used for LLM training, which includes 750M tokens for secondary pre-training and 26k Q&A pairs for fine-tuning.

## 3.3 Neural Networks

### 3.3.1 Model

**Basis Flexible Design** Xiwu is positioned as a Level 2 LLM. With the goal of minimizing costs while maintaining advanced capabilities, it has been designed to be adaptable to foundational models of Level 0 or Level 1, including LLaMA, Vicuna, Chat-GLM, Grok-1 and more in the future. This adaptability is implemented in the code repository through an 'apis' folder, where interfaces for any model have appropriate API adaptations. For more details, you can refer to the Github [35] repository.

**Current Foundation Model** The Vicuna [36] model was chosen as the foundational model after a comparative analysis, as its overall performance surpassed that of others like LLaMA-13B, LLaMA2-13B, Koala-13B, Oasst-pythia-13B, Alpaca-13B, and ChatGLM-6B. Vicuna is a Level 1 model based on LLAMA, and currently, we are using Vicuna-1.5, which corresponds to LLAMA-2. In this model, the transformer architecture is employed. Compared to Recurrent Neural Networks (RNNs), the transformer model has proven to be superior in quality for many sequence-to-sequence tasks while being more parallelizable.

**Modeling of NLP Tasks** The language modeling task is to assign a probability for the likelihood of a given word (or a sequence of words) to follow a sequence of words. A sequence of tokens are passed to the embedding layer first, followed by a positional encoding layer to account for the order of the word. Positional Encoding module injects some information about the relative or absolute position of the tokens in the sequence. The BPE (Byte-Pair Encoding) is used, the size of vocabulary is 32000. The positional encoding, RoRE and Pre-LN (layer normalization) [37] is used..

The Pre RMS Norm is as follows:

$$RMS(x) = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$$

The SwishGLU is used as activation function, there are tree trainable weight matrix in SwishGLU, the formular of SwishGLU is

$$SwishGLU = Swish_{\beta}(xW) \otimes xV$$

The  $n_{heads}$  is 32, the  $head_{dim}$  is 128, the  $FFN_{dimension}$  is 11008, the  $h_{dimension}$  is 4096.

### 3.3.2 Training Techniques

During the training process, several advanced techniques are employed to enhance the performance and efficiency of the Xiwu model. These techniques, namely BF16 and TF32 mixed-precision training, FlashAttention, FSDP (Fully Sharded Data Parallelism), CPU-Offload, and Torchrun, played a crucial role in achieving improved training outcomes. In this section, we provide a detailed explanation of these techniques and their effects.

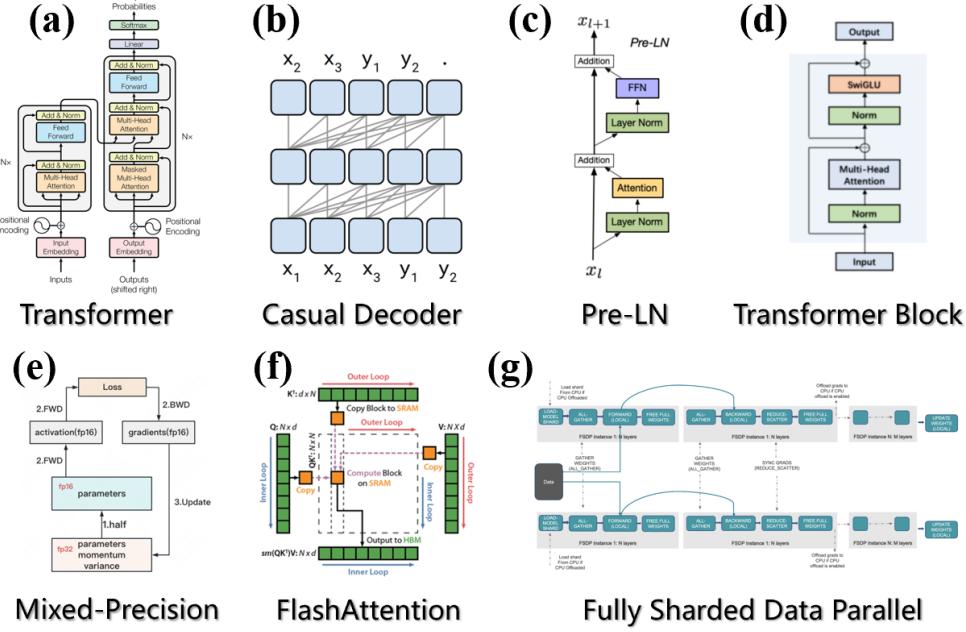


Figure 5: The illustration algorithm components and training technologies

**Mixed-precision Training** Mixed-precision training[38] involves performing computations using a combination of lower (BF16) and higher (TF32) precision formats. By utilizing BF16 for most of the computations and TF32 for certain critical operations, we were able to reduce memory usage, improve computational throughput, and accelerate the training process without sacrificing model accuracy.

**FlashAttention** FlashAttention [39] is a technique designed to enhance Flops (floating-point operations per second) utilization during self-attention computations, a critical component in transformer-based models like Xiwu. It optimizes the memory access patterns and computation flow, resulting in improved computational efficiency and reduced training time.

**FSDP** Fully Sharded Data Parallelism (FSDP)[40] is a data parallelism technique that enables efficient parallel training across multiple devices by dividing the model’s parameters into shards and assigning each shard to a different device. By distributing the model’s parameters, FSDP reduces memory consumption per device, enables larger models to fit within device memory constraints, and facilitates seamless parallelization for faster and more scalable training.

**CPU-Offload** CPU-Offload[41, 42] is a strategy used to alleviate excessive GPU memory usage during training. By offloading certain computations to the CPU, we were able to free up GPU memory for more efficient memory utilization. This approach allowed us to train Xiwu on GPUs without encountering memory limitations and improved the overall training performance.

**Torchrun for Single-Node Multi-GPU Parallelism** Torchrun is a utility that facilitates single-node multi-GPU parallelism in PyTorch[43], the framework used for developing Xiwu. By leveraging Torchrun, we achieved an effective distribution of computations across multiple GPUs within a single node, allowing for parallel training and accelerated model convergence.

**Bitsandbytes** Bitsandbytes[44] is an innovative software designed to address the substantial GPU memory requirements encountered by LLMs during inference. Employing vector quantization and mixed-precision decomposition, Bitsandbytes substantially reduces memory consumption without compromising on performance. It enables the deployment of large-scale models, such as those with 175B parameters like OPT-175B/BLOOM[45], on consumer-grade GPU servers, presenting a highly efficient solution for managing LLMs.

**Deepspeed** The Deepspeed framework[46] is designed to accelerate deep learning training by utilizing mixed precision, gradient checkpointing, and offloading memory to the host CPU. It allows for larger models and batch sizes, reducing memory requirements and improving training speed.

**ZeRO Series** Zero Redundancy Optimizer series techniques aim to reduce memory consumption during training by partitioning model weights and optimizer states across multiple devices or nodes. These techniques include ZeRO-2[47], ZeRO-3[47], and ZeRO-offload[48]. By leveraging memory optimization strategies such as activation checkpointing and optimizer state partitioning, ZeRO techniques enable training larger models that would otherwise exceed available memory limits. Adopting ZeRO techniques can potentially allow us to scale up Xiwu and achieve even better performance.

**LoRA** Low-Rank Adaptive technique[49] is a low-rank approximation technique that reduces the computational cost of large-scale language models. It achieves this by approximating the weight matrices with low-rank factors while preserving the model's representational capacity. By leveraging the inherent redundancy in model parameters, LoRA significantly reduces both the memory and compute requirements, making training and inference more efficient.

These techniques collectively contributed to the enhanced performance and efficiency of Xiwu during the training process. By leveraging mixed-precision training, FlashAttention, FSDP, CPU-Offload, and Torchrun, we achieved improved memory utilization, accelerated training speed, and scalable parallelization. These advancements enable us to train Xiwu more effectively, ultimately leading to improved performance in high-energy physics question-answering tasks.

### 3.3.3 Computing Power and Training Settings

**Computing Power** We conducted full-scale training of the Xiwu-13B model using eight A100 GPUs, each with a memory capacity of 40 GB. Additionally, the Xiwu-7B model can be trained on two A100 GPUs with a memory capacity of 80 GB each.

**Training Parameters** The following parameters were carefully selected and tuned to optimize the training process and improve the performance of Xiwu:

- Learning rate: 2e-5
- Learning rate scheduler: cosine
- Weight decay: 0.0
- Warmup ratio: 0.03
- Per-device training batch size: 2
- Gradient accumulation steps: 16

**Learning Rate** The learning rate determines the step size at which the model adjusts its weights during training. A value of 2e-5 was chosen to strike a balance between ensuring stable convergence and preventing overshooting or getting stuck in sub-optimal local minima.

**Learning Rate Scheduler** The cosine scheduler gradually reduces the learning rate as training progresses. It helps the model to converge smoothly by reducing the learning rate towards the end of training. The cosine scheduler was selected due to its proven effectiveness in stabilizing training and improving the model's generalization ability.

**Weight Decay** Weight decay is a regularization technique that adds a penalty to the loss function based on the magnitude of the model's weights. A weight decay value of 0.0 indicates that no additional regularization is applied during training, allowing the model to fully utilize the available information in the training data.

**Warmup Ratio** The warmup ratio determines the proportion of training steps dedicated to gradually increasing the learning rate from zero to its initial value. A warm-up ratio of 0.03 ensures that the model's learning rate ramps up slowly at the beginning of training, allowing the model to stabilize and avoid sudden, large weight updates.

**Per-device Training Batch Size** This parameter determines the number of samples processed in parallel on each device during training. A batch size of 2 per device strikes a balance between utilizing the available memory efficiently and maintaining reasonable training speed.

**Gradient Accumulation Steps** Gradient accumulation is a technique that helps simulate larger batch sizes by accumulating gradients over multiple smaller batches before performing weight updates. With 16 gradient accumulation steps, the model benefits from an effective utilization of computational resources while still obtaining accurate gradient estimates.

It is important to note that the presented configuration and training setup can be adjusted according to the available resources and specific requirements of different applications.

### 3.4 Just-In-Time Learning System

**Limitations of Secondary Training** Although secondary training of LLMs is effective, it is costly, inefficient, and training can lead to a degradation in model performance. For example, when dealing with unseen information, the model cannot effectively learn new information under the usual learning intensity (lower learning rates and fewer epochs). If the learning intensity is increased, the model can learn new information, but then issues such as performance degradation on other Q&A tasks and repetitive answer fragments may occur.

**Just-In-Time Learning** Just-In-Time Learning (JITL) is a learning system that utilizes a vector database for long-term memory enhancement to augment model capabilities, operating in a basic process equivalent to RAG. We call it JITL to emphasize its learning attributes. JITL is capable of rapid knowledge learning in seconds, specific knowledge updating, and unloading. These features are not present in the traditional model training approach.

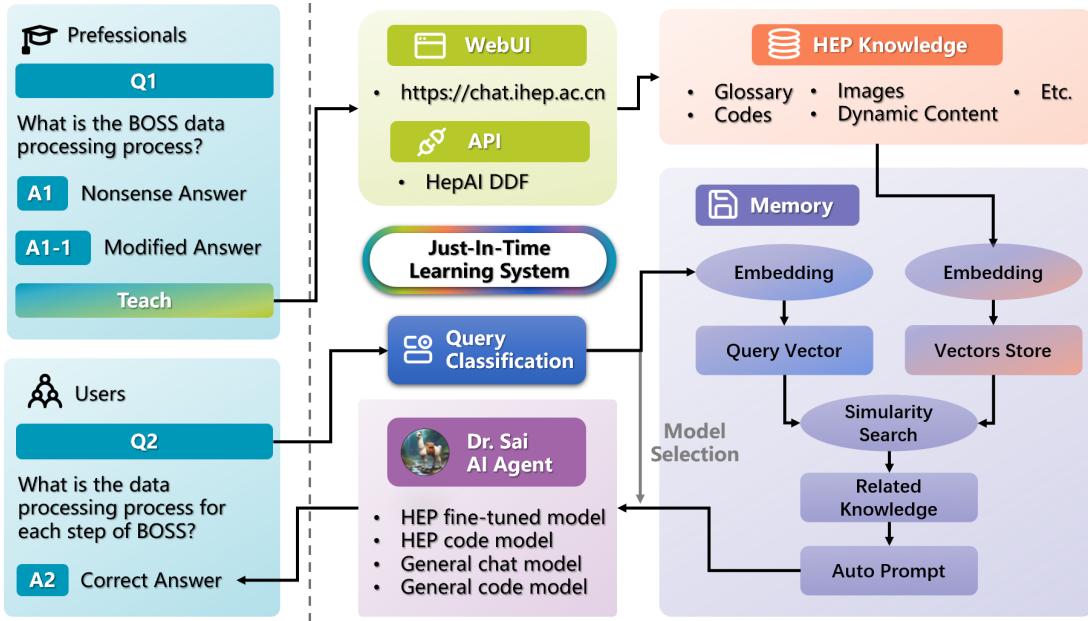


Figure 6: The just-in-time learning system

**Implementation of JITL** The architecture of the Just-in-time Learning system is illustrated in FIG. 6, where the core concept is integrating an external knowledge base as the memory part of the model. New information is saved in the knowledge base, and after retrieval, comprehensive information is outputted. The complete information flow of the JITL system is as follows:

1. When an expert user asks the model a specialized question related to HEP and receives a nonsensical answer, the expert can suggest modifications or directly edit the answer, then teach the model via WebUI or API.
2. The system categorizes and organizes the new information, storing it in a text database.
3. The text data is processed by a pre-trained embedding model to generate feature vectors, which are then stored in a vector database. These feature vectors capture the semantic information of the text, and semantically similar expressions are close to each other in the feature space.

4. When ordinary users ask the model questions, these go through a question classifier and select an appropriate model, while text data is converted into feature vectors by the embedding model. As different texts are separable in high-dimensional vector space, a simple similarity search can retrieve stored related texts.
5. The related texts and query are processed through automatic prompt engineering and then fed into a suitable LLM such as Xiwu. Due to the intent understanding and contextual learning abilities of Xiwu, the final output is an accurate answer that also includes information sources.

**Advantages** Compared to training models, Just-In-Time learning systems only store vectors after embedding into the database, making the learning process extremely fast, achieving learning in seconds. Knowledge can be dynamically updated, which is particularly effective for rapidly changing information. Knowledge can also be specifically unloaded, a feature not achievable in traditional model training processes. Compared to conventional retrieval systems, Just-In-Time learning systems accomplish semantic-level rather than keyword-level retrieval, resulting in substantial performance improvements.

**Limitations** Just-In-Time learning systems have limitations. First, the additional vector database search process can be time-consuming, leading to higher latency compared to pure LLMs. One mitigation strategy is to use GPU acceleration to speed up the search process, another approach is to train the model with information that is always correct. The second limitation is that the model doesn't truly "understand" the knowledge in the external database, meaning there's no improvement in reasoning ability. This mode is akin to open-book exams where students don't have to understand the material but can look up information when needed.

## 4 Evaluation

### 4.1 Evaluation between Models

**Evaluation Data** To assess the performance of Xiwu in comparison to other models, we conducted a human evaluation using a set of 5,100 prompts. These prompts consisted of questions related to particle physics, astrophysics, synchrotron physics, and neutron science. Examples of these questions include: "What quarks make up protons and neutrons?", "What is general relativity?", "Why can't we directly observe dark matter?", and "Why is particle therapy more expensive than X-ray treatment?". The evaluation dimensions included accuracy, clarity, and fluency. Accuracy assessed whether the answers correctly addressed the questions and contained the necessary information. Clarity evaluated the clarity and comprehensibility of the answers, including the use of accessible language and terminology. Fluency assessed the overall fluency and coherence of the answers. Human scientists participated in the evaluation process and compared the answers generated by Xiwu with those from other models, namely Vicuna-13B and ChatGPT. The evaluation results were categorized into three outcomes: Xiwu wins, Xiwu fails, and Draw.

**Comparison Results** As shown in the FIG. 7, Xiwu-13B achieved a win or draw rate of 95% when compared to the baseline model Vicuna-13B. The performance of Xiwu-13B reached 65% of ChatGPT-175B, while Vicuna achieved less than 10%. These findings indicate that fine-tuning a large-scale model with domain-specific data, even on a model with 13 billion parameters, can significantly improve the accuracy of domain-specific question answering. Fine-tuning smaller models based on general-domain knowledge has the potential to achieve or even surpass the performance of larger models in specialized domains.

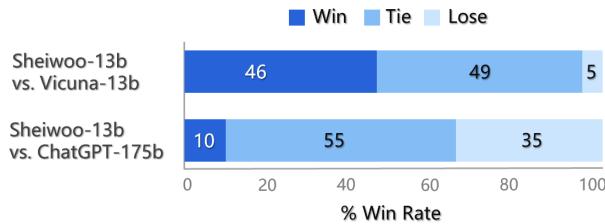


Figure 7: Human preference evaluation, Compare Xiwu-13B with Vicuna-13B and ChatGPT-175B

**Fine-tuning Cannot Learn New Knowledge** Furthermore, we also tested the ability of Xiwu to incorporate internal knowledge that was not pre-trained. Examples of such knowledge included "What is HepAI?" (a high-energy physics artificial intelligence platform we are developing) and "What is HaiGF?" (an AI application interface framework we are

developing). The results showed that a limited number of epochs were insufficient to guide Xiwu in answering these types of questions. Increasing the number of epochs not only failed to yield correct answers but also led to a decrease in the model's performance on other questions.

## 4.2 Absolute Evaluation

**Results** In the absolute evaluation, human scientists assessed the content of each Xiwu answer, categorizing them as Excellent (Xiwu provided an outstanding response), Pass (Xiwu's answer was acceptable), or Fail (Xiwu's answer was incorrect). During the absolute evaluation, it was observed that occasionally the 13-billion-parameter Xiwu model produced answers that did not understand the human intent, resulting in responses that were unrelated to the questions. Additionally, there were instances where the model repeated the same text. This behavior may be attributed to the lack of reinforcement learning from human feedback. Further investigation and incorporating reinforcement learning techniques are necessary to address these issues.

## 5 Conclusion

**First LLM for HEP** In this work, we developed Xiwu, the first large language model customized for high energy physics. Xiwu features a flexible foundational model and two distinct learning systems. The flexibility of the foundational model allows Xiwu to evolve alongside the development of open-source models. The learning system based on model training is traditional, enabling the model to effectively learn and "understand" HEP knowledge. The vector database-based just-in-time learning system enables rapid learning of new knowledge, dynamic knowledge updates, and unloading, all at a low cost.

**Results** Currently, Xiwu is an Level 2 LLM based on Vicuna. Xiwu-13B significantly outperforms the Vicuna-13B on the HEP domain Q&A test set, achieving about 65% of the performance of ChatGPT-175B. Furthermore, our just-in-time learning system enables multiple individuals to collectively teach Xiwu, demonstrating the potential for collaborative AI training. The seed fission technology we have developed has potential for wide application, capable of generating Q&A pair data in diverse fields.

**Meaning** This paper not only presents our findings but also provides the corresponding implementation code, which is readily available on GitHub [35]. We hope that our work will inspire further research and development in the application of large language models in specialized scientific fields.

## Acknowledgments

This work is Supported by the Informatization Plan of Chinese Academy of Science, Grant No. CAS-WX2022SF-0104 and "From 0 to 1" Original Innovation Project of IHEP, Grant No. E3545PU2. We would like to express our gratitude to Beijiang Liu, Yaquan Fang, Gang Li, Wuming Luo, Ye Yuan, Shengsen Sun, Yi Jiao and others who are not listed here for engaging in beneficial discussions or providing computing resources.

## 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] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [5] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

- [6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- [7] OpenAI. Introducing chatgpt, 2022.
- [8] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.
- [9] Tomáš Mikolov et al. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80(26), 2012.
- [10] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.
- [11] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.
- [12] Alex Graves and Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012.
- [13] 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.
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [15] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- [16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [18] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.
- [19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [20] Deepanway Ghosal, Yew Ken Chia, Navonil Majumder, and Soujanya Poria. Flacuna: Unleashing the problem solving power of vicuna using flan fine-tuning. *arXiv preprint arXiv:2307.02053*, 2023.
- [21] Pritam Mukherjee, Benjamin Hou, Ricardo B Lanfredi, and Ronald M Summers. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*, 309(1):e231147, 2023.
- [22] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [23] Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, and Le Sun. Toolalpac: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.
- [24] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [25] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.
- [26] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.
- [27] 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.

- [28] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrowski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.
- [29] 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. *arXiv preprint arXiv:2212.13138*, 2022.
- [30] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.
- [31] 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. *arXiv preprint arXiv:2401.07950*, 2024.
- [32] Zhengde Zhang and Yiyu Zhang. Haichat: High energy physics generative artificial intelligence chat robot service. <https://chat.ihep.ac.cn>, 2023. Accessed: 2024-04-07.
- [33] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.
- [34] Jianwen Luo and Zhengde Zhang. Hainougat: A academic document parser that preserves formulas and tables for high energy physics. <https://ai.ihep.ac.cn/m/hai-nougat>, 2024. Accessed: 2024-04-07.
- [35] Zhengde Zhang, Yiyu Zhang, Haodong Yao, Jianwen Luo, Rui Zhao, Bo Huang, Jiaomeng Zhao, Yipu Liao, Ke Li, Lina Zhao, Jun Cao, Fazhi Qi, and Changzheng Yuan. Xiwu: A basis flexible and learnable llm for high energy physics. <https://github.com/zhangzhengde0225/Xiwu>, 2024.
- [36] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023.
- [37] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.
- [38] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.
- [39] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022.
- [40] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.
- [41] Bharadwaj Pudipeddi, Maral Mesmakhosroshahi, Jinwen Xi, and Sujeeth Bharadwaj. Training large neural networks with constant memory using a new execution algorithm. *arXiv preprint arXiv:2002.05645*, 2020.
- [42] Minsoo Rhu, Natalia Gimelshein, Jason Clemons, Arslan Zulfiqar, and Stephen W Keckler. vdnn: Virtualized deep neural networks for scalable, memory-efficient neural network design. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 1–13. IEEE, 2016.
- [43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037. Curran Associates, Inc., 2019.
- [44] Dettmers Tim, Belkada Mike, Lewis adn Younes, and Zettlemoyer Luke. bitsandbytes: Highly optimized bit and byte level operations for deep learning. <https://github.com/TimDettmers/bitsandbytes>, 2023.
- [45] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [46] Microsoft. DeepSpeed: A deep learning optimization library. <https://github.com/microsoft/DeepSpeed>, 2023. Accessed: 2023-10-01.
- [47] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. *arXiv preprint arXiv:1910.02054*, 2020.

- [48] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. {Zero-offload}: Democratizing {billion-scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564, 2021.
- [49] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.