Computing Center, IHEP, CAS
National HEP Data Center

AI Progress in IHEP

Zhengde Zhang, on behalf of IHEP-CC
Beijing
Dec, 4, 2024

Outline

01 Introduction
02 HepAI platform
03 Agents based on LLMs
04 Summary

Why AI for HEP?

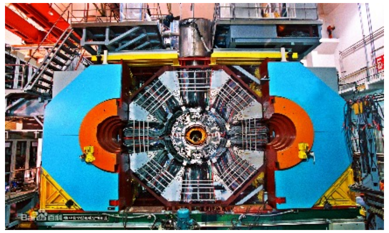
From Large Scientific Facilities to Big Data, then AI becomes a necessity.

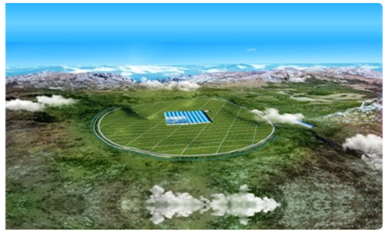
LFS of IHEP:

- Beijing Electron Positron Collider (BEPC)
- Beijing Spectrometer (BESIII) Experiment
- Jiangmen Underground Neutrino Observatory (JUNO)
- China Spallation Neutron Source (CSNS)
- Large High Altitude Air Shower Observatory (LHAASO)
- Beijing Synchrotron Radiation Facility (BSRF)
- High Energy Photon Source (HEPS)

HEP Big Data:

- The National HEP Data Center was established in 2019.
- Currently, over **40PB** of scientific data has been stored. (1PB=1024TB)
- HEPS is expected to generate 800TB/day of raw data after completion in 2025

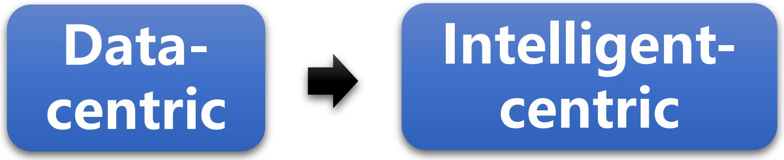

BESIII


LHAASO

CSNS

HEPS

Scientific Research Paradigm:

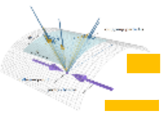
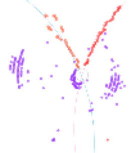
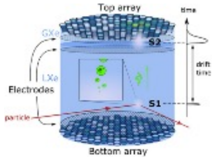

Data-centric → Intelligent-centric

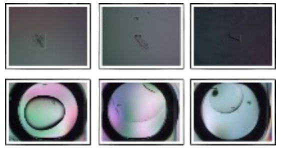
$\Omega(n^3) \rightarrow \Omega(n)$

AI offers powerful solutions for unlocking physical potential in massive datasets.

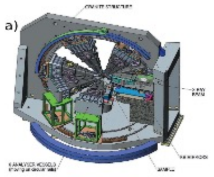
Overall Strategy for High Energy Physics AI

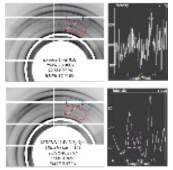
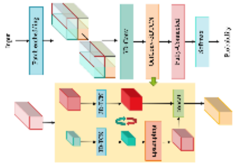
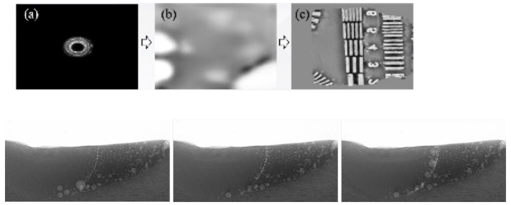
The 14th Five-Year Plan of the IHEP:



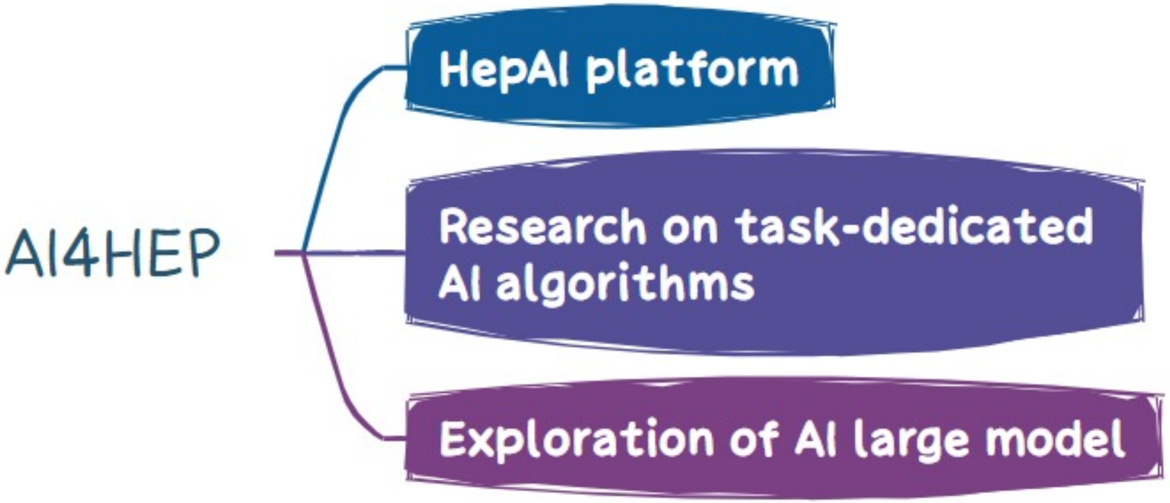
- Advanced computing technologies including AI, QC make more powerful modeling and simulation possible, which are indispensable means for major breakthroughs in HEP.
- It is necessary to strengthen the application of ML in **real-time processing, simulation, reconstruction, analysis**, etc., to enhance the ability of high energy physics exploration and new discoveries.



Our Exploration experiences:

- Prioritizing the development of **AI-Ready scientific datasets** is of utmost importance.
- Insist on the concurrent development of **software frameworks** and **specific AI applications**.
- Investigate the **large models** to propel new discoveries of HEP.



AI4HEP

HepAI platform

Research on task-dedicated AI algorithms

Exploration of AI large model

Part. 01

HepAI platform

The HepAI Platform

HepAI Platform

- Accelerate scientific research in multi-disciplinary scenarios.
- Simplify model iteration and flow.
- Serve as a **common basic infrastructure** for the development of AI algorithms and applications.

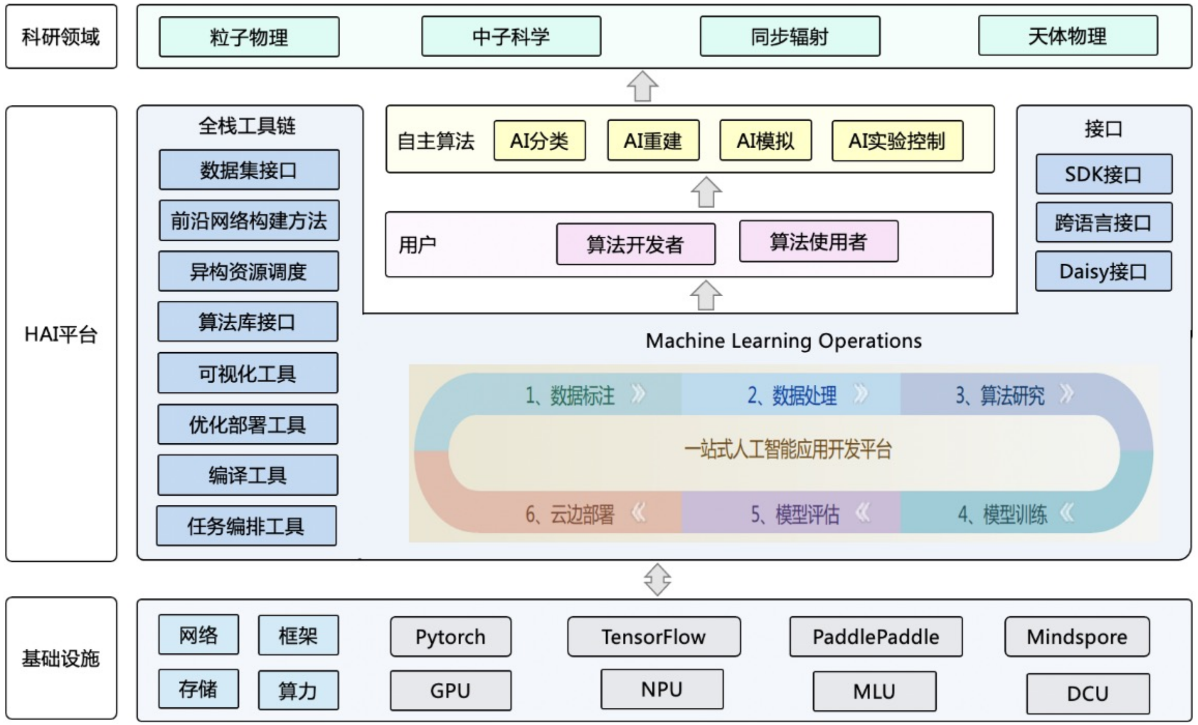
Three Key Elements of AI:

Data **Algorithms** **Comp. Power**

- HepAI Core Codes and Framework（80%）
- 10 AI algorithms.
- 4 AI datasets.
- Heterogeneous computing resources including GPU, NPU, and DCU.

Goal: Make AI4HEP **simpler and more advanced!**

Portal site: https://ai.ihep.ac.cn
Open source: https://code.ihep.ac.cn/zdzhang/hai



科研领域
粒子物理 中子科学 同步辐射 天体物理

HAI平台

全栈工具链
数据集接口
前沿网络构建方法
异构资源调度
算法库接口
可视化工具
优化部署工具
编译工具
任务编排工具

自主算法 AI分类 AI重建 AI模拟 AI实验控制

用户 算法开发者 算法使用者

接口
SDK接口
跨语言接口
Daisy接口

Machine Learning Operations

1、数据标注 2、数据处理 3、算法研究
一站式人工智能应用开发平台
6、云边部署 5、模型评估 4、模型训练

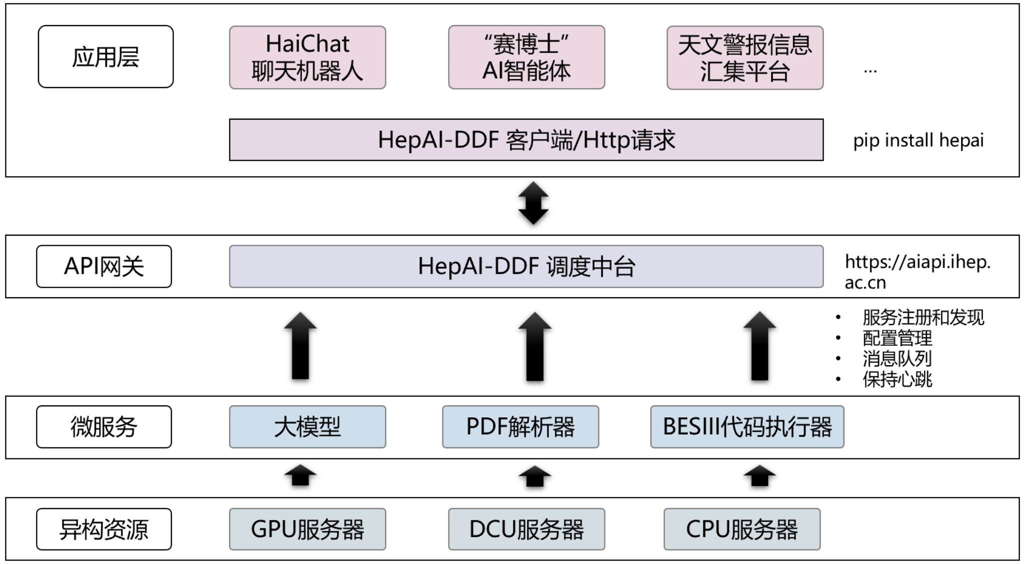
基础设施
网络 框架 Pytorch TensorFlow PaddlePaddle Mindspore
存储 算力 GPU NPU MLU DCU

The architecture of HepAI platform

Key Core Tech. HaiDDF Upgrade to v2

HepAI Distributed Deployment Framework

HaiDDF2, short for High-energy AI Distributed Deployment Framework, is a key core technology of HepAI platform. It provides crucial technology for unifying access to services such as AI models, scientific tools, vector databases, and knowledge graphs deployed on heterogeneous resources.



应用层
HaiChat 聊天机器人
"赛博士" AI智能体
天文警报信息汇集平台
...
HepAI-DDF 客户端/Http请求
pip install hepai

API网关
HepAI-DDF 调度中台
https://aiapi.ihep.ac.cn

- 服务注册和发现
- 配置管理
- 消息队列
- 保持心跳

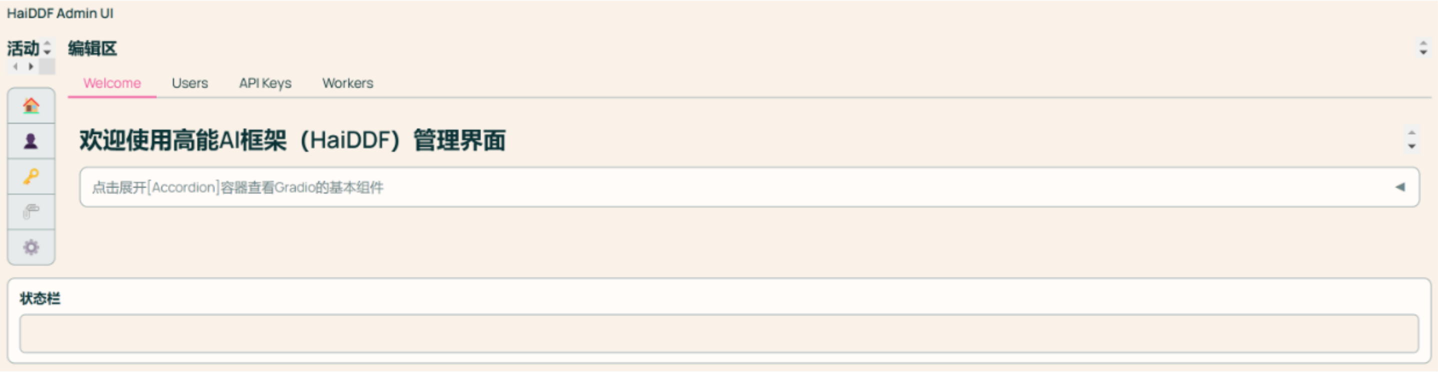
微服务
大模型
PDF解析器
BESIII代码执行器

异构资源
GPU服务器
DCU服务器
CPU服务器

Features :
- Unified API Gateway with API-KEY authentication
- Support for unified access to models and algorithms deployed on heterogeneous resources like CPUs, GPUs, and DCUs
- Support for services including AI models, scientific tools, vector databases, and knowledge graphs
- Automatic heartbeat monitoring and load balancing scheduling
- Supports both streaming and non-streaming outputs
- **Remote Models**! Deploy models on a remote server and use the get_remote_model method to obtain remote model instances, allowing you to call them just like local models.

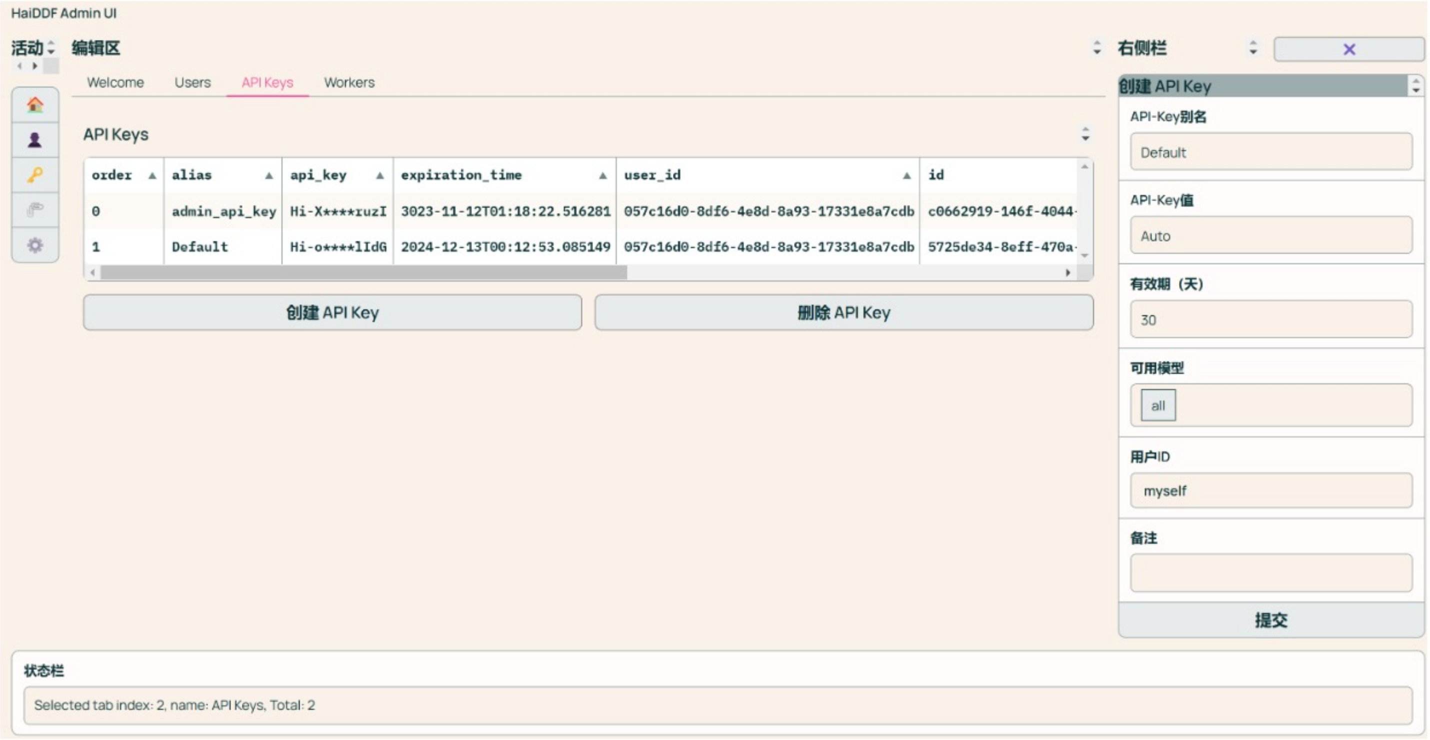
HaiDDF2

Admin UI



欢迎使用高能AI框架（HaiDDF）管理界面

点击展开[Accordion]容器查看Gradio的基本组件



API Keys

order	alias	api_key	expiration_time	user_id	id
0	admin_api_key	Hi-X****ruzI	3023-11-12T01:18:22.516281	057c16d0-8df6-4e8d-8a93-17331e8a7cdb	c0662919-146f-4044-
1	Default	Hi-o****lIdG	2024-12-13T00:12:53.085149	057c16d0-8df6-4e8d-8a93-17331e8a7cdb	5725de34-8eff-470a-

创建 API Key

API-Key别名: Default

API-Key值: Auto

有效期（天）: 30

可用模型: all

用户ID: myself

备注:

Selected tab index: 2, name: API Keys, Total: 2

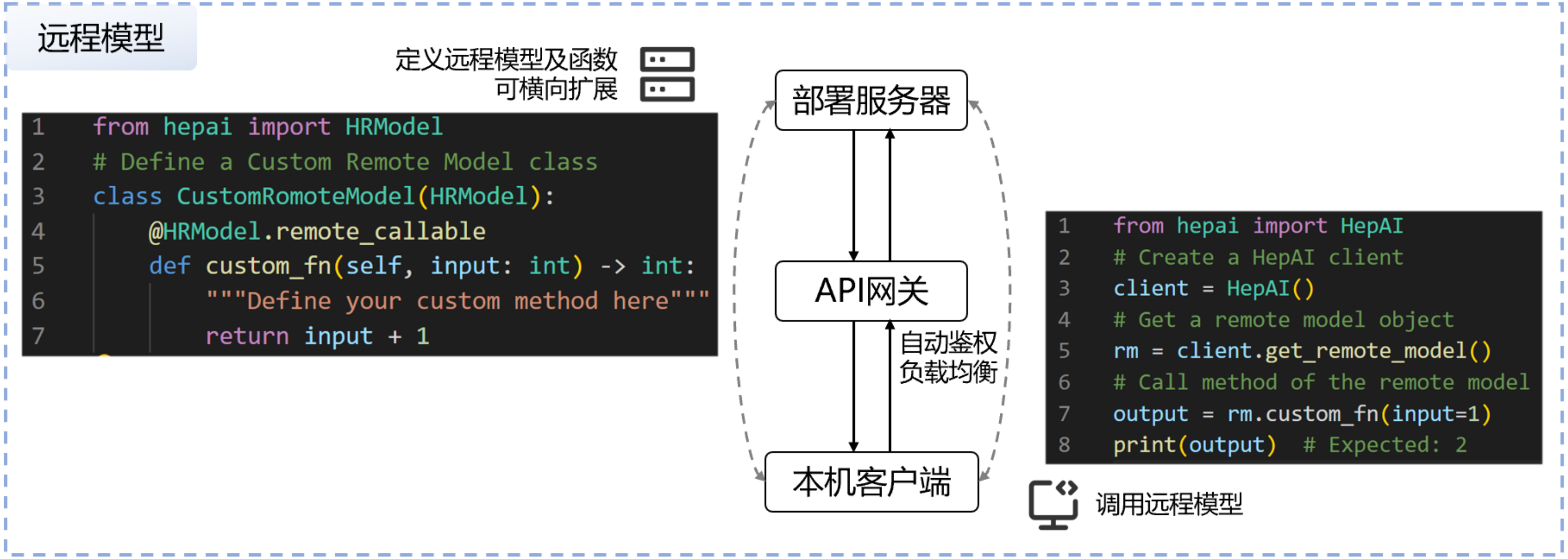
- Admin
- App Admin
- Database
- Authentication
- Remote Models

HaiDDF2

Python Remote Model !

By deploying models and any other software programs onto cloud servers via DDF Workers, and pairing them with the HepAI client, it enables low-latency, distributed invocation of remote models and more.

```
pip install hepai -U
```



远程模型

定义远程模型及函数
可横向扩展

```python
from hepai import HRModel
# Define a Custom Remote Model class
class CustomRomoteModel(HRModel):
    @HRModel.remote_callable
    def custom_fn(self, input: int) -> int:
        """Define your custom method here"""
        return input + 1
```

部署服务器

API网关

自动鉴权
负载均衡

本机客户端

```python
from hepai import HepAI
# Create a HepAI client
client = HepAI()
# Get a remote model object
rm = client.get_remote_model()
# Call method of the remote model
output = rm.custom_fn(input=1)
print(output)  # Expected: 2
```

调用远程模型

HaiChatv2 Upgrade to v3

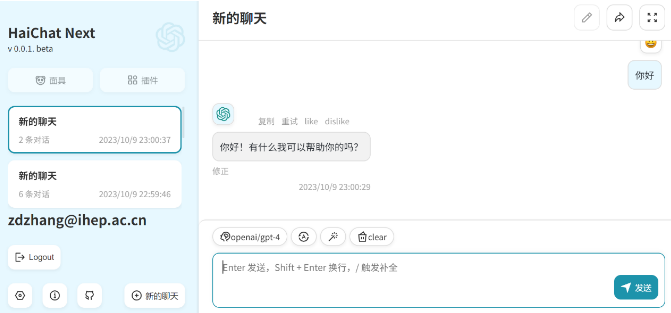
https://ai.ihep.ac.cn

- Generative AI system improve the daily work efficiency.

Old mode: When encountering a problem, researchers search for information or ask a mentor.
New mode: Researchers consult the LLM first.

- Translation, proofreading, and providing solutions, etc.
- Guiding algorithm and software development

Available LLMs:
- HEP·Xiwu
- Qwen
- Baichuan
- LLaMA
- ChatGPT
- GPT-4



HaiChat Next
v 0.0.1. beta

新的聊天
2 条对话 2023/10/9 23:00:37

新的聊天
6 条对话 2023/10/9 22:59:46

zdzhang@ihep.ac.cn

新的聊天

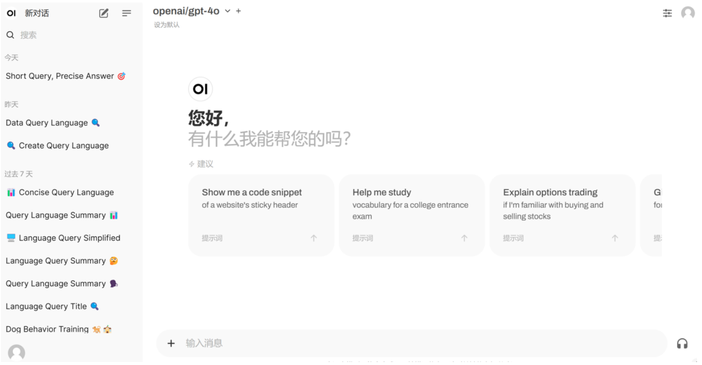
你好

你好！有什么我可以帮助你的吗？

2023/10/9 23:00:29

Enter 发送，Shift + Enter 换行，/ 触发补全

HaiChatv2



openai/gpt-4o
设为默认

今天
Short Query, Precise Answer 🎯

昨天
Data Query Language 🔍
🔍 Create Query Language

过去 7 天
📊 Concise Query Language
Query Language Summary 📊
🖥️ Language Query Simplified
Language Query Summary 🤔
Query Language Summary 🗣️
Language Query Title 🔍
Dog Behavior Training 🐕 🏠

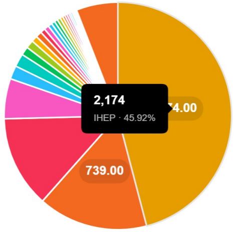
您好，
有什么我能帮您的吗？

建议
Show me a code snippet of a website's sticky header
Help me study vocabulary for a college entrance exam
Explain options trading if I'm familiar with buying and selling stocks

HaiChatv3

New Features:
- System Prompt Configuration
- Text-to-Speech Functionality
- Like Button Feature
- Multi-Model Output Comparison
- Message Editing
- Continue Generation
- Automatic Generation of Titles with Emoji
- Chat History Download
- Chat Title Search

☐ Total users: 3022 → 4850
☐ Daily Active Users: 800 (de-duplicated)

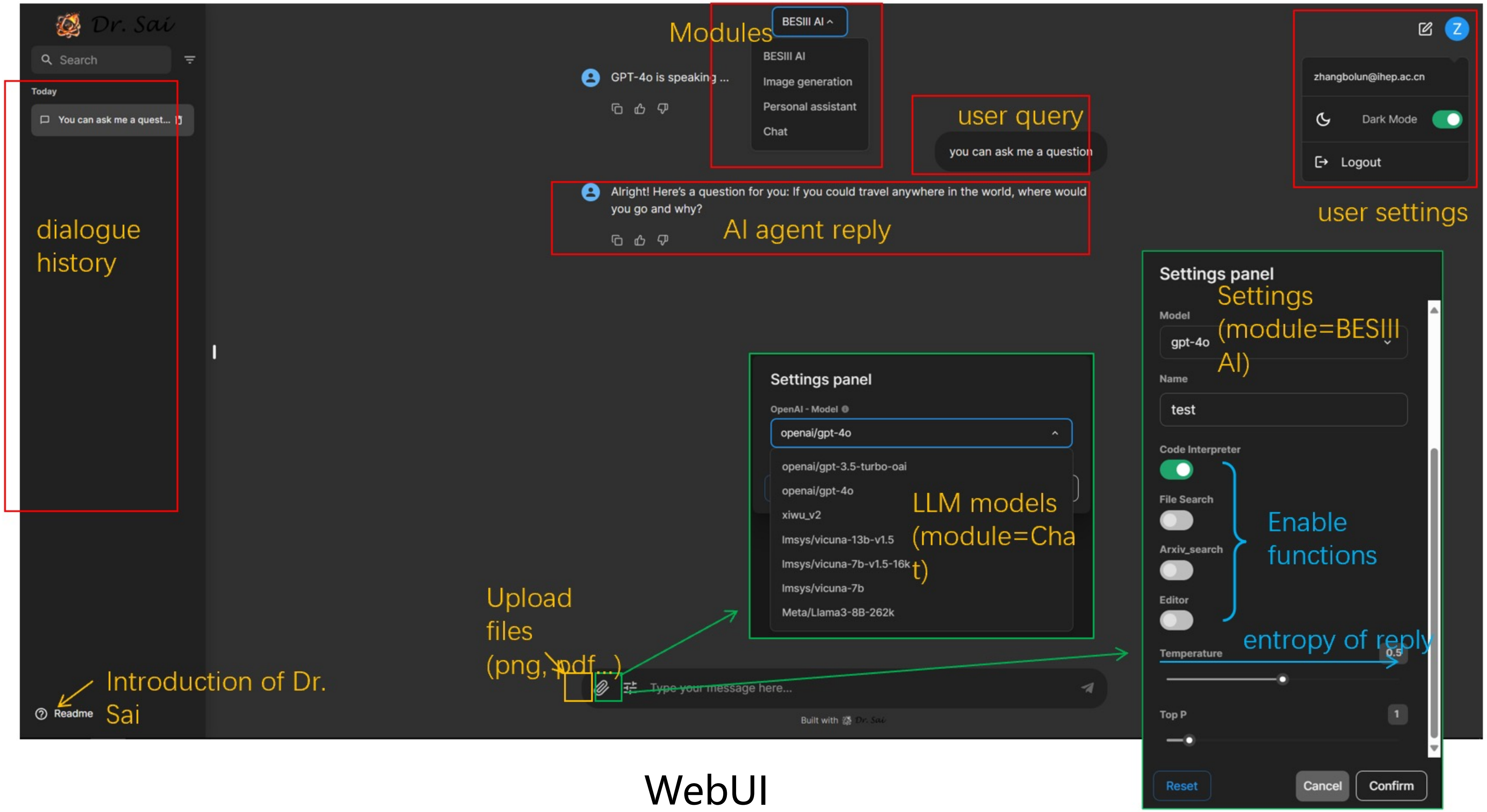


Preliminary Dr.Sai v1.0 was online

Dr.Sai v1.0.0 was available on July 1, 2024 https://drsai.ihep.ac.cn

(Based on Chainlit)

- Four parts:
- BESIII AI Assistant
    - Equipped with BESIII internal knowledge
    - Automatic physical analysis by task decomposition, code/text generation, calling BOSS to execute code
- Personal Assistant
    - Allow individuals to upload knowledge
- Image Generation
    - Drawing based on AI
- Chatbot
    - Pure LLM: Xiwu, LLaMA3, GPT etc
    - Supports PDF and image input



GPT-4o is speaking ...

you can ask me a question

Alright! Here's a question for you: If you could travel anywhere in the world, where would you go and why?

Settings panel

OpenAI - Model: openai/gpt-4o

Settings panel

Model: gpt-4o

Name: test

Code Interpreter: ●

File Search: ○

Arxiv_search: ○

Editor: ○

Temperature: 0.5

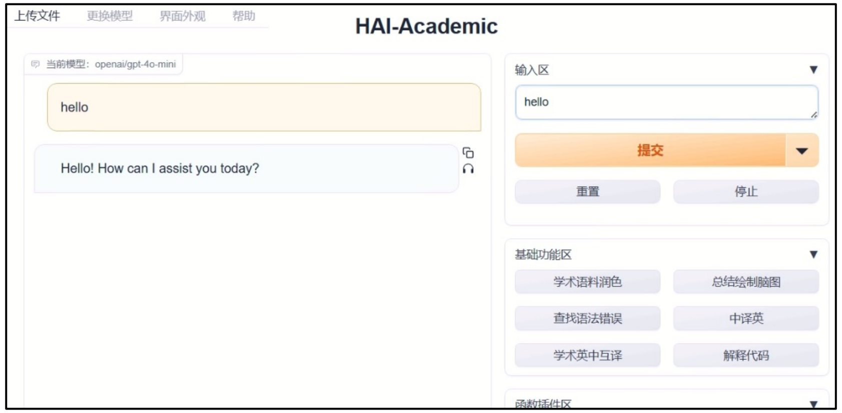
Top P: 1

zhangbolun@ihep.ac.cn

WebUI

Intergrated Applications

HaiAcademic



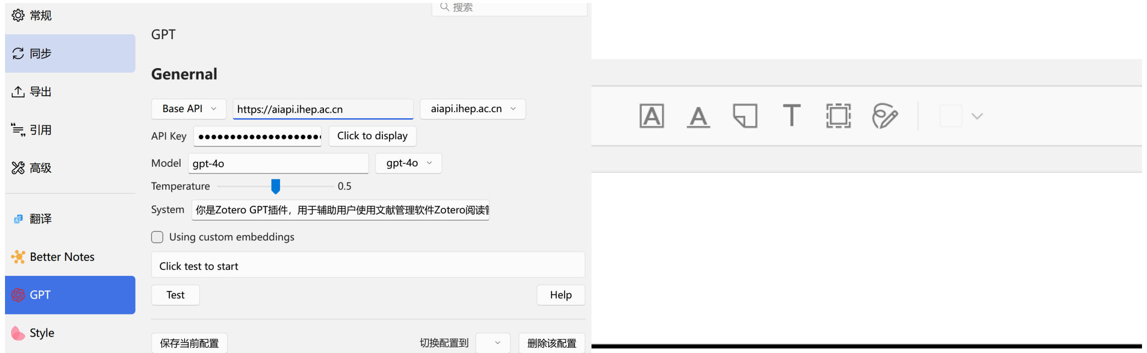
HAI-Academic

当前模型：openai/gpt-4o-mini

hello

Hello! How can I assist you today?

Support Zotero GPT Plugin



GPT

Genernal

Base API: https://aiapi.ihep.ac.cn (aiapi.ihep.ac.cn)

API Key: ••••••••••••••••••• Click to display

Model: gpt-4o (gpt-4o)

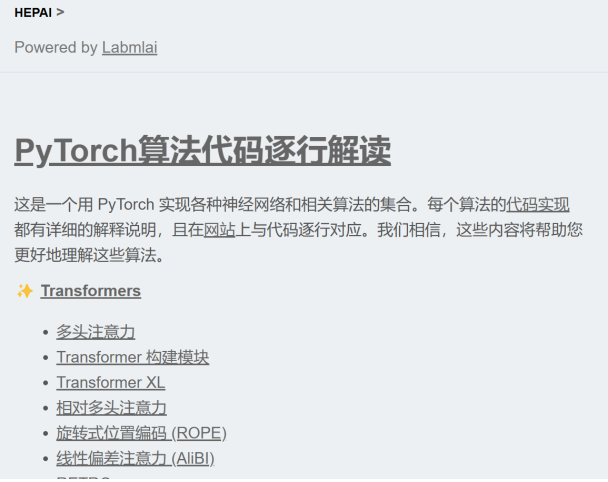
Temperature: 0.5

System: 你是Zotero GPT插件，用于辅助用户使用文献管理软件Zotero阅读

☐ Using custom embeddings

Click test to start

60+Deep Learning Paper Codes, Implementations


HEPAI >

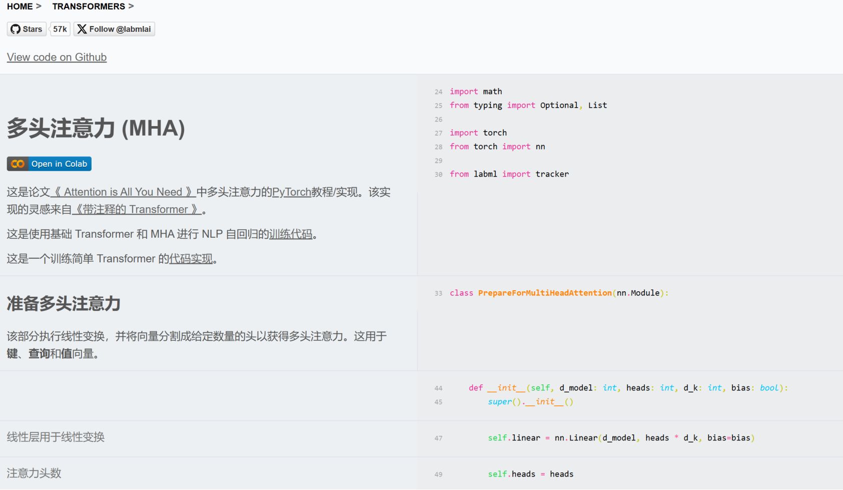
Powered by Labmlai

PyTorch算法代码逐行解读

这是一个用 PyTorch 实现各种神经网络和相关算法的集合。每个算法的代码实现都有详细的解释说明，且在网站上与代码逐行对应。我们相信，这些内容将帮助您更好地理解这些算法。

✨ Transformers

- 多头注意力
- Transformer 构建模块
- Transformer XL
- 相对多头注意力
- 旋转式位置编码 (ROPE)
- 线性偏差注意力 (AliBI)


HOME > TRANSFORMERS >

View code on Github

多头注意力 (MHA)

这是论文《 Attention is All You Need 》中多头注意力的PyTorch教程/实现。该实现的灵感来自《带注释的 Transformer 》。

这是使用基础 Transformer 和 MHA 进行 NLP 自回归的训练代码。

这是一个训练简单 Transformer 的代码实现。

```python
import math
from typing import Optional, List

import torch
from torch import nn

from labml import tracker
```

准备多头注意力

该部分执行线性变换，并将向量分割成给定数量的头以获得多头注意力。这用于**键**、**查询**和**值**向量。

```python
class PrepareForMultiHeadAttention(nn.Module):
```

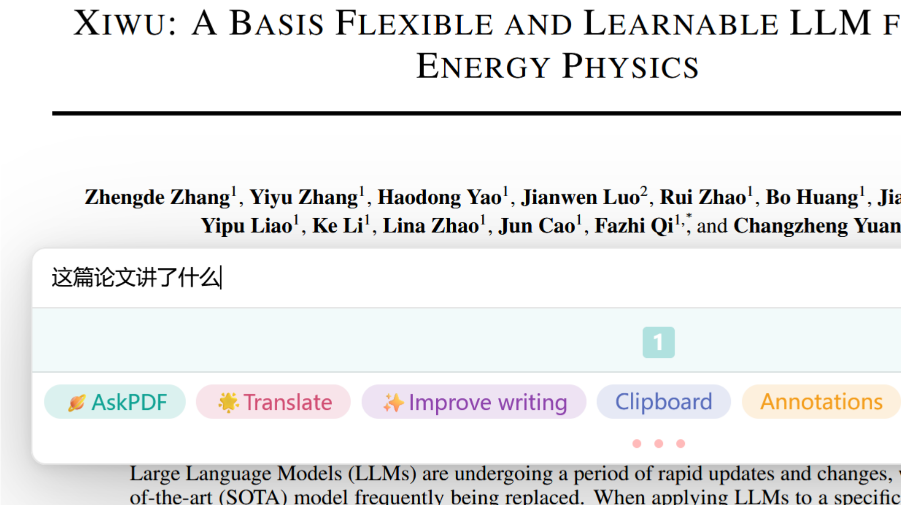
```python
    def __init__(self, d_model: int, heads: int, d_k: int, bias: bool):
        super().__init__()
```

线性层用于线性变换

```python
        self.linear = nn.Linear(d_model, heads * d_k, bias=bias)
```

注意力头数

```python
        self.heads = heads
```


XIWU: A BASIS FLEXIBLE AND LEARNABLE LLM F
ENERGY PHYSICS

Zhengde Zhang[1], Yiyu Zhang[1], Haodong Yao[1], Jianwen Luo[2], Rui Zhao[1], Bo Huang[1], Jia
Yipu Liao[1], Ke Li[1], Lina Zhao[1], Jun Cao[1], Fazhi Qi[1,*] and Changzheng Yuan

这篇论文讲了什么

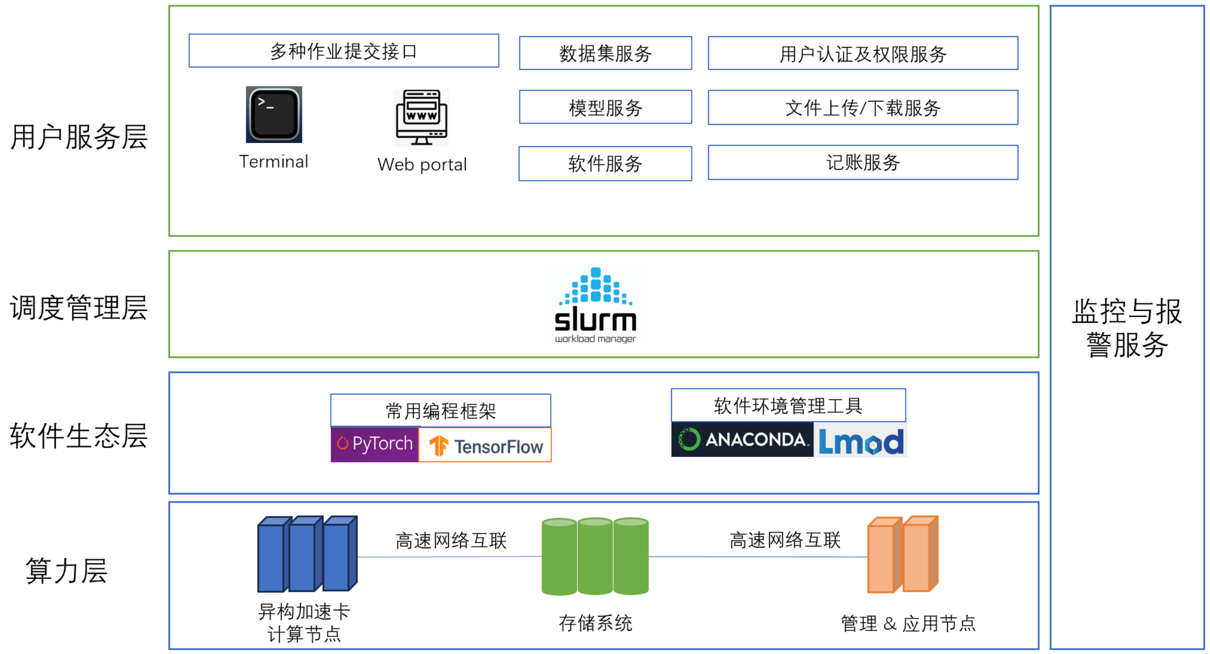
Large Language Models (LLMs) are undergoing a period of rapid updates and changes,
of-the-art (SOTA) model frequently being replaced. When applying LLMs to a specific

AI Computing Power

New AI Computing Power Infrastucture

☐ Initial investment of 9 million RMB (2024.03)

☐ Raise funds from other channels



用户服务层

多种作业提交接口

Terminal

Web portal

数据集服务

用户认证及权限服务

模型服务

文件上传/下载服务

软件服务

记账服务

调度管理层

slurm workload manager

软件生态层

常用编程框架

PyTorch TensorFlow

软件环境管理工具

ANACONDA Lmod

算力层

异构加速卡计算节点

高速网络互联

存储系统

高速网络互联

管理 & 应用节点

监控与报警服务

Accelerator Cards

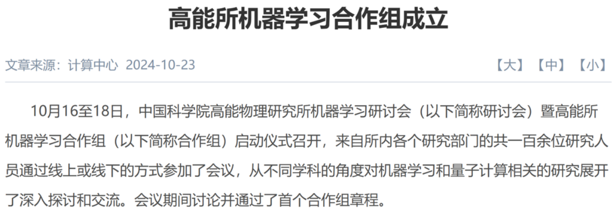
- 2 nodes of NVIDIA GPU
  - 8 * A800 (80GB)
  - 8 * L40 (48GB)
- 4 nodes of Sugon DCU (Domestic)
  - 32 * K100AI (64GB)

Storage

- 200TB flash memory

ML Collaboration Group

- ☐ Establishment of the Machine Learning Collaboration Group of IHEP
- ☐ From October 16 to 18, 2024, the IHEP, CAS, held the Machine Learning Symposium and the inauguration ceremony of the Machine Learning Collaboration Group. Over 100 researchers from various departments within the institute participated in the meeting either online or in person. They engaged in in-depth discussions and exchanges on research related to machine learning and quantum computing from diverse disciplinary perspectives. During the conference, the first charter of the Collaboration Group was also discussed and approved.



高能所机器学习合作组成立

文章来源：计算中心 2024-10-23 【大】【中】【小】

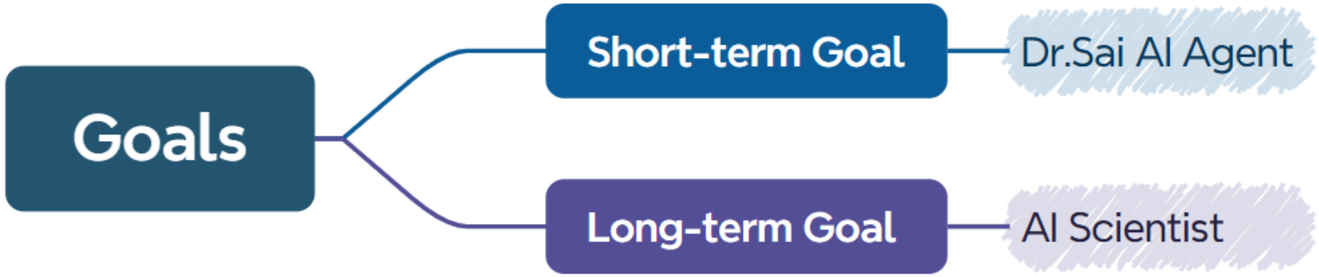
10月16至18日，中国科学院高能物理研究所机器学习研讨会（以下简称研讨会）暨高能所机器学习合作组（以下简称合作组）启动仪式召开，来自所内各个研究部门的共一百余位研究人员通过线上或线下的方式参加了会议，从不同学科的角度对机器学习和量子计算相关的研究展开了深入探讨和交流。会议期间讨论并通过了首个合作组章程。

Part. 02

Exploration of AI large model

Dr.Sai: An AI agents system for BESIII experiment

Dr.Sai Project in IHEP



Research Agent

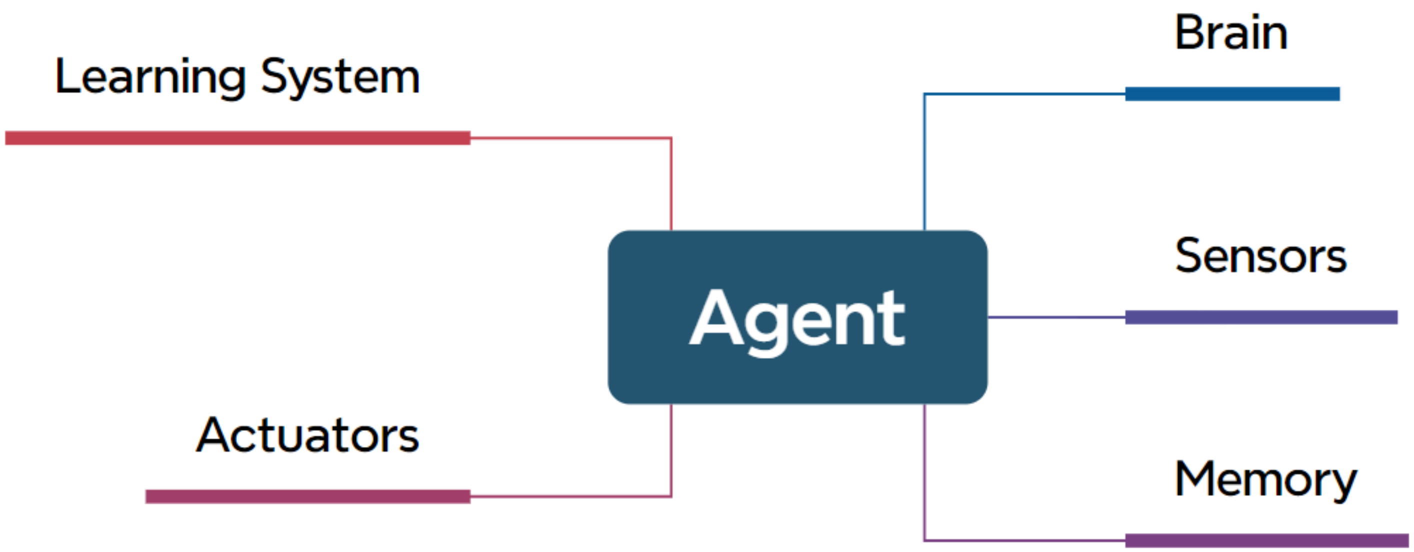
- Literature review
- Research topic selection
- Data processing
- Physics analysis
- Result interpretation
- Article writing

AI Scientist

- Understand particle data
- Master basic physics laws
- Discover patterns from data
- Independently choose topics and complete research

What is AI Agent?

An AI agent refers to a **system** or **software** that can make autonomous decisions or perform actions on behalf of its users based on its knowledge, programming, environment, and inputs.



A LLM ≠ A person

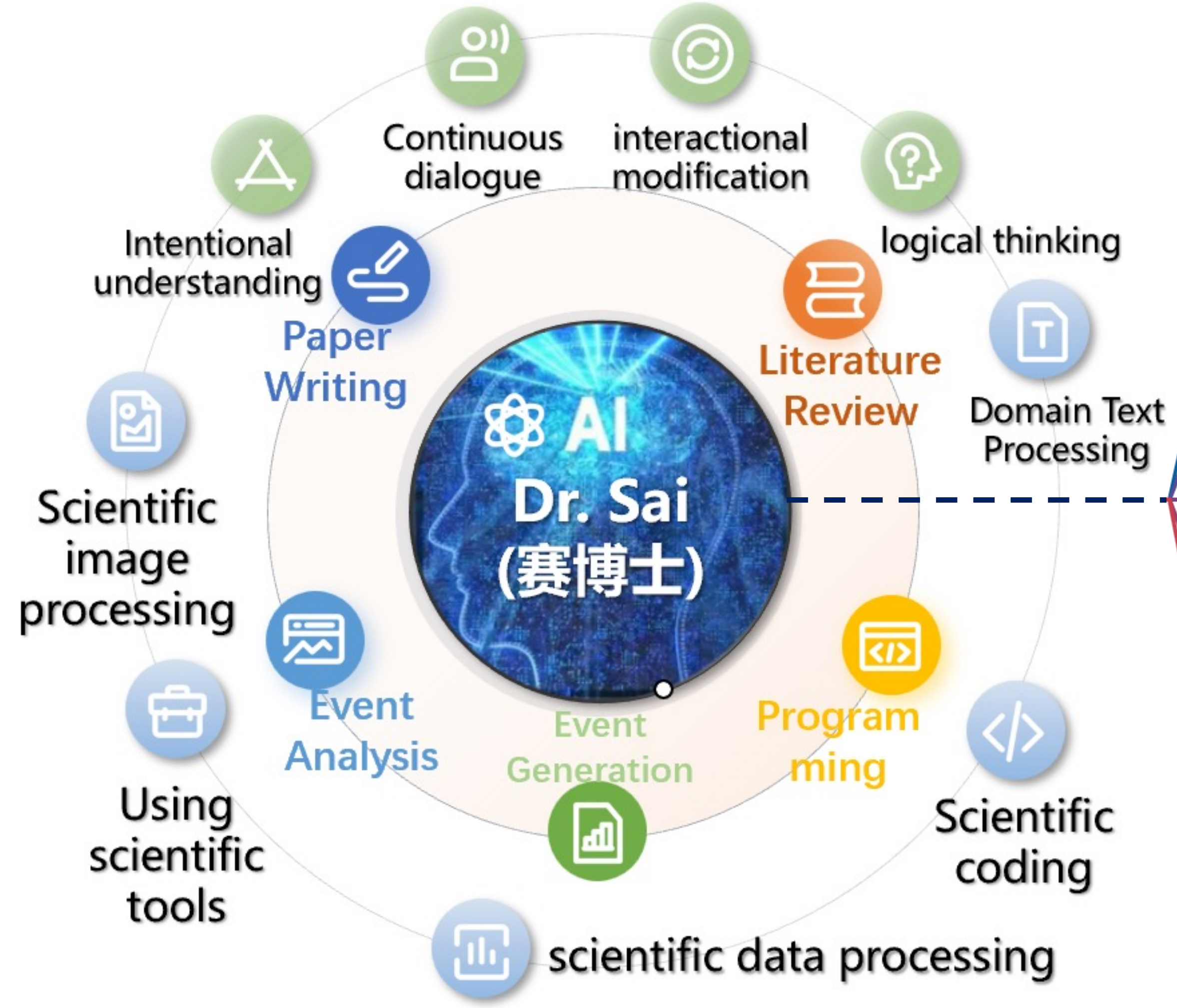
A agent ≈ A person

The LLM is the **processing core** (the Brain) of agent.

Overview of Dr.Sai Agent

Let the large model conduct particle physics research

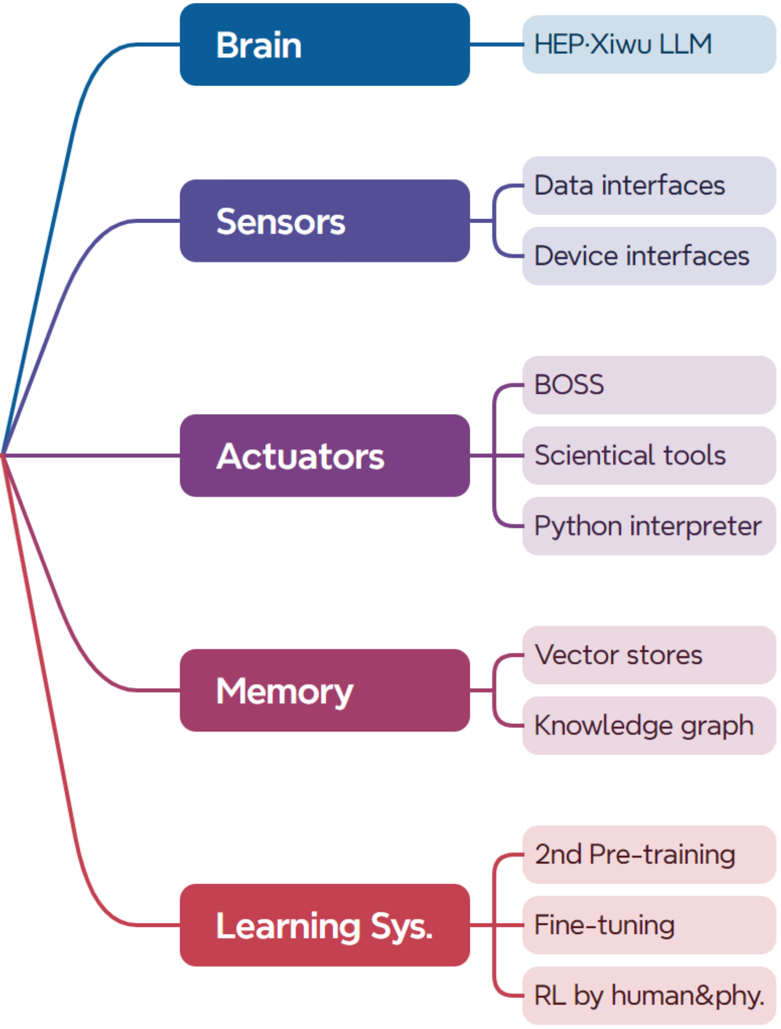
Essential: Modeling the research process.



AI
Dr. Sai
(赛博士)

Continuous dialogue
interactional modification
logical thinking
Intentional understanding
Paper Writing
Literature Review
Domain Text Processing
Scientific image processing
Event Analysis
Event Generation
Program ming
Scientific coding
Using scientific tools
scientific data processing

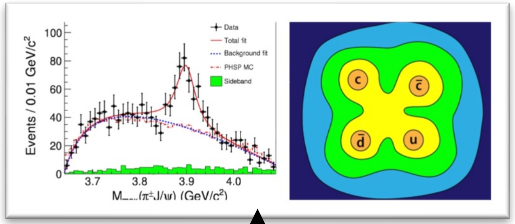
Tasks & Required capabilities



- **Brain**
  - HEP·Xiwu LLM
- **Sensors**
  - Data interfaces
  - Device interfaces
- **Actuators**
  - BOSS
  - Sciential tools
  - Python interpreter
- **Memory**
  - Vector stores
  - Knowledge graph
- **Learning Sys.**
  - 2nd Pre-training
  - Fine-tuning
  - RL by human&phy.

Components

Rediscover
Zc(3900)

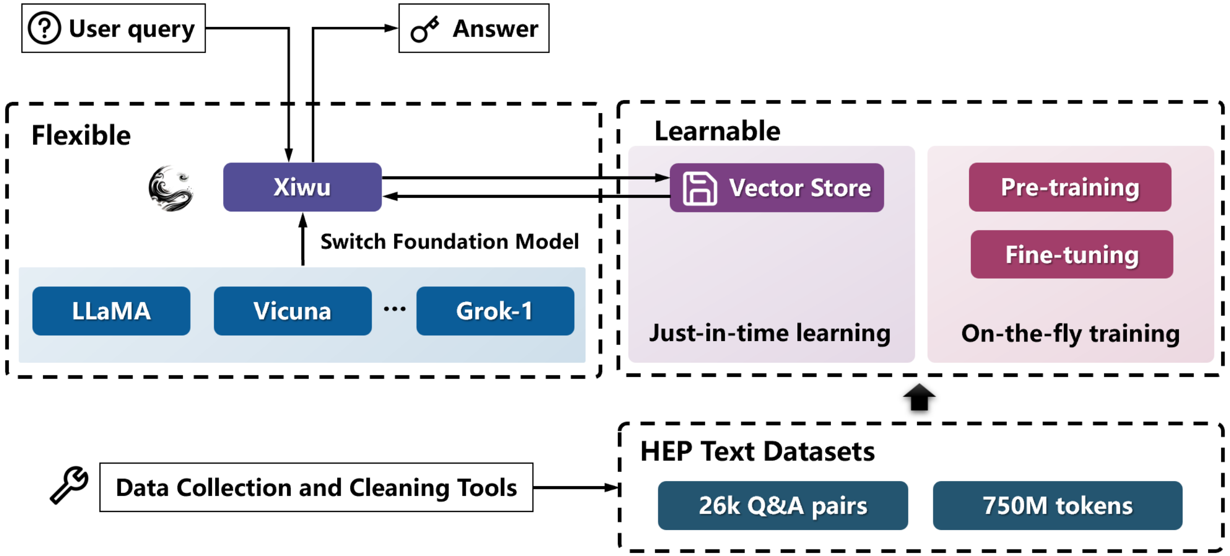


Applications

The Brain of Dr.Sai – Xiwu LLM

Xiwu(溪悟): A Basis Flexible and Learnable LLM for High Energy Physics [1]

- Currently, Xiwu is based on LLaMA3-8B. Historically, LLaMA, LLaMA2 (7B, 13B) etc.
- Secondary pre-training and fine-tuning.
- Significantly better than the base model in HEP Q&A and internal code generation.
- New version based on LLaMA3-70B and Qwen2-72B is in training.



User query

Answer

Flexible

Xiwu

Switch Foundation Model

LLaMA

Vicuna

…

Grok-1

Learnable

Vector Store

Just-in-time learning

Pre-training

Fine-tuning

On-the-fly training

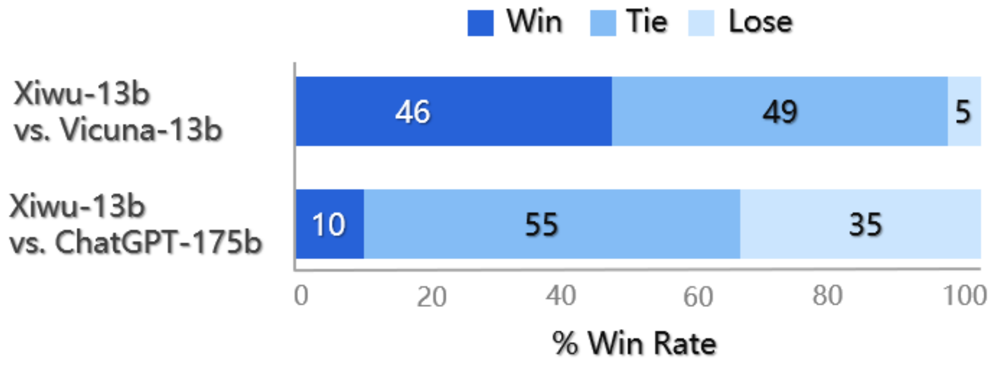
HEP Text Datasets

26k Q&A pairs

750M tokens

Data Collection and Cleaning Tools

Test Results



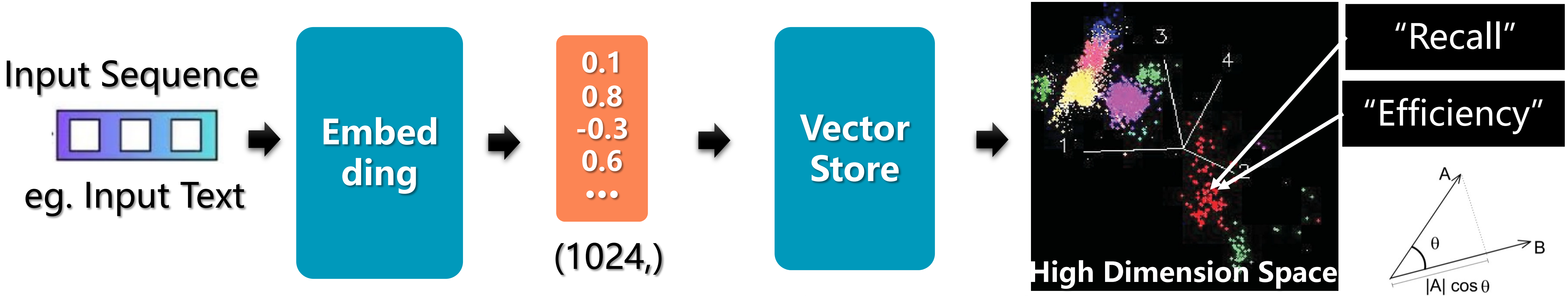
---

[1] arXiv:2404.08001

https://github.com/zhangzhengde0225/Xiwu

The **Memory** of Dr.Sai – VS & KG

- Save HEP knowledge into **Vector Store** and **Knowledge Graph**.
- Retrieve them to enhance the model's generation accuracy and reduce hallucination.



Input Sequence

eg. Input Text

Embedding

0.1
0.8
-0.3
0.6
…

(1024,)

Vector Store

High Dimension Space

"Recall"

"Efficiency"

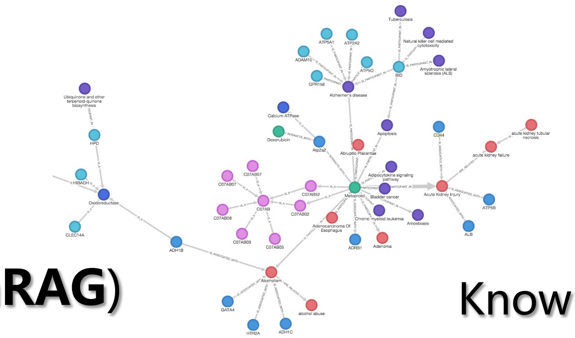
(Based on **BGE-M3** Model from BAAI)

(Based on **LLaMA Index**)

$$\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}|cos\theta$$

RAG (Retrieval-Augmented Generation) based on vector store is done.

- self-reflection could be realized.



Knowledge Graph

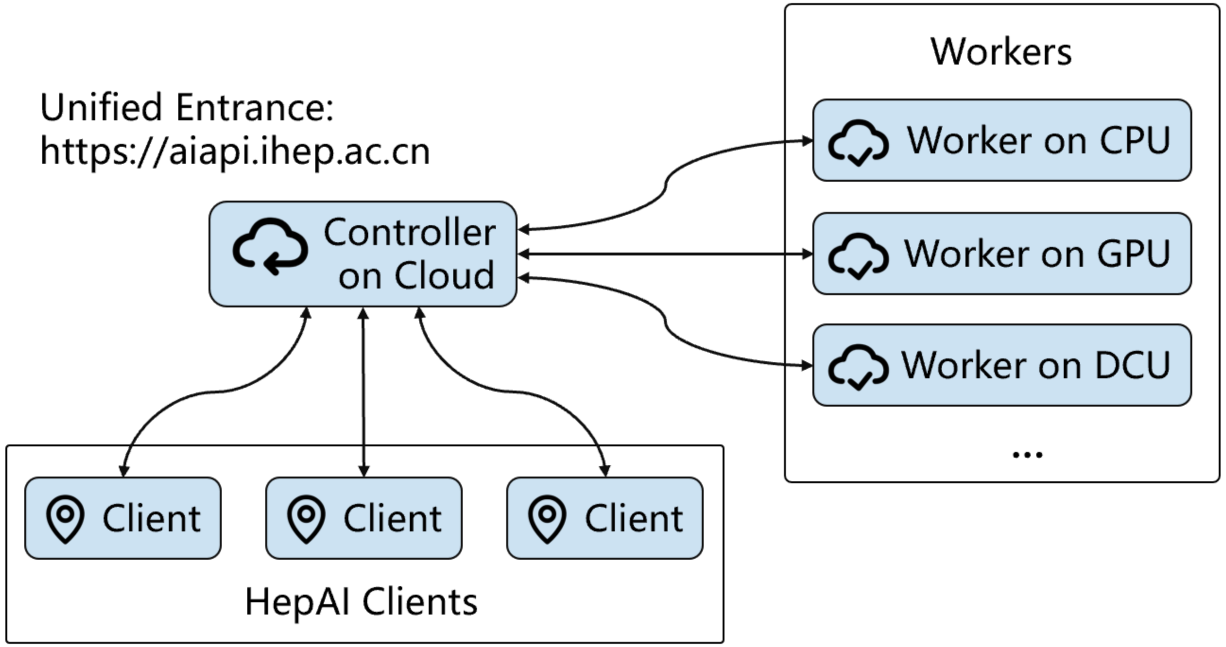
DBSCAN

- Unsupervised clustering

Knowledge Graph is studing (Based **GraphRAG**)

The **Actuators** of Dr.Sai – HepAI DDF

- The Distributed Deployment Framework (HepAI-DDF) is developed.
  - Featured with flexibility, cross-language, cross-platform, hererogeneous.
  - Allows Dr. Sai to easily scale its actuator components.
- BOSS (BESIII Offline Software System) actuator enables BESIII code execution and result retrieval.
- No longer a Q&A assistant; It can perform operations.



Unified Entrance:
https://aiapi.ihep.ac.cn

Controller on Cloud

Workers

Worker on CPU

Worker on GPU

Worker on DCU

...

Client

Client

Client

HepAI Clients

BOSS 7.1.0

For executing physical analysis code.

Daisy

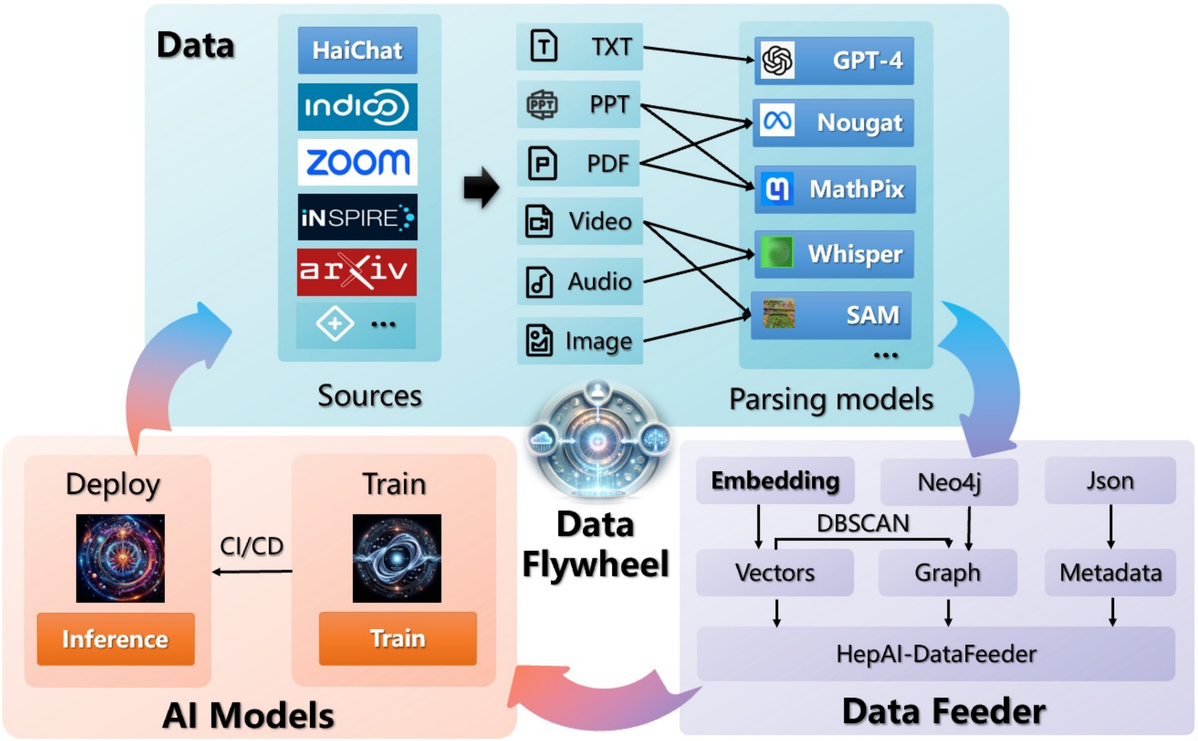
For invoking scientific tools

Python Interp.

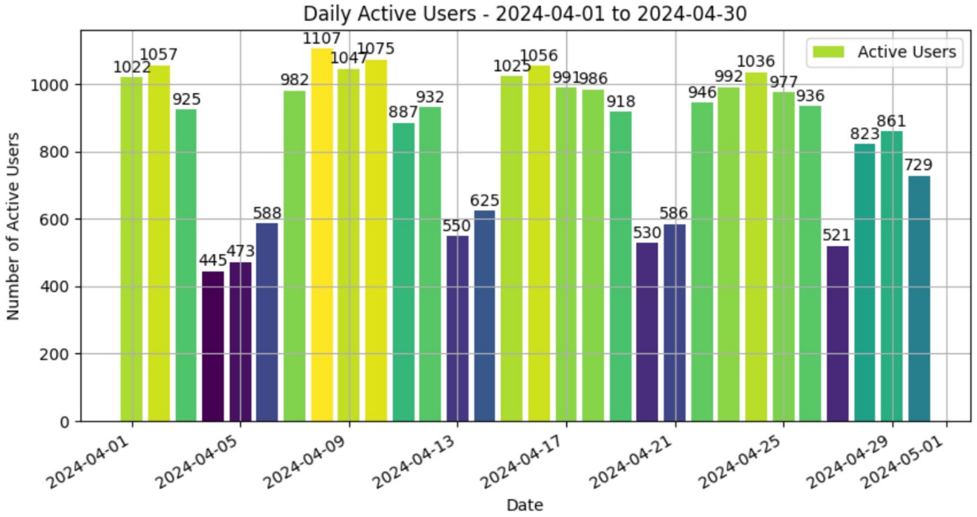
For interpreting general Python code.

The Sensors of Dr.Sai – Data Flywheel

- The "data flywheel" enables continuous model iteration and evolution.
  - Data flywheel, i.e. **data-driven flywheel effect**
  - Improves models by constructing circular data pathways
  - Attracts more users, generating more data, further enhance the models.



- **HaiChat** service
  - AIGC service based on LLM
  - 4000+ users
  - Daily active users exceed 900.
  - Real needs from HEP users

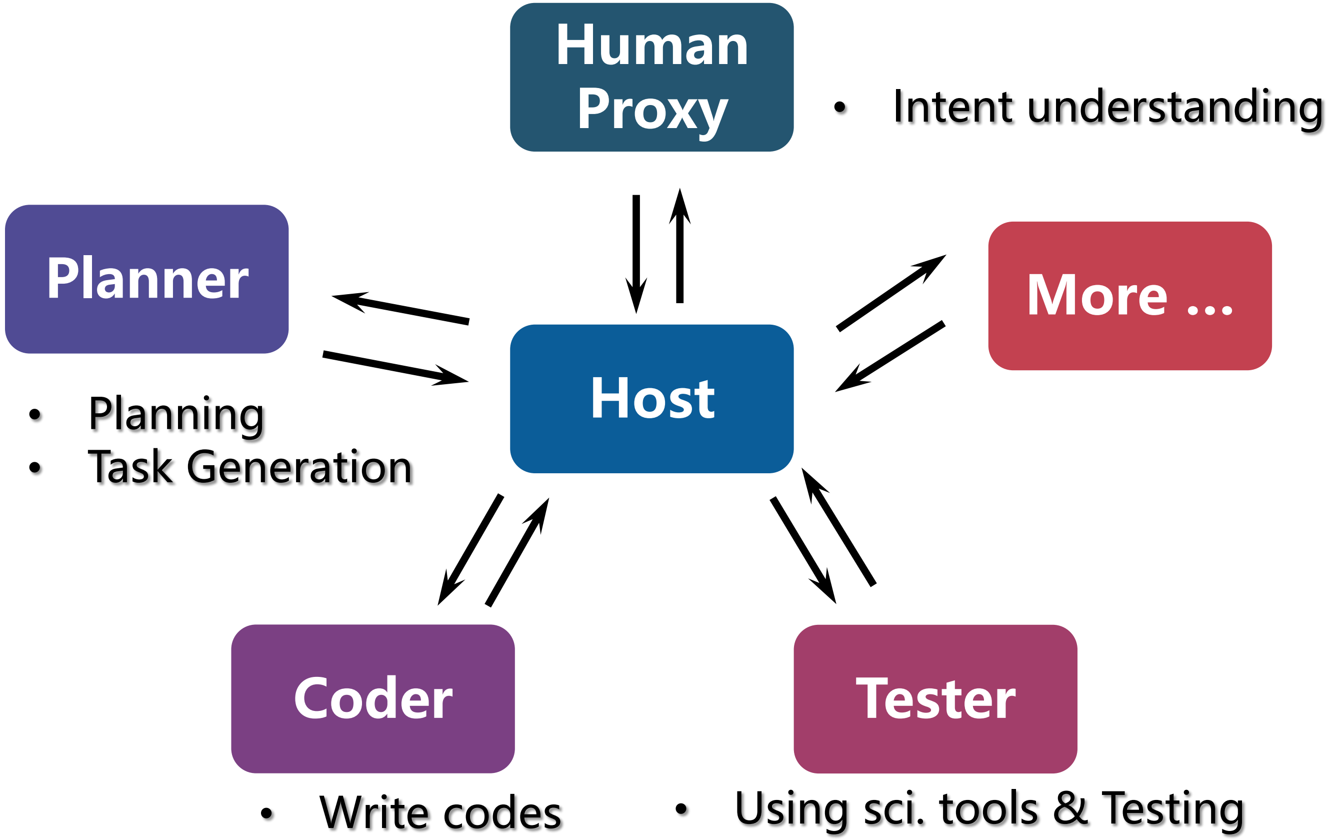


Multi-Agent Collaborative System

Dr. Sai's multi-agent collaborative system - handling complex tasks

- Based on AutoGen framework.
- Each agent is equipped with specific knowledge, tools, and LLM.
- A Host agent is introduced to manage group chats, making it easy to expand with more agents
- A human proxy agent is intrdroduced to allow humans intervene at any time.

Autogen: arXiv.2308.08155)



- Intent understanding

- Planning
- Task Generation

- Write codes

- Using sci. tools & Testing

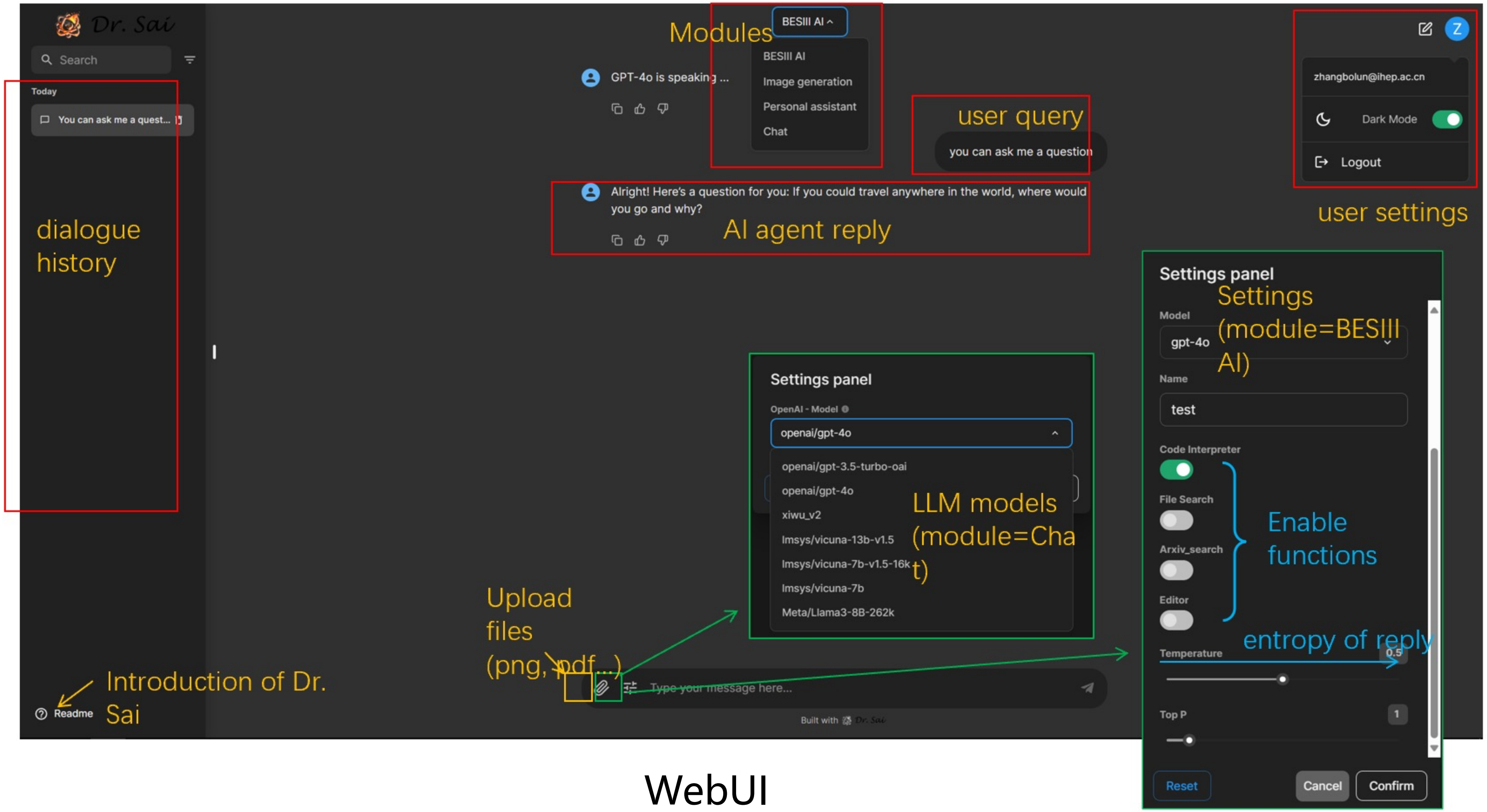
Workerflow:
Msg → host → **LLM**
Planner → **RAG** → **LLM** → host
Host → **LLM**

User Interface of Dr.Sai

Dr.Sai v1.0.0 was available on July 1, 2024 https://drsai.ihep.ac.cn

(Based on Chainlit)

- Four parts:
- BESIII AI Assistant
    - Equipped with BESIII internal knowledge
    - Automatic physical analysis by task decomposition, code/text generation, calling BOSS to execute code
- Personal Assistant
    - Allow individuals to upload knowledge
- Image Generation
    - Drawing based on AI
- Chatbot
    - Pure LLM: Xiwu, LLaMA3, GPT etc
    - Supports PDF and image input



GPT-4o is speaking ...

you can ask me a question

Alright! Here's a question for you: If you could travel anywhere in the world, where would you go and why?

Settings panel

OpenAI - Model: openai/gpt-4o

Settings panel

Model: gpt-4o

Name: test

Temperature: 0.5

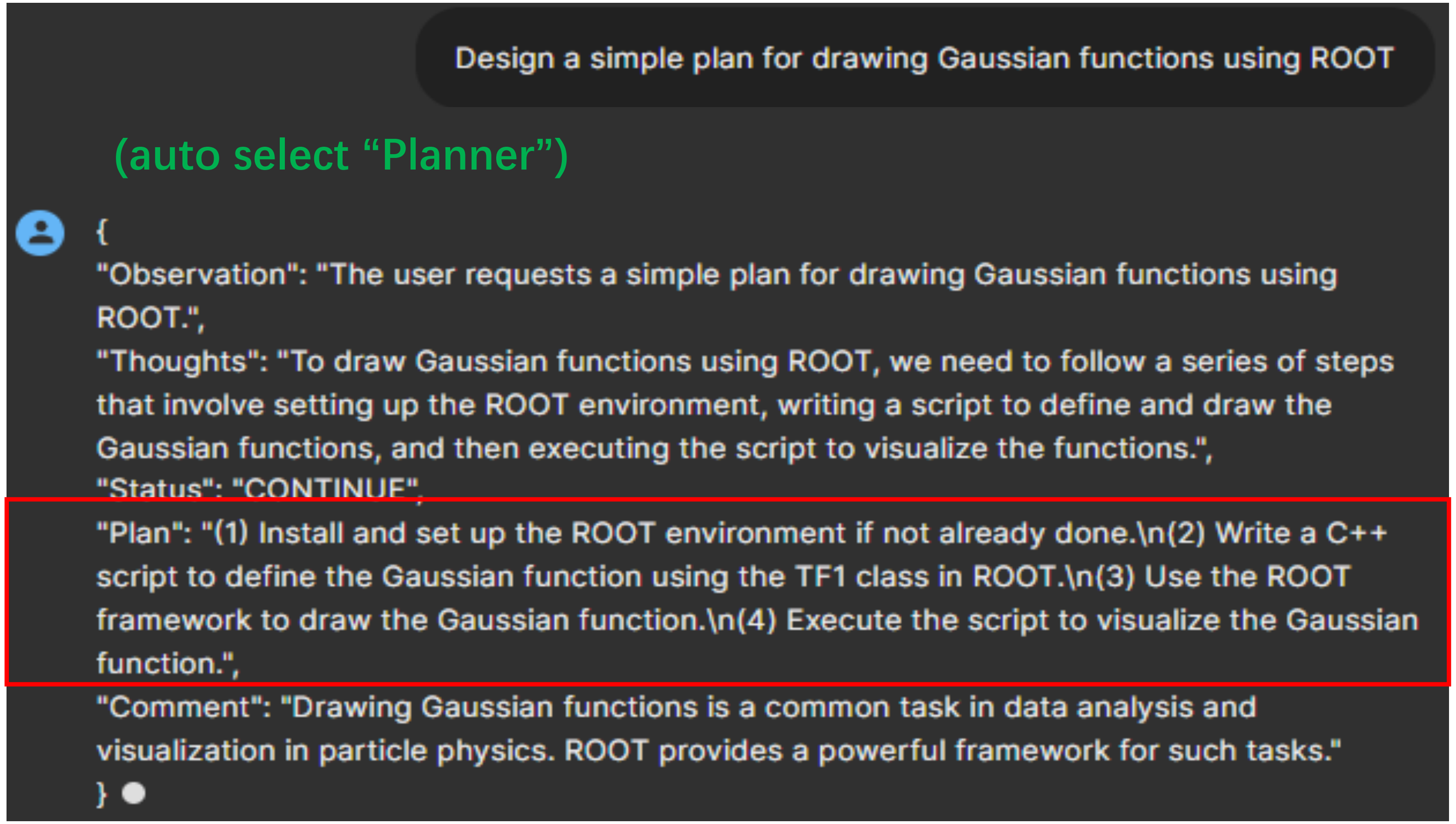
Top P: 1

WebUI

Example of Single Agent

Planner

- Decompose complex task into multiple sub-tasks



Design a simple plan for drawing Gaussian functions using ROOT

(auto select "Planner")

```
{
"Observation": "The user requests a simple plan for drawing Gaussian functions using
ROOT.",
"Thoughts": "To draw Gaussian functions using ROOT, we need to follow a series of steps
that involve setting up the ROOT environment, writing a script to define and draw the
Gaussian functions, and then executing the script to visualize the functions.",
"Status": "CONTINUE",
"Plan": "(1) Install and set up the ROOT environment if not already done.\n(2) Write a C++
script to define the Gaussian function using the TF1 class in ROOT.\n(3) Use the ROOT
framework to draw the Gaussian function.\n(4) Execute the script to visualize the Gaussian
function.",
"Comment": "Drawing Gaussian functions is a common task in data analysis and
visualization in particle physics. ROOT provides a powerful framework for such tasks."
}
```

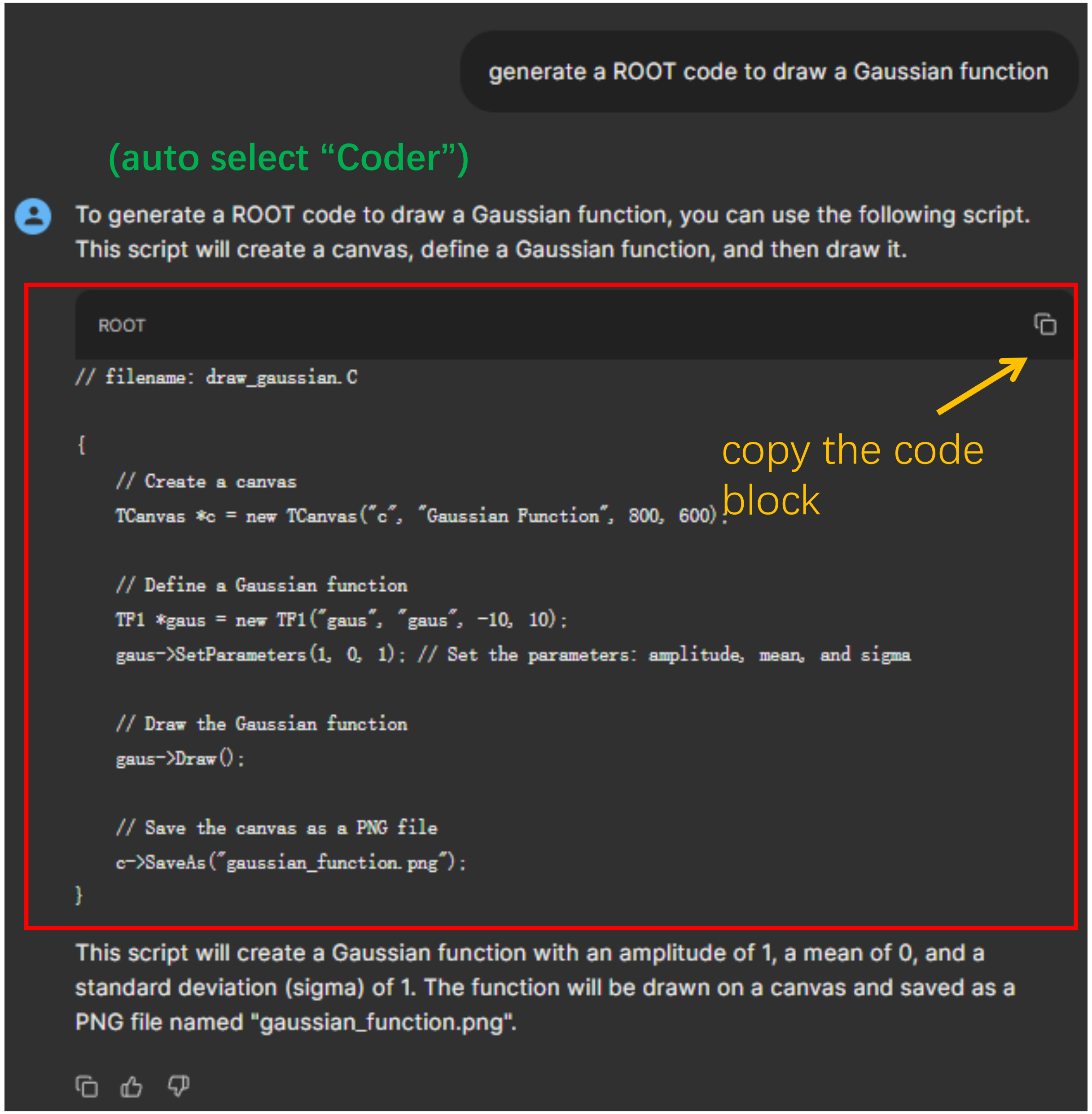
Example

Example of Single Agent

Coder

- Generate internal BESIII codes

Example



generate a ROOT code to draw a Gaussian function

(auto select “Coder”)

To generate a ROOT code to draw a Gaussian function, you can use the following script. This script will create a canvas, define a Gaussian function, and then draw it.

ROOT

```
// filename: draw_gaussian.C

{
    // Create a canvas
    TCanvas *c = new TCanvas("c", "Gaussian Function", 800, 600);

    // Define a Gaussian function
    TF1 *gaus = new TF1("gaus", "gaus", -10, 10);
    gaus->SetParameters(1, 0, 1); // Set the parameters: amplitude, mean, and sigma

    // Draw the Gaussian function
    gaus->Draw();

    // Save the canvas as a PNG file
    c->SaveAs("gaussian_function.png");
}
```

copy the code block

This script will create a Gaussian function with an amplitude of 1, a mean of 0, and a standard deviation (sigma) of 1. The function will be drawn on a canvas and saved as a PNG file named "gaussian_function.png".

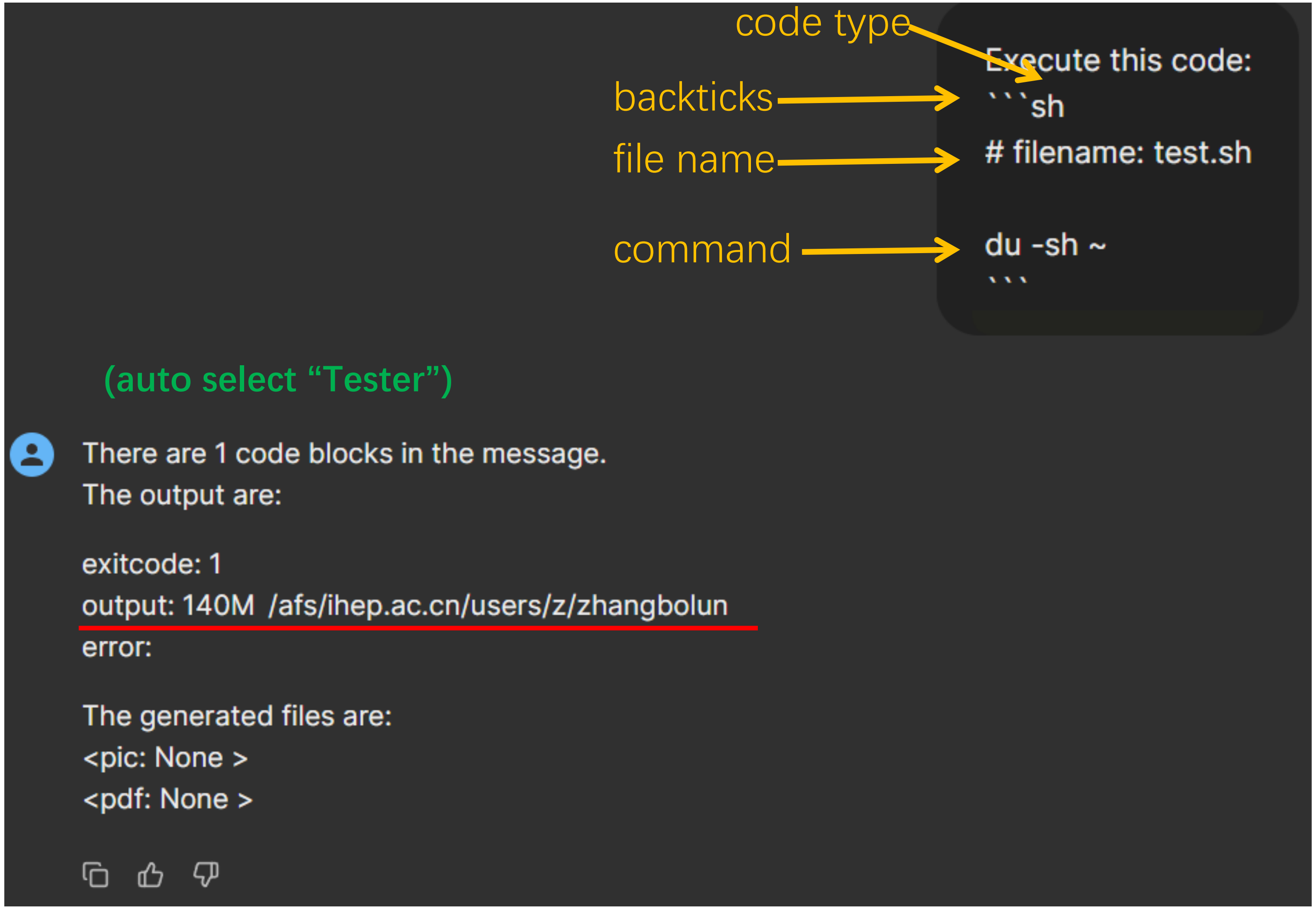
Note: Demonstrated the ability to generate ROOT code, using BOSS code internally.

Example of Single Agent

Tester

- Excute analysis code in BESIII environment
- Run on remote HPC server
- Retrieve the results

Example



~~~
Execute this code:
```sh
# filename: test.sh

du -sh ~
```
~~~

```
There are 1 code blocks in the message.
The output are:

exitcode: 1
output: 140M  /afs/ihep.ac.cn/users/z/zhangbolun
error:

The generated files are:
<pic: None >
<pdf: None >
```

Example of Multi-Agent System

- **Host** agent select suitable speaker
- **Coder** generate domain code
- **Tester** Call BOSS (BESIII Offiline Software System) to excute
- Draw a signal histogram

Example

Summary

- ☐ **HepAI Platform**
  - ☐ The core component **HaiDDF**v1 → v2. General Remote Model!
  - ☐ Generative AI system **HaiChat**v2 → v3.
  - ☐ New AI Apps. (**HaiAcademic**, **Zotero Plugin, DL Codes tutorials**) have been intergrated
  - ☐ New AI Computing Power Infrastructure (**48 GPU&DCU cards**)
  - ☐ IHEP **ML Collaboration Group** has been officially established.

- ☐ **AI Agents based on LLMs**
  - ☐ The **HEP•Xiwu** LLM has been trained to enhance its domain-specific capabilities.
  - ☐ Initial version of **Dr.Sai Agents** for physics analysis based on LLM was launched.
  - ☐ The feasibility study has been successfully concluded, and we have now entered the formal research phase.

Contact

HepAI platform website:
https://ai.ihep.ac.cn

- **We are seeking more collaborations.**
- **Recruiting international students, postdoctoral fellows, staff, and talents through recruitment programs.**

- **Welcome to discuss any interests related to AI for High Energy Physics!**

- Fazhi Qi (齐法制)
- qfz@ihep.ac.cn

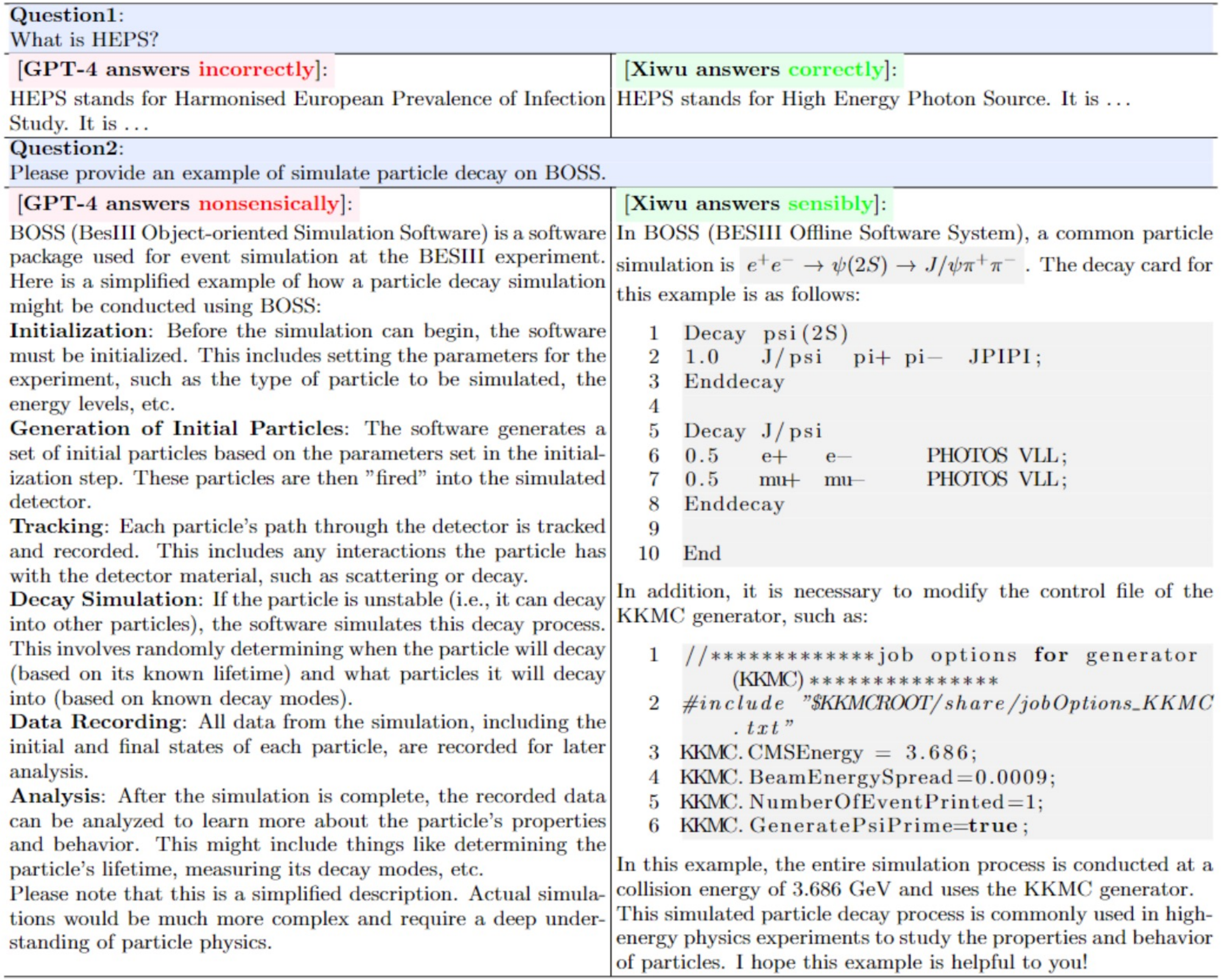
- Zhengde Zhang (张正德)
- zdzhang@ihep.ac.cn

- Computing Center, Institute of High Energy Physics, CAS, Beijing

Thank you for listening

Backup

The limitations of general LLMs



Question1:
What is HEPS?

[GPT-4 answers incorrectly]:

HEPS stands for Harmonised European Prevalence of Infection Study. It is . . .

[Xiwu answers correctly]:

HEPS stands for High Energy Photon Source. It is . . .

Question2:
Please provide an example of simulate particle decay on BOSS.

[GPT-4 answers nonsensically]:

BOSS (BesIII Object-oriented Simulation Software) 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:

Initialization: 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.

Generation of Initial Particles: 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.

Tracking: 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.

Decay Simulation: 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).

Data Recording: All data from the simulation, including the initial and final states of each particle, are recorded for later analysis.

Analysis: 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.

[Xiwu answers sensibly]:

In BOSS (BESIII Offline Software System), a common particle simulation is $e^+e^- \to \psi(2S) \to J/\psi\pi^+\pi^-$. The decay card for this example is as follows:

```
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
```

In addition, it is necessary to modify the control file of the KKMC generator, such as:

```
//**************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;
```

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!

- **Hallucination:**

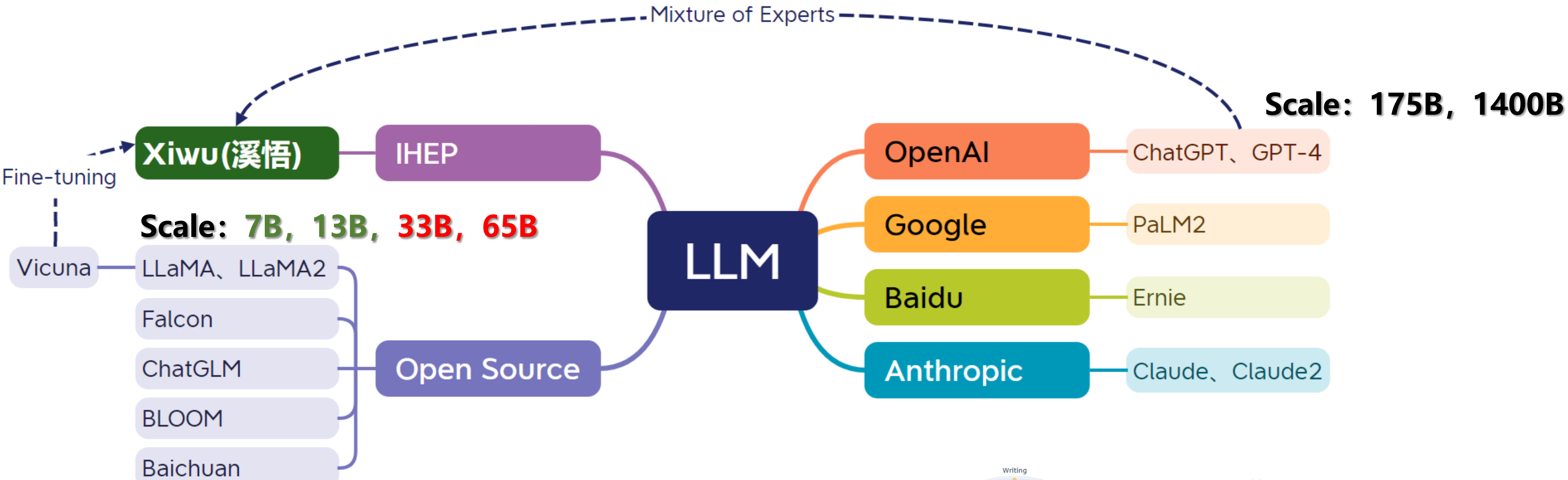
Produce content that is nonsensical or untruthful in
relation to certain sources[3, 4],
which is particularly detrimental in scientific fields

- **Lack of rapid learning ability**

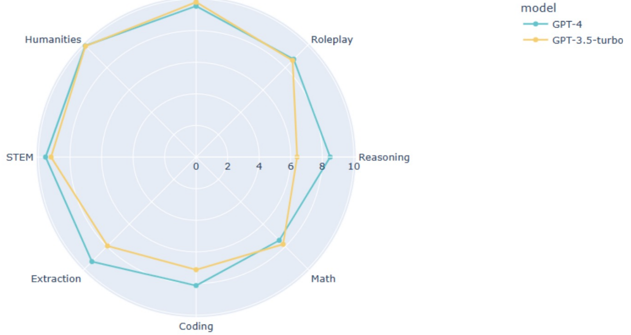
- **Solutions**
- No.1 **Fine-tuning LLM**
- No.2 **External Knowledge**

HEP•Xiwu: Algorithm

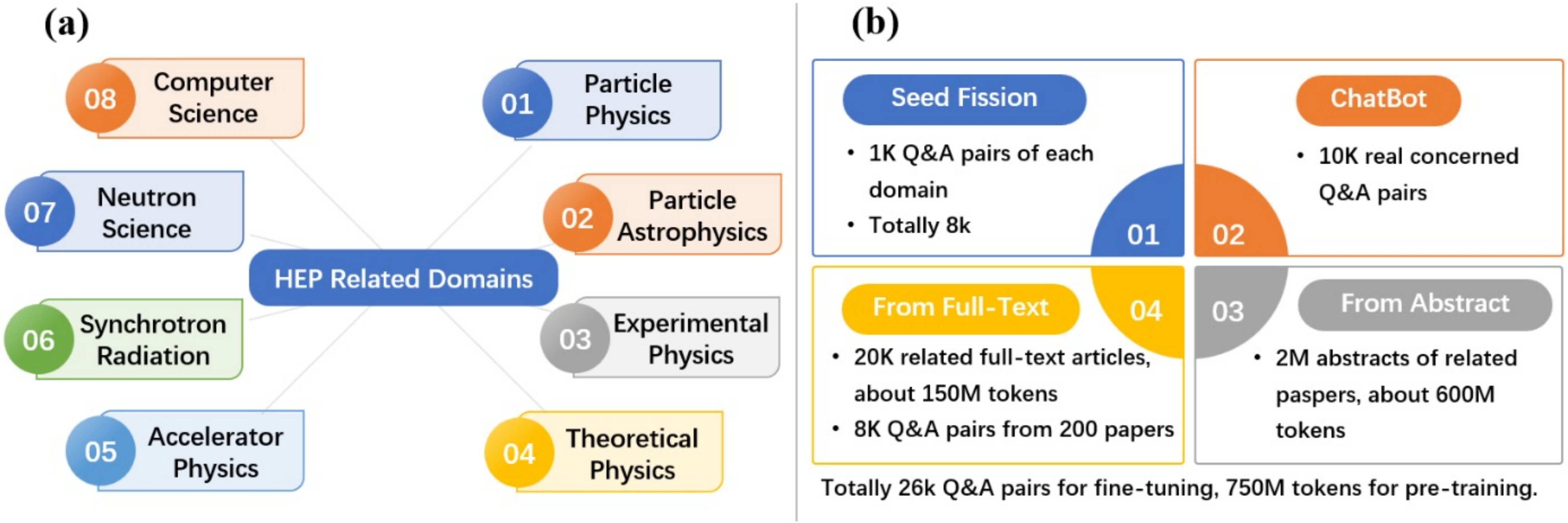
Xiwu: a customized LLM for High Energy Physics



Evaluation Matrix：Writing, Coding, Extraction, Math, Reasoning, Humanities and STEM.



HEP•Xiwu: Data



(a)

HEP Related Domains

- 01 Particle Physics
- 02 Particle Astrophysics
- 03 Experimental Physics
- 04 Theoretical Physics
- 05 Accelerator Physics
- 06 Synchrotron Radiation
- 07 Neutron Science
- 08 Computer Science

(b)

01 **Seed Fission**

- 1K Q&A pairs of each domain
- Totally 8k

02 **ChatBot**

- 10K real concerned Q&A pairs

03 **From Abstract**

- 2M abstracts of related paspers, about 600M tokens

04 **From Full-Text**

- 20K related full-text articles, about 150M tokens
- 8K Q&A pairs from 200 papers

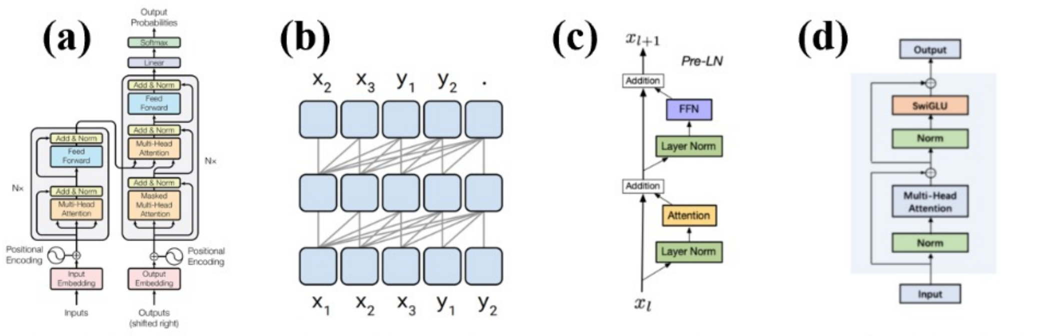
Totally 26k Q&A pairs for fine-tuning, 750M tokens for pre-training.

(a) HEP Related Domains; (b) Four methods employed to collect and clean up data

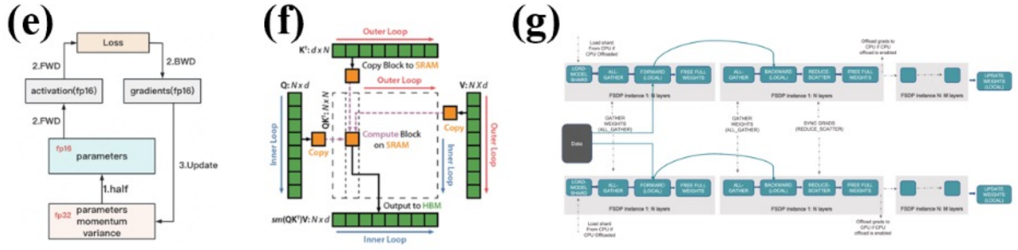
Collected, cleaned and trained dataset on July 2023: **5K Q&A pairs**
Collected, cleaned and **NOT** trained dataset on Sept. 2023: **21K Q&A pairs + 750M tokens**

New source: open Indico-Conferences-2168, pdf, ppt & mp4, about 100GB on disk.

HEP•Xiwu: Training techniques and Results



Transformer Casual Decoder Pre-LN Transformer Block



Mixed-Precision FlashAttention Fully Sharded Data Parallel

The illustration algorithm components and training technologies

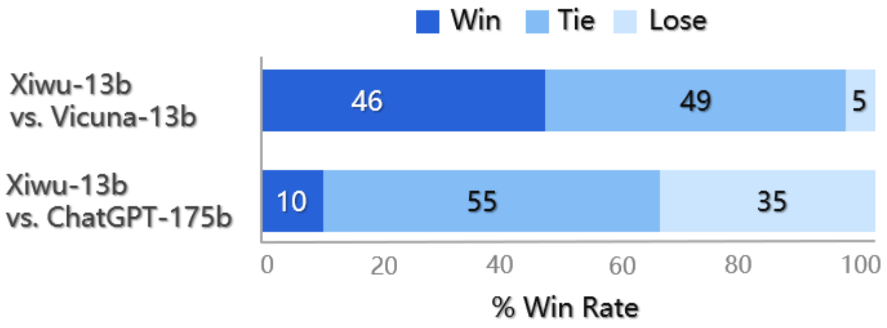
⬇

- 8*A100(40G) GPU，Xiwu-7B & 13B √
- 40*V100(32G) GPU，Xiwu-33B，Failed

Advanced training technologies for LLM：

- Transformers with Self-Attension
- Casual Decoder
- Pre-LN
- Transformer Blocks with multi-head attesion
- Mixed-Precision for fast training
- FlashAttention improves the utilization rate of the GPU's FLOPs.
- Low-Rank Adaptation (LORA) significantly reduces the need for fine-tuning.

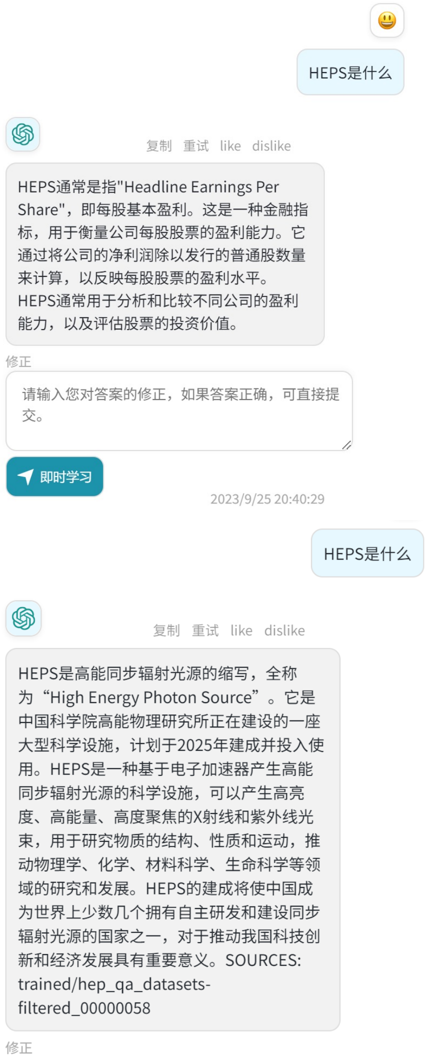
High Energy Physics 100 Question Test Results：



The Xiwu significantly outperforms the baseline model!

Just-In-Time Learning System

Pure LLM cannot handle dynamically changing information and are difficult to train on unprecedented data. ➡ Vector store as a memory module



HEPS是什么

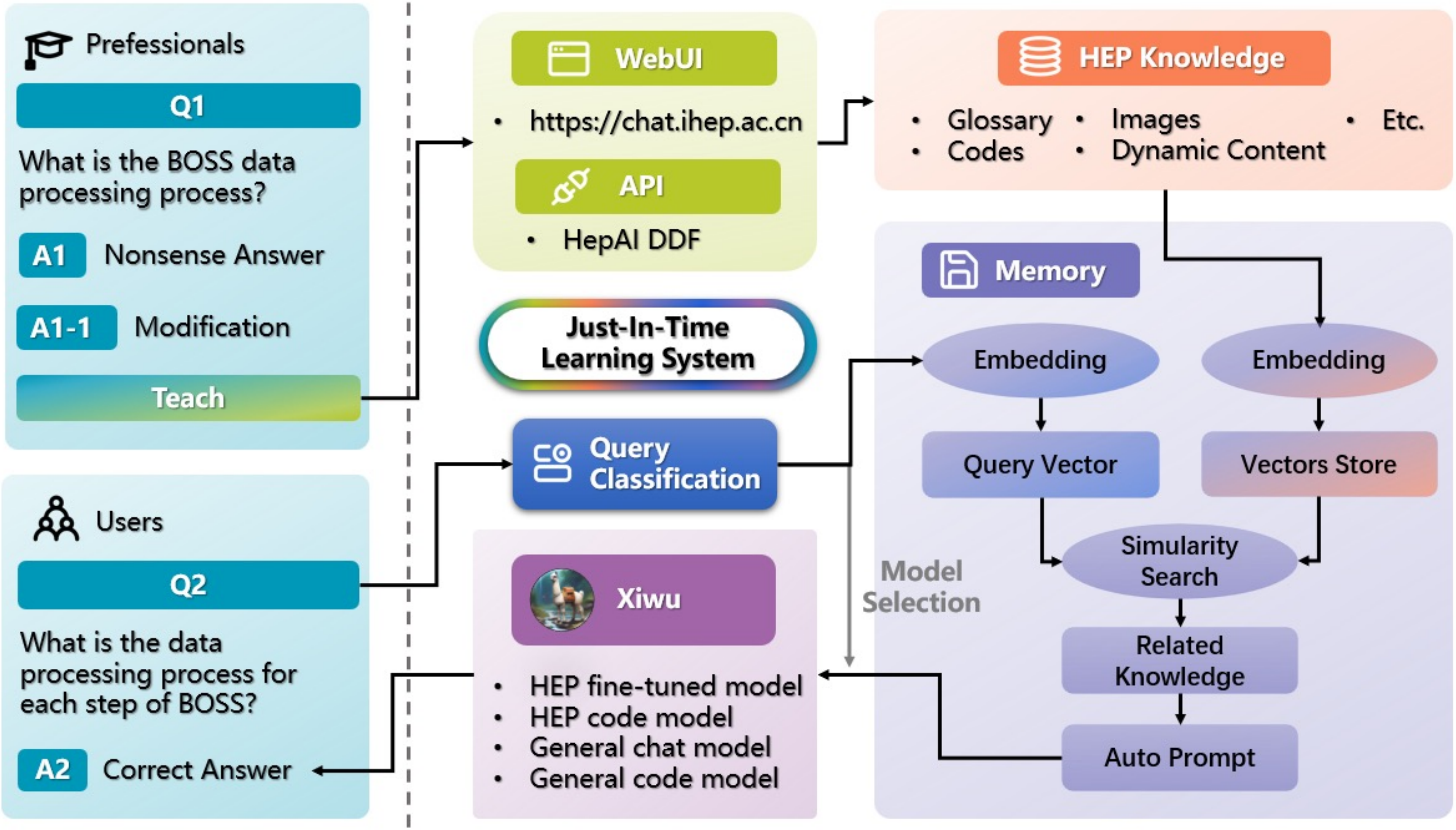
HEPS通常是指"Headline Earnings Per Share"，即每股基本盈利。这是一种金融指标，用于衡量公司每股股票的盈利能力。它通过将公司的净利润除以发行的普通股数量来计算，以反映每股股票的盈利水平。HEPS通常用于分析和比较不同公司的盈利能力，以及评估股票的投资价值。

请输入您对答案的修正，如果答案正确，可直接提交。

2023/9/25 20:40:29

HEPS是什么

HEPS是高能同步辐射光源的缩写，全称为“High Energy Photon Source”。它是中国科学院高能物理研究所正在建设的一座大型科学设施，计划于2025年建成并投入使用。HEPS是一种基于电子加速器产生高能同步辐射光源的科学设施，可以产生高亮度、高能量、高度聚焦的X射线和紫外线光束，用于研究物质的结构、性质和运动，推动物理学、化学、材料科学、生命科学等领域的研究和发展。HEPS的建成将使中国成为世界上少数几个拥有自主研发和建设同步辐射光源的国家之一，对于推动我国科技创新和经济发展具有重要意义。SOURCES: trained/hep_qa_datasets-filtered_00000058



Professionals

Q1

What is the BOSS data processing process?

A1 Nonsense Answer

A1-1 Modification

Teach

Users

Q2

What is the data processing process for each step of BOSS?

A2 Correct Answer

WebUI
- https://chat.ihep.ac.cn

API
- HepAI DDF

Just-In-Time Learning System

Query Classification

Xiwu
- HEP fine-tuned model
- HEP code model
- General chat model
- General code model

HEP Knowledge
- Glossary
- Codes
- Images
- Dynamic Content
- Etc.

Memory

Embedding

Embedding

Query Vector

Vectors Store

Simularity Search

Related Knowledge

Auto Prompt

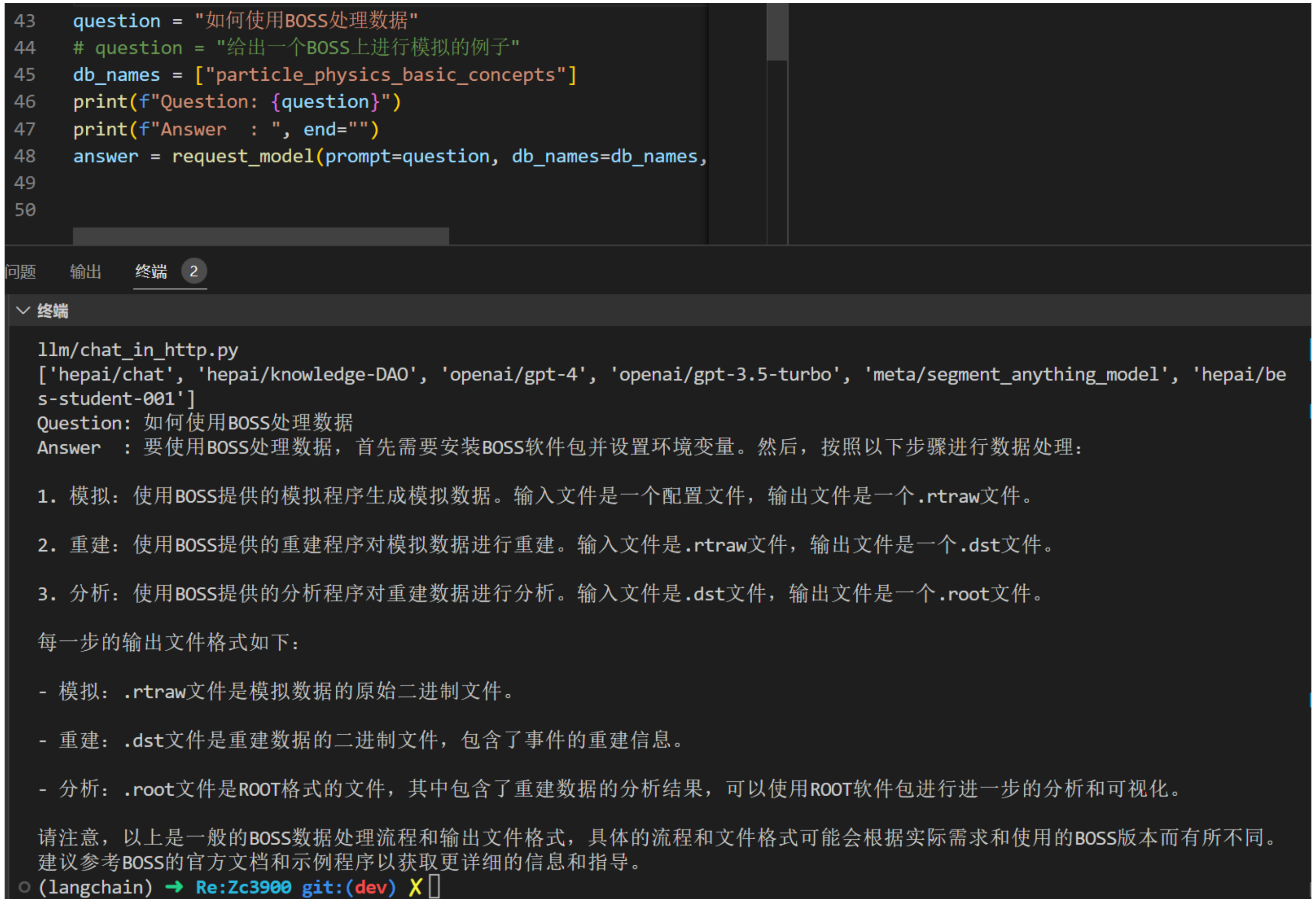
Model Selection

Xiwu vs GPT-4

- Q: **How to use BOSS to process BES data?**

- Ans. from Xiwu：

Correct →

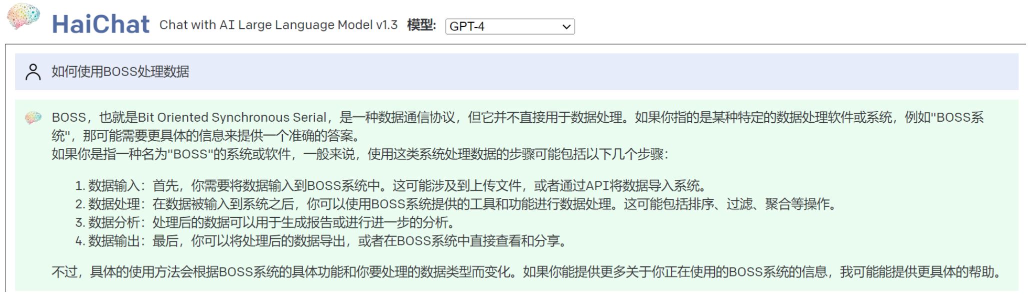


```python
question = "如何使用BOSS处理数据"
# question = "给出一个BOSS上进行模拟的例子"
db_names = ["particle_physics_basic_concepts"]
print(f"Question: {question}")
print(f"Answer  : ", end="")
answer = request_model(prompt=question, db_names=db_names,
```

```
llm/chat_in_http.py
['hepai/chat', 'hepai/knowledge-DAO', 'openai/gpt-4', 'openai/gpt-3.5-turbo', 'meta/segment_anything_model', 'hepai/be
s-student-001']
Question: 如何使用BOSS处理数据
Answer  : 要使用BOSS处理数据，首先需要安装BOSS软件包并设置环境变量。然后，按照以下步骤进行数据处理：

1. 模拟：使用BOSS提供的模拟程序生成模拟数据。输入文件是一个配置文件，输出文件是一个.rtraw文件。

2. 重建：使用BOSS提供的重建程序对模拟数据进行重建。输入文件是.rtraw文件，输出文件是一个.dst文件。

3. 分析：使用BOSS提供的分析程序对重建数据进行分析。输入文件是.dst文件，输出文件是一个.root文件。

每一步的输出文件格式如下：

- 模拟：.rtraw文件是模拟数据的原始二进制文件。

- 重建：.dst文件是重建数据的二进制文件，包含了事件的重建信息。

- 分析：.root文件是ROOT格式的文件，其中包含了重建数据的分析结果，可以使用ROOT软件包进行进一步的分析和可视化。

请注意，以上是一般的BOSS数据处理流程和输出文件格式，具体的流程和文件格式可能会根据实际需求和使用的BOSS版本而有所不同。
建议参考BOSS的官方文档和示例程序以获取更详细的信息和指导。
○ (langchain) ➜ Re:Zc3900 git:(dev) ✗ ▯
```

- Ans. from GPT-4：

↓ Nonsensical



如何使用BOSS处理数据

BOSS，也就是Bit Oriented Synchronous Serial，是一种数据通信协议，但它并不直接用于数据处理。如果你指的是某种特定的数据处理软件或系统，例如"BOSS系统"，那可能需要更具体的信息来提供一个准确的答案。
如果你是指一种名为"BOSS"的系统或软件，一般来说，使用这类系统处理数据的步骤可能包括以下几个步骤：

1. 数据输入：首先，你需要将数据输入到BOSS系统中。这可能涉及到上传文件，或者通过API将数据导入系统。
2. 数据处理：在数据被输入到系统之后，你可以使用BOSS系统提供的工具和功能进行数据处理。这可能包括排序、过滤、聚合等操作。
3. 数据分析：处理后的数据可以用于生成报告或进行进一步的分析。
4. 数据输出：最后，你可以将处理后的数据导出，或者在BOSS系统中直接查看和分享。

不过，具体的使用方法会根据BOSS系统的具体功能和你要处理的数据类型而变化。如果你能提供更多关于你正在使用的BOSS系统的信息，我可能能提供更具体的帮助。

- Xiwu excels in HEP-specific question answering and generating scientific codes.
- Xiwu-MoE combines domain knowledge with general capabilities for a more comprehensive approach.

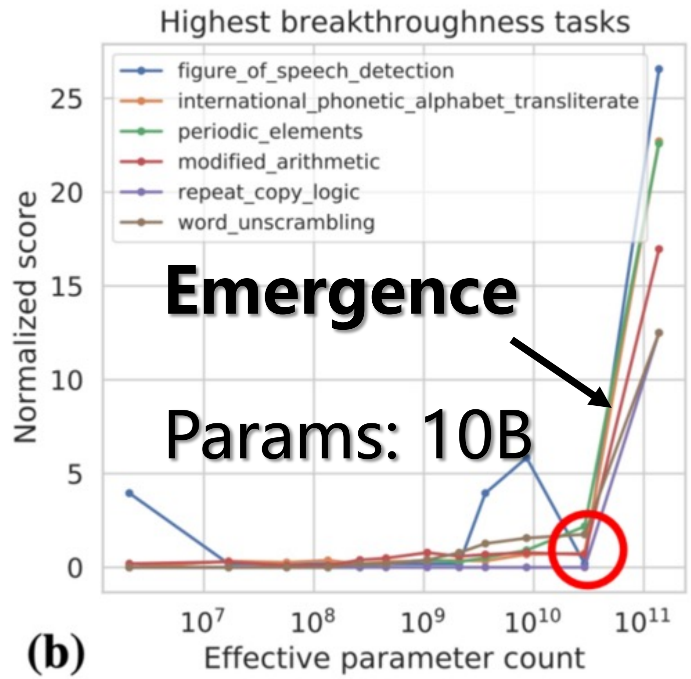
Part. 03

Exploration of AI large model

The first pathway:
Research Agent based on Xiwu

Particle Physics Large Model

GPT-4 has surpassed dedicated AI models in all downstream tasks of natural language processing (NLP), indeed showing sparks of artificial general intelligence (AGI).



Emergence

Params: 10B

(b)

Why and how LLM achieves such remarkable intelligence?

- The large amount of data forces neural networks to learn generic and useful “**neural circuits**”, while the large size of models provide enough **redundancy** and **diversity** for the neural circuits to specialize and fine-tune to specific tasks.
- The huge size of the model could have several other benefits, such as making gradient descent more effective by connecting different minima or by simply enabling smooth fitting of high-dimensional data.

Idea: **40PB HEP data** ➡ Leverage large-scale unsupervised pre-training methodologies to enable AI to grasp the **global rules** from all data, and employ Reinforcement Learning by Physical Feedback (**RLPF**) to facilitate its emergence.

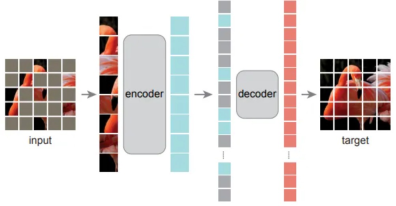
Difficulty 1: How is scientific data pre-trained?

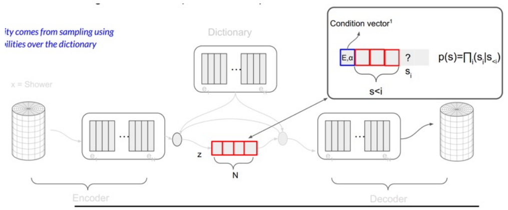
Characteristics of HEP data:

- Floating-point numbers with physical meaning.
- Non-sequential.

Pre-training Methods:

- Next word prediction ❌
- Mask Auto Encoding (MAE) ▬
- Vector Quantized Variational Auto-Encoder (VQ-VAE) ▬
- Graph + Transformer ▬
- Cross-modal supervision ▬ May be





Seeking collaboration

Particle Physics Large Model

- Cross-modal self-supervised method i.e. **text as supervisory signal**
- We are trying to develop a model that can handle multiple HEP tasks.

JetClass

Inputs：17 features

- part_px
- part_py
- part_pz
- part_energy:粒子的能量
- part_pt: 粒子的横向动量
- part_eta:粒子的赝快度，是一种描述粒子方向的量
- part_phi:粒子的方位角，即在探测器横截面上的角度
- part_deta:粒子赝快度与喷注赝快度的差
- part_dphi:
- part_d0val:粒子的横向冲量的测量值
- part_d0err:粒子的横向冲量的误差
- part_dzval:粒子的纵向冲量的测量值
- part_dzerr:粒子的纵向冲量的误差
- part_charge:粒子的电荷
- part_isChargedHadron:是否为带电强子
- part_isNeutralHadron：是否为中性强子
- part_isPhoton：是否为光子
- part_isElectron：是否为电子
- part_isMuon：是否为μ子

Out：10 classes

[label_QCD, label_Hbb, label_Hcc, label_Hgg, label_H4q, label_Hqql, label_Zqq, label_Wqq, label_Tbqq, label_Tbl]

Toplandscape

Inputs：7 features

- [part_pt_log, 1.7, 0.7]
- [part_e_log, 2.0, 0.7]
- [part_logptrel, -4.7, 0.7]
- [part_logerel, -4.7, 0.7]
- [part_deltaR, 0.2, 4.0]
- [part_deta, null]
- [part_dphi, null]

Out：2 classes

jet_isTop → a top jet
jet_isQCD

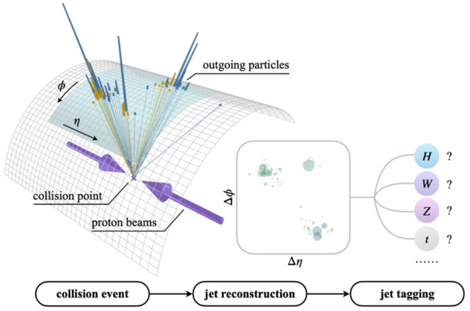
QuarkGloun

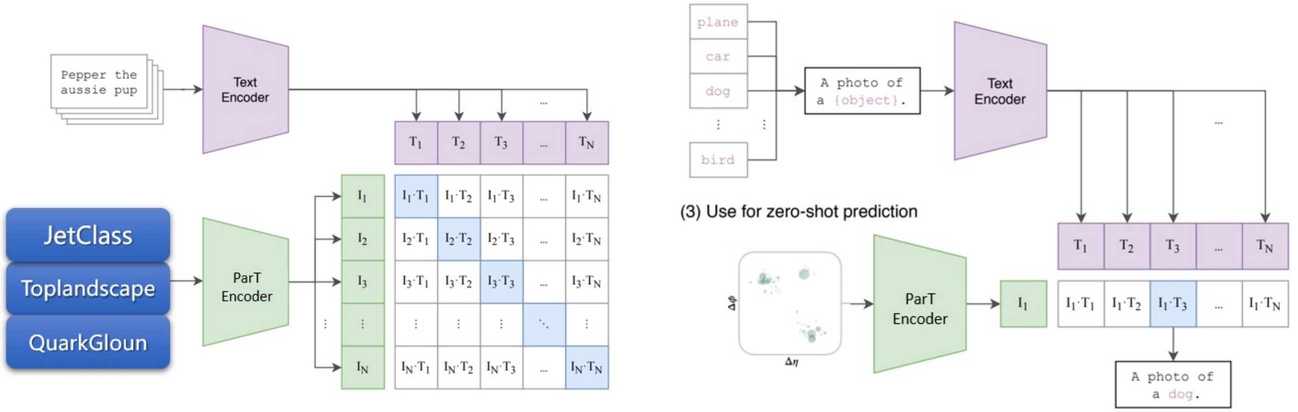
Inputs：13 features

- [part_pt_log, 1.7, 0.7]
- [part_e_log, 2.0, 0.7]
- [part_logptrel, -4.7, 0.7]
- [part_logerel, -4.7, 0.7]
- [part_deltaR, 0.2, 4.0]
- [part_charge, null]
- [part_isCHad, null]
- [part_isNHad, null]
- [part_isPhoton, null]
- [part_isElectron, null]
- [part_isMuon, null]
- [part_deta, null]
- [part_dphi, null]

Out：2 classes

jet_isQ
jet_isG

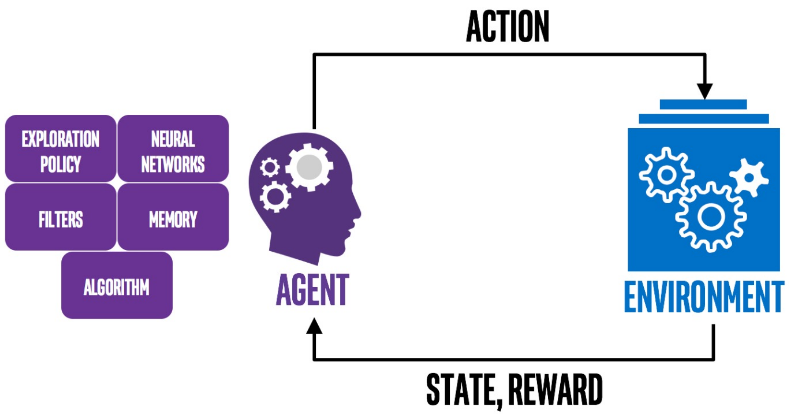




Difficulty 2: How to align large models with physics?

Possibility 1: Reinforcement learning with interactive environments.

- A reinforcement learning system that uses the BOSS program as the environment
- Enhance abilities such as scientific code programming in the context of reinforcement learning.



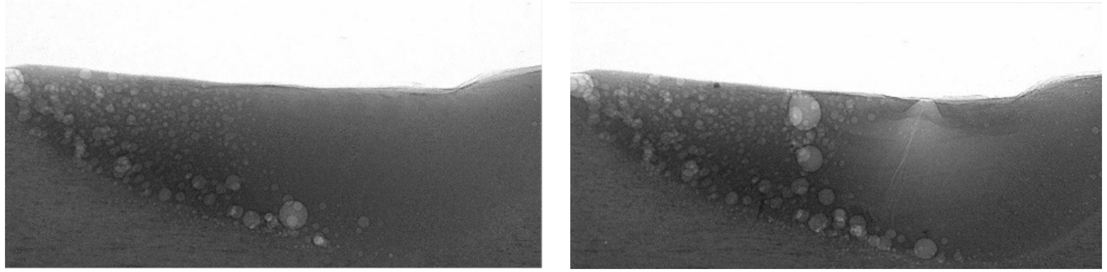
Possibility 2: Reinforcement learning by Physical Feedback (RLPF)

- In the evolution from GPT-3 to ChatGPT, Reinforcement Learning from Human Feedback (RLHF) plays a crucial role in alignment. The performance of the 175 billion parameter model without RLHF is even worse than the 7 billion parameter model with RLHF.
- The establishment of **Reinforcement Learning from Physical Feedback (RLPF)** mechanism can achieve alignment between the model and physical principles. (Requires further thought)

Example 1: Defects real-time feedback

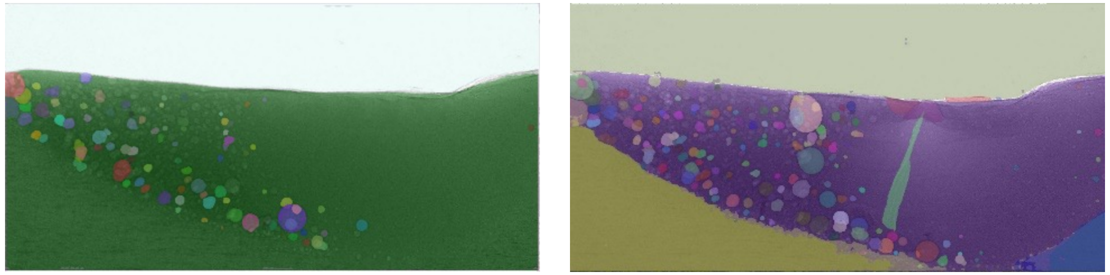
Defects real-time feedback of additive manufacturing

(Dr. B.B.Zhang, Z.D.Zhang)



- Complex
- Low resolution
- dynamic

- Defect Detection and Tracking



- Defects such as bubbles and melt pools directly affect material properties
- Laser intensity and powder feeding rate can control the distribution of bubbles and melt pools

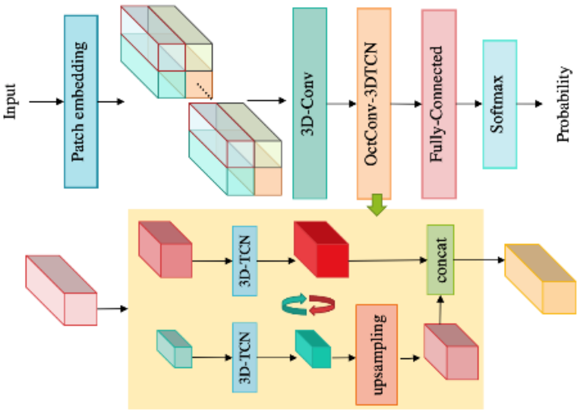
Old Mode: The laser intensity and powder delivery rate are determined based on experience.
New Mode: Based on AI real-time feedback of defect information, the laser intensity and powder delivery rate are **dynamically controlled**.

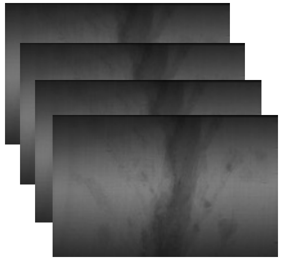
Example 4: Image data compression

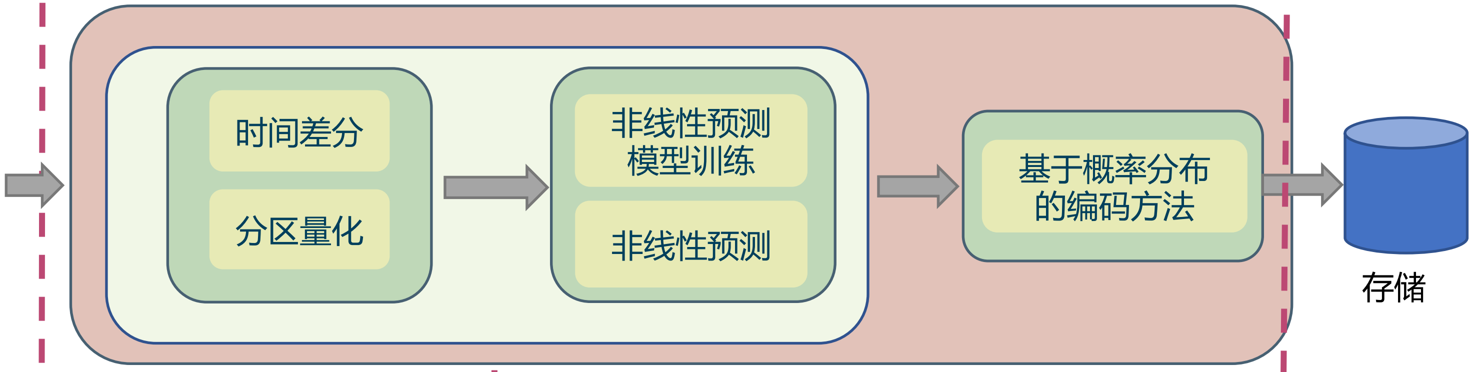
lossless compression for light source images

(Dr. S.Y.Fu)

- Improving compression rate through convolutional neural network prediction probability and Huffman coding
- Saving **58%~66%** storage space
- Equivalent to reducing several hundred PB of data storage per year for HEPS







时间差分

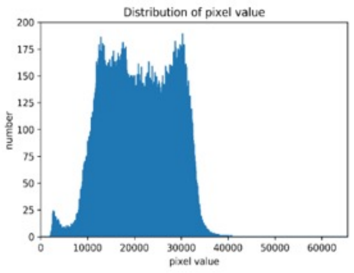
分区量化

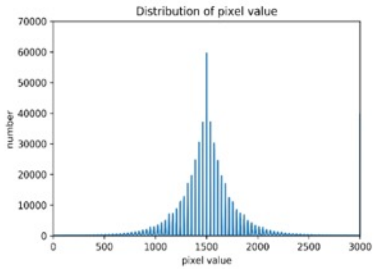
非线性预测
模型训练

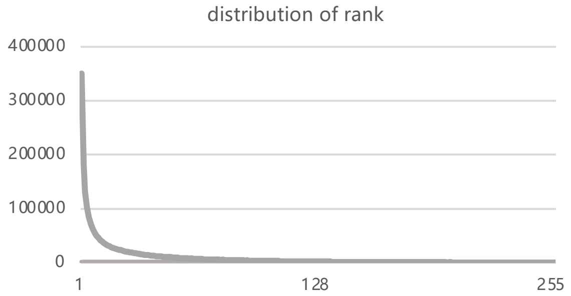
非线性预测

基于概率分布
的编码方法

存储







The Roadmap of Developing Large Models for HEP

Applications

Text and image tasks

- Domain knowledge Q&A
- Scientific code generation
- Image Processing
- Research Agent

Scientific tasks

- Jet tagging
- Shower simulation etc.

…

System Emergence

- Propose issues
- Validate issues
- Make new discoveries

Algorithms

Large Language & Image Model

+

Large Model for HEP Scientific Data

BESIII

HEPS

SNS

…

AI Scientist

$$\mathcal{L}_{StandModel} = \mathcal{L}_{gauge} + \mathcal{L}_{Higgs} + \mathcal{L}_{fermion} + \mathcal{L}_{Yukawa}$$

Data Level

Domain Text

Domain Image

2023

Scientific Data

JetClass, JetImages etc. datasets

2028

- Multimodal fusion of Text, image, and scientific data
- Virtual world data loop

2035

Computing Power

GPU

GPU → NPU

NPU、DCU