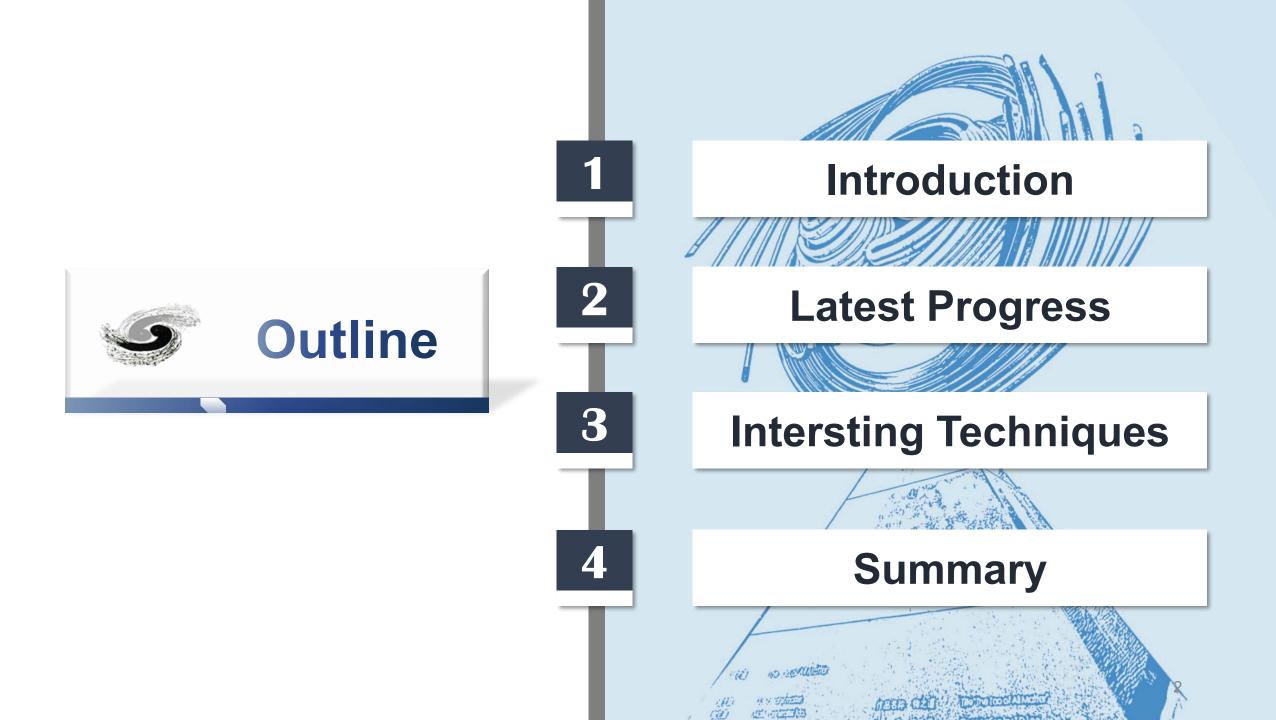


Computing Center & Experimental Physics Center, IHEP, CAS
National High Energy Physics Scientific Data Center (NHEPSDC)



Dr. Sai Al Agents for Particle Physics Analysis

Zhengde Zhang, on behalf of Dr.Sai Group Quantum Computing and Machine Learning Workshop March, 21, 2025



Physical analysis: a key link in scientific discovery





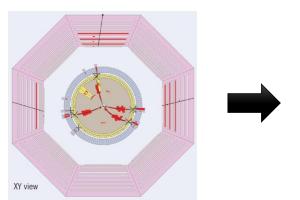
Data Processing Flow for High Energy Physics Experiments

(Collider as an Example)

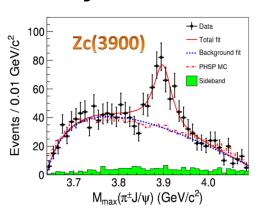


Physical results Selection and statistical analysis extraction

Collision data



Physical results



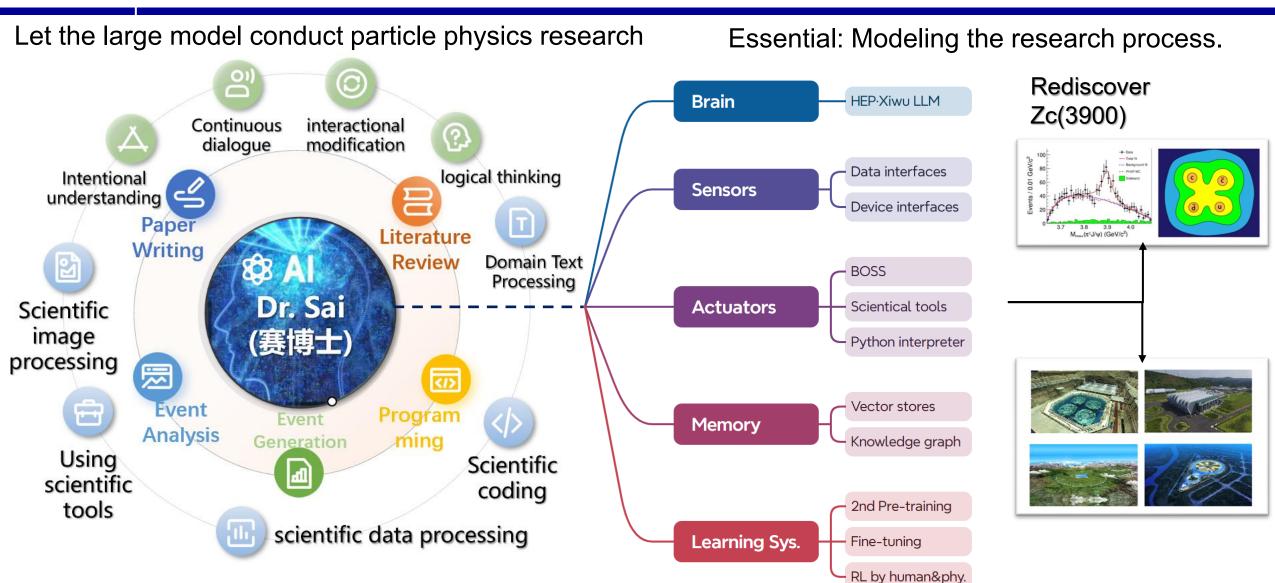
For the first time, mankind has discovered a particle composed of four quarks.

Complex tasks!

- Math + Physics domain expertise required
- Equivalent lines of code for windows/macOS
- A small mission requires: several people + several years

Dr.Sai Agents for Particle Physics Analysis





Components

Latest Progress (2025.07)



Online Demo:

https://ihepbox.ihep.ac.cn/ihepbox/index.php/s/DsaHtndLiu52R5w

Enhance Domain Capabilities



89.82%

Dr.Sai-Host

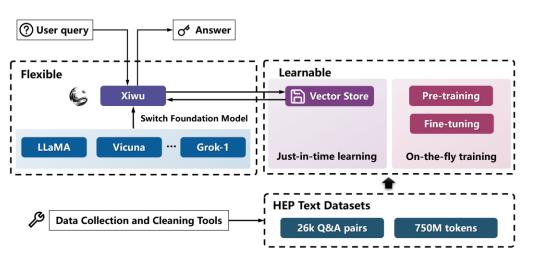
72.85%

DeepSeek-671B

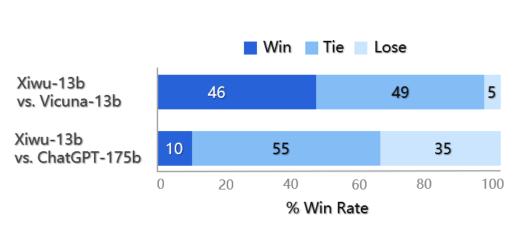
Brain

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

- ☐ The **Brain** of Dr.Sai Agent.
- □ Currently, Xiwu is based on **DeepSeek-R1-Distill-Llama-70B**, Historically, LLaMA1~3.
- Secondary pre-training and fine-tuning techniques.
- Trained on GPU-A100 previously, now trained on DCU-K100 (Domestic GPU)
- □ Significantly better than the base model in HEP Q&A and internal code generation.
- New Model: **xLAM-2-32b-fc-r**, for funcition calling



Test Results



¹ arXiv:2404.08001

Dr.Sai Agent - RAG

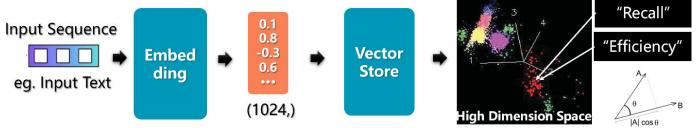


Memory

Vector Databases and Knowledge Graphs

- Retrieval Augmented Generation (RAG)
 - ☐ The most cost-effective and promising solution for hallucination reduction
 - Store private data without retraining
- ☐ Current approach: vector storage
 - **□** Embedding (using BAAI's BGE-M3 model):
 - Convert input data into high dimensional spatial vectors
 - RAG system (LLaMA Index).

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



(Based on LLaMA Index)

? What is HEPS?



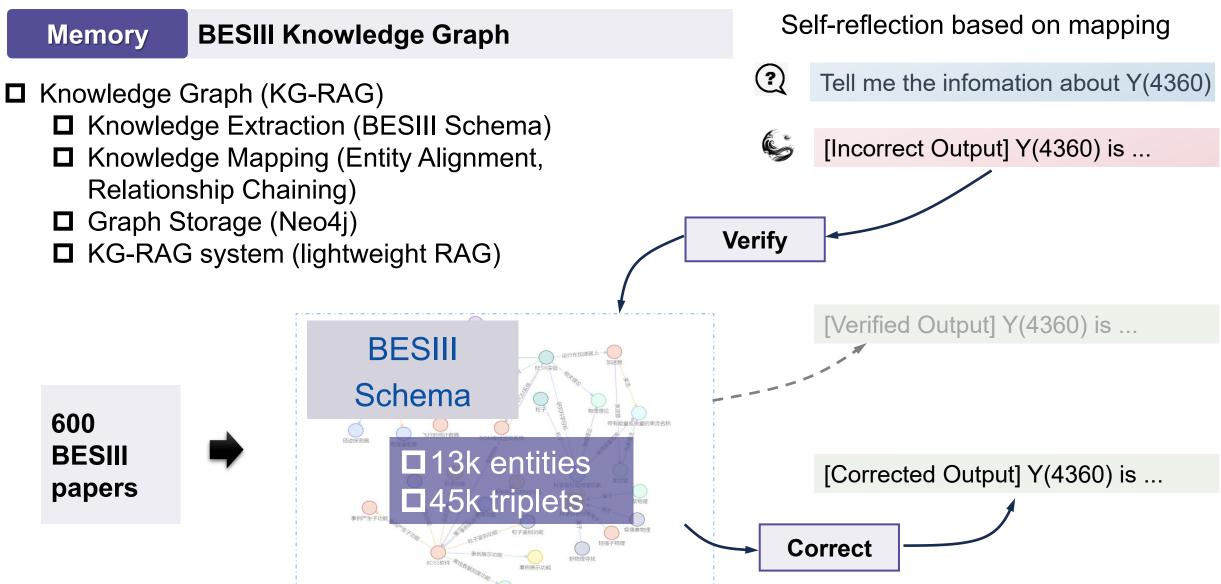
HEPS stands for High Energy Photon Source. It is ...

source: /docs/ihep/facilities.pdf

- Provide answers based on user intent and information
- □ Providing sources of information

Dr.Sai Agent - Knowledge Graph





From Answers to Action



Actuator

HepAI-DDF

High-performance AI Distributed Deployment Framework

- Core Features
- Unified Access: Seamless integration of AI models, scientific tools, vector databases, and knowledge graphs across heterogeneous resources (CPU/GPU/DCU).
- Smart Scheduling: Automated heartbeat monitoring + dynamic load balancing.
- **Unified API Gateway**: Standardized service interfaces with API-KEY authentication for enhanced security.
- **Remote Model Invocation**: Groundbreaking support for remote server model deployment with seamless local calling experience.
- Adapted MCP: Model Context Protocol, Access to 2600+ MCP Server.



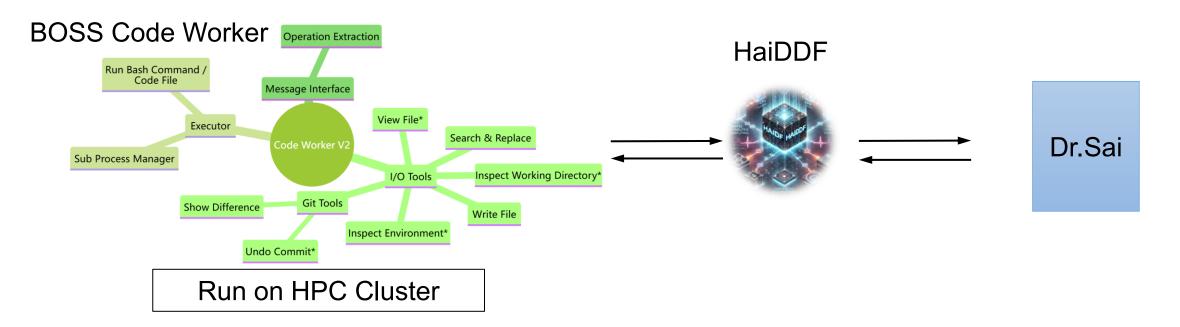
- □ 50+ Models
- 2 local tools
- N MCP tools
- In one server

From Answers to Action



Actuator

BOSS Code Worker



- I/O Tools (files and data)
- Message interface
- Executor

From **Q&A chatbot** to **expert assistant**

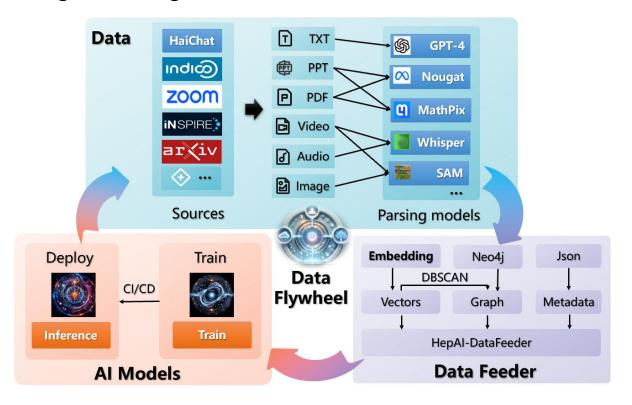
Dr.Sai Agent – Data Fly-wheel System



Sensors

Data Flywheel System

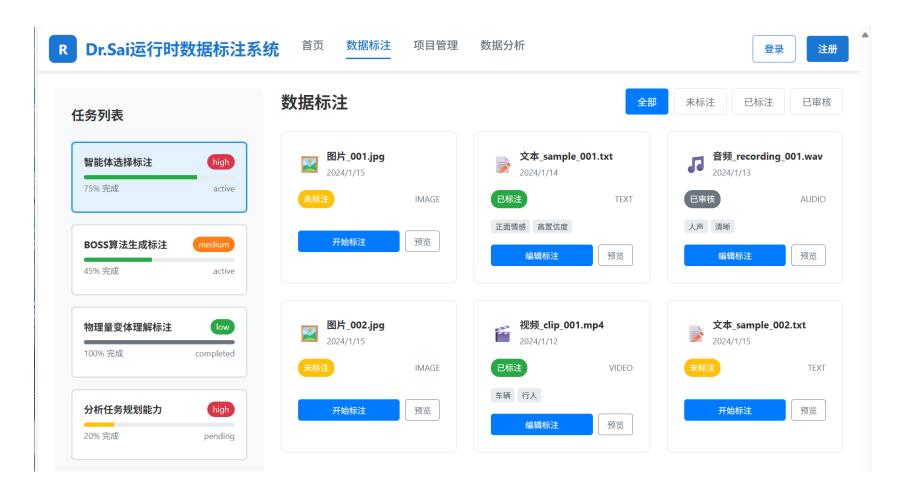
- The "data flywheel" enables continuous model iteration and evolution.
 - Data flywheel, i.e. data-driven flywheel effect
 - Improves models by constructing circular data pathways
 - Attracts more users, generating more data, further enhance the models.



Runtime Snapshot

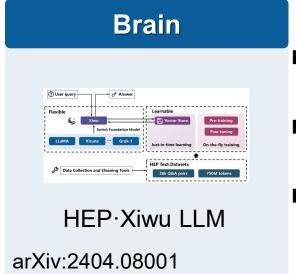


- Runtime snapshots save configuration information, detailed conversations of Dr. Sai.
- It can be further processed into an AI-Ready dataset to feed back to train Dr. Sai's base model.



Dr.Sai Single Agent





- Based on LLaMA3 and DeepSeek
- Trained on A800 and domestic DCU
- Enhanced domain capabilities

Memory



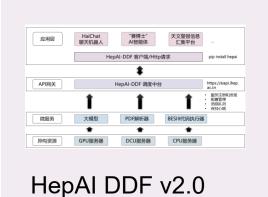
Vector Store



Knowledge Graph

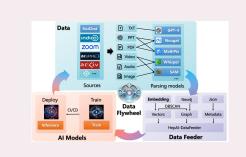
- Vector Store
 - BAAI BGE-M3
 - LLaMA Index
- Knowledge Graph
 - LightRAG

Actuators



- BESIII Actuator
 - BOSS Worker
- Daisy Framework
 - HEPSCT
 - HEPSPtycho etc
- Unified access of models, tools etc

Sensors



Data Flywheel System

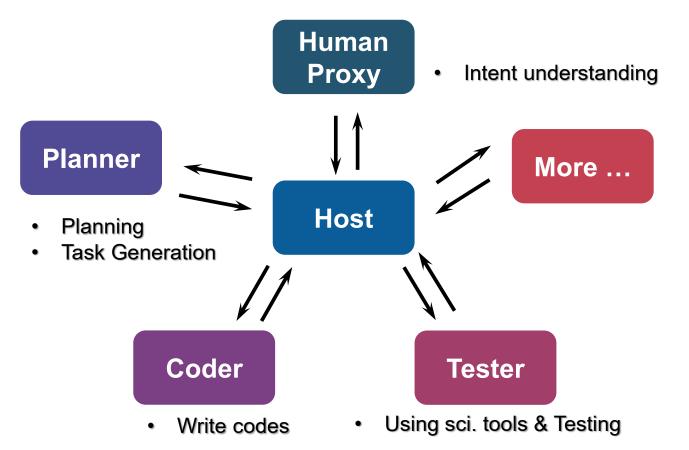
- Data Flywheel
 - HaiNougat
 - HaiChat
 - Whisper, SAM
 - ArXiv, Indico
- Domas Framework
 - Scientifc metadata

Dr.Sai Multi-Agent Collaborative System



Multi agent dynamic task processing system supporting human-machine collaboration

- Multi-Agent Collaboration (**MAC**) is implemented based on <u>AutoGen</u> 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.



Technological breakthrough

Hierarchical task model: The task tree system automatically decomposes complex targets and tracks the status and dependencies of subtasks in real-time.

Domain customization process: Optimize interaction logic for physical analysis scenarios.

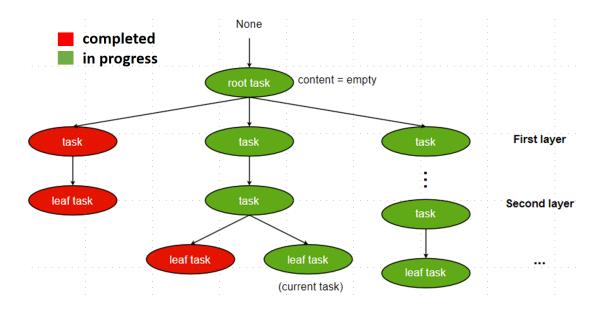
Dr.Sai Tech - HTM



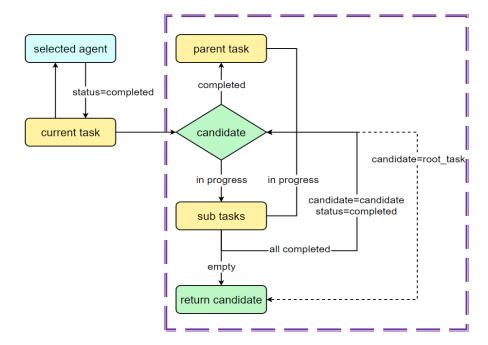
HTM

Hierarchical Task Model

- Independent task system (simulating questioning behavior)
 - Using task tree to manage complex physical analysis tasks
 - Ensure task backtracking closed-loop logic through fractal design
 - Implement task tree addition, deletion, modification, and query



任务树结构



Dr.Sai Agent - Interface





- Dr.Sai v2.x (based on MagenticUI)
- **■** Features:
 - Data Awareness and Progress Details Sidebar
 - Rich human-computer interaction
 - expected to go live on 2025.12

https://drsai.ihep.ac.cn

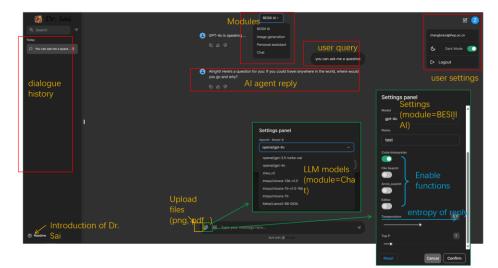
Current version:

Dr.Sai v2.0 (2025.07)

(Based on OpenwebUI)



Dr.Sai v1.0 (July, 2024) (Based on Chainlit)







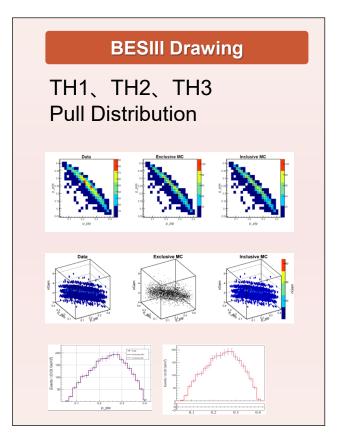
- **■** Features:
 - Automatic authentication
 - Guardian Process Manager Tasks

Dr.Sai Agent – Mapping



- BOSS Modular Code Coding Technology
- Solve the problem that **BOSS code** is long and the modeling ability to grow code is insufficient, using "template + mapping" to achieve.





Dr.Sai Agent – Particle Spelling variant DB

- Database of spelling variants of particle physics variables
- Resolving spelling differences in domain object description symbols

A database for compatible spelling variants and standardization of particle names

Manual Data Annotation with <u>PDG</u> and <u>Particle</u> APIs

Usage: compatible with different spellings during IO, generate data for ML

```
P+ Proton $\overline{p}$

$pbar

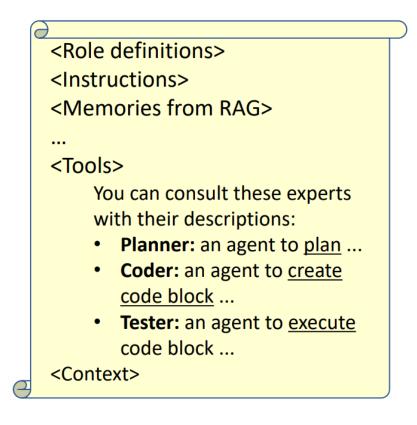
$p_bar
```

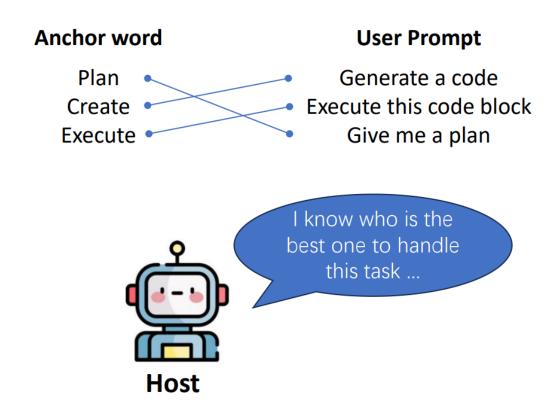
```
"name": "pbar",
"mcid": -2212,
"programmatic name": "p_bar",
"latex name": "\\overline{p}"
"evtgen_name": "anti-p-",
"html name": "p̅",
"unicode name": null,
"aliases": [
  "\\bar{p}",
  "anti-proton"
"typo": [
  "p_-",
  "pbar(0)",
  "p^-(0)",
  "p^- 0",
  "p^-(1)"
```

Dr.Sai Agent – Anchor word



- Anchor words.
- Anchor words can enhance the Host agent's ability to allocate tasks.



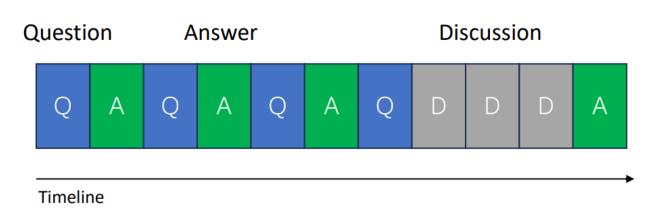


Dr.Sai Agent – MOM



- Message Offloading Mechanism (MOM) in Multi-Intelligent Body Collaboration
- Solving the problem of distracted model attention in long context scenarios

- ➤ Host: sees all the QA pairs
- Experts: sees the latest QA pair and global info
- Global info: key info extracted from chat history

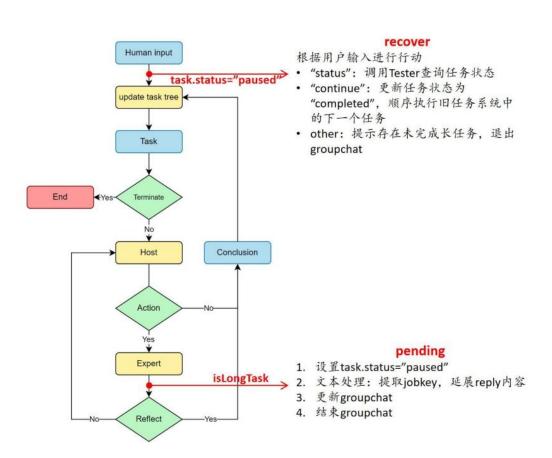


A typical conversation

Dr.Sai Tech - Daemon



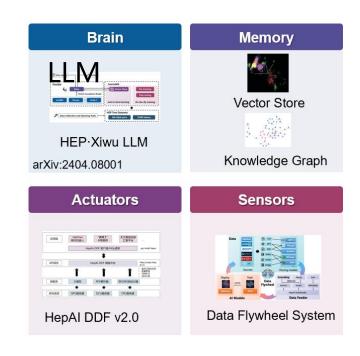
- **Daemon** takes over the concurrent tasks of the task queue
- Solve the problem of front-end, back-end and multi-component communication of intelligences in long task scenarios (days or even weeks) of physical analysis

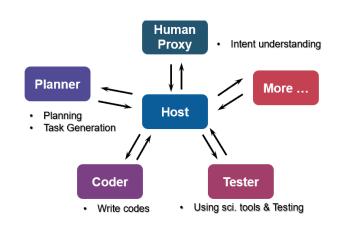


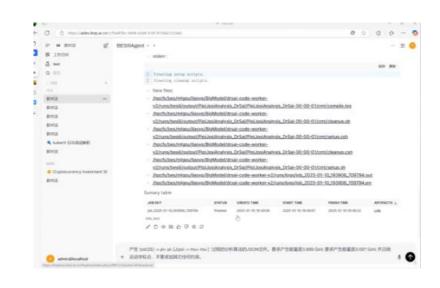
Expected messages Q: <task1> A: The jobs are submitted, please hold on. Enter key words: ["status", "continue"] to proceed. >> exit groupchat Q: status A: <job status from Tester> >> exit groupchat Q: continue A: <answer to task2> (groupchat continue) Q: <other query> A: You have unfinished long tasks! >> exit groupchat

Summary









Single Agent

Multi-Agents

User Interface

Start from data, generate simulation code and analysis code, execute tasks, create plots, and analyze to obtain physical results.

Summary and outlook





- We expect high-energy physics AI scientists, who have currently achieved Dr. Sai v2.0.0.
- **Six major modules**: brain, perception layer, execution layer, memory layer, multi-agent collaboration, human-computer interaction interface
- Domain adaptation technology: Mapping, Code Worker, DSL, Spelling Variant, Anchor word, MOM, HTM, Daemon, Snashot
- Realize automatic analysis on 10 decay channels, with a success rate of approximately 30% for unoptimized new decay channels.

Outlook



- Dr. Sai upgrade
 - Upgraded new frontend for Human-Computer interaction
 - Support for more decay modes
 - Training stronger base models
- How far from practical? Need engineering realization.
- How far from full autonomous scientific discovery? Limited by the capabilities of the base model.
- Can it be generalized to other experiments or scientific studies? Completely feasible!

Related Links



- Al Services
 - HepAl Platform: https://ai.ihep.ac.cn
 - HaiDDF Modeling Service:: https://aiapi.ihep.ac.cn
 - Dr.Sai Intelligence Body: https://drsai.ihep.ac.cn
- Open Source Project
 - HepAl open source Python library: https://github.com/zhangzhengde0225/hepai
 - OpenDr.Sai: https://code.ihep.ac.cn/hepai/drsai
 - HEP-Xiwu LLM: https://github.com/zhangzhengde0225/Xiwu

- Contact Us:
- Zhengde Zhang (zdzhang@ihep.ac.cn)
- Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences

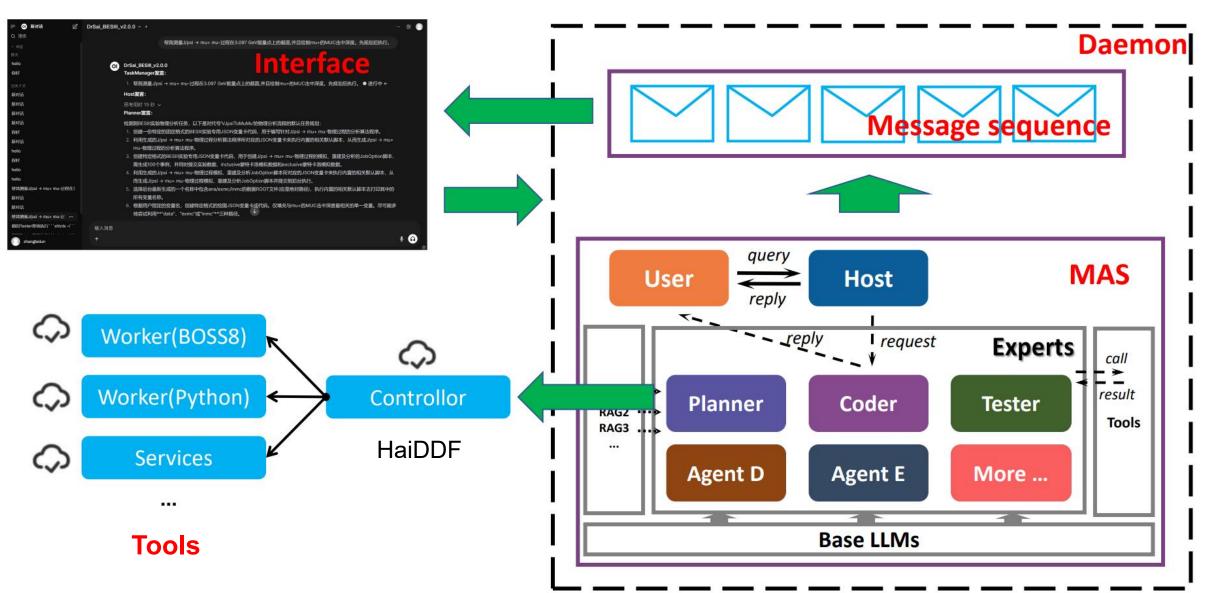




Backup

Dr.Sai Architecture





语言模型和科学模型需要深度融合



- 智能体为科学模型提供训练数据
- ■科学模型与语言模型通过新 "Language" 融合

"The limits of my language mean the limits of my world."

我的语言之局限,即我的世界之局限

——Wittgenstein

维特根斯坦



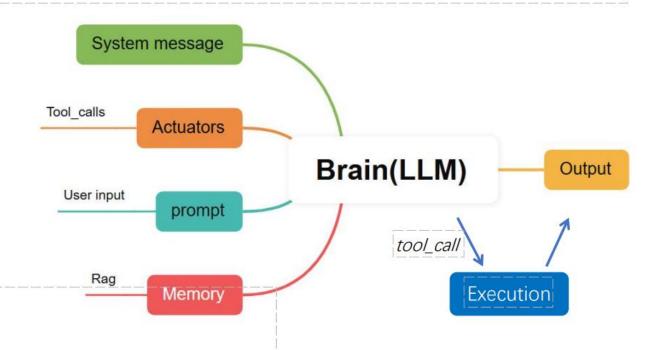
从LLM到多智能体系统



什么是Agent (智能体)?

An entity that can <u>perceive its environment</u>, <u>make decisions</u>, and <u>take actions</u> in order to achieve certain goals or sets of goals.

- Perception: multimodal -> text
- Decision: process text
- Action: reply or tool_call



- ➤ Agent = LLM + tools
- DIY reply_function to design the LLM output procedure

Enhance Domain Capabilities



- Please provide an example of simulate particle decay on BOSS (请给出在BOSS (BESIII离线软件系统)中常见的粒子模拟示例)
- **GPT-4:**

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.

In BOSS (BESIII Offline Software System), a common particle simulation is:

```
e^+e^-	o\psi(2S)	o J/\psi\pi^+\pi^-
```

📞 Xiwu :

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:

In this example, the entire simulation process is conducted at a collision energy of 3.686 GeV and uses the KKMC generator.

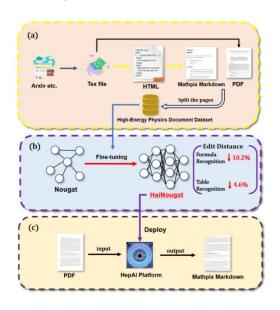
This simulated particle decay process is commonly used in high-energy physics experiments to study the properties and behavior of particles. I hope this example is helpful to you!

母 **结论**:GPT-4生成错误概念,提供看似正确但无用的方案;溪悟正确解释BOSS概念,提供 衰变配置与生成器示例代码。

PDF Parser

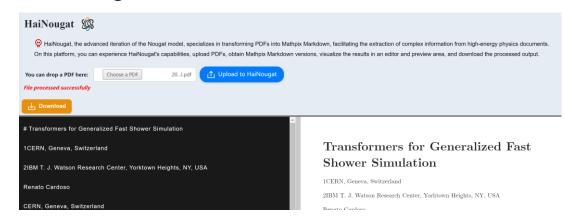


HaiNougat: A More Accurate PDF parser for High Energy Physics



- Approch:
 - Collected and cleaned 100,000 paper images and Al Ready data in markdown format of HEP
 - HaiNougat model was fine tuned and trained based on Nougat

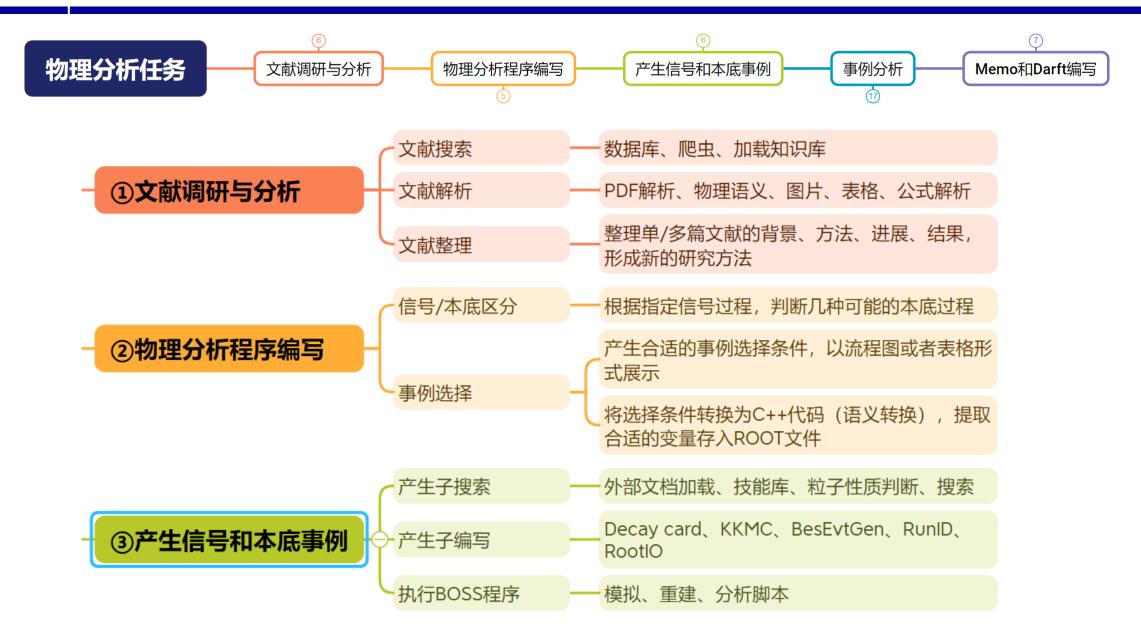
HaiNougat UI https://chat.ihep.ac.cn:5557/



- Results
 - PDF papers → Markdown (Machine readable)
 - Preserve formulas, tables, and images
 - Speed: 8 pages per batch, 20 seconds (RTX 4090)

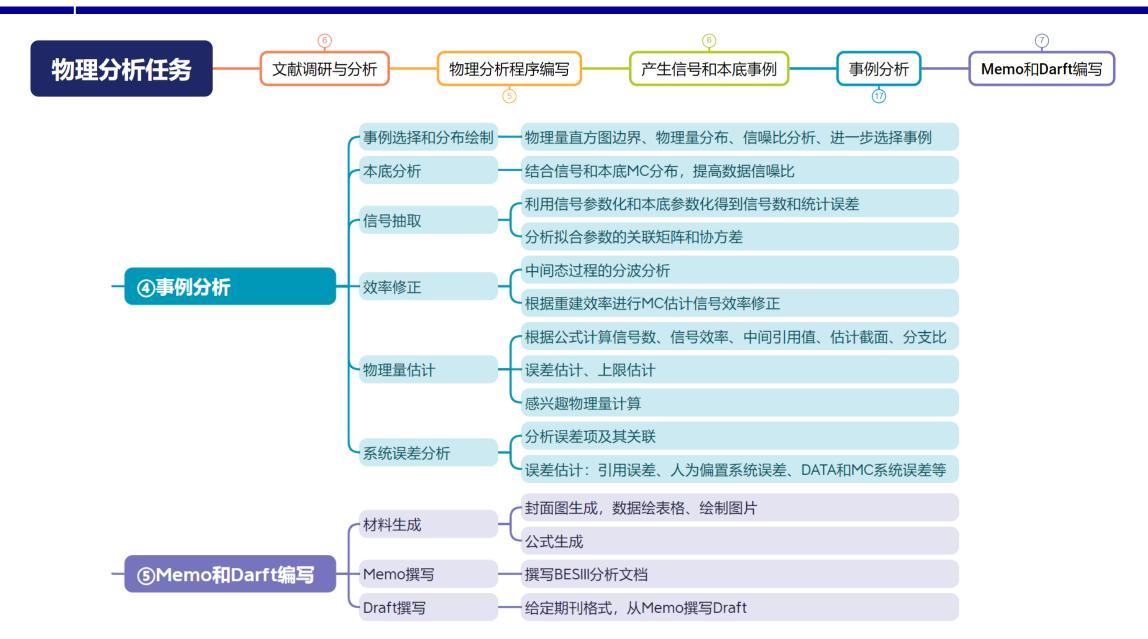
物理分析任务





物理分析任务





Latest Progress (2025.07)



0 Unclear task request

- 1 Well-defined sub-tasks (Planner)
- 2 Generate BESIII simulation & analysis programs (Coder)
 - Retrieve template
 - Generate variable card
 - ◆ Template + variable card →
 - ◆ Config: decay card
 - Script: Sim+Rec
 - ◆ Script: Ana
- 4 Execute tasks (Actuator)
 - ◆ Task management system
 - ◆ Sim→Electronic signals.rtraw
 - ◆ Rec →Track information.dst
 - ◆ Ana→ Particle level information.root (momentum,
 - energy, mass etc)
- 5 Retrieve and analyze results
 - mc.root, data.root: utilize MC signal shapes as templates to identify genuine signals
 - **E** Example: using $\pi^+\pi^-$ J/ψ signal profile, determining the existence of Zc(3900)

Dr.Sai: Automated & Collaborative Workflow for BESIII Analysis

Online Demo:

https://ihepbox.ihep.ac.cn/ihepbox/index.php/s/DsaHtndLiu52R5w