

**BES III**

Computing Center, IHEP, CAS  
National HEP Data Center

# Dr.Sai: An AI agents system for BESIII experiment

Zhengde ZHANG, on behalf of Dr.Sai project

July, 20, 2024

Prague, Czech Republic

# Outline

01 Introduction of BESIII experiment & AI Agent

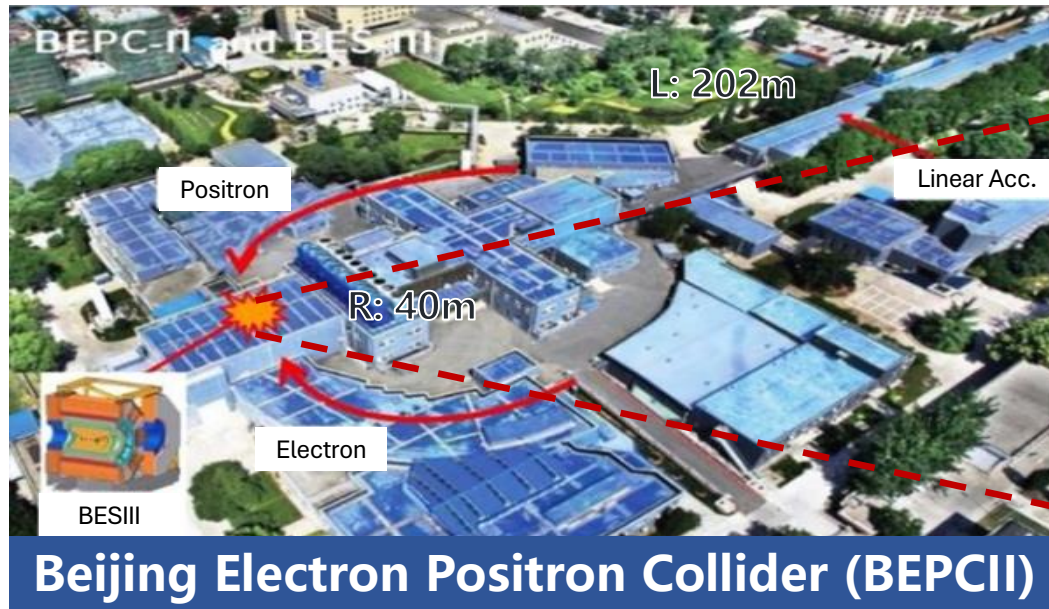
02 Overview and Components of Dr.Sai

03 Demonstrations

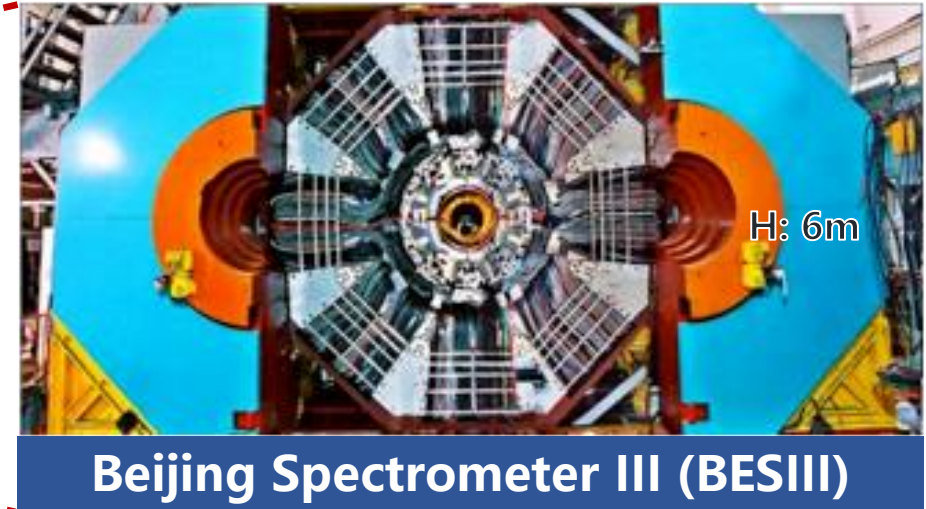
04 Summary



# The BESIII Experiment



Detector  
Installed  
on BEPC



## BESIII

- Rich physics
- Over 600 published articles so far
- 16 countries, over 600 scientists

## Data

- The largest data sample in the world in the charm energy region
- 2-5.6GeV, resonance peak and scanning

## Utilize to the fullest extent

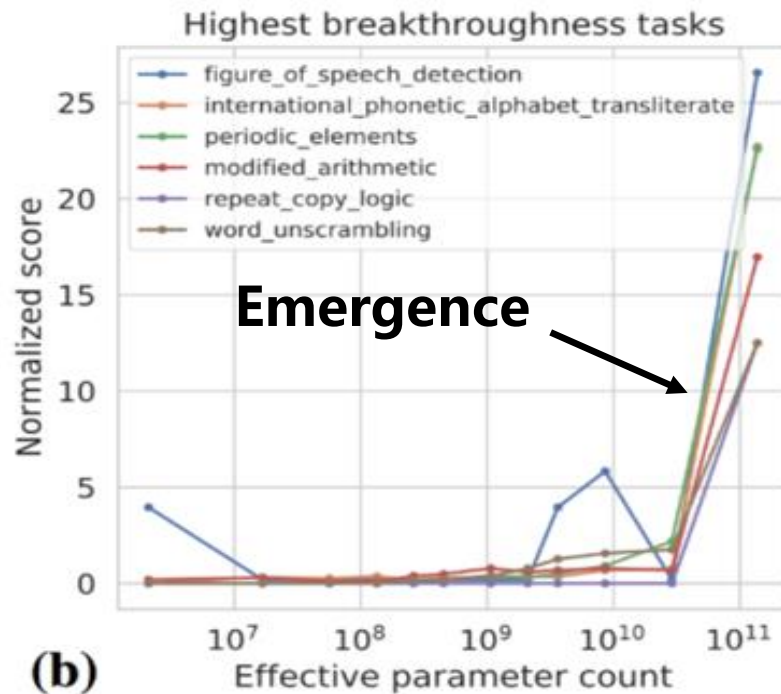
- The physics potential is yet to be fully uncovered
- AI can help fully unlock the potential of massive high-dimensional data.

# Generative AI achieve significant progress



GPT-4 surpasses specialized AI models in all downstream natural language processing tasks, potentially paving the way towards general artificial intelligence.

(10.48550/arXiv.2303.12712)

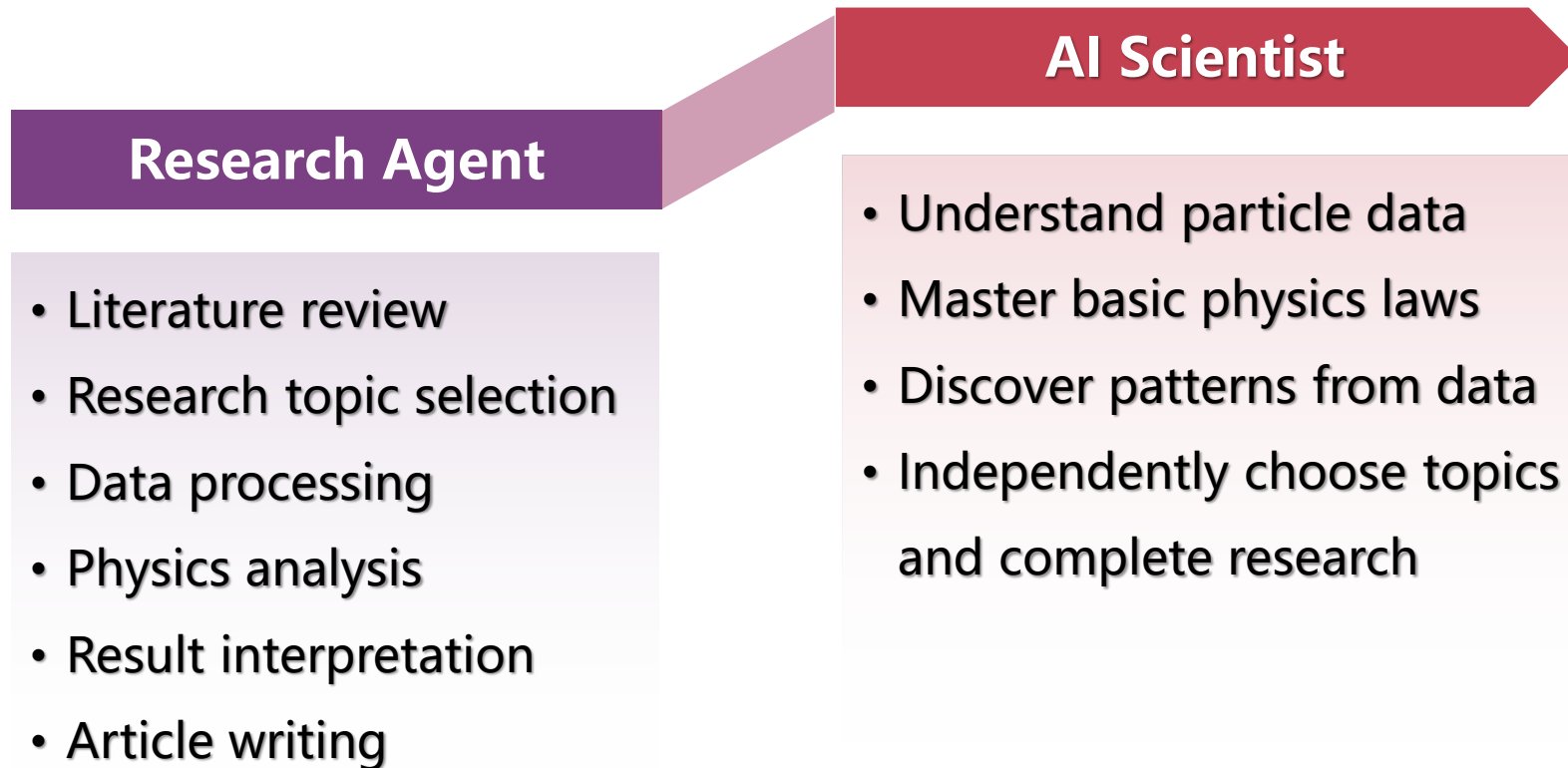
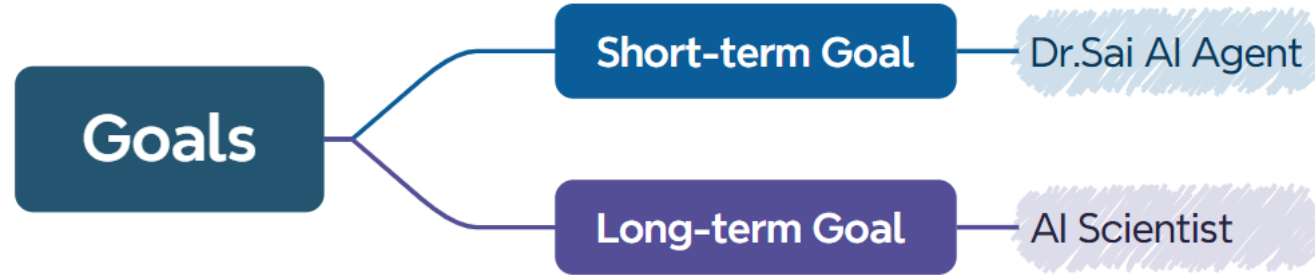


## Impressive Capabilities of LLM:

- General Intent Understanding Capability
- Powerful Code Generation Ability
- Intelligent Interaction Correction Ability
- Moderate Inference Ability

LLM: Large Language Model

AI is expected to undertake higher-level scientific tasks, providing "semi-finished products" for scientific discoveries.



# 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  $\neq$  A person

**A agent  $\approx$  A person**

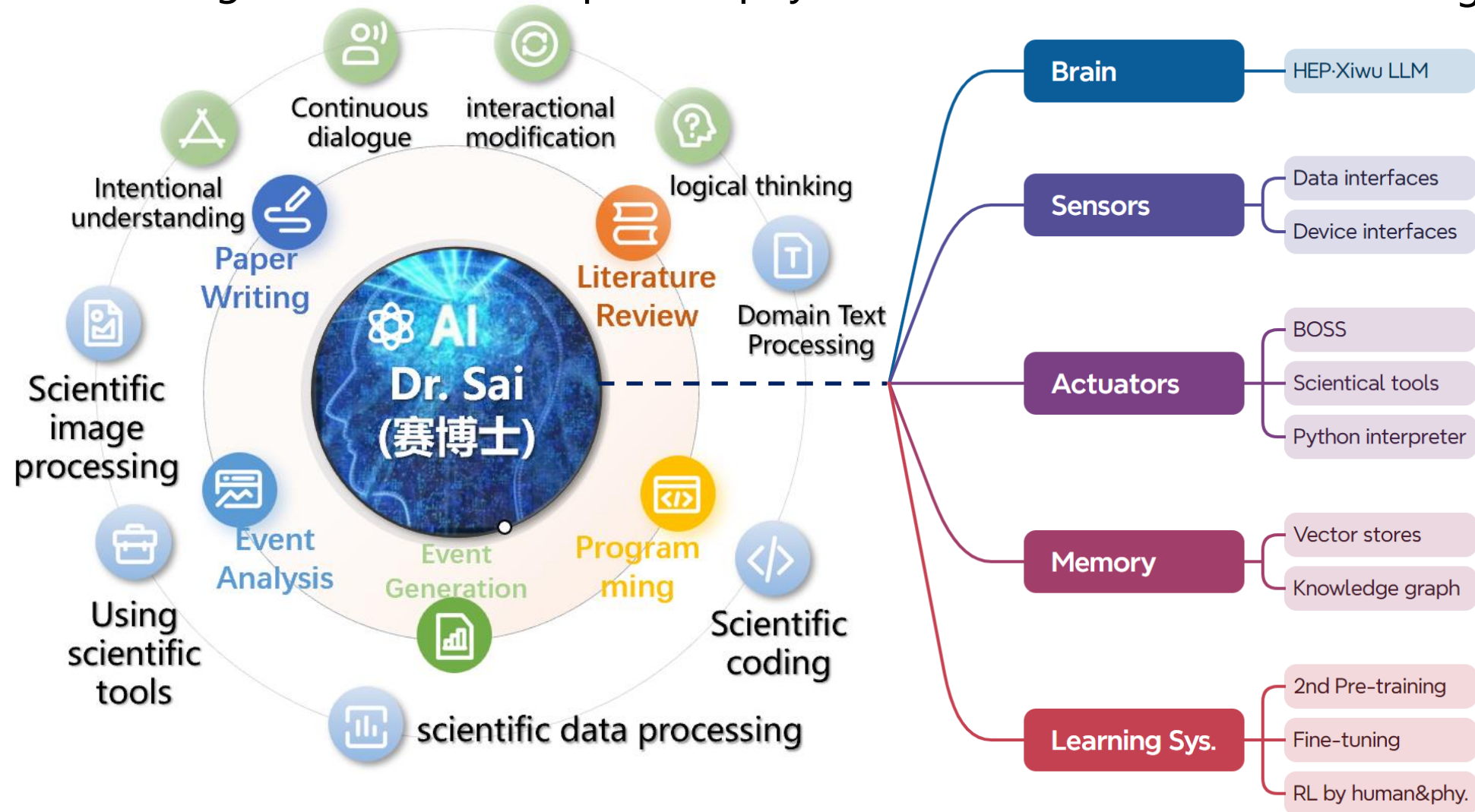
The LLM is the **processing core** (the Brain) of agent.



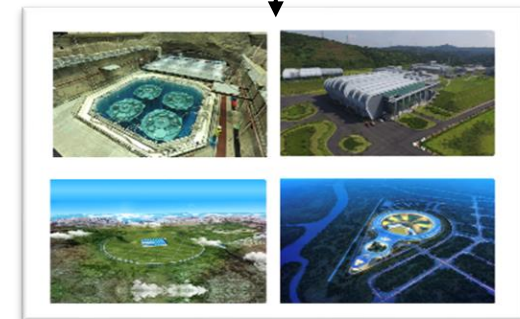
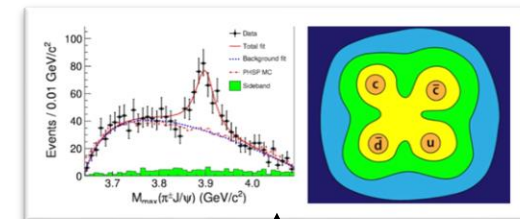
# Overview of Dr.Sai Agent for BESIII



Let the large model conduct particle physics research    Essential: Modeling the research process.



## Rediscover Zc(3900)



Tasks & Required capabilities

Components

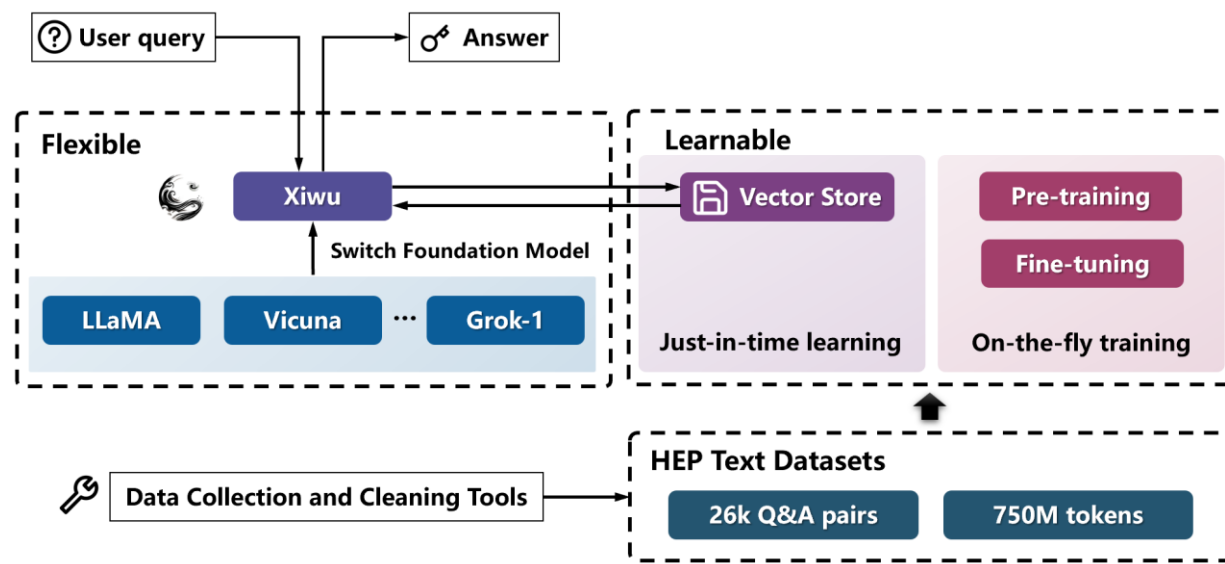
Applications 7

# The Brain of Dr.Sai – Xiwu LLM

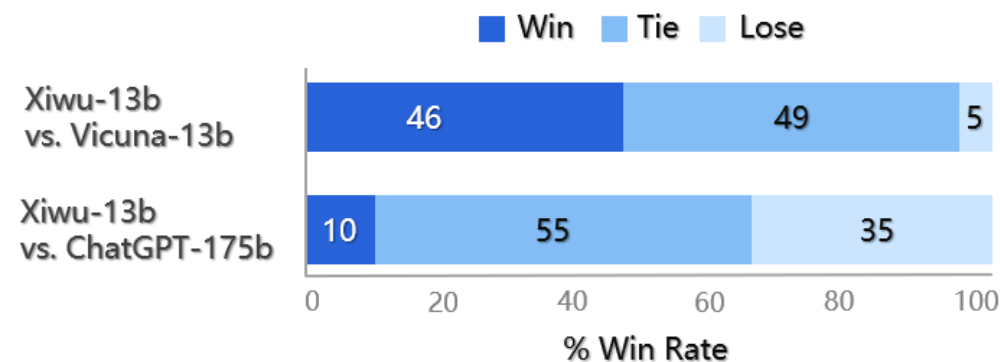


## Xiwu(溪悟): A Basis Flexible and Learnable LLM for High Energy Physics <sup>1</sup>

- 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.



## Test Results



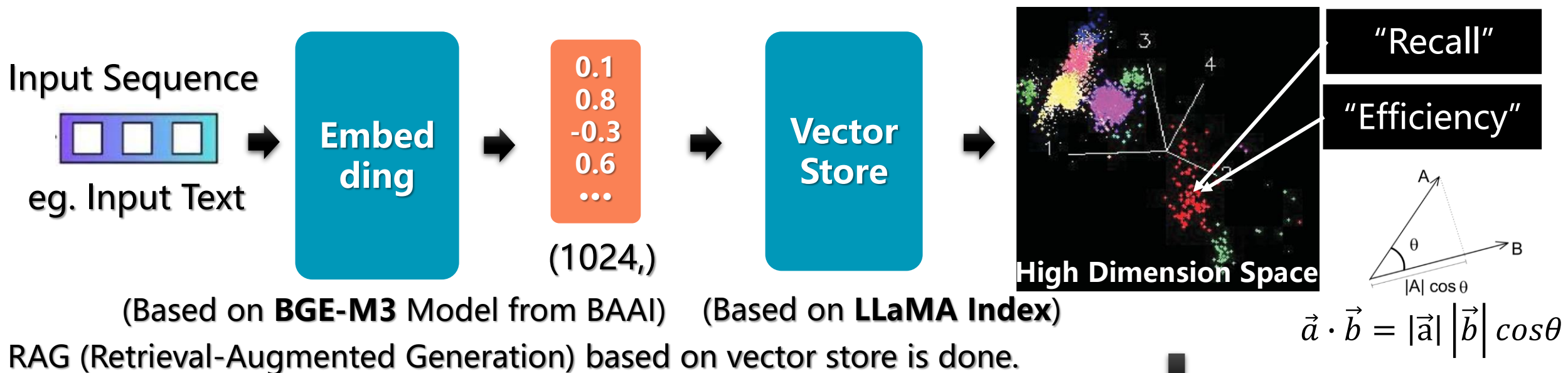
<sup>1</sup> [arXiv:2404.08001](https://arxiv.org/abs/2404.08001)



# The **Memory** of Dr.Sai – VS & KG

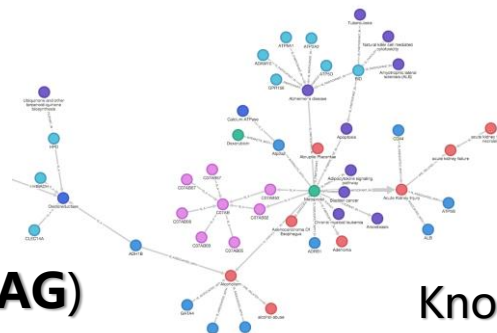


- Save HEP knowledge into **Vector Store** and **Knowledge Graph**.
- Retrieve them to enhance the model's generation accuracy and reduce hallucination.



- self-reflection could be realized.

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



Knowledge Graph

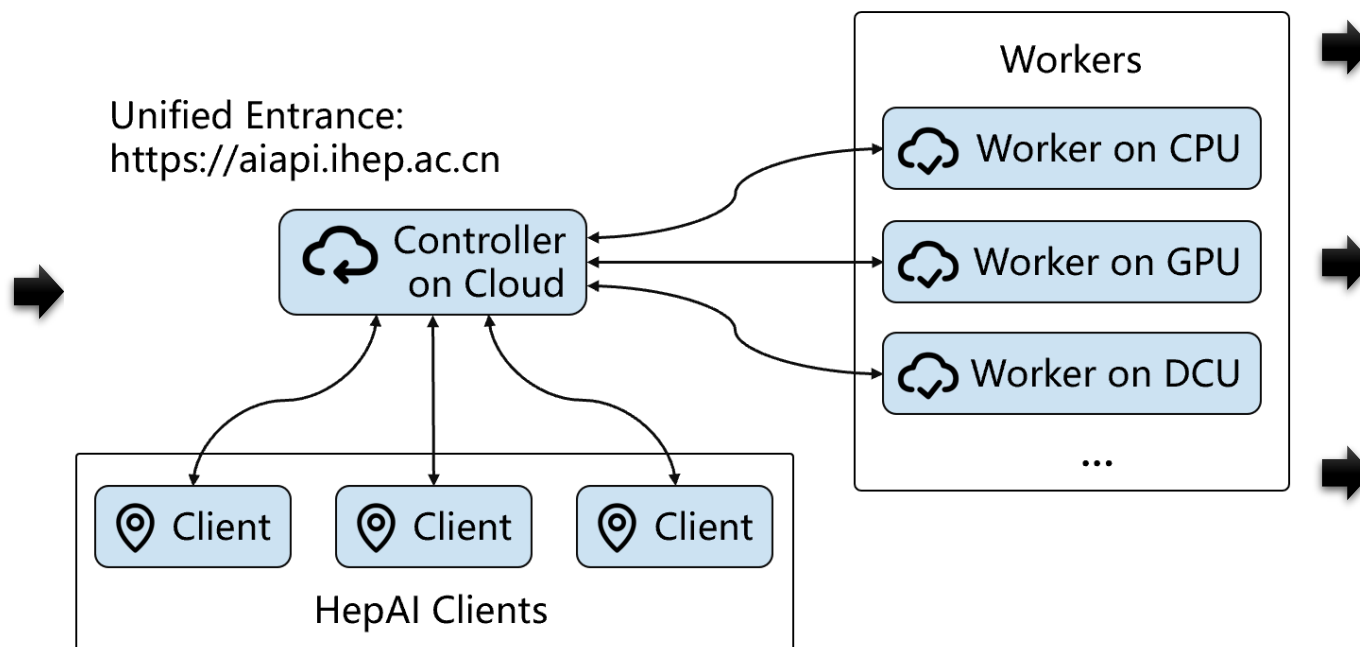
**DBSCAN**

- Unsupervised clustering

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



- The Distributed Deployment Framework (HepAI-DDF) is developed.
  - Featured with flexibility, cross-language, cross-platform, heterogeneous.
  - 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.



**BOSS 7.1.0**

For executing physical analysis code.

**Daisy**

For invoking scientific tools

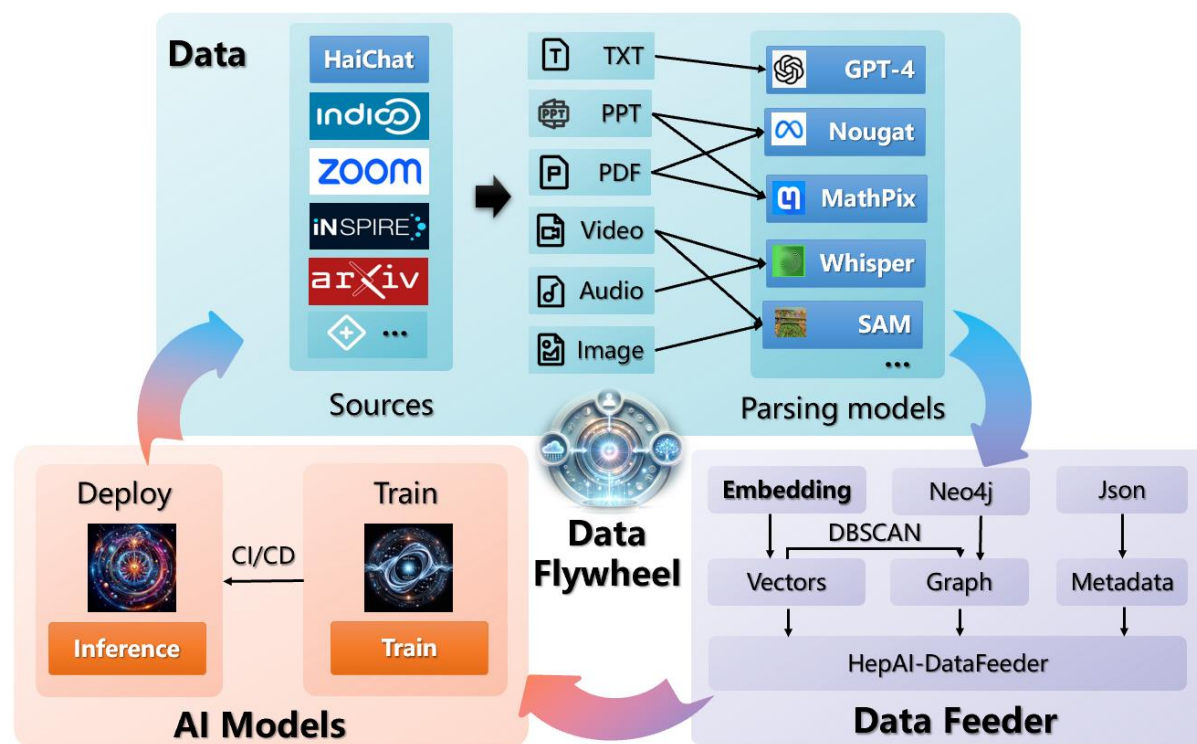
**Python Interp.**

For interpreting general Python code.

# The **Sensors** of Dr.Sai – Data Flywheel

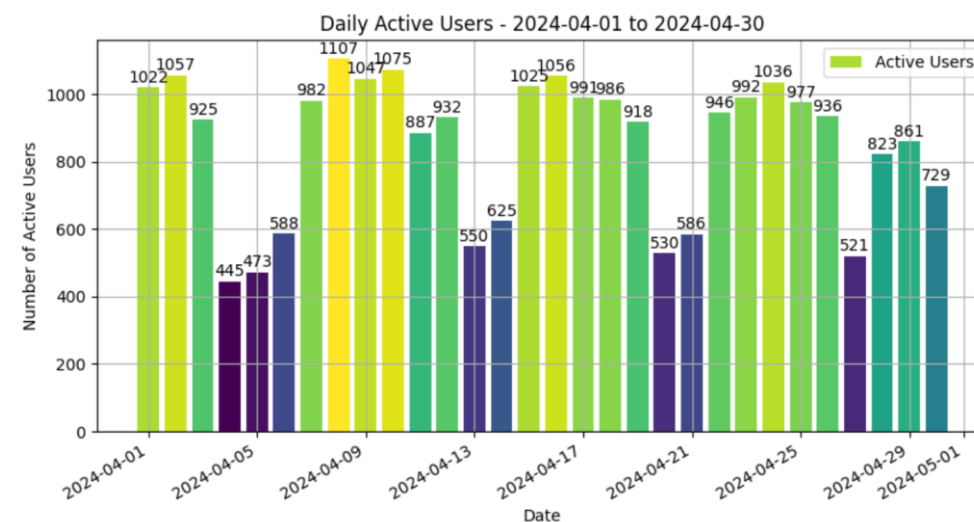


- The "data flywheel" enables continuous model iteration and evolution.
  - Data flywheel, i.e. **data-driven flywheel effect**
  - Improves models by constructing circular data pathways
  - Attracts more users, generating more data, further enhance the models.



- **HaiChat service**

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





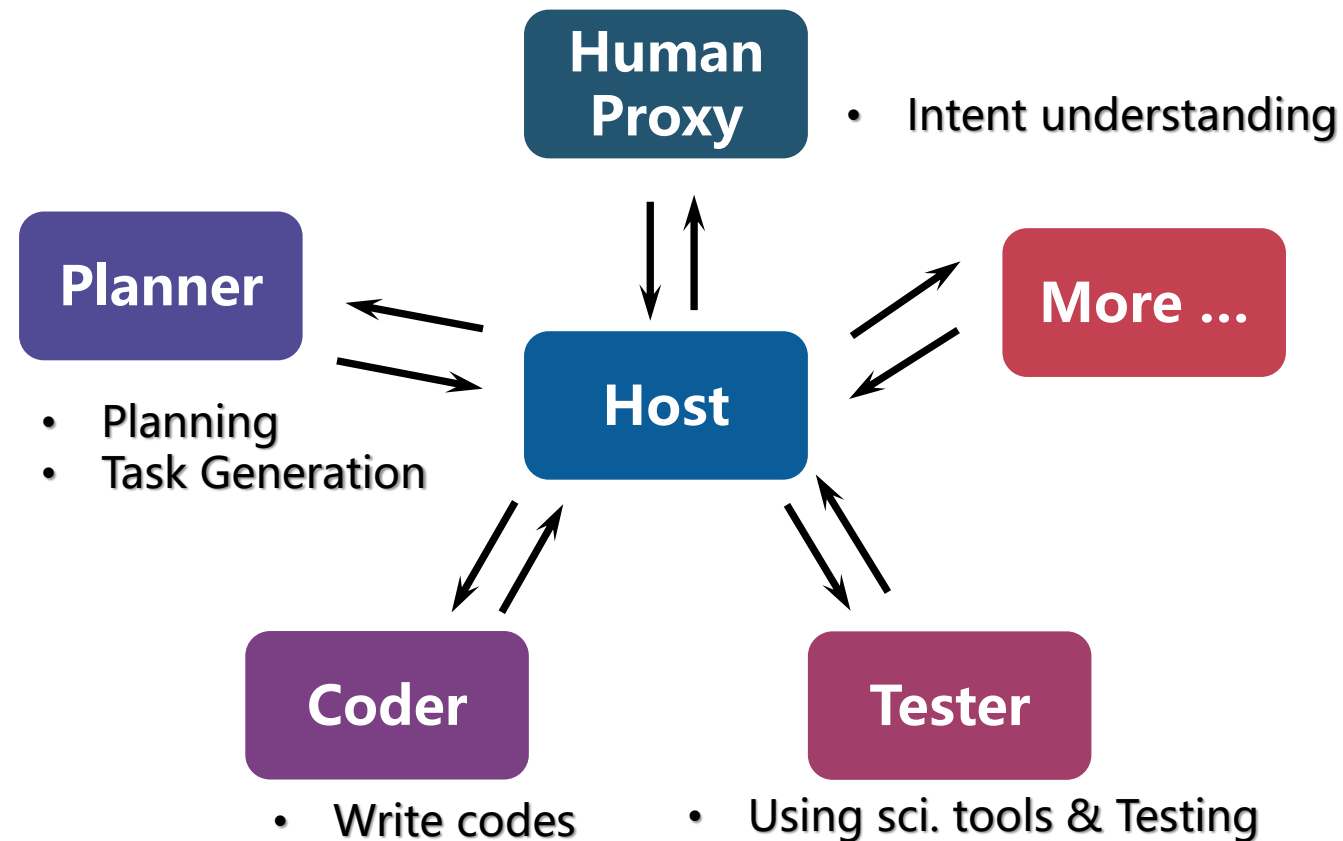
# Multi-agent collaborative system



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 introduced to allow humans intervene at any time.

Autogen: arXiv.2308.08155)



# WebUI of Dr.Sai



Dr.Sai v1.0.0 is 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

The screenshot shows the Dr.Sai WebUI interface. Key components are highlighted and labeled:

- dialogue history**: A red box highlights the chat history on the left side of the interface.
- Modules**: A red box highlights the 'Modules' dropdown menu, which includes options like 'BESIII AI', 'Image generation', 'Personal assistant', and 'Chat'.
- user query**: A red box highlights the input field where a user enters a question.
- AI agent reply**: A red box highlights the AI's response to the user's query.
- user settings**: A red box highlights the user settings panel, including the user's email, dark mode toggle, and logout button.
- Settings panel**: A green box highlights the settings panel, which includes options for the model (e.g., 'openai/gpt-4o'), name, code interpreter, file search, arxiv search, editor, temperature, and top p.
- LLM models (module=Chat)**: A green box highlights the list of LLM models available for selection.
- Upload files (png, pdf...)**: A green box highlights the file upload button in the chat input area.
- Enable functions**: A blue box highlights the 'Code Interpreter', 'File Search', and 'Arxiv\_search' options in the settings panel.
- entropy of reply**: A blue box highlights the 'Temperature' slider in the settings panel.

Additional labels include 'Introduction of Dr. Sai' pointing to a 'Readme' link and 'WebUI' at the bottom center.

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", 300, 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

The screenshot shows a chat message with a code block. The code block is titled "code type" and contains the following text:

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

Yellow arrows point from the labels "code type", "backticks", "file name", and "command" to their respective parts in the code block.

Below the code block, the chat message says: "(auto select "Tester")"

A user icon is followed by the text: "There are 1 code blocks in the message. The output are:"

The output is displayed as follows:

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

The output line is underlined in red.

Below the output, the chat message says: "The generated files are:"

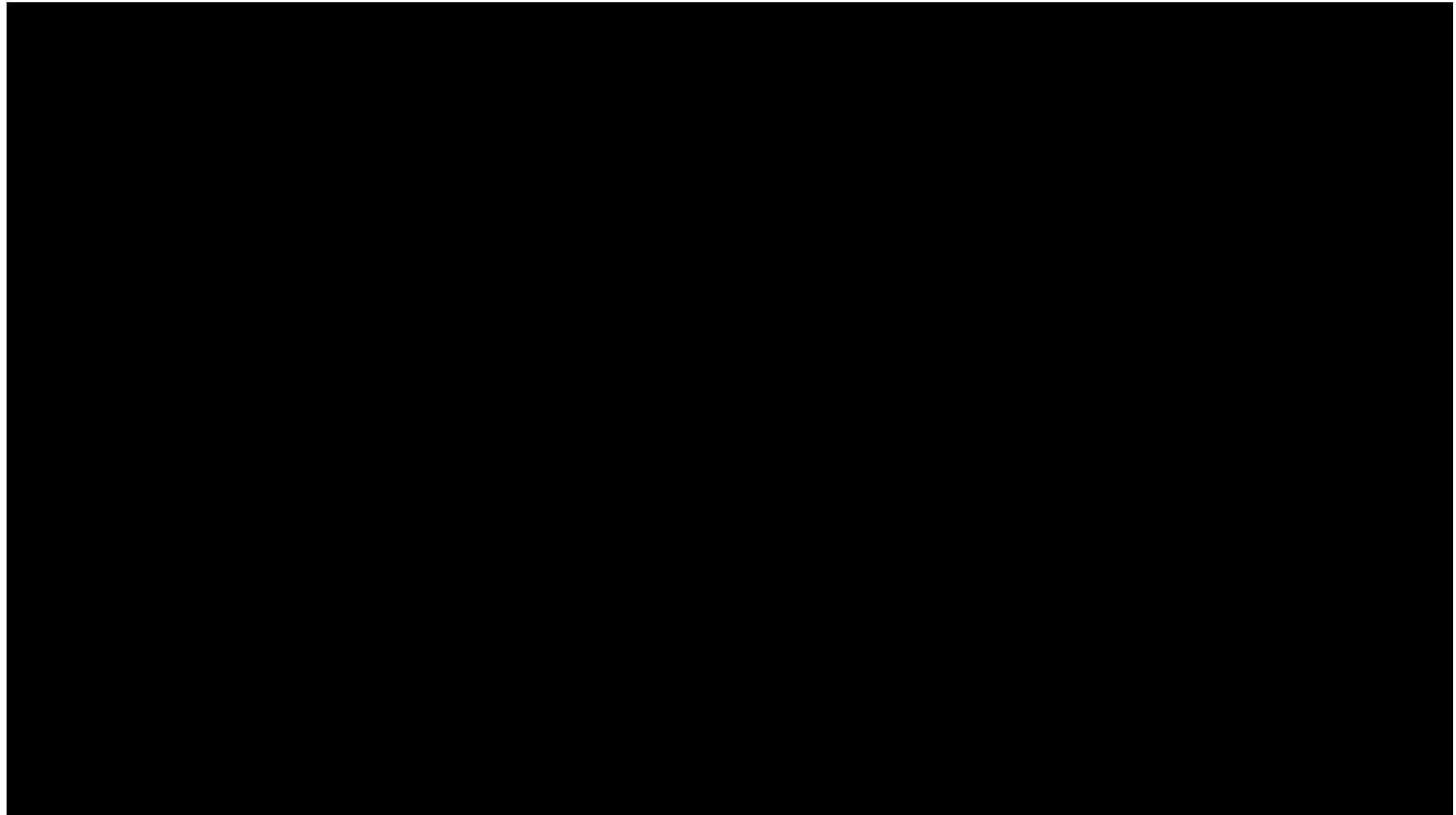
```
<pic: None >  
<pdf: None >
```

At the bottom of the chat message, there are three icons: a copy icon, a thumbs up icon, and a thumbs down icon.

# Example of Multi-Agent System



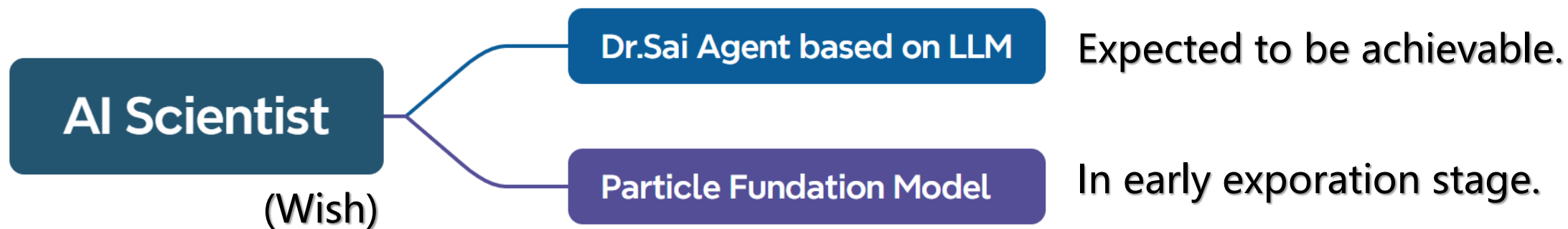
- **Host** agent select suitable speaker
- **Coder** generate domain code
- **Tester** Call BOSS (BESIII Offline Software System) to excute
- Draw a signal histogram



Example



# Summary and Outlook



- A customized LLM (**HEP•Xiwu**) is developed, enhancing domain-specific capabilities.
- Single agent equipped with an LLM, **memory, sensors and executors** has been developed, demonstrating its ability to handle some HEP tasks.
- **Dr. Sai** – a multi-agent collaborative system has been launched (v1.0), preliminarily demonstrating its ability to automate physics analysis processes.
- Dr. Sai is a new system, which still needs further improvement.
- Dr. Sai processes BESIII data through BOSS, it cannot exhibit “emergence” of LLM; there is a need to develop a **Large Model** for scientific data (or **Foundation Model**).

# Some links



HepAI platform: <https://ai.ihep.ac.cn>

Dr.Sai Agents: <https://drsai.ihep.ac.cn>

HEP·Xiwu LLM: <https://github.com/zhangzhengde0225/Xiwu>

- **Welcome to discuss any interests related to AI for HEP !**
  - Zhengde Zhang (zdzhang@ihep.ac.cn)
  - Computing Center, Institute of High Energy Physics, CAS, Beijing



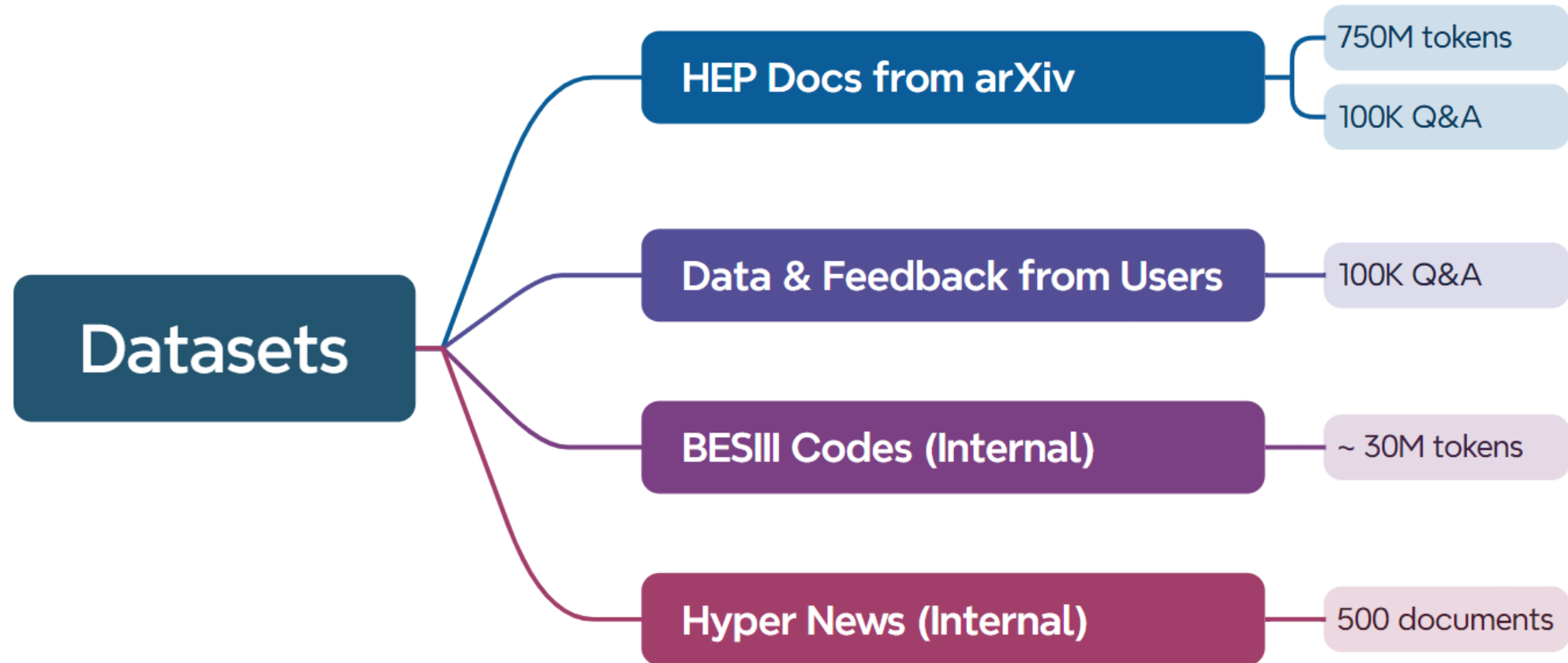
~~Thank you for listening~~



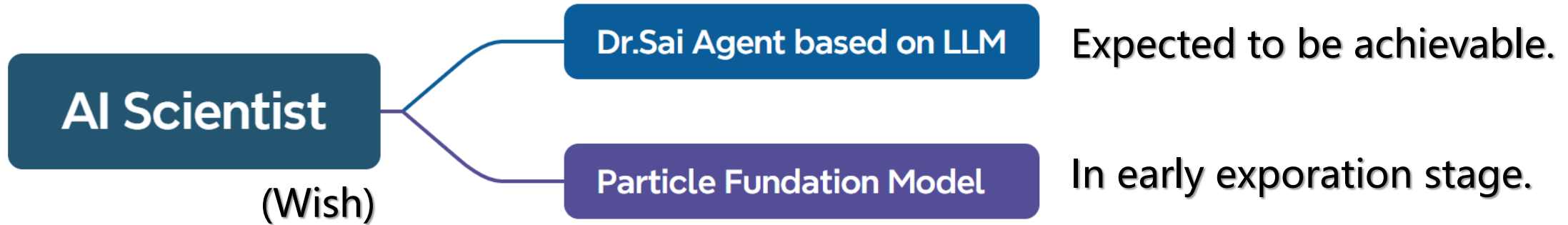


# Backup

# Datasets



# Summary and Outlook

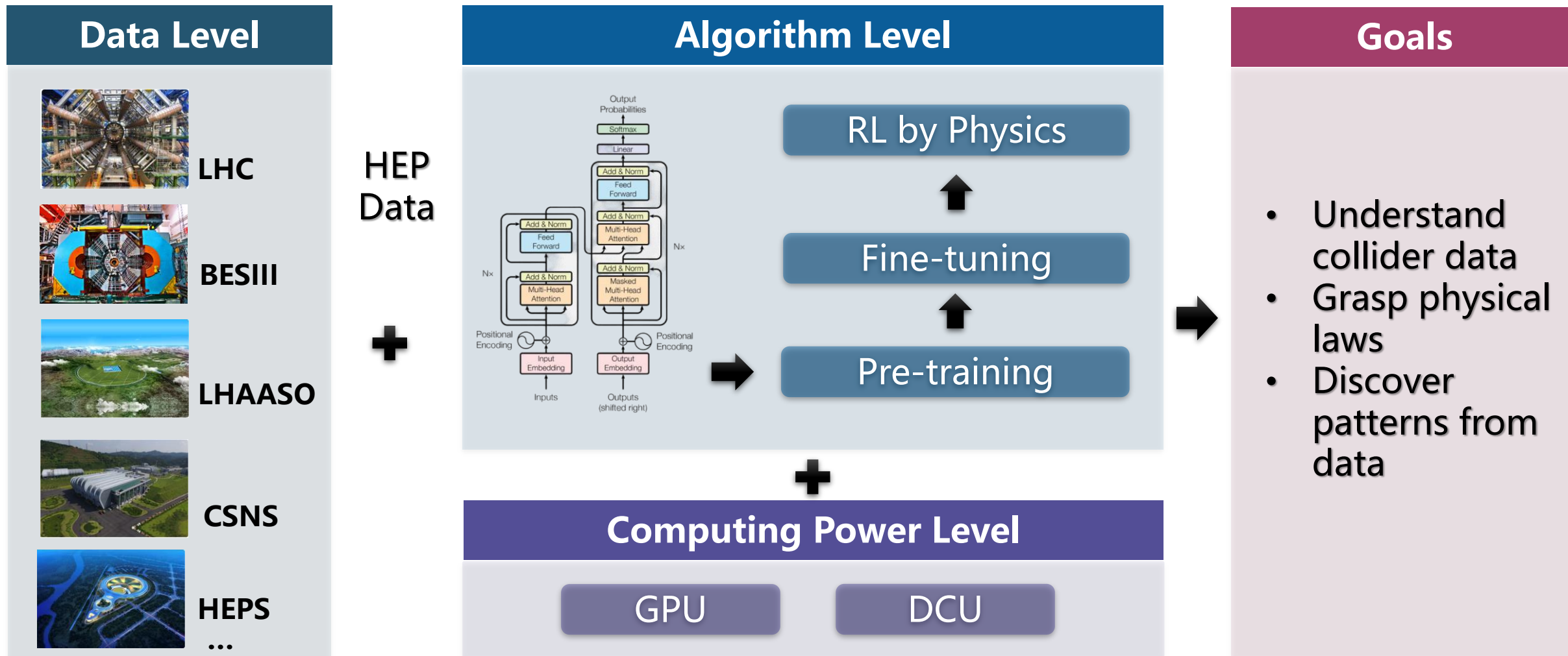


- One of the main criteria for categorizing scientific research paradigms is the use of **research tools**. The adoption of new tools inevitably enhances research efficiency and stimulates the generation of new findings.
- We aim to leverage advanced technologies to **drive fundamental research, accelerate scientific progress, and benefit human society**.
- Gazing at the stars, while keeping our feet on the ground (rooted in conventional methods, with an eye on innovation).

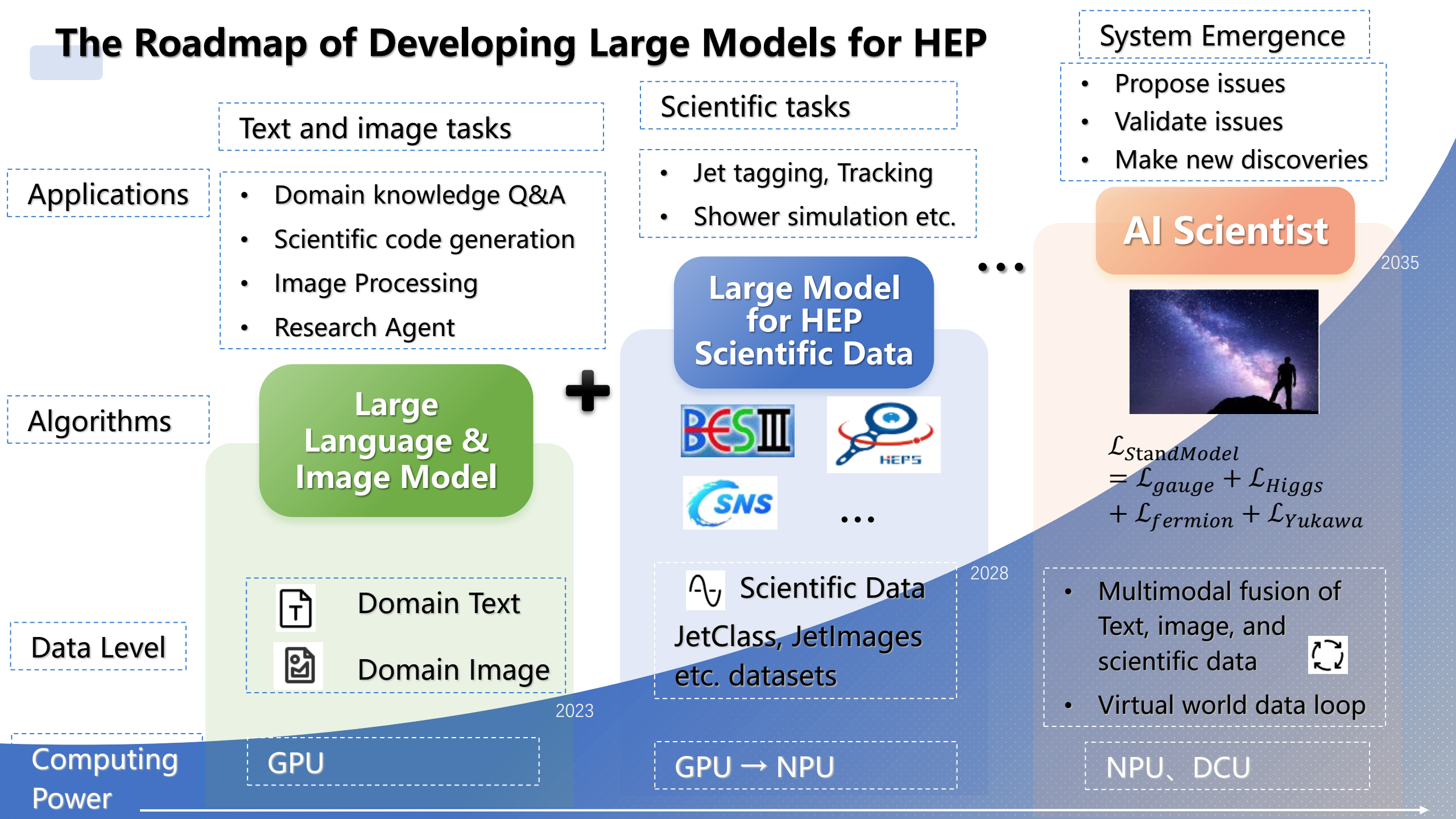
# Outlook - Particle Foundation Model?



Trained on a large amount of particle physics data, capable of handling multiple tasks simultaneously, and is expected to exhibit emergent phenomena in physics-related tasks.



# The Roadmap of Developing Large Models for HEP



Applications

Text and image tasks

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

Algorithms

Large Language & Image Model



Domain Text



Domain Image

Data Level

Computing Power

GPU

Scientific tasks

- Jet tagging, Tracking
- Shower simulation etc.

Large Model for HEP Scientific Data



...

Scientific Data  
JetClass, JetImages  
etc. datasets

2028

GPU → NPU

System Emergence


- Propose issues
- Validate issues
- Make new discoveries

AI Scientist



2035

$$\begin{aligned} \mathcal{L}_{\text{StandModel}} &= \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{Higgs}} \\ &+ \mathcal{L}_{\text{fermion}} + \mathcal{L}_{\text{Yukawa}} \end{aligned}$$

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

NPU, DCU



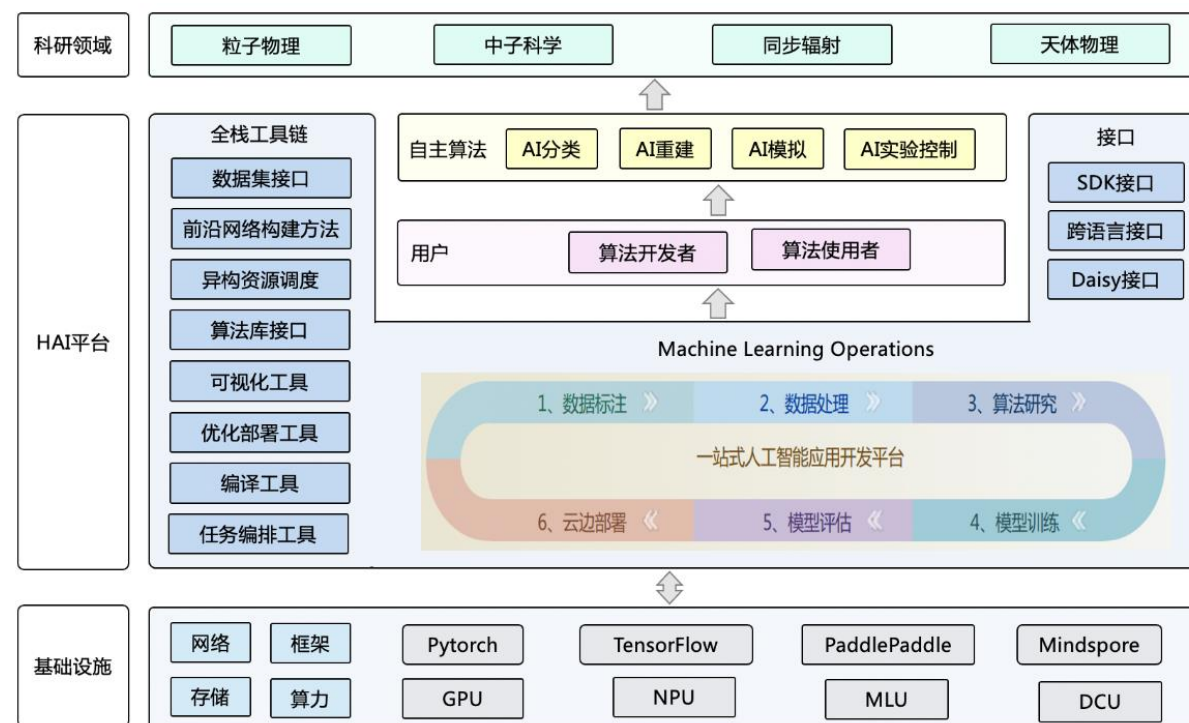
# HepAI Platform



- 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.

- HepAI Core Codes and Framework (50%)
- 10 AI algorithms.
- 4 AI datasets.
- Heterogeneous computing resources including GPU, NPU, and DCU.

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



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

The architecture of HepAI platform

# About US



## Project leaders



C. Z. Yuan



Z. D. Zhang

(Sorted by surname)

## Mentors



G. Li



K. Li



F. Z. Qi



Y. Y. Zhang



L. N. Zhao

## BESIII Members



P. Huang



J. K. Jiao



M. R. Li



Y. P. Liao



Z. J. Shang



B. L. Zhang

## non-BESIII Members



S. Y. Chen



J. F. Li



Q. Luo



Y. H. Pang



Q. R. Sun



H. F. Wang



D. B. Xiong



R. P. Yang



F. Y. Jiang

# Scientific Paradigm is Shifting



HEP is currently transitioning from a data paradigm to an intelligent paradigm.

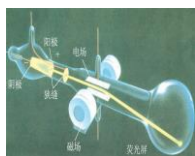
- The main criterion for dividing scientific paradigms is the use of research tools.

## Experiment

Describes and records natural phenomena. experimentation or empirical induction



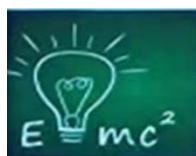
Leaning Tower of Pisa



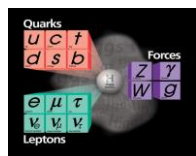
Cathode Ray

## Theoretic

Simplifies natural phenomena into mathematical models



Relativity



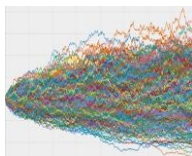
Standard Model

## Computing

Uses computers to simulate expr. of complex phenomena



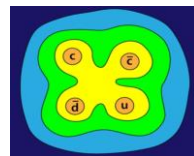
Nuclear Test



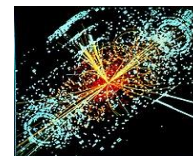
Monte Carlo Sim.

## Data

Analyzes big data to draw conclusions.



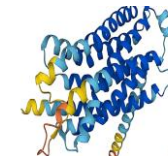
Zc(3900)



Higgs

## Intelligent

Uses AI to train and optimize uncertain systems, discovering patterns in data.



Protein



Agent

1<sup>st</sup> Parad.

2<sup>nd</sup> Parad.

3<sup>rd</sup> Parad.

4<sup>th</sup> Parad. (Current)

5<sup>th</sup> Parad.

- Different paradigms are **not substitutive** but complementary, working together synergistically.