# LLM-based physics analysis assistant at BESIII - ‘Dr.Sai’

Yipu Liao (廖一朴)
Inistitute of High Energy Physics, CAS, Beijing
On behalf of **Dr.Sai** working group

2025.7.17
第六届 粒子物理前沿研讨会
@ Jilin University, Changchun

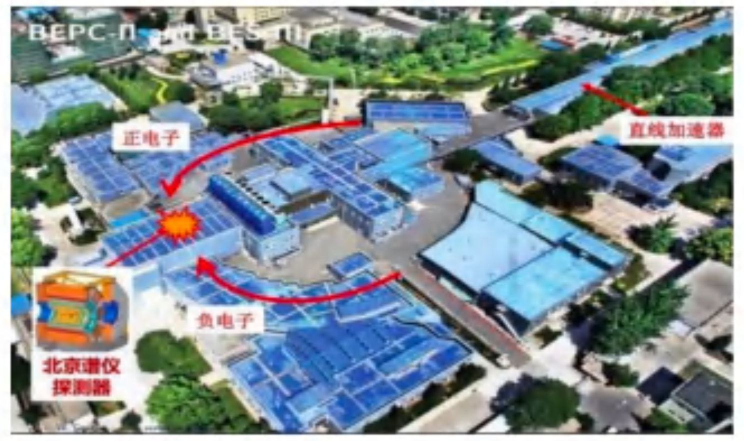
# Outline

1. Motivation
2. Dr.Sai project
3. Dr.Sai-BESIII
4. Prospect
5. Summary

# Challenges in HEP

- Data magnitude: **~10 PB**
- Growth rate: accelerated with new technology
  - BEPCII → BEPCII-U (May 17, 2025): **Luminosity increase ~300% at 2.35 GeV**
  - HEPS (2025): **~800 TB / day**

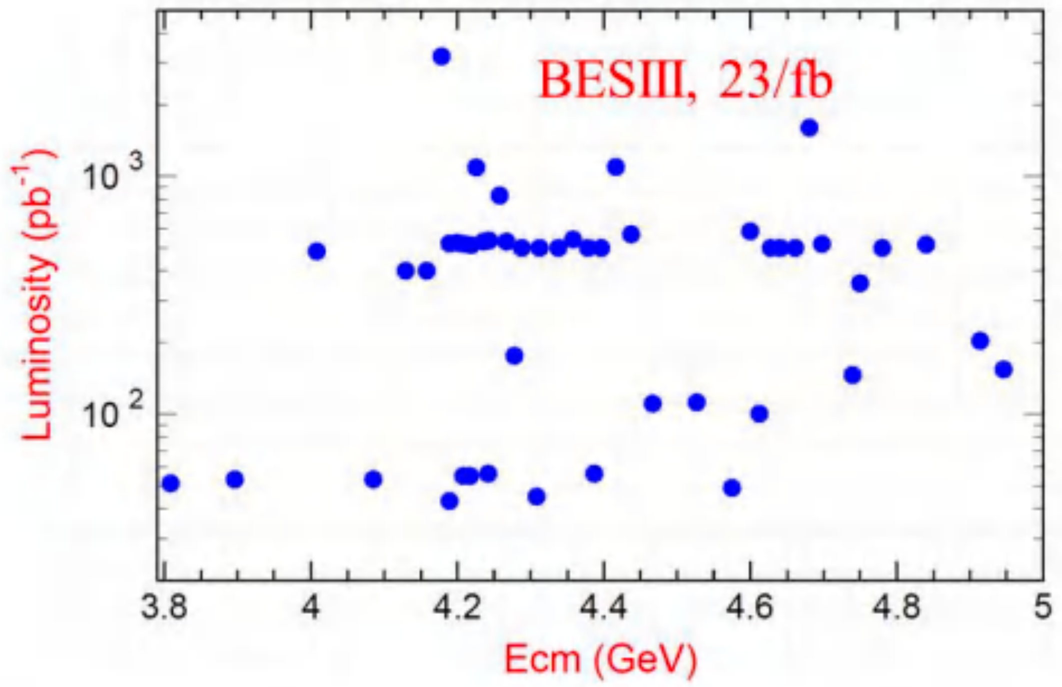
**Particle physics**

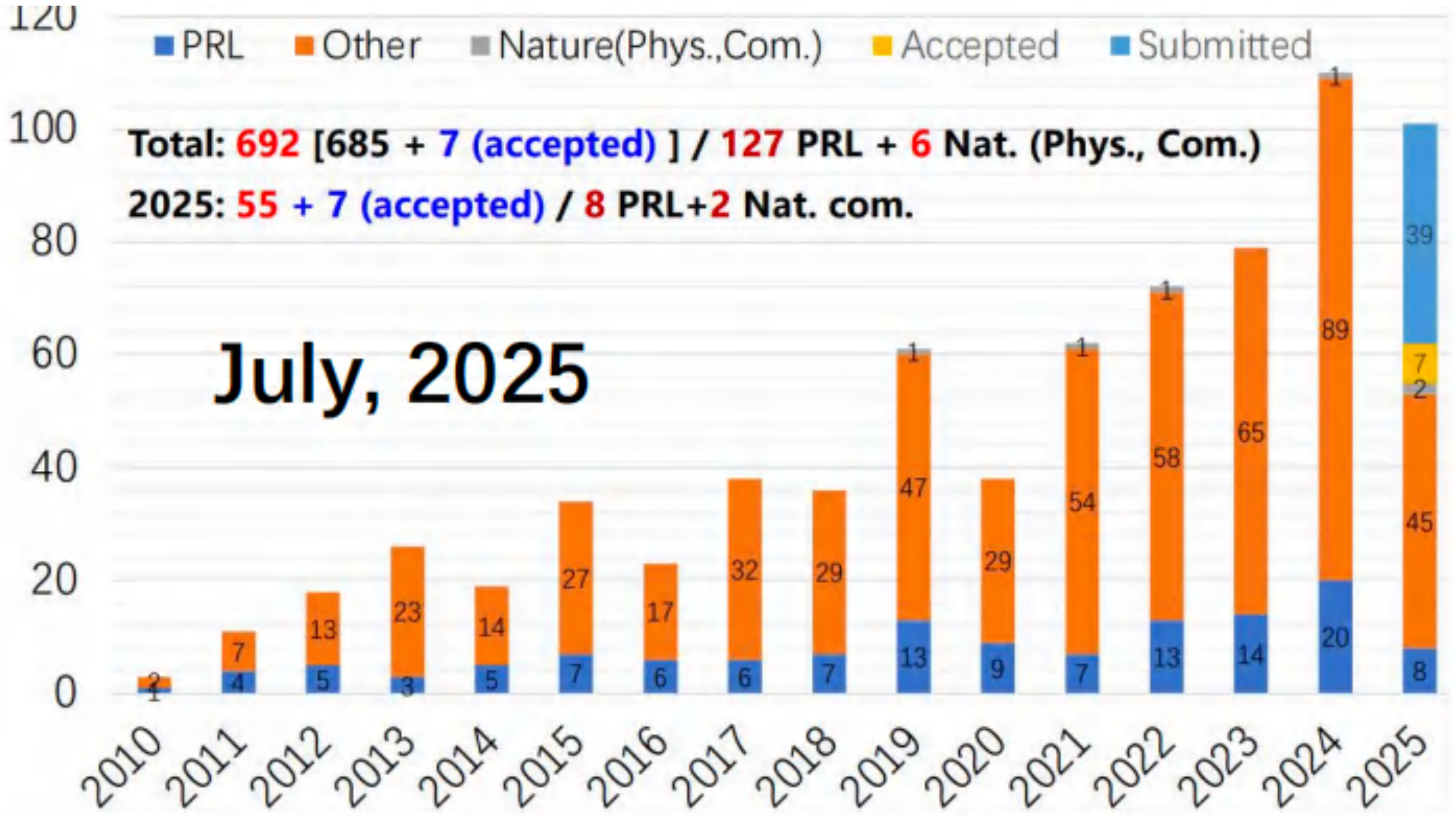


**Neutron science**

**Astrophysics**

**Photon science**





**Total: 692 [685 + 7 (accepted) ] / 127 PRL + 6 Nat. (Phys., Com.)**

**2025: 55 + 7 (accepted) / 8 PRL+2 Nat. com.**

July, 2025

- ~700 physics results with ~700 people in 16 years
  - One result normally took ~3 years
- HEP need a more efficient tool to investigate data

# Challenges in HEP

**Data process workflow at HEP experiment** (take Hadron collider as an example)

**Collision, MC generator**

- Accelerator control
- Initial-state-radiation
- Parton showering
- Hadronization
- NP-correction
- Pileup
- …

**Trigger, Simulation**

- Data acquisition
- Fast reconstruction
- Data input/output
- Online monitoring
- Detector geometry
- Detector noise
- Calibration
- Multi-scattering
- …

**Reconstruction**

- Track & vertex Finding and fitting
- Clusterization & reconstruction of jet
- Jet tagging
- Kinematic fit
- Detector calibration
- …

**Statistical analysis**

- Event selection
- Optimizations
- Background analysis
- Injection test
- Reweighting
- Correlation corrections
- …

**Extract physics variables**

- Systematic uncertainty
- Fitting
- Uncertainty propagation
- Radiation and VP corrections
- …

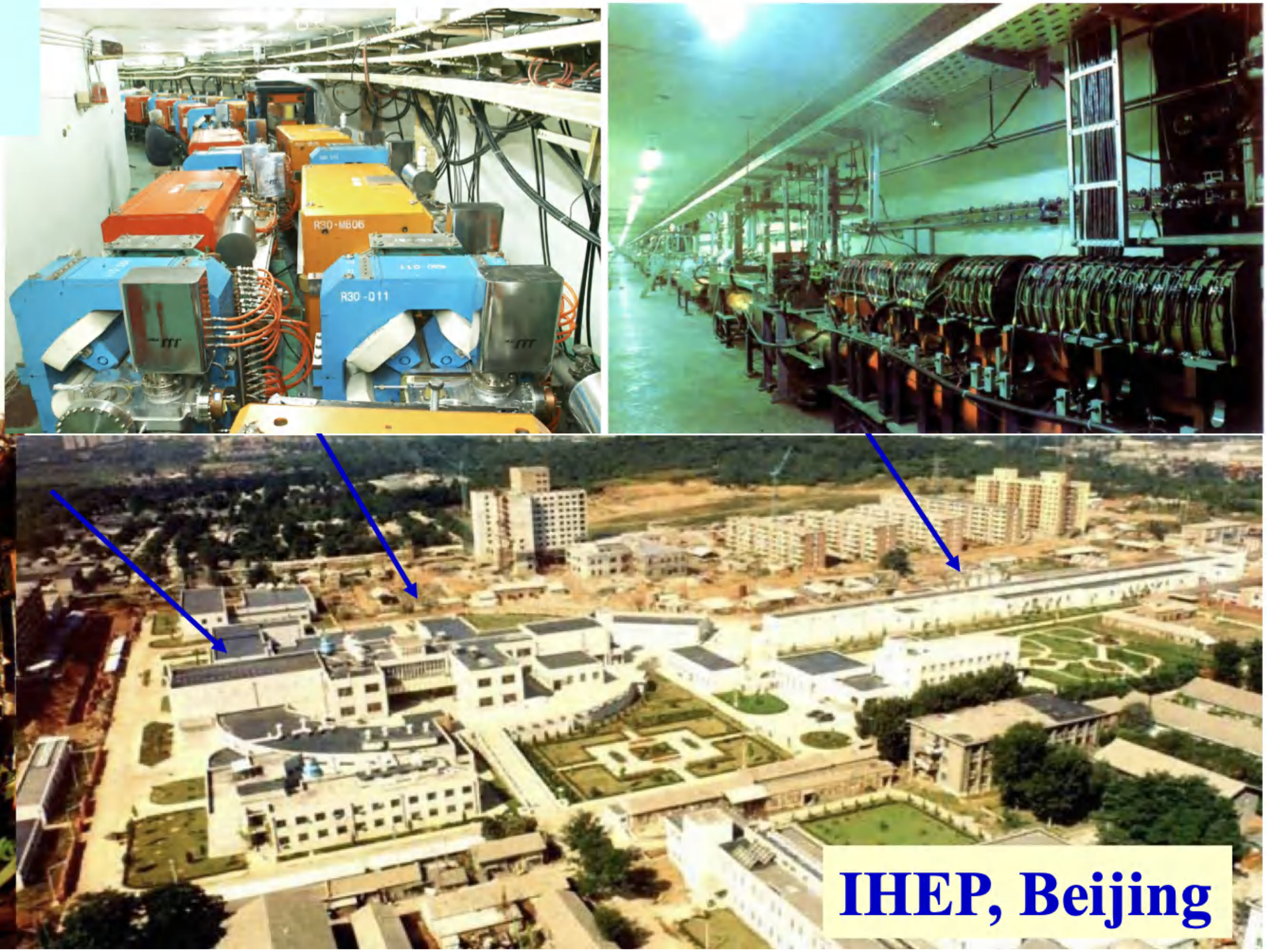
- **Too complicated, similar lines of code as windows/macOS**
- **One small task need: several people + several years!**

# Beijing Electron-Positron Collider (BEPC II)

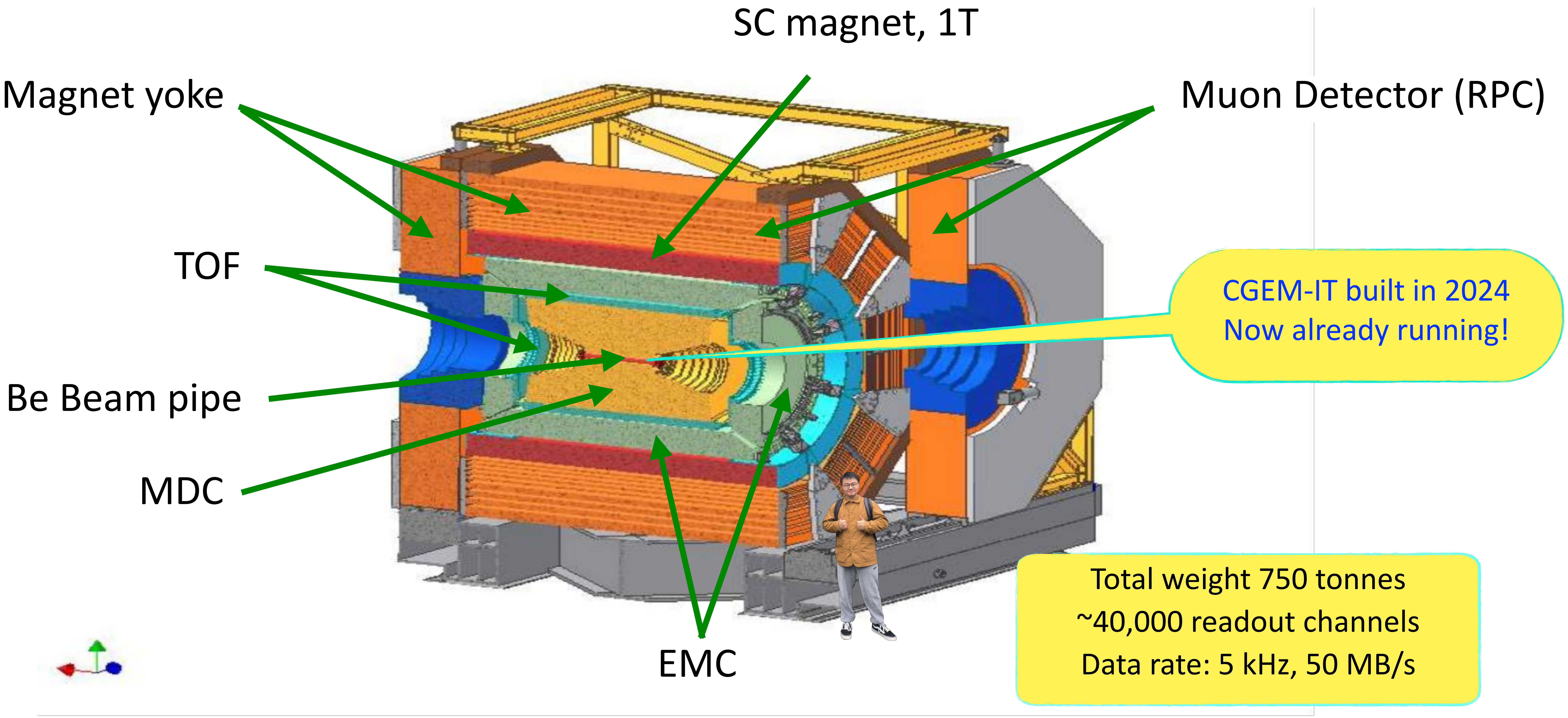
Ground breaking: 1984
CM energy : 2 - 5 GeV
Major upgrade: 2004
Energy upgrade: 2024

1989-2005 (BEPC): $L_{peak}=1.0\times10^{31}$ /cm$^2$s
2008-now (BEPCII): $L_{peak}=1.0\times10^{33}$/cm$^2$s (Apr. 5, 2016)

World unique $e^+e^-$ accelerator in τ-charm energy region



# The BESIII experiment



CGEM-IT built in 2024
Now already running!

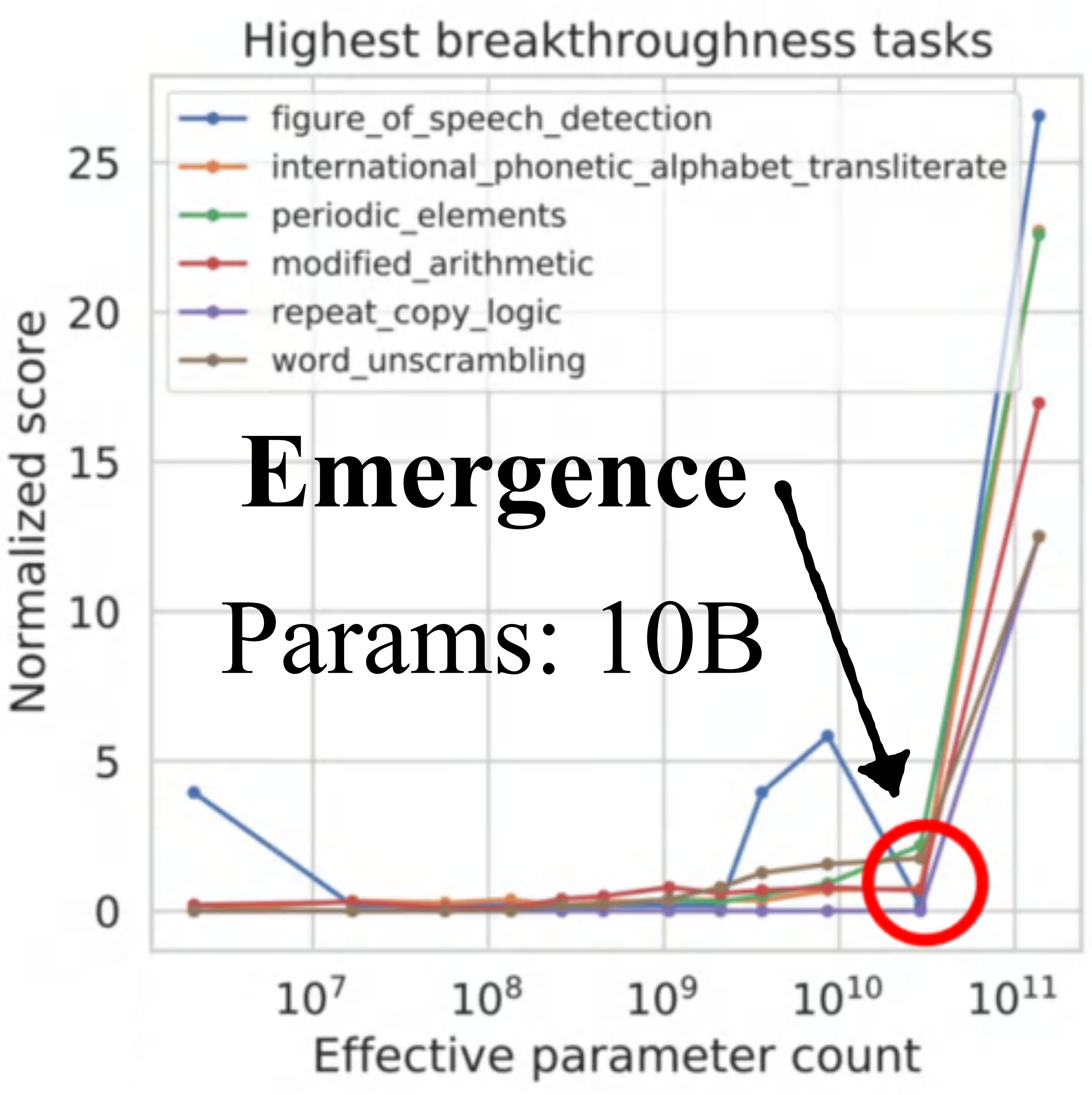
Total weight 750 tonnes
~40,000 readout channels
Data rate: 5 kHz, 50 MB/s

Has been in full operation since 2008, all sub-detectors are in good status!

**A lot of data to be analyzed → Use the more intelligent automatic workflow**

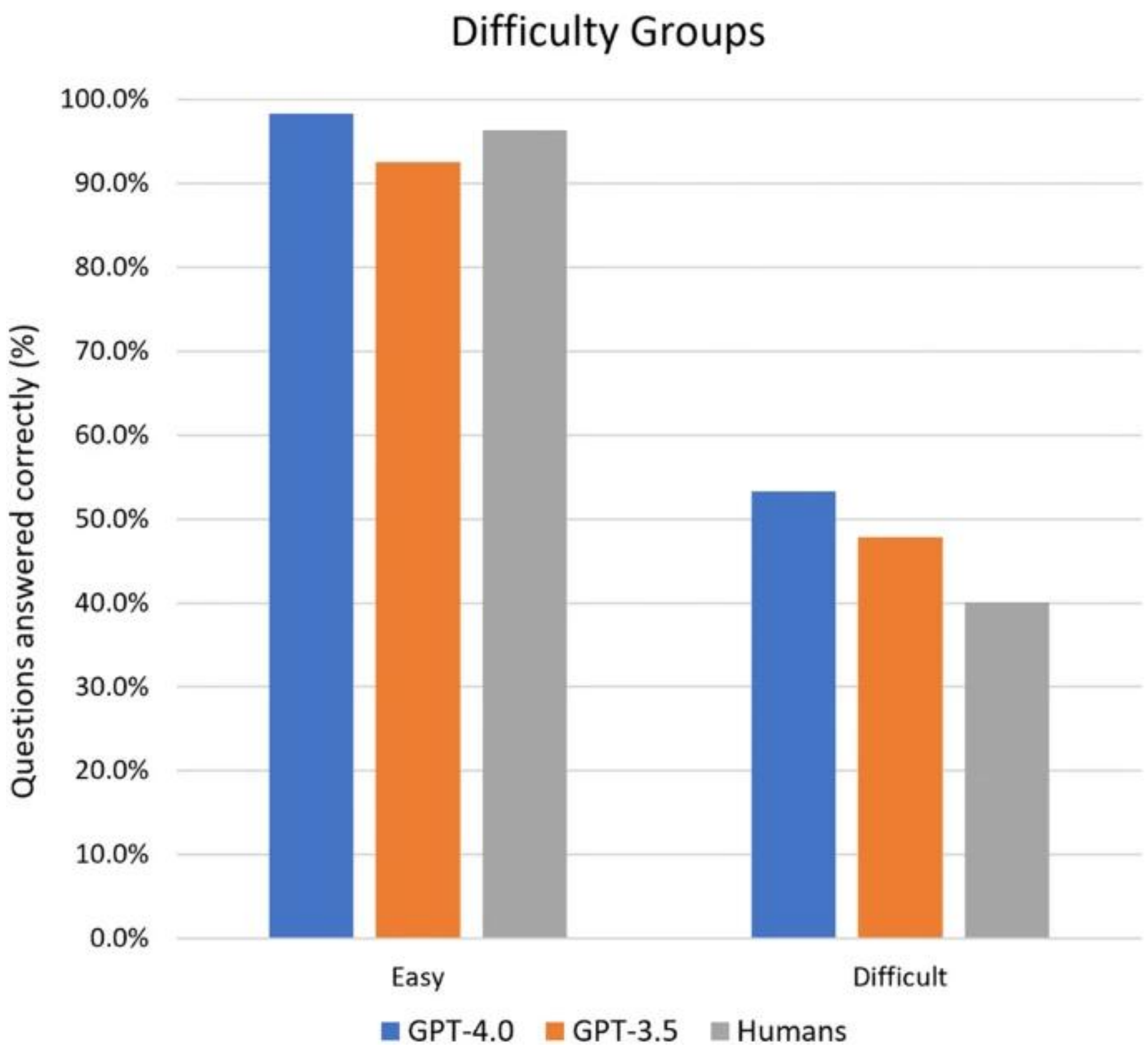
# What is Large Language Model (LLM)

- **Large language models (LLMs), normally build on Transformer architecture (Deep Learning) and based on the next-token prediction**
- **It has demonstrated impressive performance in text / code generation**
  - GPT4o, Gemini, LLaMa3 ...
  - Could be used for HEP studies
  - Game changer
- A foundation model (large, computing intensive) + fine tuning for each task individually (smaller data set)



**Emergence**

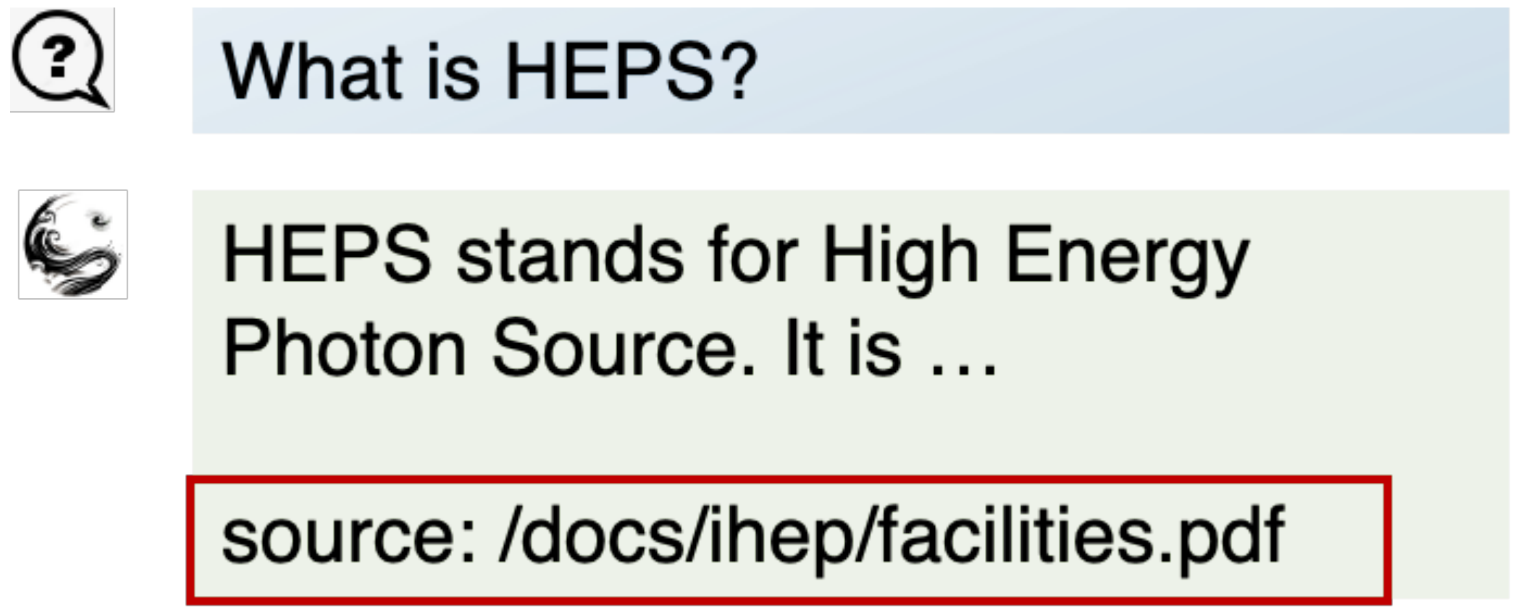
Params: 10B



Scientific Reports volume13, Article number: 18562 (2023)

# Retrieval Augment Generation (RAG)

- RAG is the most cheap & **promising** solution to reduce hallucinations (幻觉)
- Store private data so no need for retraining the model

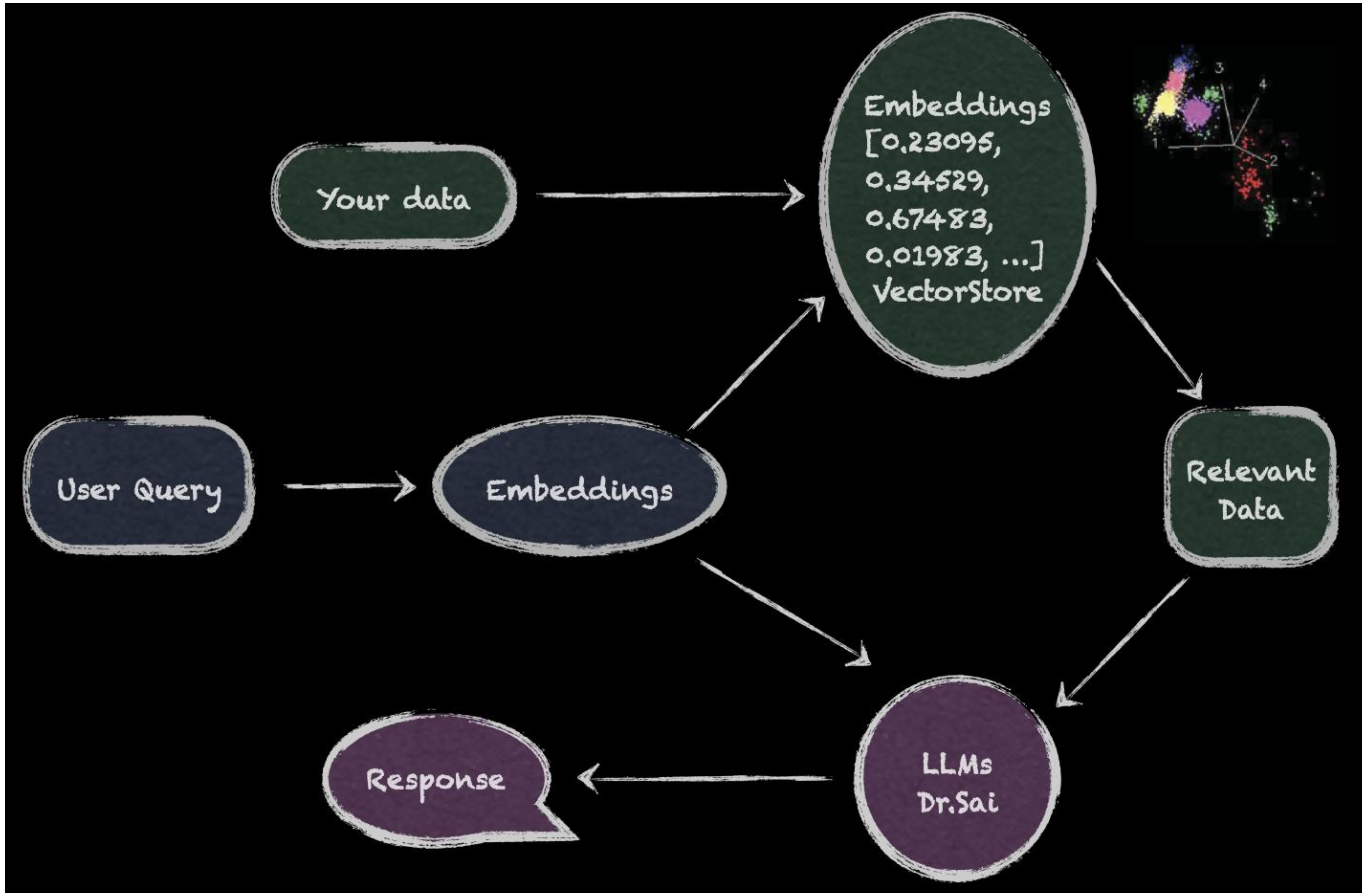


What is HEPS?

HEPS stands for High Energy Photon Source. It is …

source: /docs/ihep/facilities.pdf

☐ Provide answers based on user intent and information
☐ Provide **sources** of information

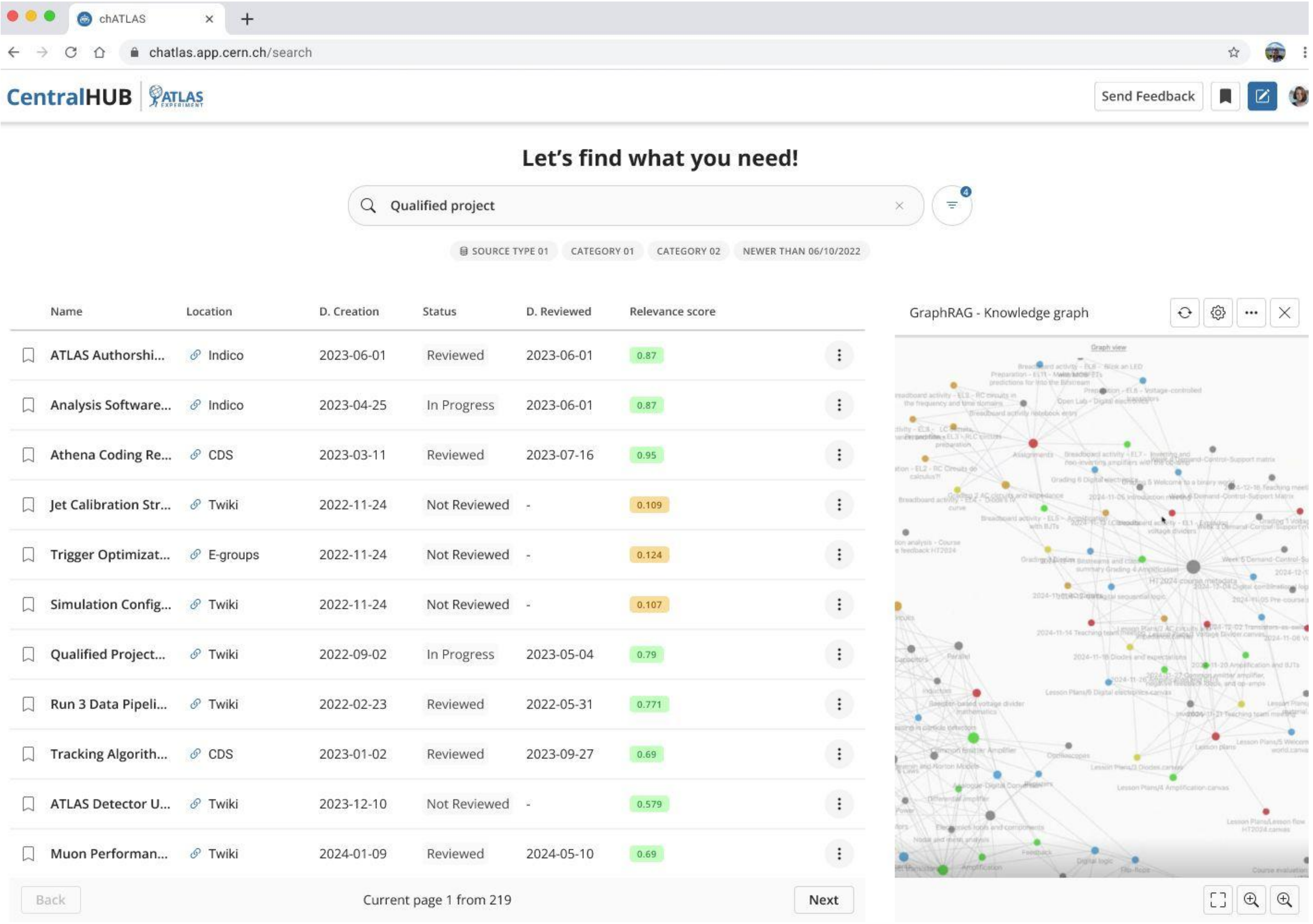


# Some applications based on RAG

** Focus on HEP*

chATLAS

- An AI assistant of the ATLAS Collaboration with decades of institutional documentation (wiki, docs, indico, …)
- Focus on the search system and questions about ATLAS



Let's find what you need!

Qualified project

SOURCE TYPE 01 · CATEGORY 01 · CATEGORY 02 · NEWER THAN 06/10/2022

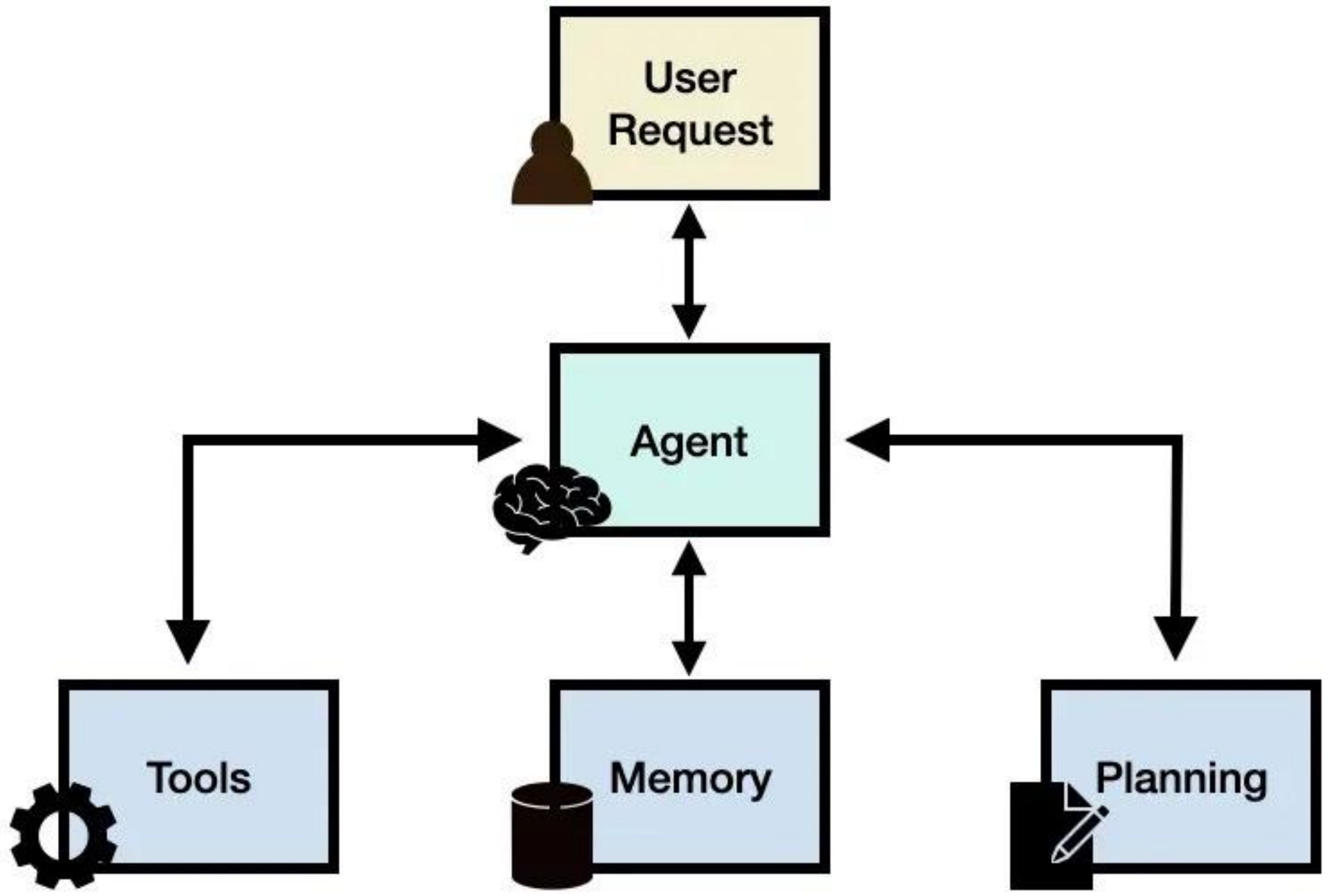
| Name | Location | D. Creation | Status | D. Reviewed | Relevance score |
|---|---|---|---|---|---|
| ATLAS Authorshi... | Indico | 2023-06-01 | Reviewed | 2023-06-01 | 0.87 |
| Analysis Software... | Indico | 2023-04-25 | In Progress | 2023-06-01 | 0.87 |
| Athena Coding Re... | CDS | 2023-03-11 | Reviewed | 2023-07-16 | 0.95 |
| Jet Calibration Str... | Twiki | 2022-11-24 | Not Reviewed | - | 0.109 |
| Trigger Optimizat... | E-groups | 2022-11-24 | Not Reviewed | - | 0.124 |
| Simulation Config... | Twiki | 2022-11-24 | Not Reviewed | - | 0.107 |
| Qualified Project... | Twiki | 2022-09-02 | In Progress | 2023-05-04 | 0.79 |
| Run 3 Data Pipeli... | Twiki | 2022-02-23 | Reviewed | 2022-05-31 | 0.771 |
| Tracking Algorith... | CDS | 2023-01-02 | Reviewed | 2023-09-27 | 0.69 |
| ATLAS Detector U... | Twiki | 2023-12-10 | Not Reviewed | - | 0.579 |
| Muon Performan... | Twiki | 2024-01-09 | Reviewed | 2024-05-10 | 0.69 |

Current page 1 from 219

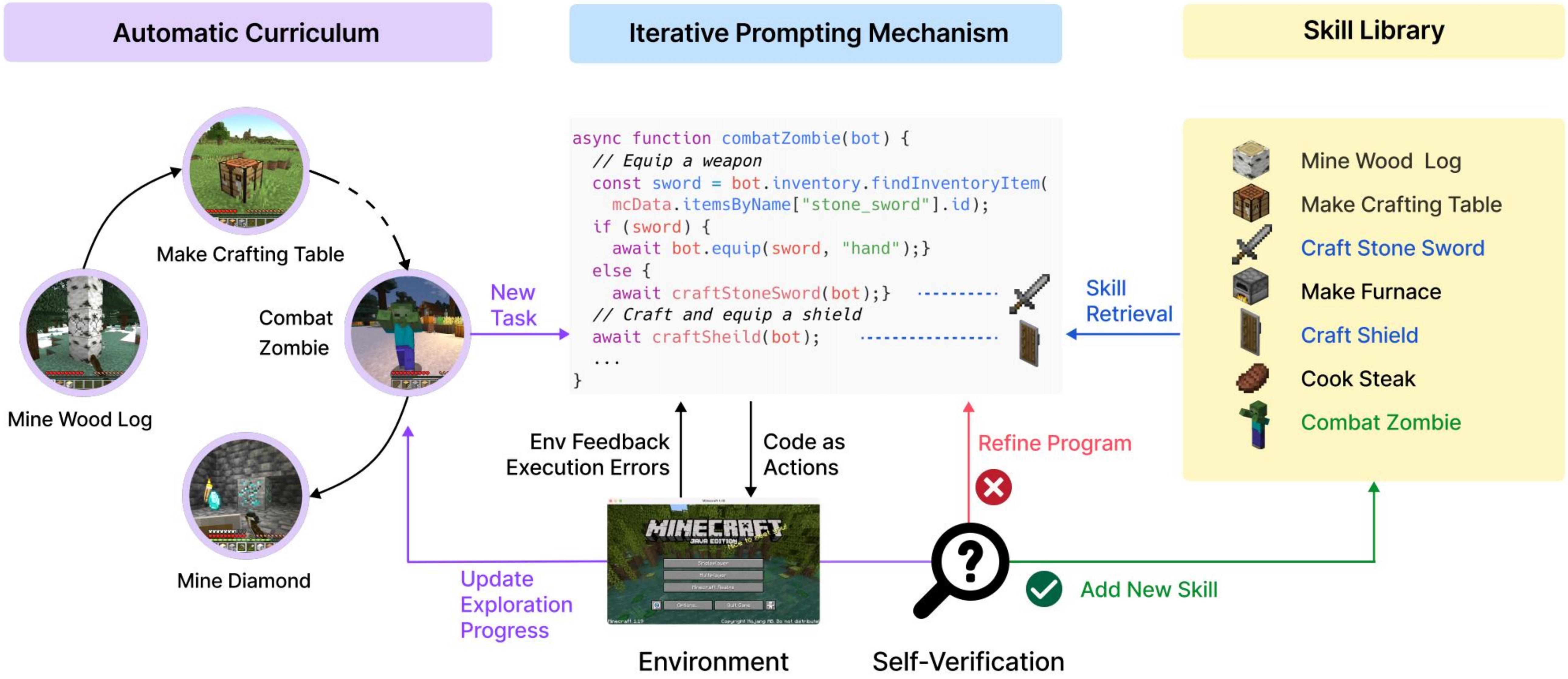
# Agent(ic): LLM with tools

A LLM agent framework consists of:

- User request: a user question or request
- **Agent: a LLM-based agent core**
- **Tools: manage tools except for text/ image reply; more professional tools to finish specific tasks**
- Sensor: text/video/image interface
- Planning: choose method to reply
- Memory: manage the old behavior and reply history



* Figures cited

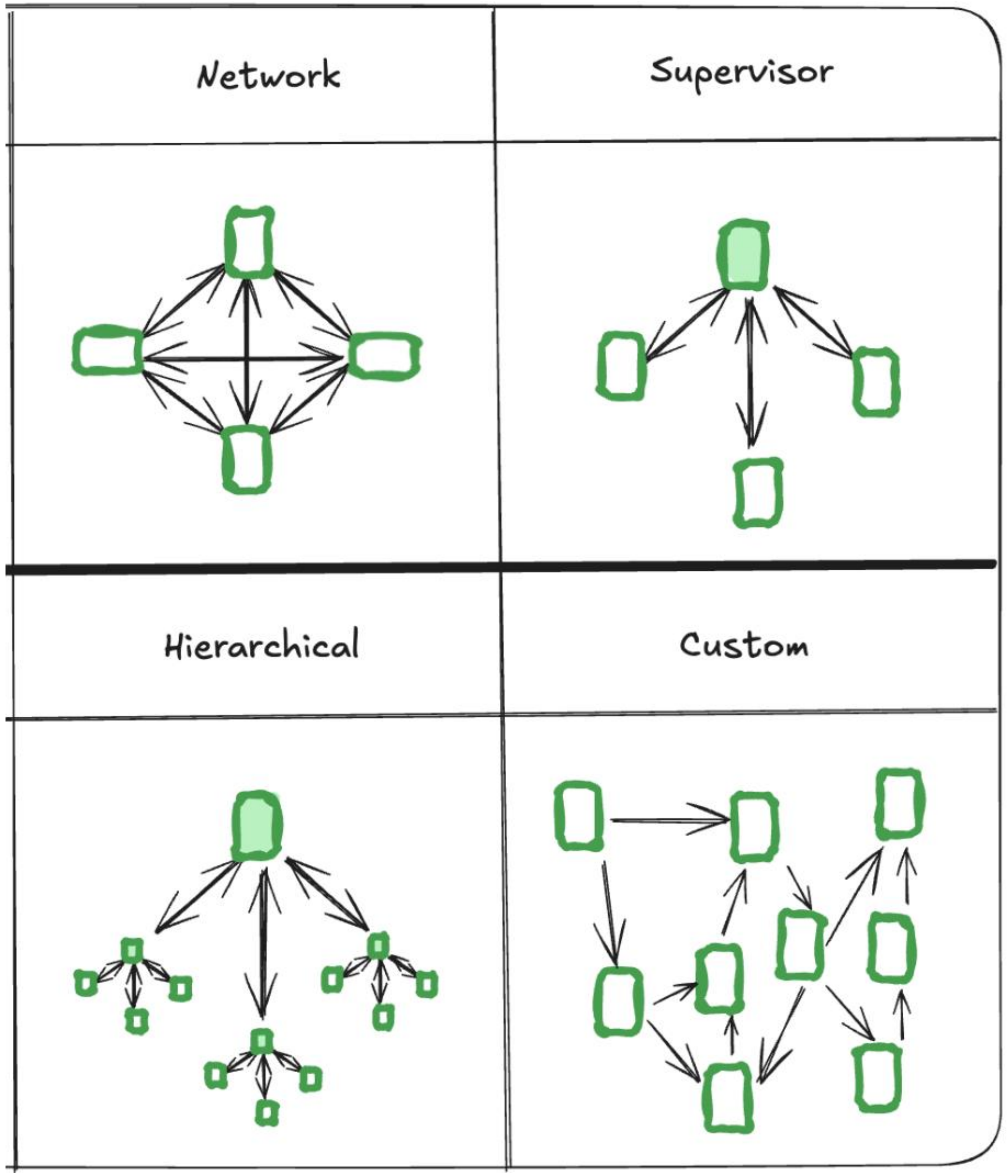


```
async function combatZombie(bot) {
  // Equip a weapon
  const sword = bot.inventory.findInventoryItem(
    mcData.itemsByName["stone_sword"].id);
  if (sword) {
    await bot.equip(sword, "hand");}
  else {
    await craftStoneSword(bot);}
  // Craft and equip a shield
  await craftSheild(bot);
  ...
}
```

Figure 2: VOYAGER consists of three key components: an automatic curriculum for open-ended exploration, a skill library for increasingly complex behaviors, and an iterative prompting mechanism that uses code as action space.

* VOYAGER: an agent of Minecraft

# Multi-Agent System (MAS)

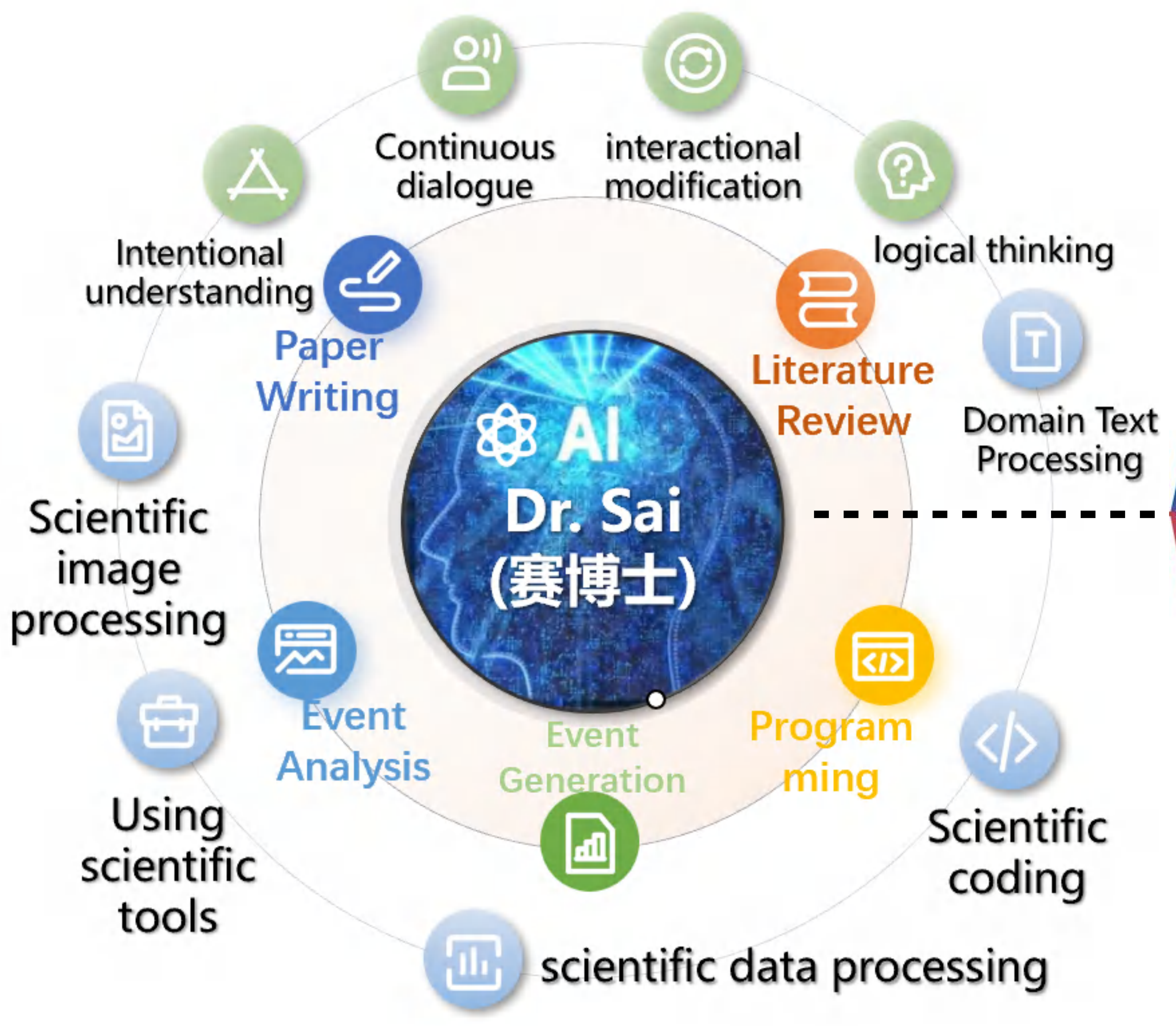


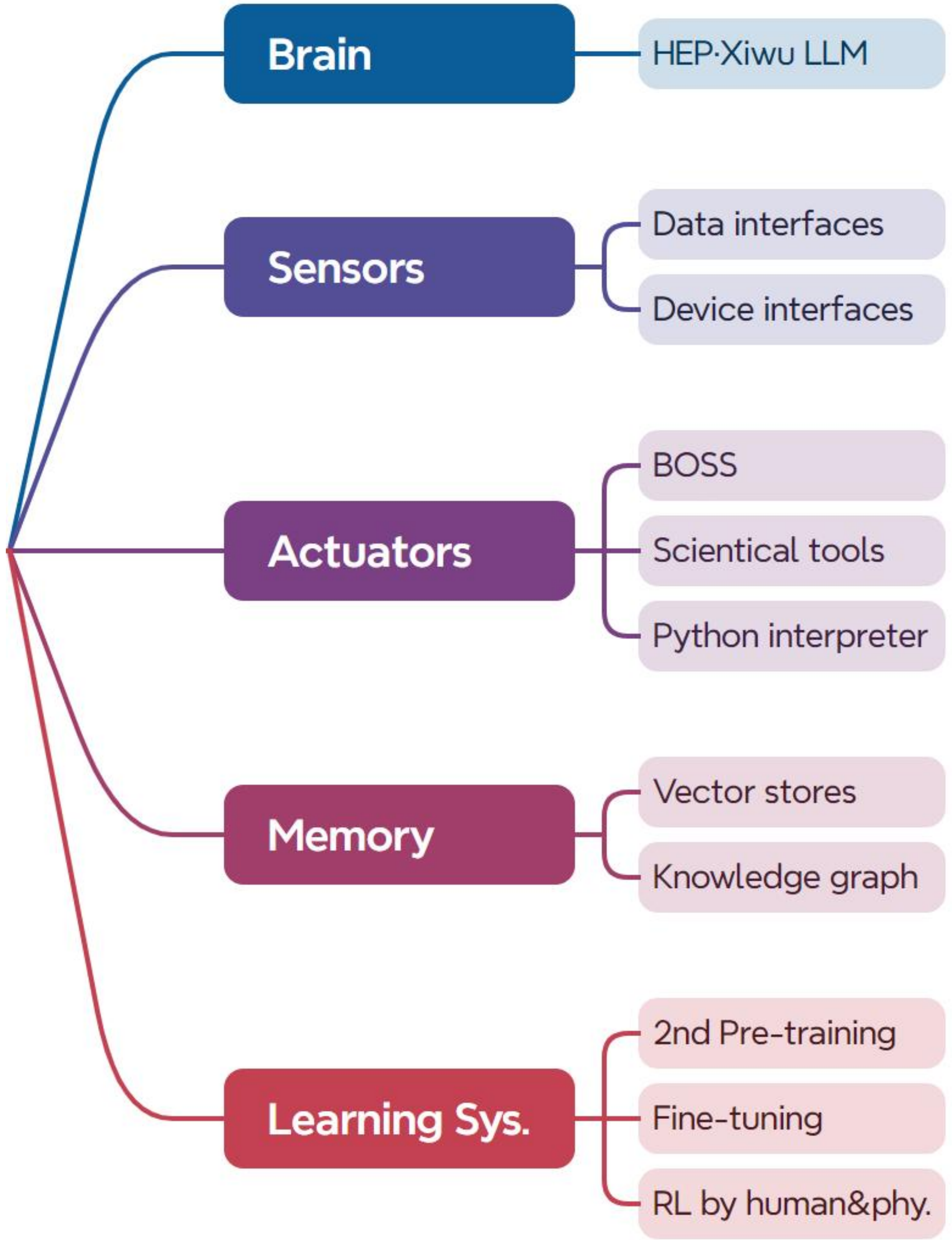
For different types of tasks, agents will have different prompts and tools, so a multi-agent system (MAS) is needed.

Some architectures of MAS:

- **Round-table**: fixed-order speaking
- **Network**: free connection among agents
- **Supervisor**: each agent is connected with a supervising agent, who decides to call others
- **Hierarchical**: Supervising agent as intermediate layer
- **Custom**: each agent talks with designated agents
- **Talk-reasoner**: Supervising agent interact with human

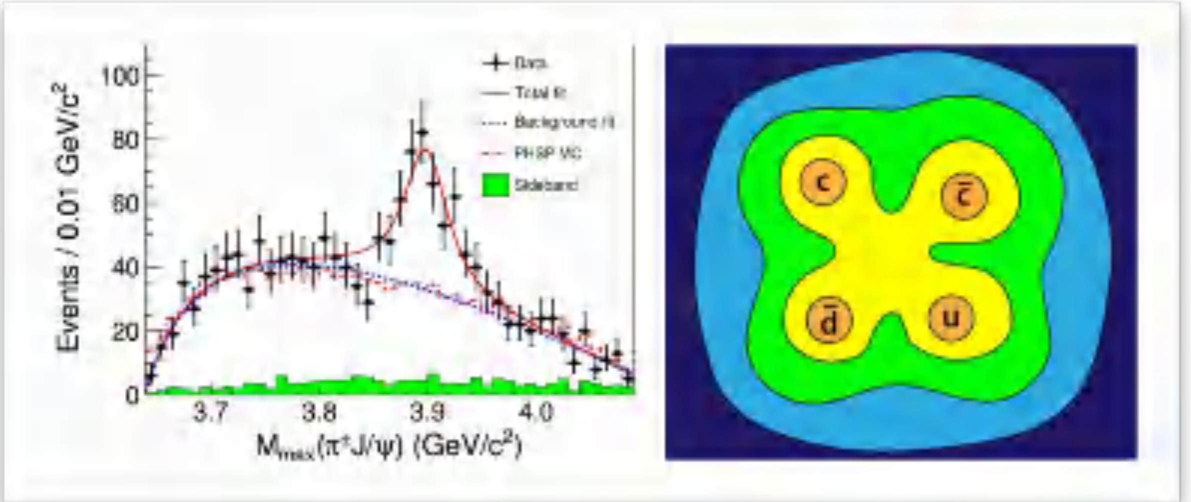
# Overview of Dr.Sai project





- **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.

BESIII

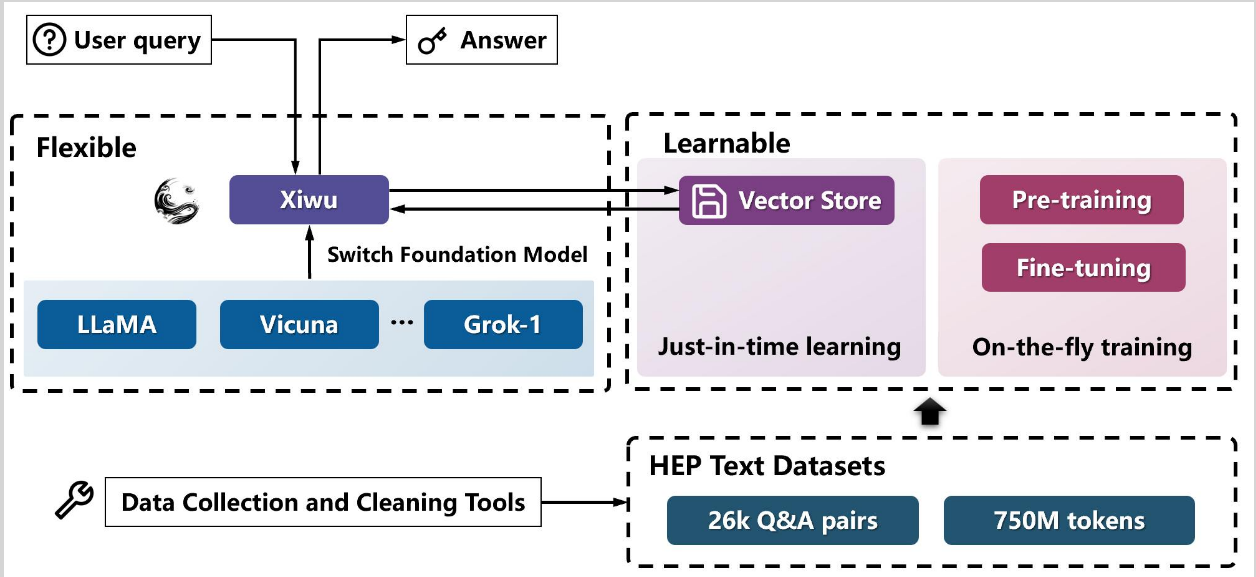


JUNO, HEPS, LHAASO

**Tasks & Required capabilities** **Components** **Applications**

# Dr.Sai single agent

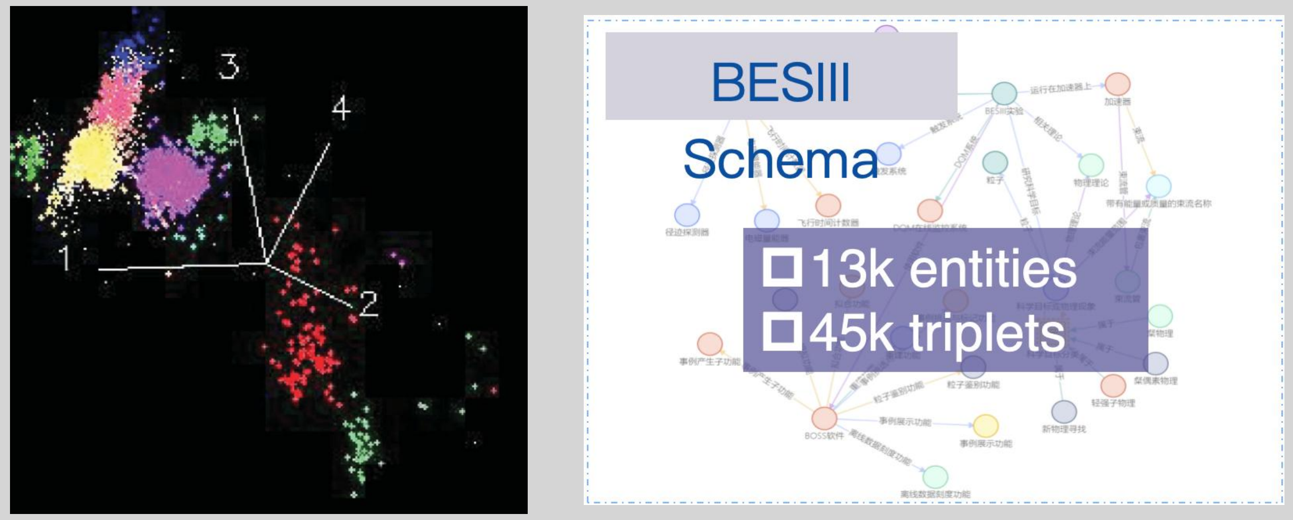
## Brain



HEP Xiwu LLM

- Based on Llama3-8B, 70B, and DeepSeek
- Trained on A800 and DCU from Sugon
- Enhanced domain capabilities

## Memory

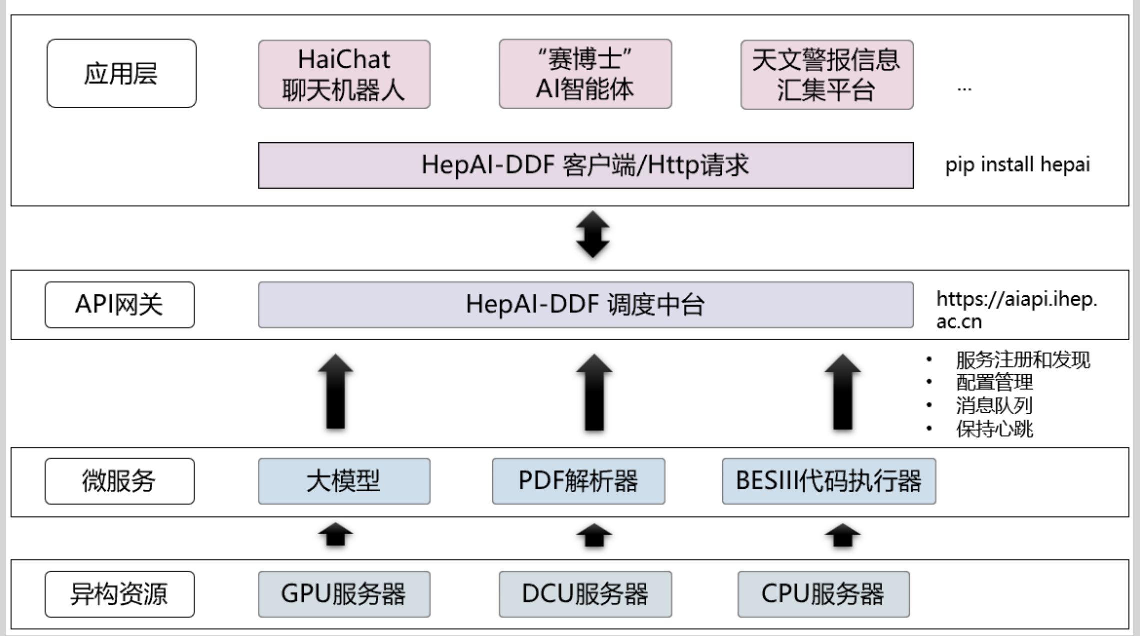


Vector store

Knowledge graph

- Vector store
  - BAAI BGE-M3
  - Llama-Index
- Knowledge graph
  - LightRAG

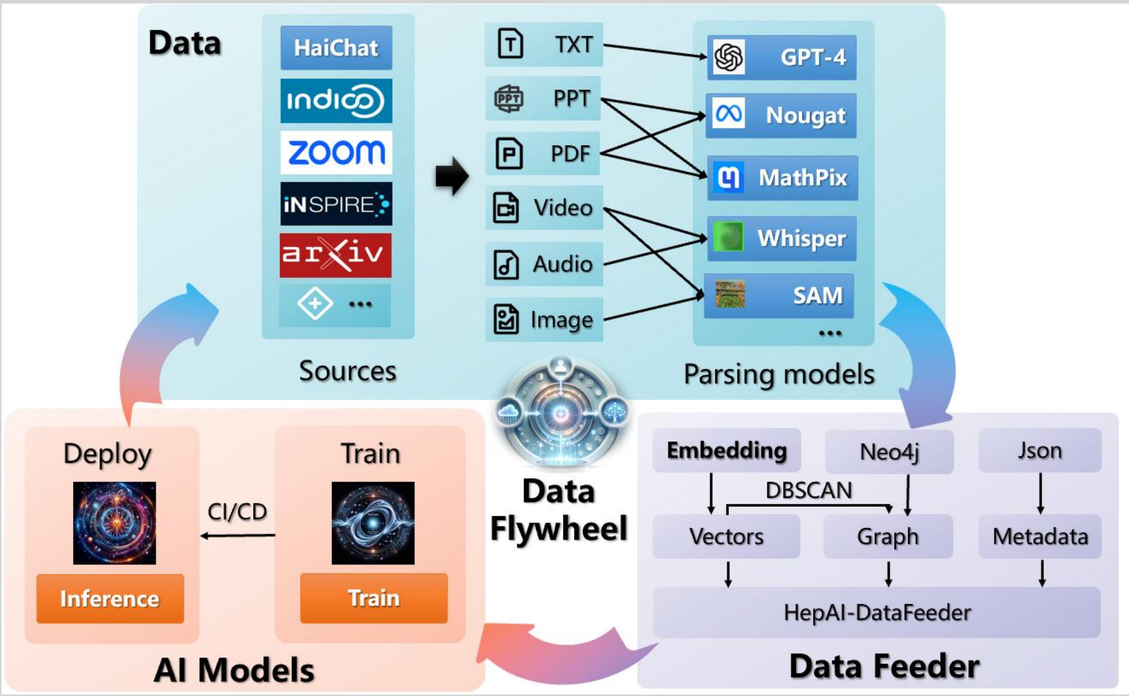
## Actuators



HepAI DDF v2.0

- BESIII actuator
  - BOSS worker
- Daisy framework
  - HEPSCT, ptycho
- Unified access of models, tools

## Sensors



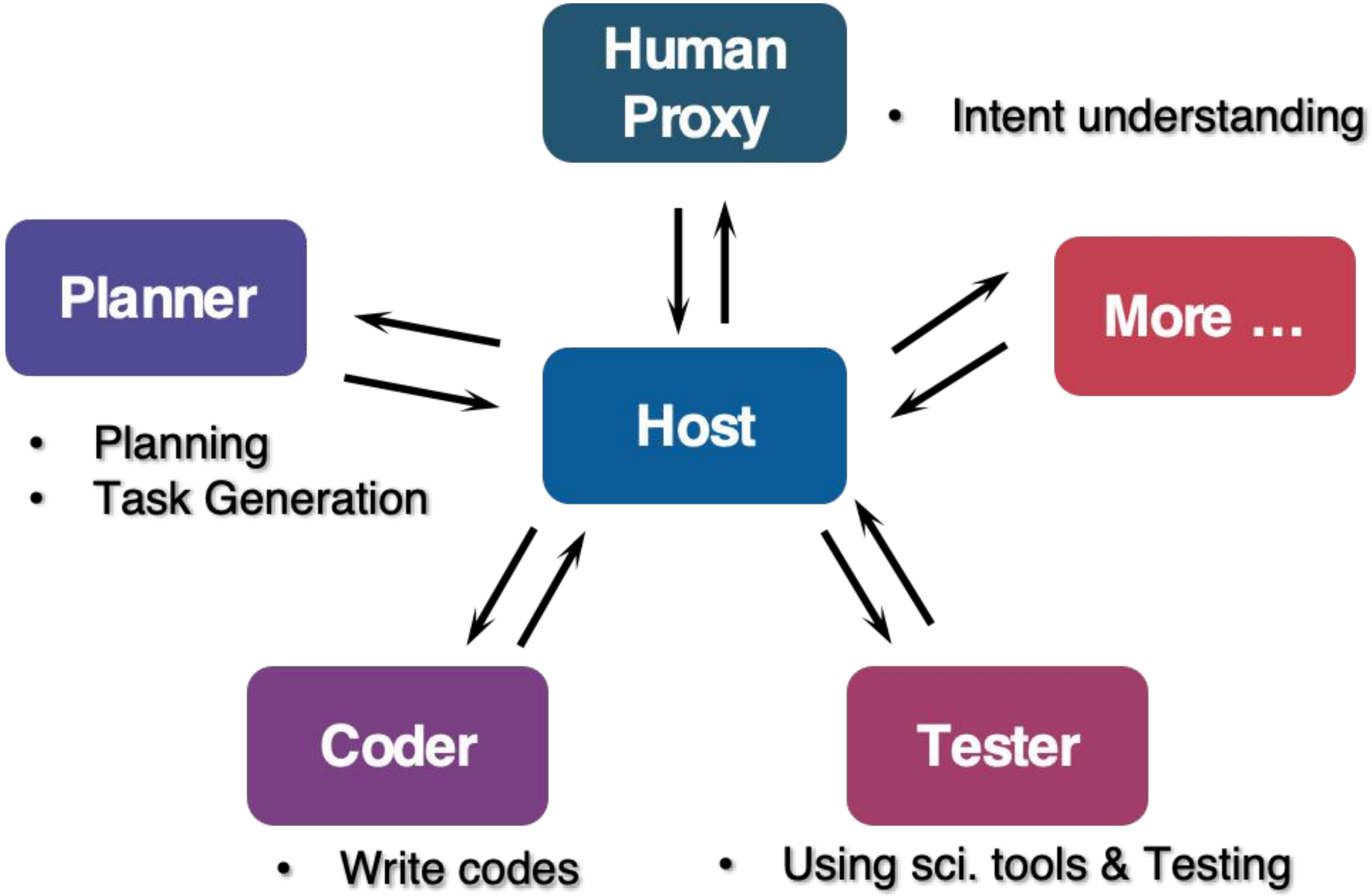
Data Flywheel System

- Data flywheel
  - HaiNougat
  - HaiChat
  - Whisper, SAM
  - arXiv, indico
- DOMAS framework
  - Scientific metadata

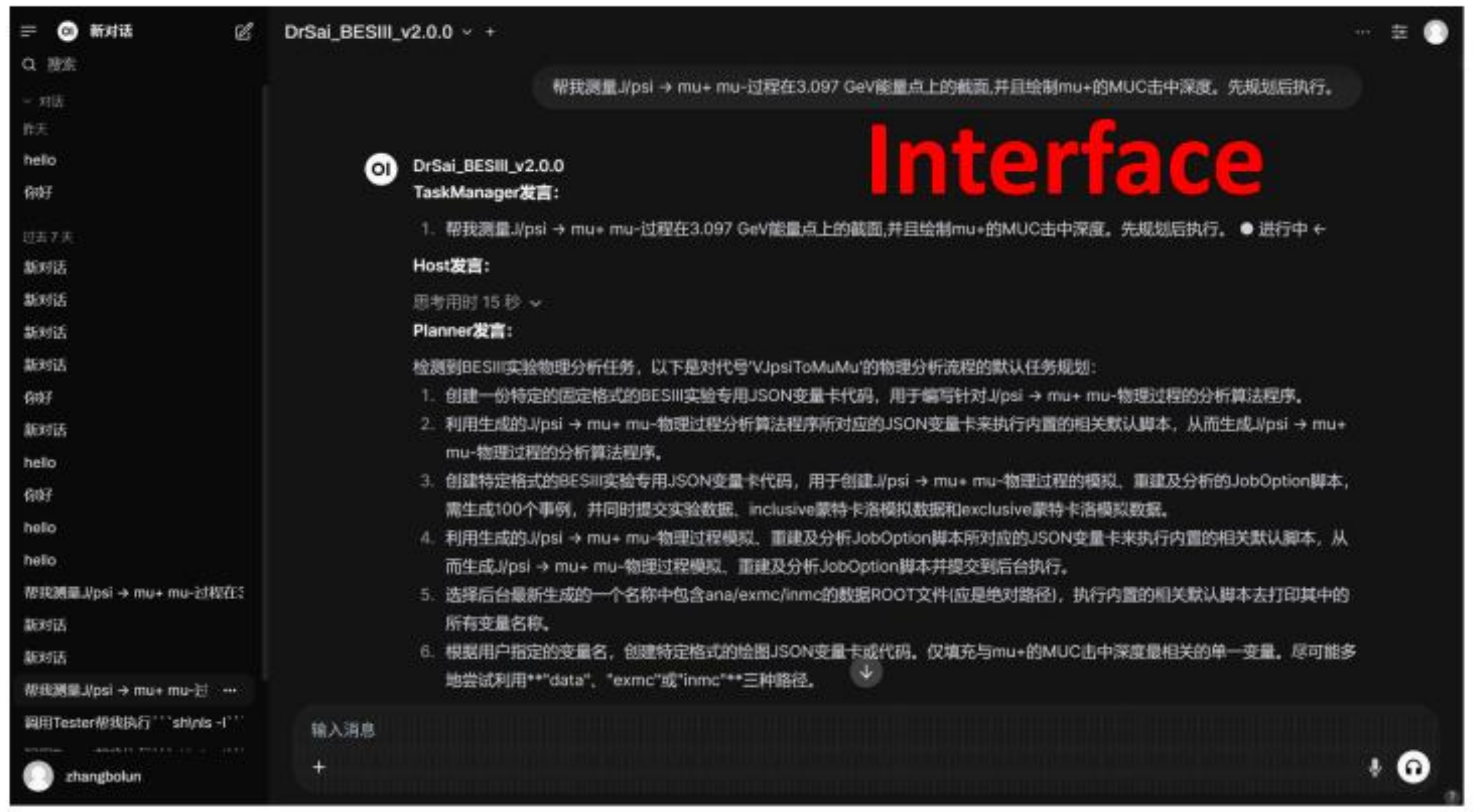
# Dr.Sai-BESIII

**Dr.Sai-BESIII** is an application of Dr.Sai framework, focus on the automated workflow of the **BESIII** experiment

- Talker-Reasoner architecture 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.



# Workflow of Dr.Sai-BESIII



帮我测量J/psi → mu+ mu-过程在3.097 GeV能量点上的截面,并且绘制mu+的MUC击中深度。先规划后执行。

DrSai_BESIII_v2.0.0

**TaskManager发言：**

1. 帮我测量J/psi → mu+ mu-过程在3.097 GeV能量点上的截面,并且绘制mu+的MUC击中深度。先规划后执行。 ● 进行中 ←

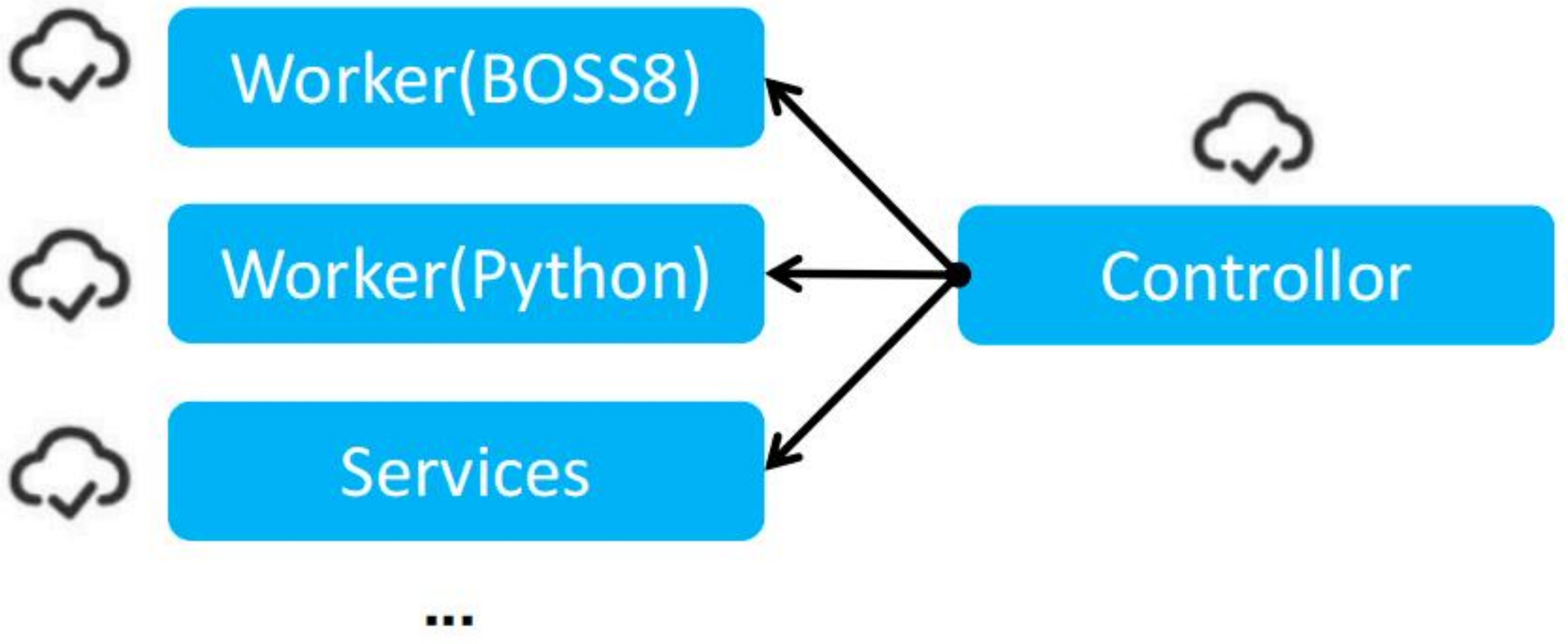
**Host发言：**

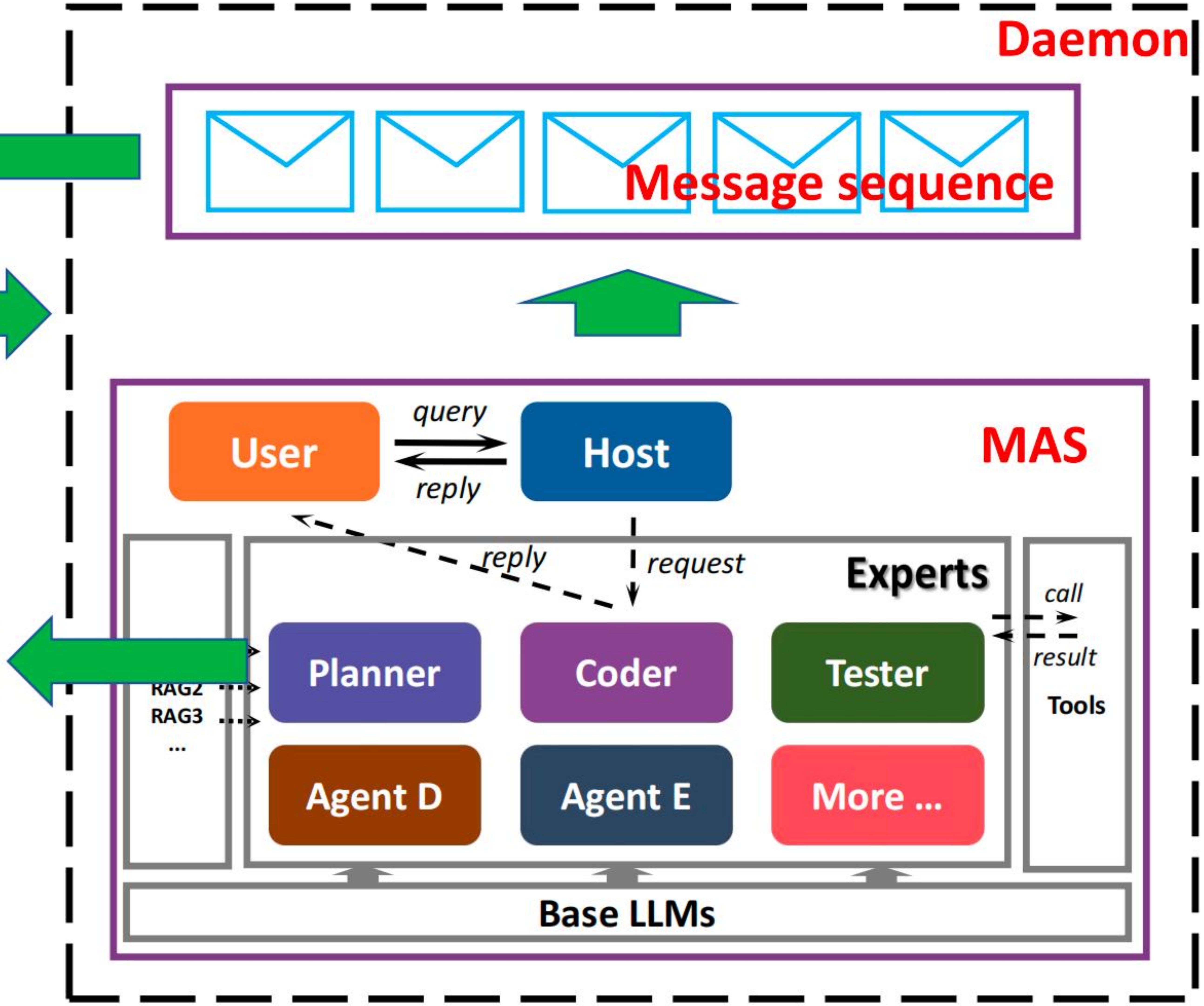
思考用时 15 秒

**Planner发言：**

检测到BESIII实验物理分析任务，以下是对代号'VJpsiToMuMu'的物理分析流程的默认任务规划：

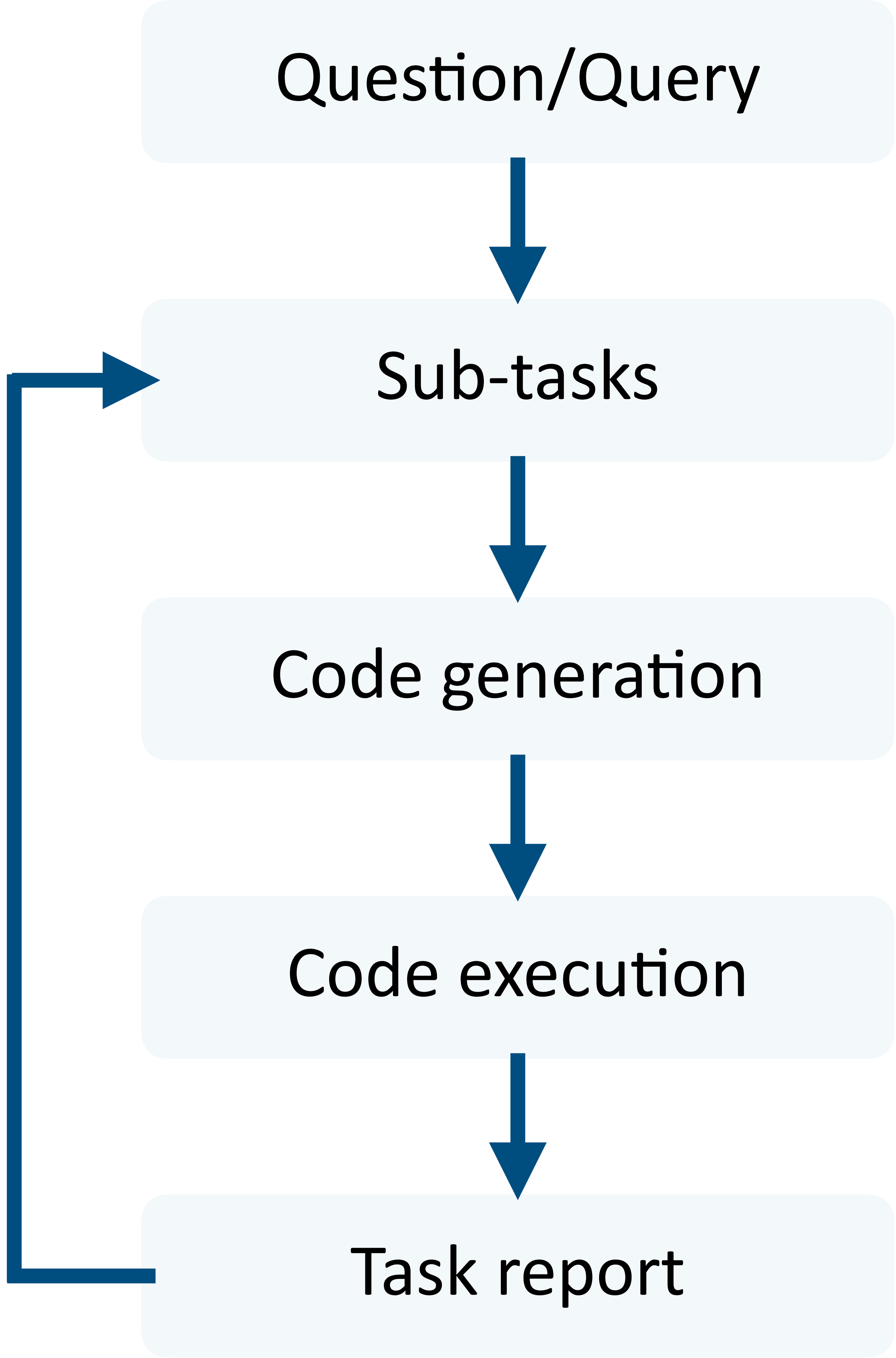
1. 创建一份特定的固定格式的BESIII实验专用JSON变量卡代码，用于编写针对J/psi → mu+ mu-物理过程的分析算法程序。
2. 利用生成的J/psi → mu+ mu-物理过程分析算法程序所对应的JSON变量卡来执行内置的相关默认脚本，从而生成J/psi → mu+ mu-物理过程的分析算法程序。
3. 创建特定格式的BESIII实验专用JSON变量卡代码，用于创建J/psi → mu+ mu-物理过程的模拟、重建及分析的JobOption脚本，需生成100个事例，并同时提交实验数据、inclusive蒙特卡洛模拟数据和exclusive蒙特卡洛模拟数据。
4. 利用生成的J/psi → mu+ mu-物理过程模拟、重建及分析JobOption脚本所对应的JSON变量卡来执行内置的相关默认脚本，从而生成J/psi → mu+ mu-物理过程模拟、重建及分析JobOption脚本并提交到后台执行。
5. 选择后台最新生成的一个名称中包含ana/exmc/inmc的数据ROOT文件(应是绝对路径)，执行内置的相关默认脚本去打印其中的所有变量名称。
6. 根据用户指定的变量名，创建特定格式的绘图JSON变量卡或代码。仅填充与mu+的MUC击中深度最相关的单一变量。尽可能多地尝试利用**"data"、"exmc"或"inmc"**三种路径。





# One example of Dr.Sai-BESIII

- ■ Multi agent collaboration
- ■ Professional domain abilities
  - ■ BESIII Knowledge
  - ■ Physics analysis planning
  - ■ BOSS coding
  - ■ Perform BESIII jobs
  - ■ Result analysis

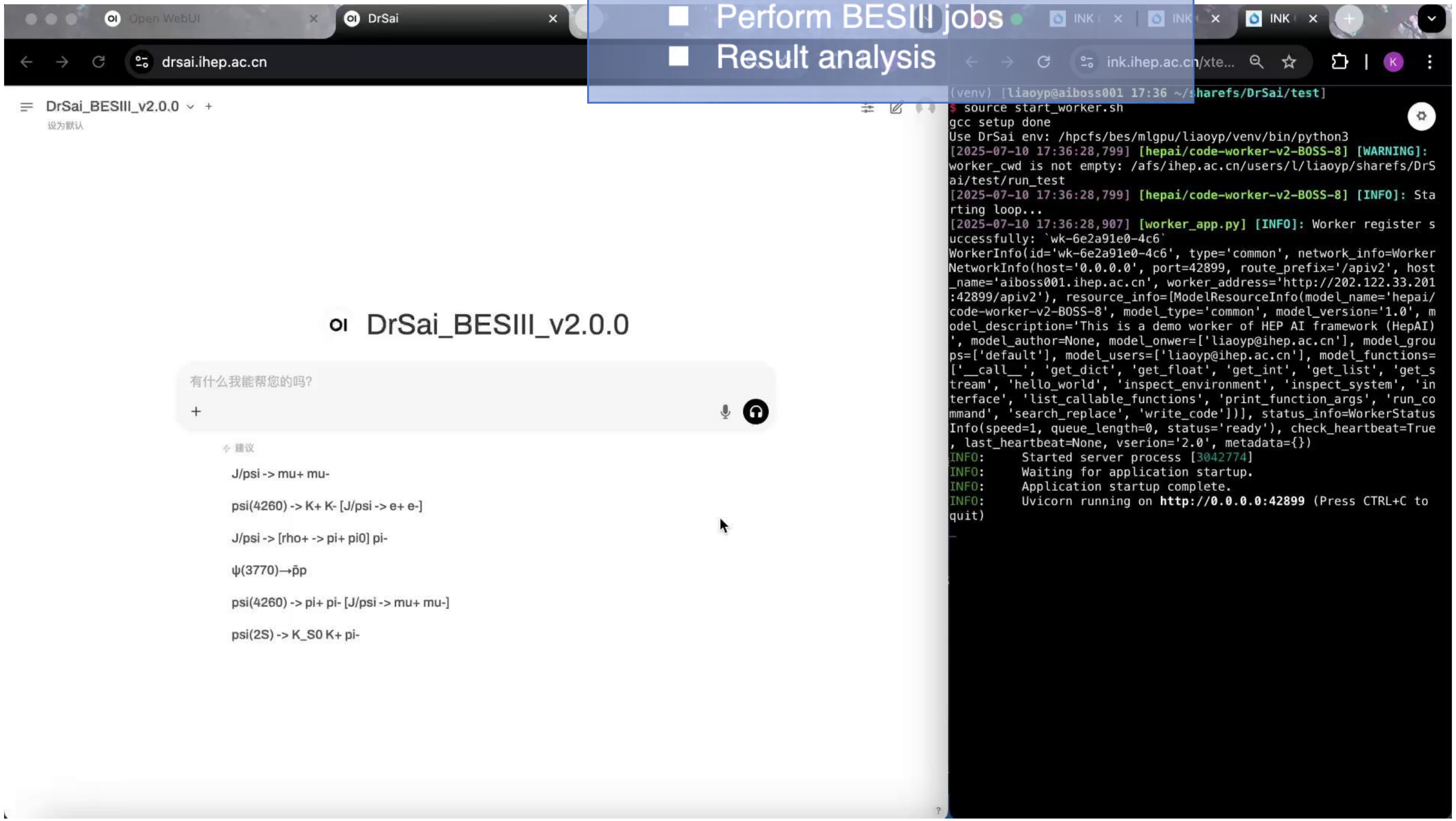

Question/Query

Sub-tasks

Code generation

Code execution

Task report



DrSai_BESIII_v2.0.0

有什么我能帮您的吗?

建议

J/psi -> mu+ mu-

psi(4260) -> K+ K- [J/psi -> e+ e-]

J/psi -> [rho+ -> pi+ pi0] pi-

ψ(3770)→p̄p

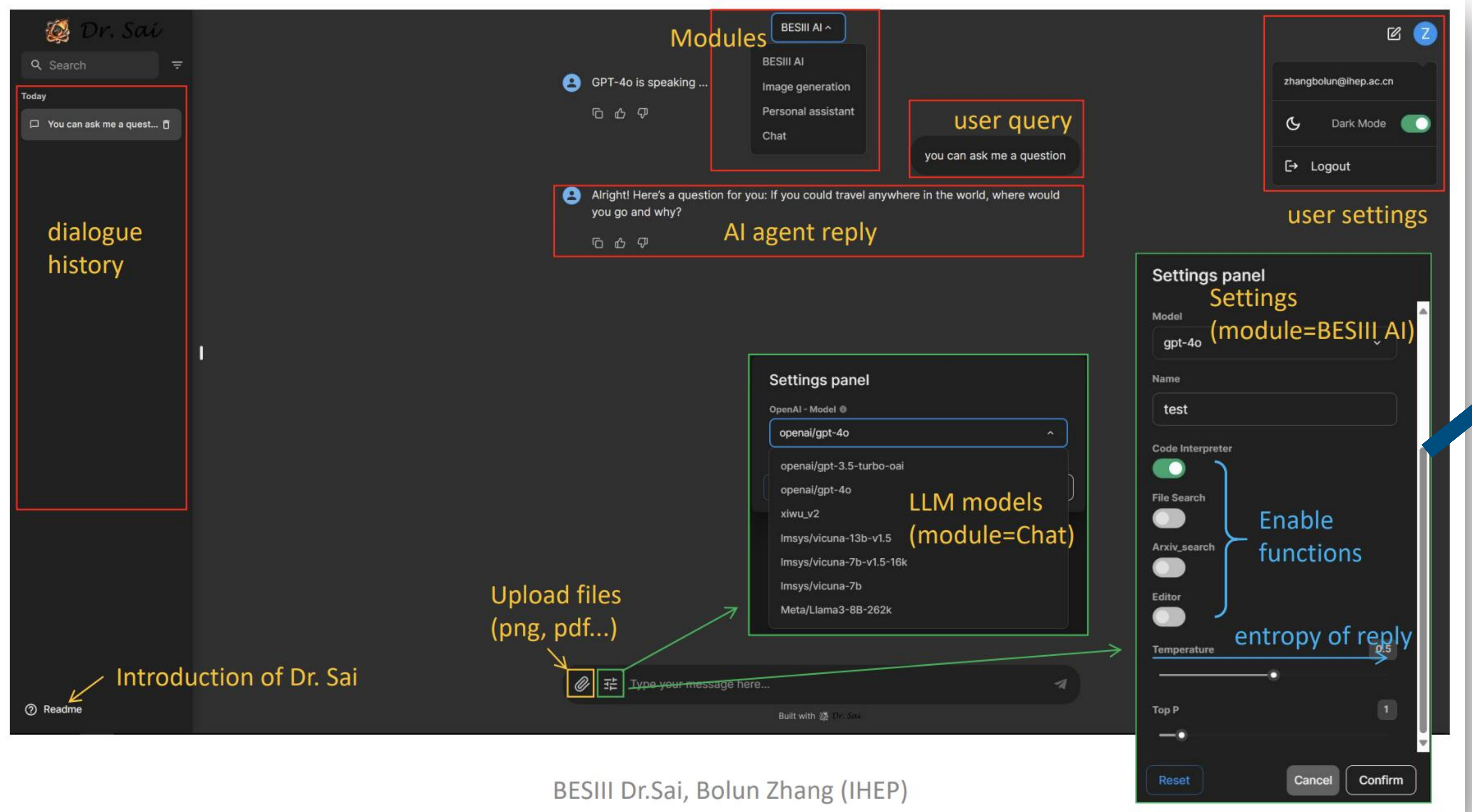
psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]

psi(2S) -> K_S0 K+ pi-

```
(venv) [liaoyp@aiboss001 17:36 ~/sharefs/DrSai/test]
$ source start_worker.sh
gcc setup done
Use DrSai env: /hpcfs/bes/mlgpu/liaoyp/venv/bin/python3
[2025-07-10 17:36:28,799] [hepai/code-worker-v2-BOSS-8] [WARNING]:
worker_cwd is not empty: /afs/ihep.ac.cn/users/l/liaoyp/sharefs/DrS
ai/test/run_test
[2025-07-10 17:36:28,799] [hepai/code-worker-v2-BOSS-8] [INFO]: Sta
rting loop...
[2025-07-10 17:36:28,907] [worker_app.py] [INFO]: Worker register s
uccessfully: `wk-6e2a91e0-4c6`
WorkerInfo(id='wk-6e2a91e0-4c6', type='common', network_info=Worker
NetworkInfo(host='0.0.0.0', port=42899, route_prefix='/apiv2', host
_name='aiboss001.ihep.ac.cn', worker_address='http://202.122.33.201
:42899/apiv2'), resource_info=[ModelResourceInfo(model_name='hepai/
code-worker-v2-BOSS-8', model_type='common', model_version='1.0', m
odel_description='This is a demo worker of HEP AI framework (HepAI)
', model_author=None, model_onwer=['liaoyp@ihep.ac.cn'], model_grou
ps=['default'], model_users=['liaoyp@ihep.ac.cn'], model_functions=
['__call__', 'get_dict', 'get_float', 'get_int', 'get_list', 'get_s
tream', 'hello_world', 'inspect_environment', 'inspect_system', 'in
terface', 'list_callable_functions', 'print_function_args', 'run_co
mmand', 'search_replace', 'write_code'])], status_info=WorkerStatus
Info(speed=1, queue_length=0, status='ready'), check_heartbeat=True
, last_heartbeat=None, vserion='2.0', metadata={})
INFO:     Started server process [3042774]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:42899 (Press CTRL+C to
quit)
```

# Interface of Dr.Sai-BESIII

## Dr.Sai v1.0 UI based on Chainlit (2024.07)



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

openai/gpt-3.5-turbo-oai
openai/gpt-4o
xiwu_v2
lmsys/vicuna-13b-v1.5
lmsys/vicuna-7b-v1.5-16k
lmsys/vicuna-7b
Meta/Llama3-8B-262k

Settings panel

Model: gpt-4o

Name: test

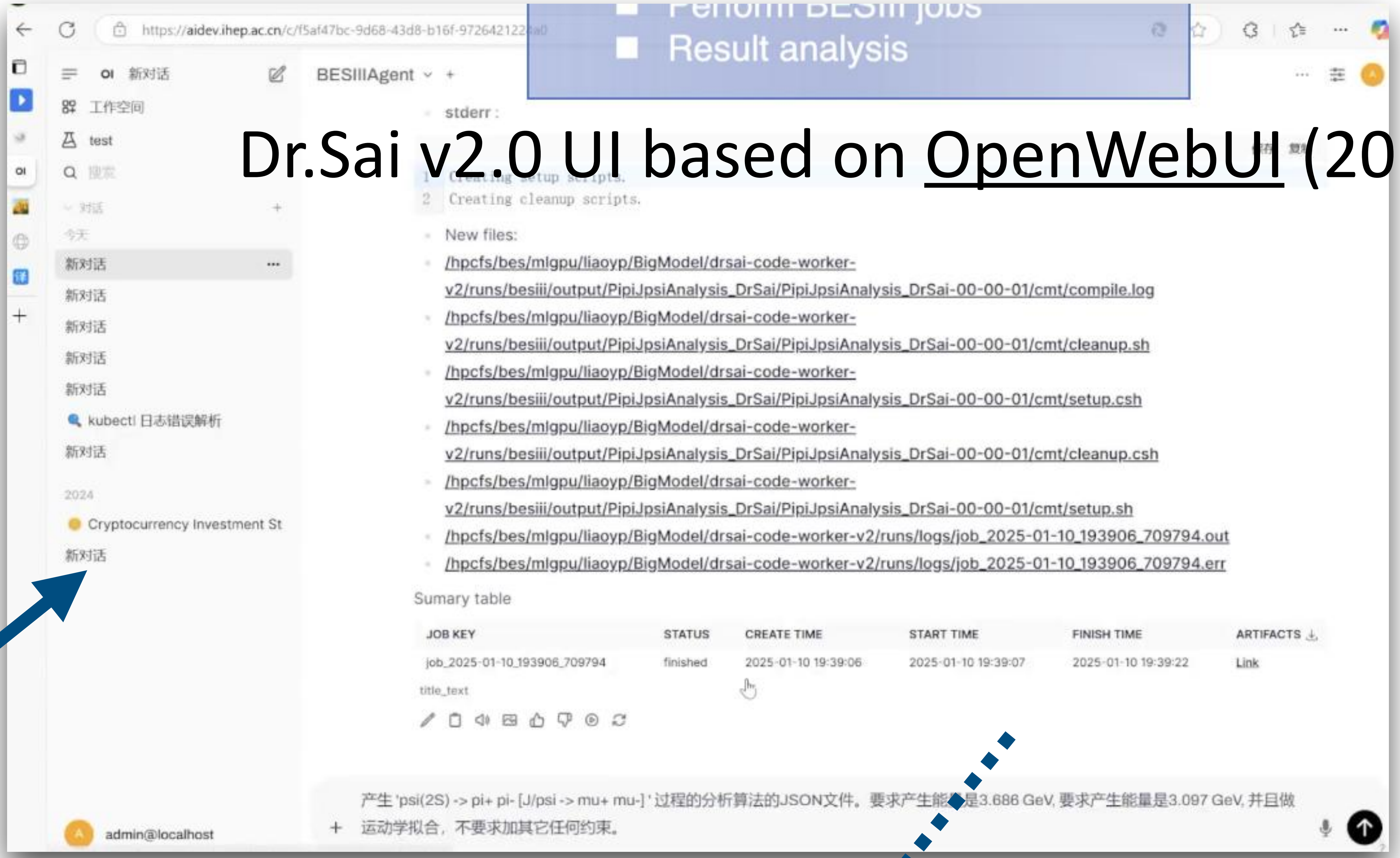
Code Interpreter ●
File Search ○
Arxiv_search ○
Editor ○

Temperature 0.5

Top P 1

BESIII Dr.Sai, Bolun Zhang (IHEP)

## Dr.Sai v2.0 UI based on OpenWebUI (2025.07)



stderr :

```
2 Creating cleanup scripts.
```

New files:

- /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai-code-worker-v2/runs/besiii/output/PipiJpsiAnalysis_DrSai/PipiJpsiAnalysis_DrSai-00-00-01/cmt/compile.log
- /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai-code-worker-v2/runs/besiii/output/PipiJpsiAnalysis_DrSai/PipiJpsiAnalysis_DrSai-00-00-01/cmt/cleanup.sh
- /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai-code-worker-v2/runs/besiii/output/PipiJpsiAnalysis_DrSai/PipiJpsiAnalysis_DrSai-00-00-01/cmt/setup.csh
- /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai-code-worker-v2/runs/besiii/output/PipiJpsiAnalysis_DrSai/PipiJpsiAnalysis_DrSai-00-00-01/cmt/cleanup.csh
- /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai-code-worker-v2/runs/besiii/output/PipiJpsiAnalysis_DrSai/PipiJpsiAnalysis_DrSai-00-00-01/cmt/setup.sh
- /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai-code-worker-v2/runs/logs/job_2025-01-10_193906_709794.out
- /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai-code-worker-v2/runs/logs/job_2025-01-10_193906_709794.err

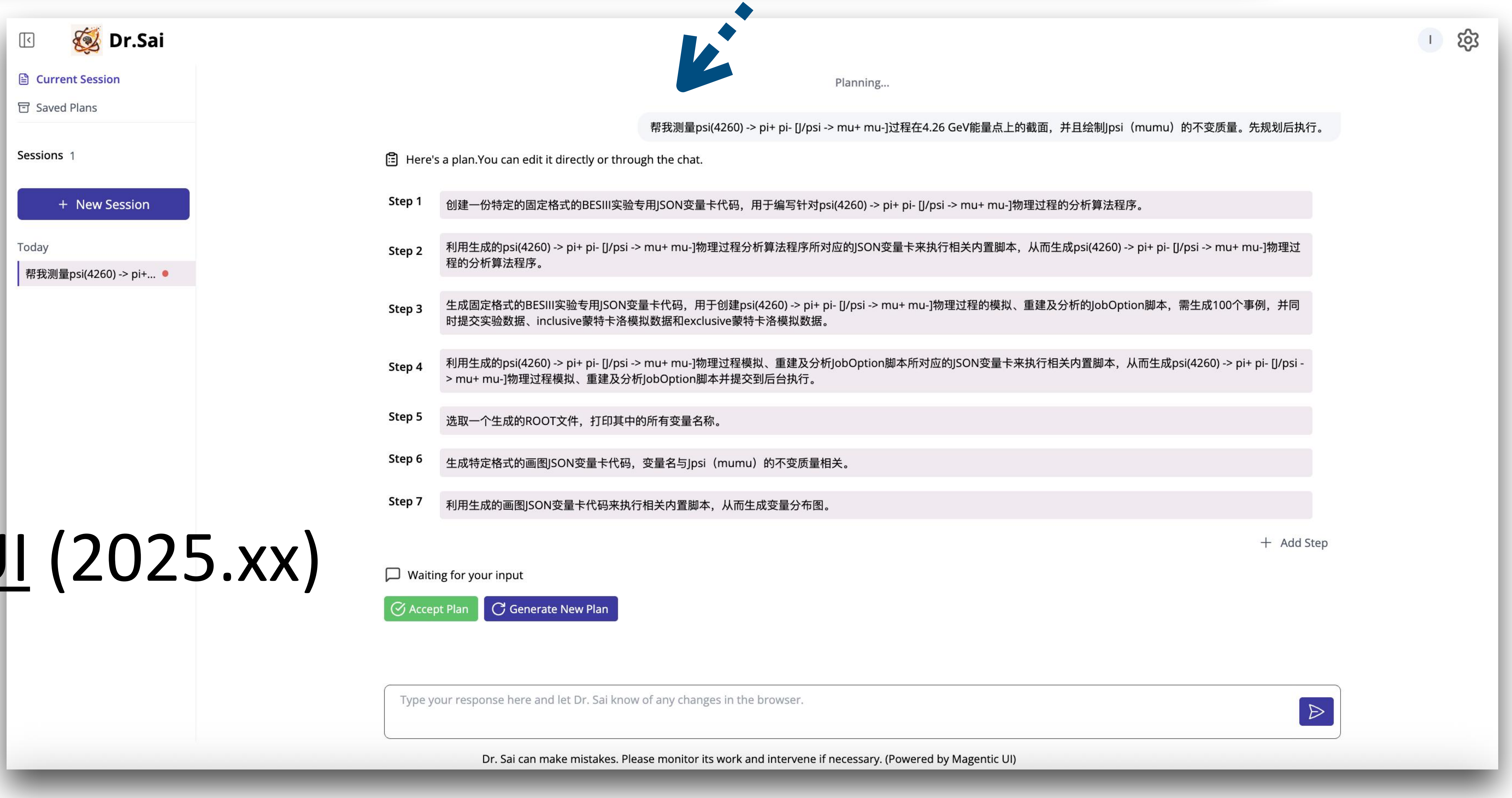
Sumary table

| JOB KEY | STATUS | CREATE TIME | START TIME | FINISH TIME | ARTIFACTS |
|---|---|---|---|---|---|
| job_2025-01-10_193906_709794 | finished | 2025-01-10 19:39:06 | 2025-01-10 19:39:07 | 2025-01-10 19:39:22 | Link |

title_text

产生 'psi(2S) -> pi+ pi- [J/psi -> mu+ mu-]' 过程的分析算法的JSON文件。要求产生能量是3.686 GeV, 要求产生能量是3.097 GeV, 并且做运动学拟合，不要求加其它任何约束。

## Dr.Sai v2.x (Planned) UI based on Magentic-UI (2025.xx)


Dr.Sai
Current Session
Saved Plans
Sessions 1
New Session
Today
帮我测量psi(4260) -> pi+...
Add Step
Accept Plan
Generate New Plan
Type your response here and let Dr. Sai know of any changes in the browser.

Planning...

帮我测量psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]过程在4.26 GeV能量点上的截面，并且绘制Jpsi（mumu）的不变质量。先规划后执行。

Here's a plan.You can edit it directly or through the chat.

Step 1 创建一份特定的固定格式的BESIII实验专用JSON变量卡代码，用于编写针对psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程的分析算法程序。

Step 2 利用生成的psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程分析算法程序所对应的JSON变量卡来执行相关内置脚本，从而生成psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程的分析算法程序。

Step 3 生成固定格式的BESIII实验专用JSON变量卡代码，用于创建psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程的模拟、重建及分析的JobOption脚本，需生成100个事例，并同时提交实验数据、inclusive蒙特卡洛模拟数据和exclusive蒙特卡洛模拟数据。

Step 4 利用生成的psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程模拟、重建及分析JobOption脚本所对应的JSON变量卡来执行相关内置脚本，从而生成psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程模拟、重建及分析JobOption脚本并提交到后台执行。

Step 5 选取一个生成的ROOT文件，打印其中的所有变量名称。

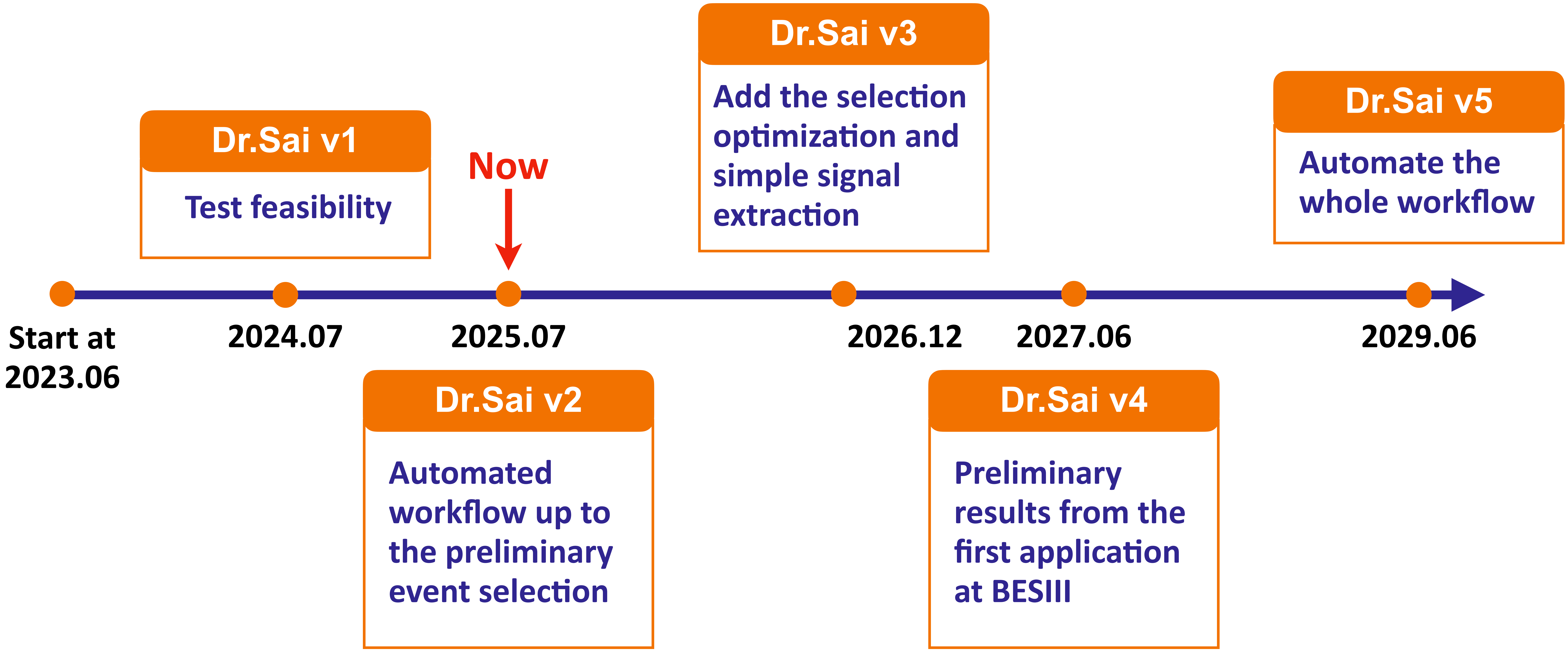
Step 6 生成特定格式的画图JSON变量卡代码，变量名与Jpsi（mumu）的不变质量相关。

Step 7 利用生成的画图JSON变量卡代码来执行相关内置脚本，从而生成变量分布图。

Waiting for your input

Dr. Sai can make mistakes. Please monitor its work and intervene if necessary. (Powered by Magentic UI)

# Prospect of Dr.Sai-BESIII



Start at 2023.06

**Dr.Sai v1**

Test feasibility

2024.07

Now

2025.07

**Dr.Sai v2**

Automated workflow up to the preliminary event selection

**Dr.Sai v3**

Add the selection optimization and simple signal extraction

2026.12

2027.06

**Dr.Sai v4**

Preliminary results from the first application at BESIII

**Dr.Sai v5**

Automate the whole workflow

2029.06

# In the end

- LLM could be very helpful for HEP, and also for other domains!
  - Not just generate draft code/text, but also can be used to **automate the analysis**
- **Dr.Sai** - a multi-agent collaborative system has been developed, and one of the application **Dr.Sai-BESIII** preliminarily demonstrates its ability to automate physics analysis processes

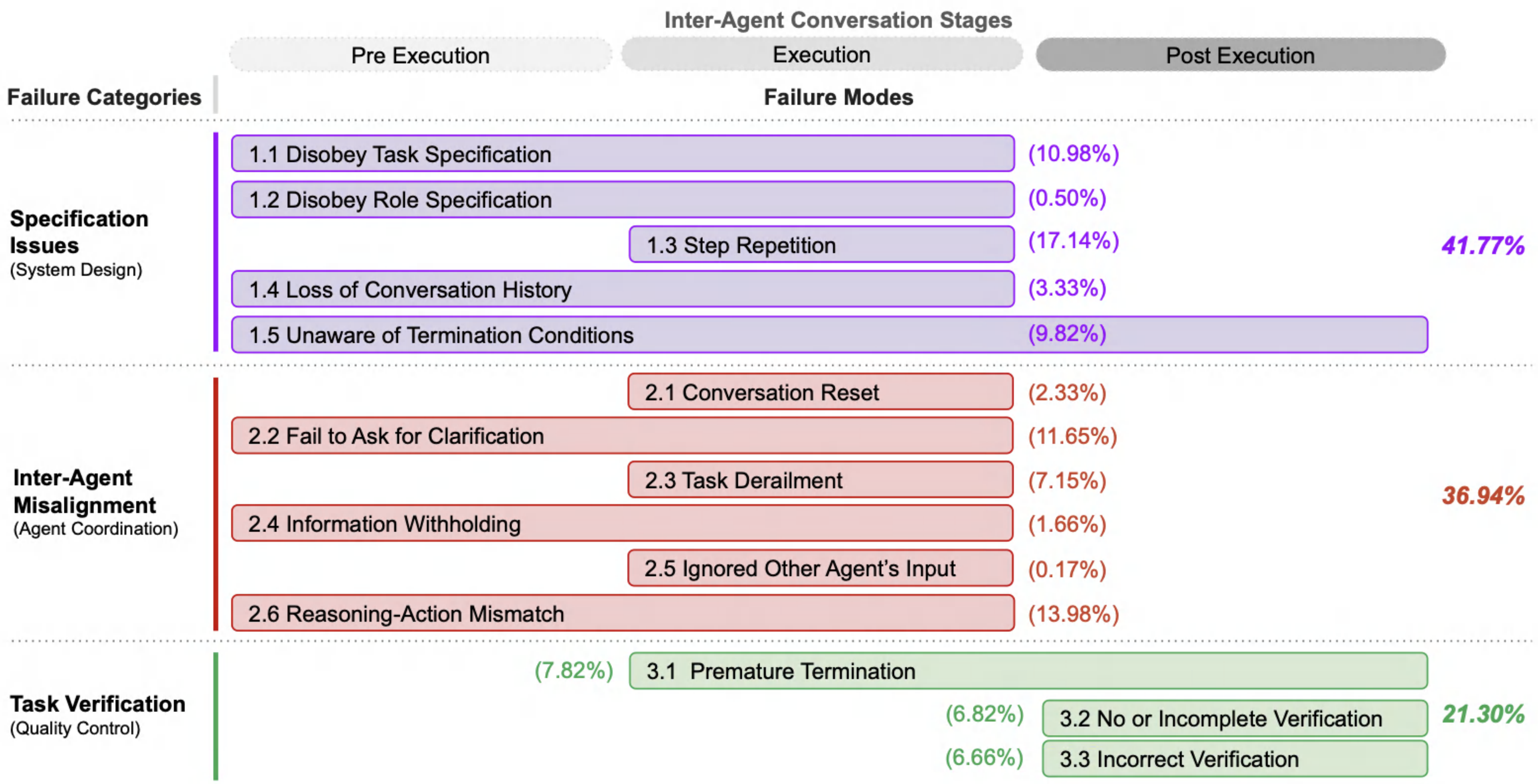
- Dr.Sai is a new system, which still needs further improvement
  - DSL development; Benchmark design; Better HEP-LLM …
  - In the future, use Dr.Sai to **quickly explore hundreds of physical processes**, which is expected to be used to search for new hadron states or new decays

Yipu Liao (廖一朴)
IHEP, CAS, Beijing
liaoyp@ihep.ac.cn

# Thank you for listening!

# LLM nowadays: Multi-Agent System

**Multi-Agent System (MAS)**



Inter-Agent Conversation Stages

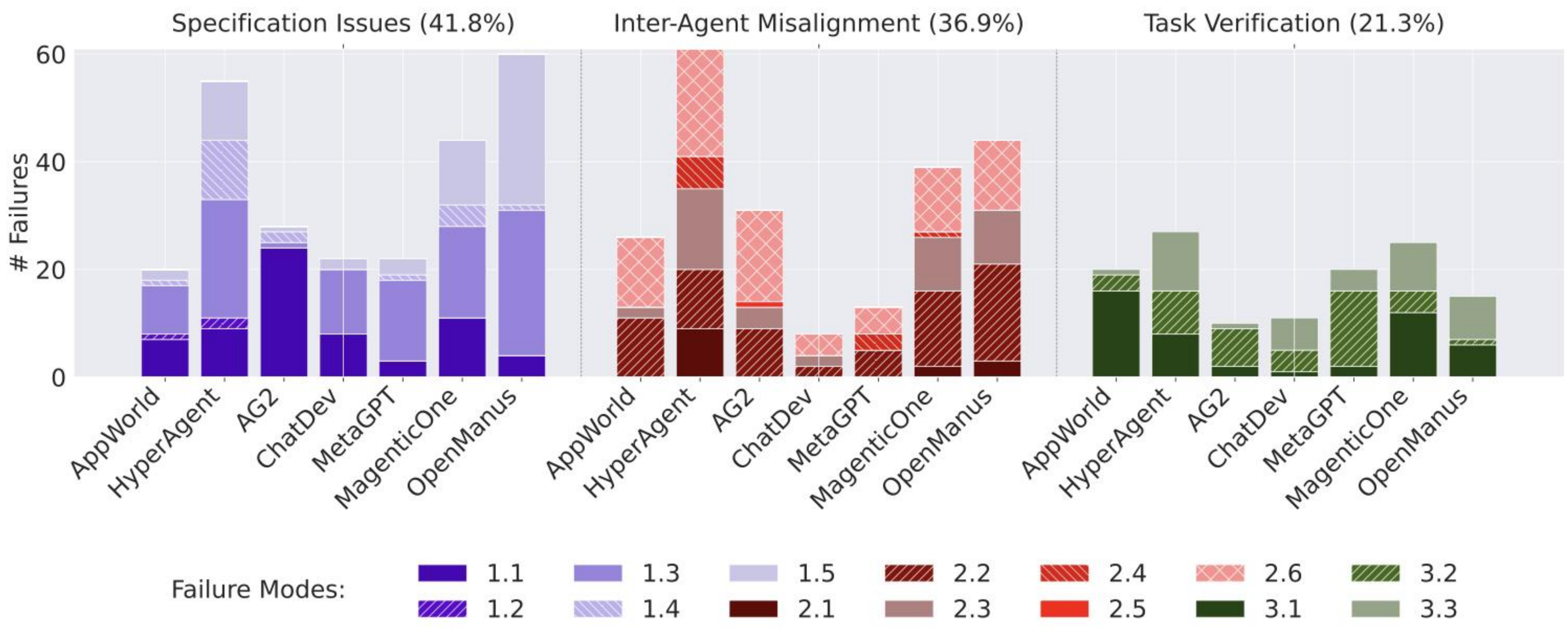
Pre Execution | Execution | Post Execution

**Failure Categories** | **Failure Modes**

**Specification Issues** (System Design) — 41.77%
- 1.1 Disobey Task Specification (10.98%)
- 1.2 Disobey Role Specification (0.50%)
- 1.3 Step Repetition (17.14%)
- 1.4 Loss of Conversation History (3.33%)
- 1.5 Unaware of Termination Conditions (9.82%)

**Inter-Agent Misalignment** (Agent Coordination) — 36.94%
- 2.1 Conversation Reset (2.33%)
- 2.2 Fail to Ask for Clarification (11.65%)
- 2.3 Task Derailment (7.15%)
- 2.4 Information Withholding (1.66%)
- 2.5 Ignored Other Agent's Input (0.17%)
- 2.6 Reasoning-Action Mismatch (13.98%)

**Task Verification** (Quality Control) — 21.30%
- (7.82%) 3.1 Premature Termination
- (6.82%) 3.2 No or Incomplete Verification
- (6.66%) 3.3 Incorrect Verification



However, a lot of questions appear in the MAS (Still developing):

1. **Specification issues** (System design)

   Wrong system design decisions, and poor or ambiguous prompt specifications.

2. **Inter-agent misalignment** (Agent coordination)

   Incorrect or redundant reasoning of agents

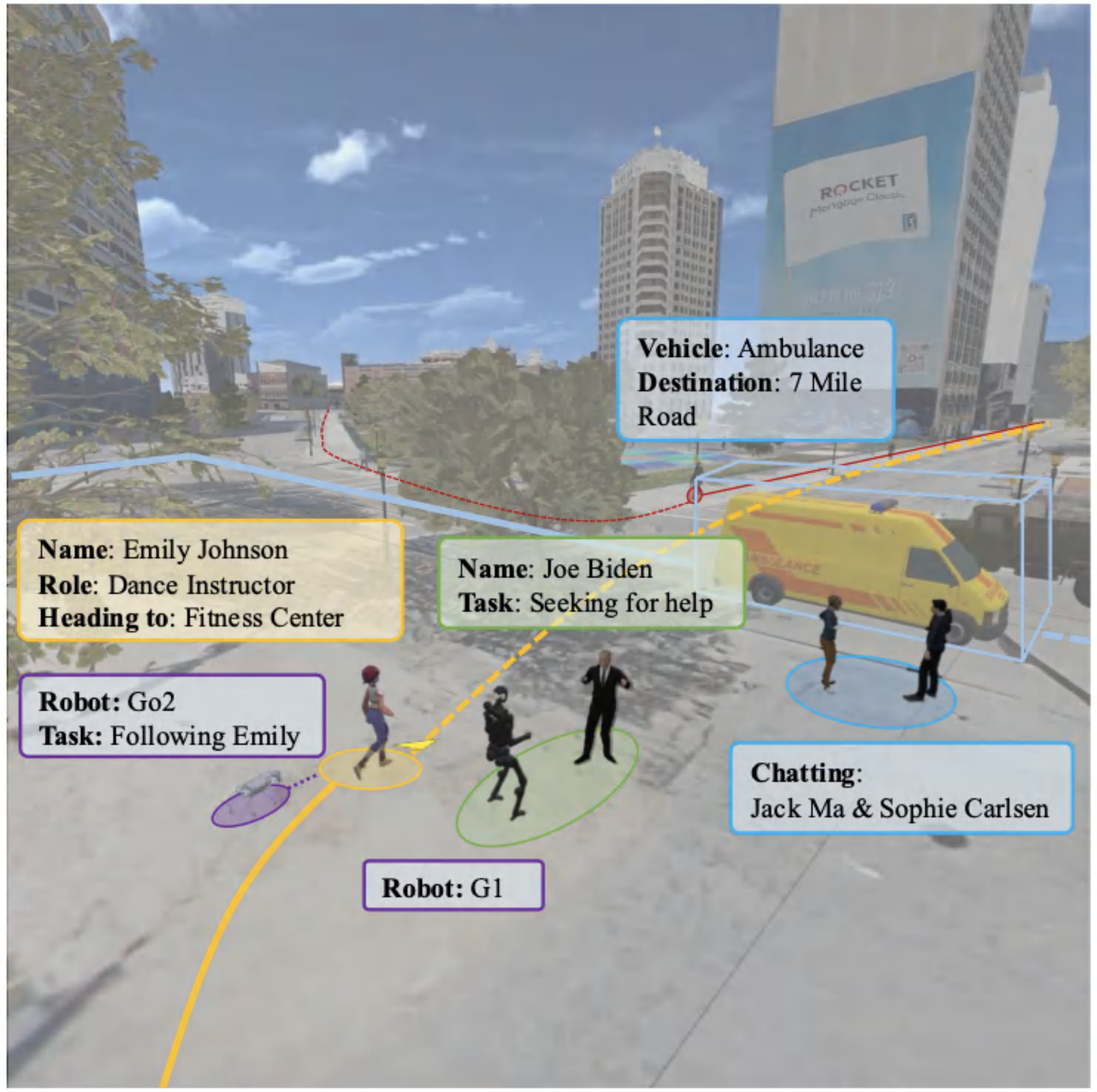
3. **Task verification** (Quality control)

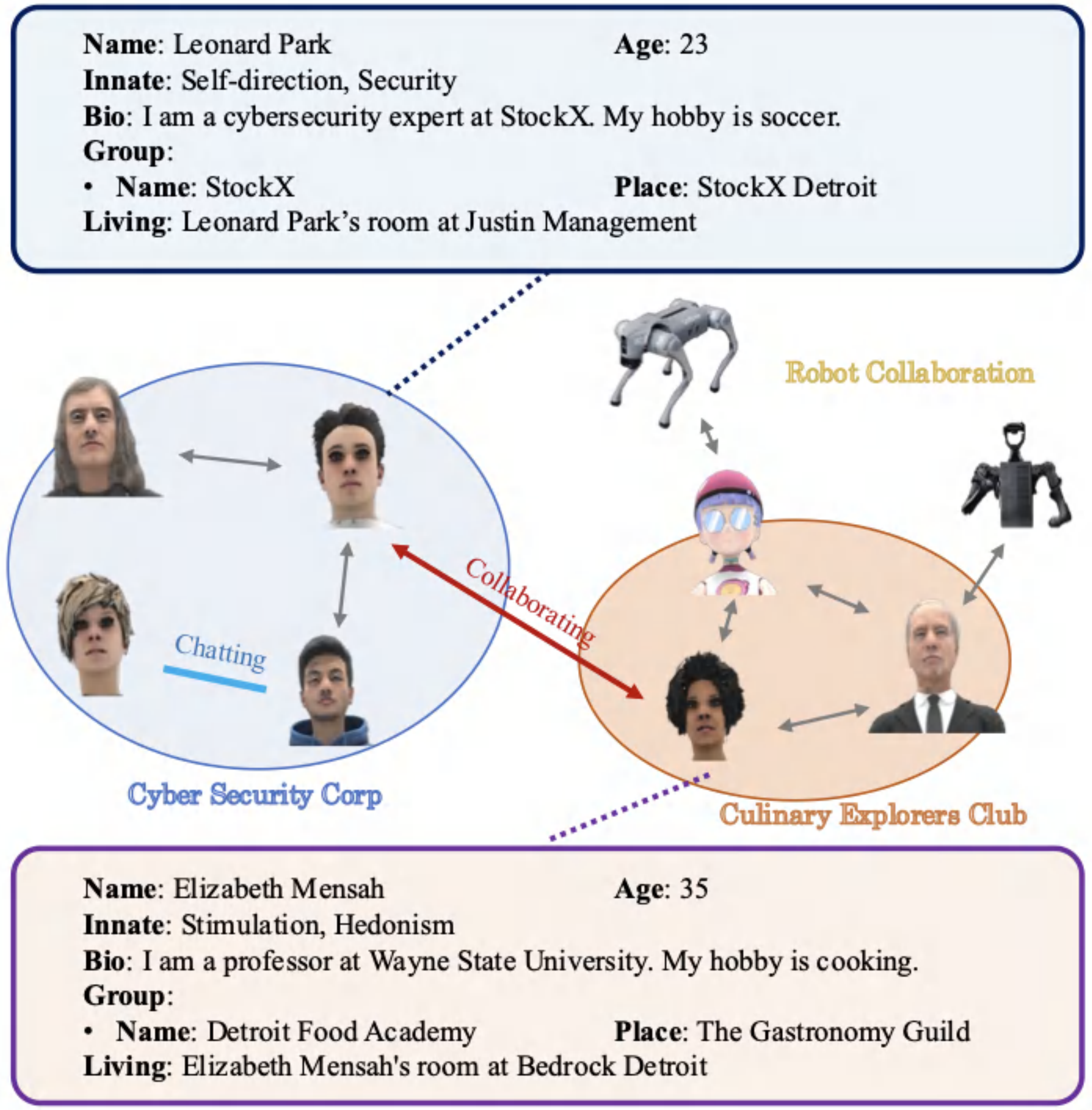
   Difficult to determine whether the task result is correct

* Why Do Multi-Agent LLM Systems Fail?

# Some applications based on MAS

* Virtual Community An Open World for Humans, Robots, and Society


**Vehicle**: Ambulance
**Destination**: 7 Mile Road
**Name**: Emily Johnson
**Role**: Dance Instructor
**Heading to**: Fitness Center
**Name**: Joe Biden
**Task**: Seeking for help
**Robot:** Go2
**Task:** Following Emily
**Chatting**:
Jack Ma & Sophie Carlsen
**Robot:** G1



**Name**: Leonard Park **Age**: 23
**Innate**: Self-direction, Security
**Bio**: I am a cybersecurity expert at StockX. My hobby is soccer.
**Group**:
- **Name**: StockX **Place**: StockX Detroit

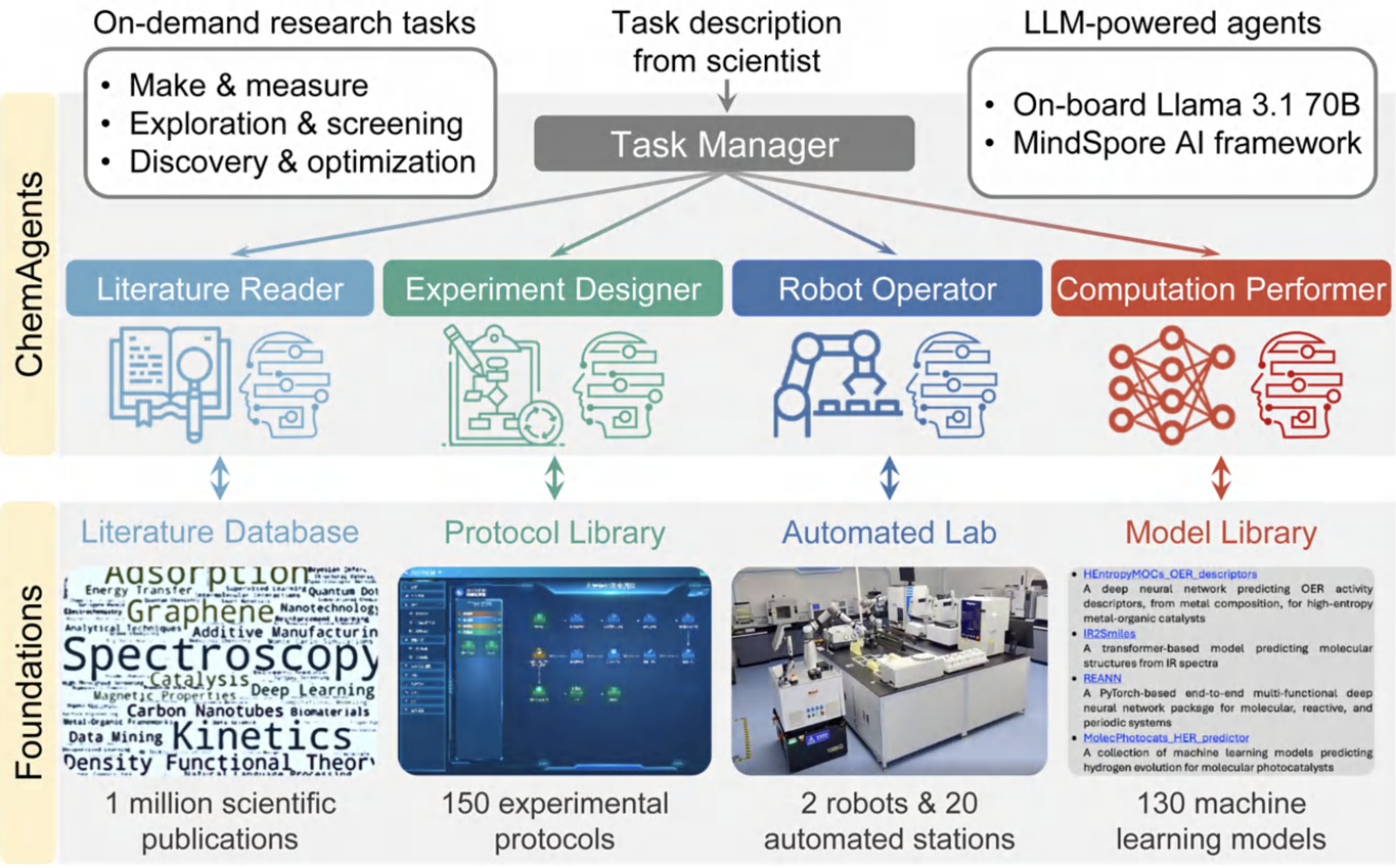
**Living**: Leonard Park's room at Justin Management

**Name**: Elizabeth Mensah **Age**: 35
**Innate**: Stimulation, Hedonism
**Bio**: I am a professor at Wayne State University. My hobby is cooking.
**Group**:
- **Name**: Detroit Food Academy **Place**: The Gastronomy Guild

**Living**: Elizabeth Mensah's room at Bedrock Detroit

- Real geospatial data and physics engine
- Agents with different identities and Robots to form virtual communities with World Model (3D LLM)
- Focus on community interaction and task execution (such as campaigning and voting)

# Some applications based on MAS

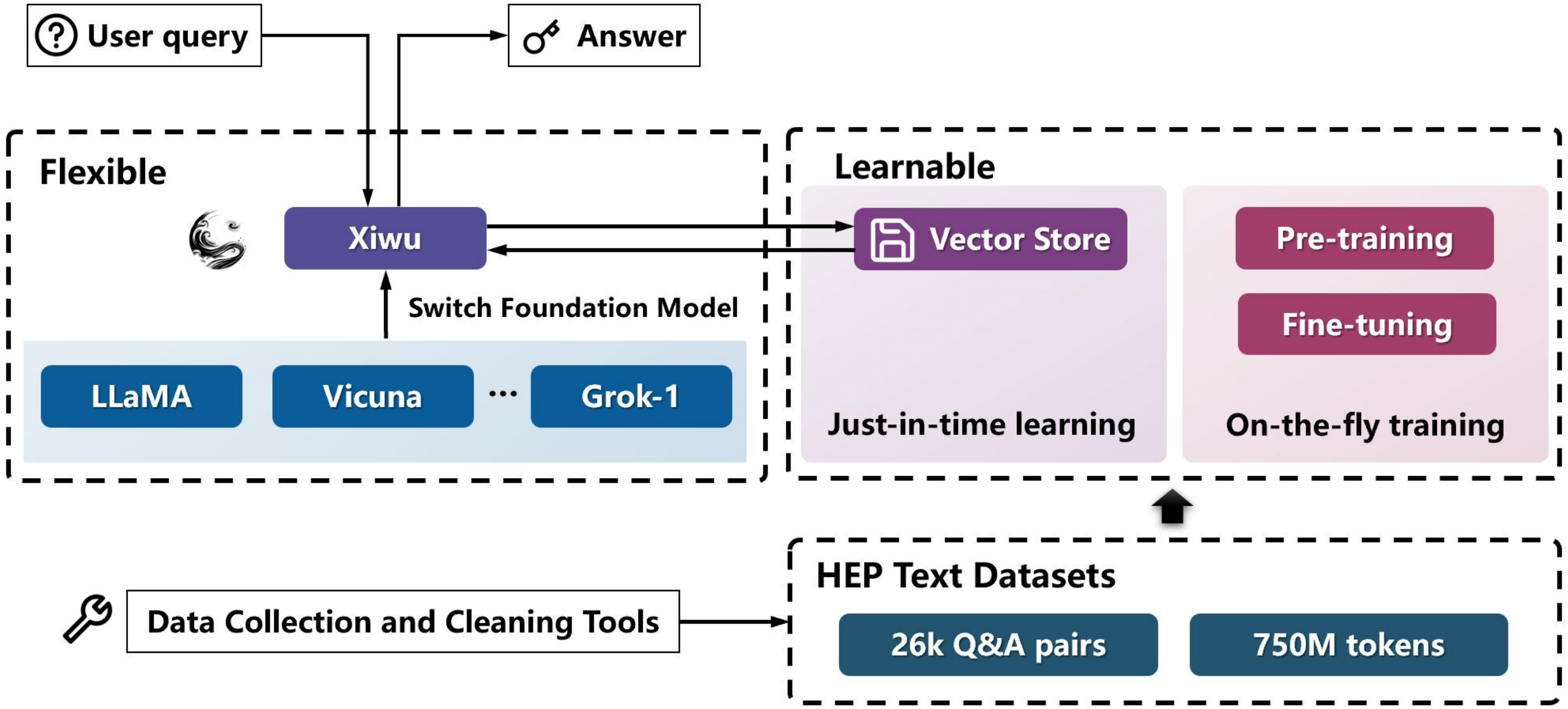


- HEntropyMOCs_OER_descriptors
  A deep neural network predicting OER activity descriptors, from metal composition, for high-entropy metal-organic catalysts
- IR2Smiles
  A transformer-based model predicting molecular structures from IR spectra
- REANN
  A PyTorch-based end-to-end multi-functional deep neural network package for molecular, reactive, and periodic systems
- MolecPhotocats_HER_predictor
  A collection of machine learning models predicting hydrogen evolution for molecular photocatalysts

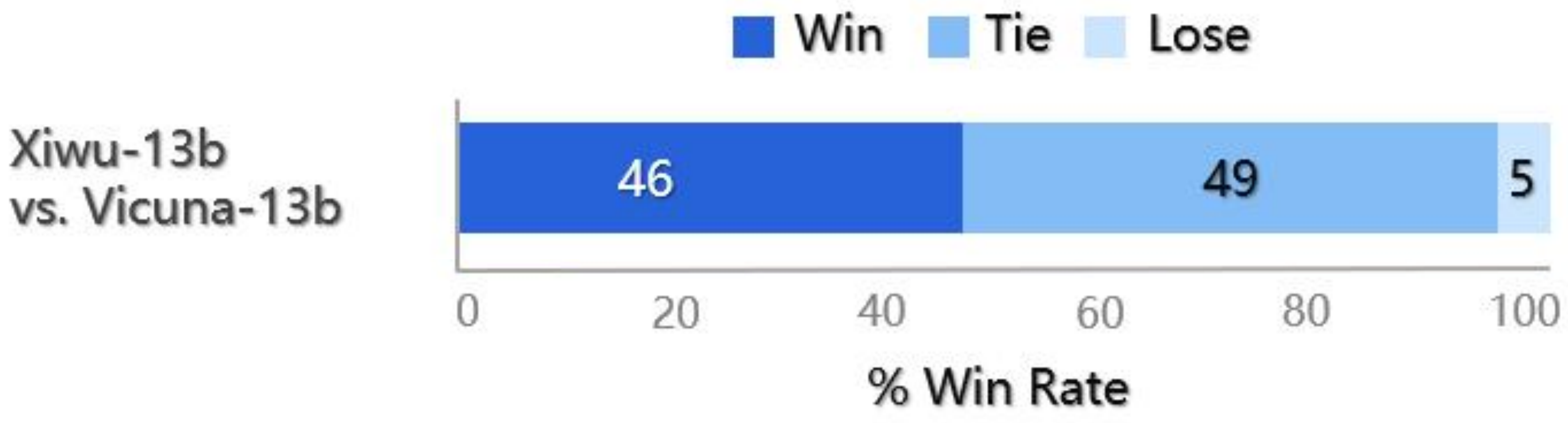
- A Multi-Agent AI (USTC) + automated Lab for Chemistry
- Chemical focus
- Reader + Planner + Robot Operator + Executor + DSL

# Xiwu (溪悟): a HEP-specific fine-tuned LLM

- The **brain** of Dr.Sai to enhance the domain capabilities
- Currently, Xiwu is based on LLaMA3, will be upgrade to **DeepSeek**, Historically, LLaMA, LLaMA2
- Secondary pre-training and fine-tuning techniques
- Trained on GPU-A100 previously, now trained on DCU-K100 (GPU from Sugon)
- Significantly better than the base model in **HEP Q&A** and **internal code generation** (at that time)



Test results



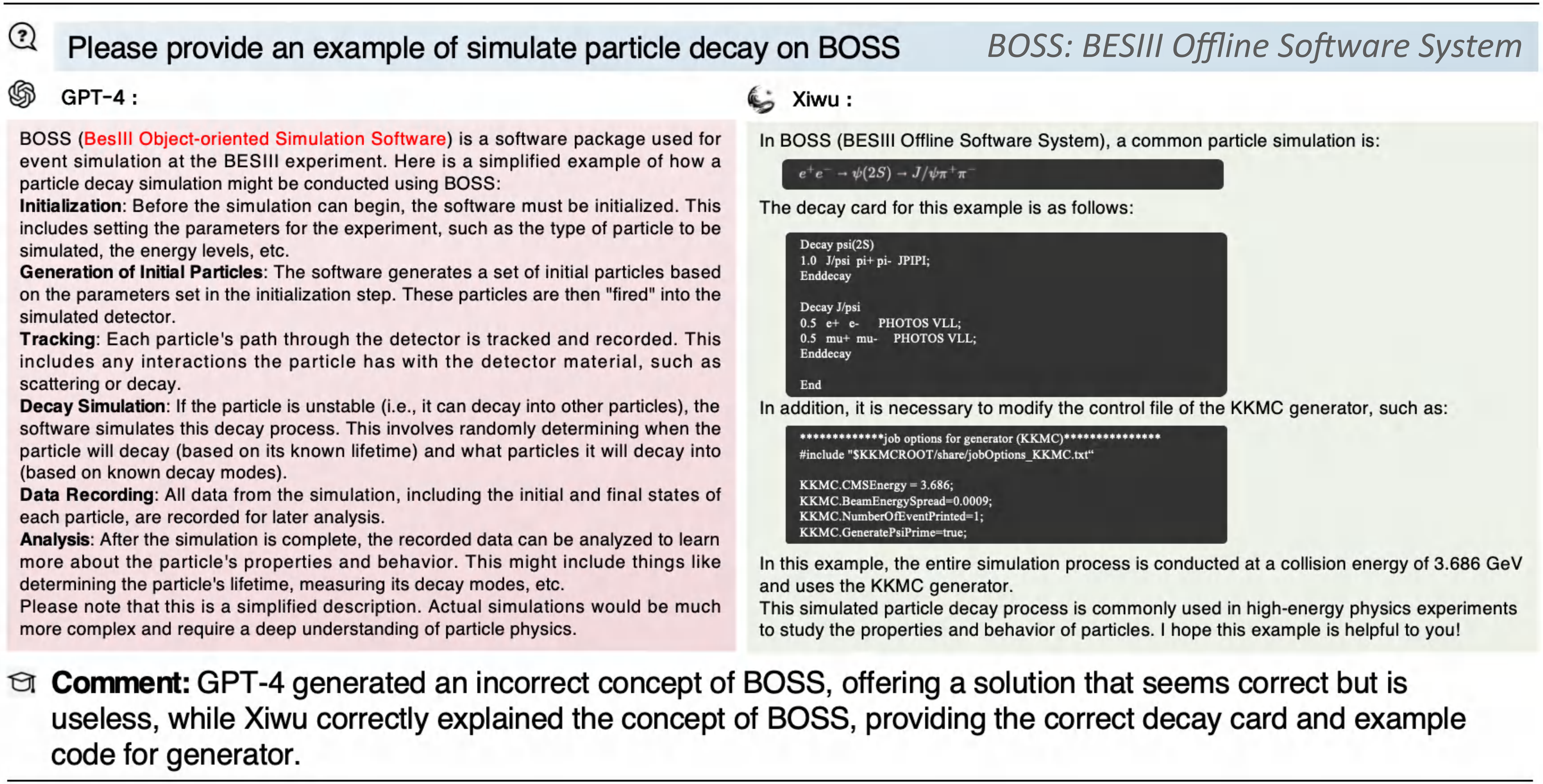
* arXiv:2404.08001

* *Preliminary trial work*

https://github.com/zhangzhengde0225/Xiwu

# Xiwu (溪悟): a HEP-specific fine-tuned LLM

Brain



Please provide an example of simulate particle decay on BOSS *BOSS: BESIII Offline Software System*

**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.

**Xiwu :**

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!

**Comment:** 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.

* arXiv:2404.08001 * *Preliminary trial work* https://github.com/zhangzhengde0225/Xiwu

# Current training data for Xiwu (溪悟)

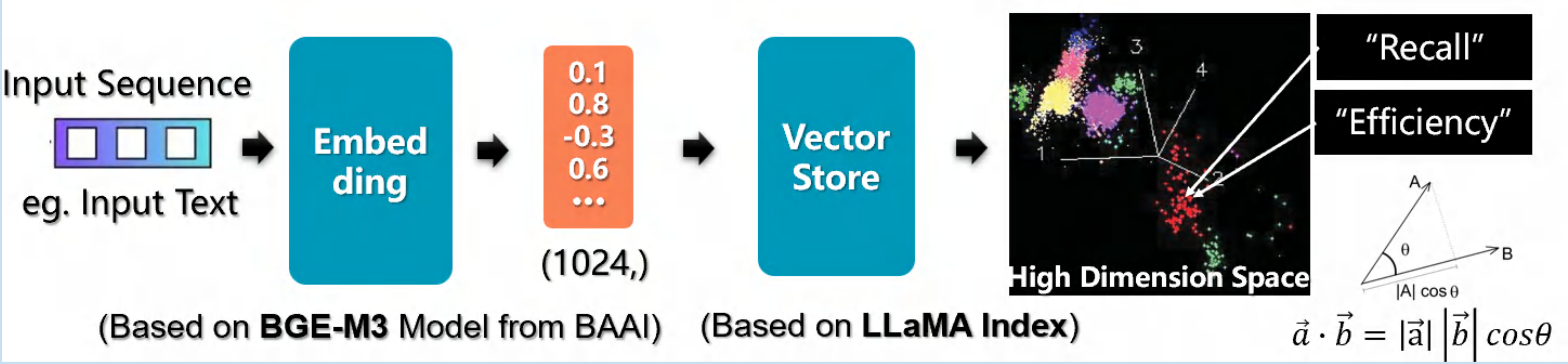
Brain

- Cleaned and evaluated papers arXiv
  - PDF files parser: *HaiNougat*, advanced iteration of the Nougat model
- Good quality chat history from HepAI platform
  - The data is cleaned by human or AI (GPT4)
  - 180k Question-Answer pairs in 3 months
- **BESIII internal data**
  - Internal memo/notes, parsed by HaiNougat
  - Question-Answer pairs from internal paper review
  - BESIII Offline Software System (BOSS) source code
  - BESIII public webpages and internal webpages

https://github.com/ai4hep/hai-nougat

# Memory component of Dr.Sai

Memory component is based on **vector store** and **knowledge graph**

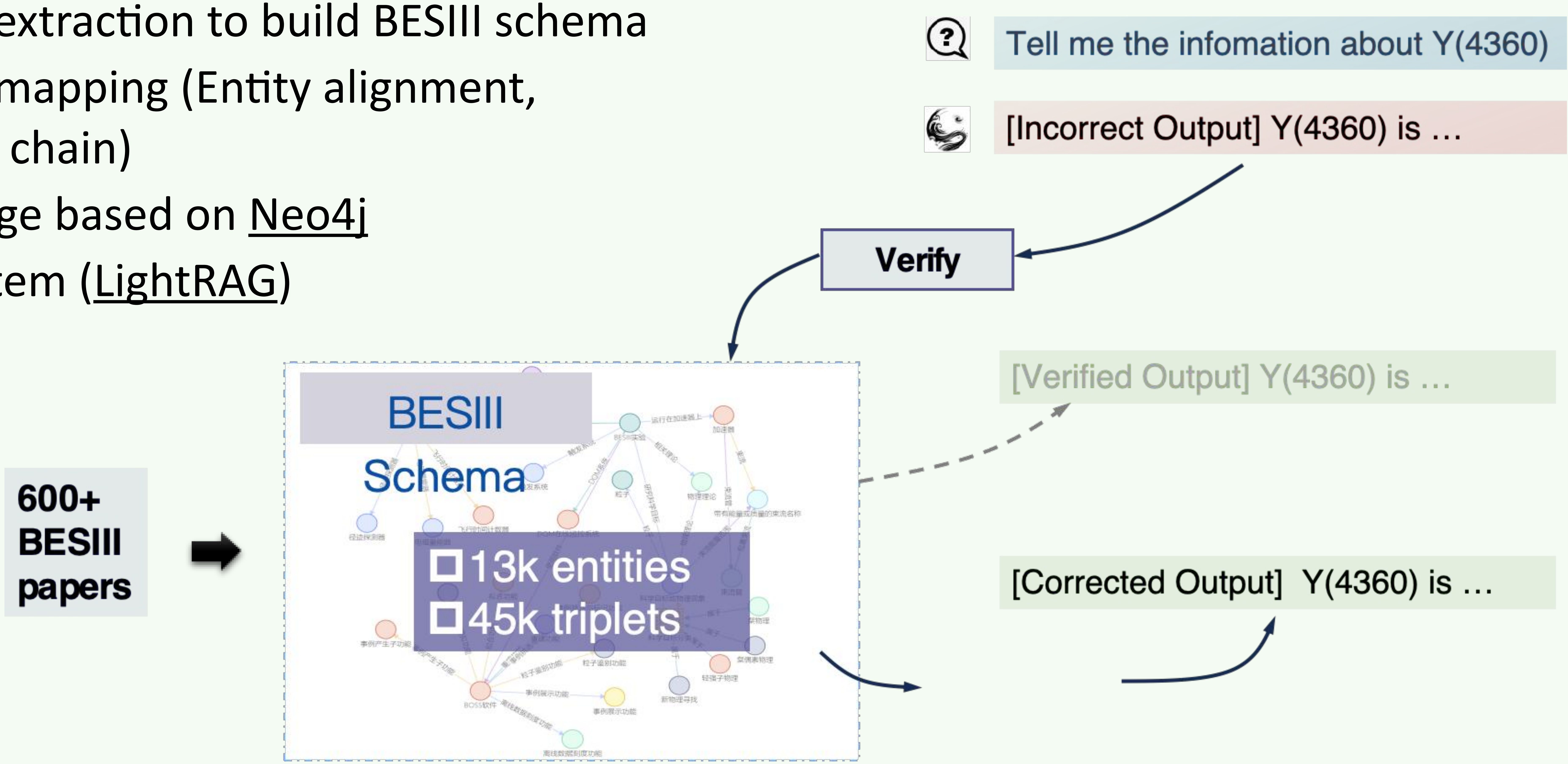

Input Sequence
eg. Input Text
Embed ding
0.1
0.8
-0.3
0.6
•••
(1024,)
Vector Store
3
4
1
2
High Dimension Space
"Recall"
"Efficiency"
A
B
θ
|A| cos θ
(Based on **BGE-M3** Model from BAAI)
(Based on **LLaMA Index**)
$\vec{a} \cdot \vec{b} = |\vec{a}| \left|\vec{b}\right| cos\theta$

- Embeddings (BGE-M3 model from BAAI)
- RAG system (LlamaIndex)
- Unified metadata and support of fragment segmentation and full-text search

# Memory component of Dr.Sai

Memory component is based on **vector store** and **knowledge graph**

- Knowledge extraction to build BESIII schema
- Knowledge mapping (Entity alignment, relationship chain)
- Graph storage based on Neo4j
- KG-RAG system (LightRAG)



Tell me the infomation about Y(4360)

[Incorrect Output] Y(4360) is …

**Verify**

**600+ BESIII papers**

BESIII Schema

☐13k entities

☐45k triplets

[Verified Output] Y(4360) is …

[Corrected Output]  Y(4360) is …

*Y(4360): an exotic particle with 4.36 GeV/c2 mass*

# From answer to action: HepAI-DDF

Actuators

**High-performance AI Distributed Deployment Framework**

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

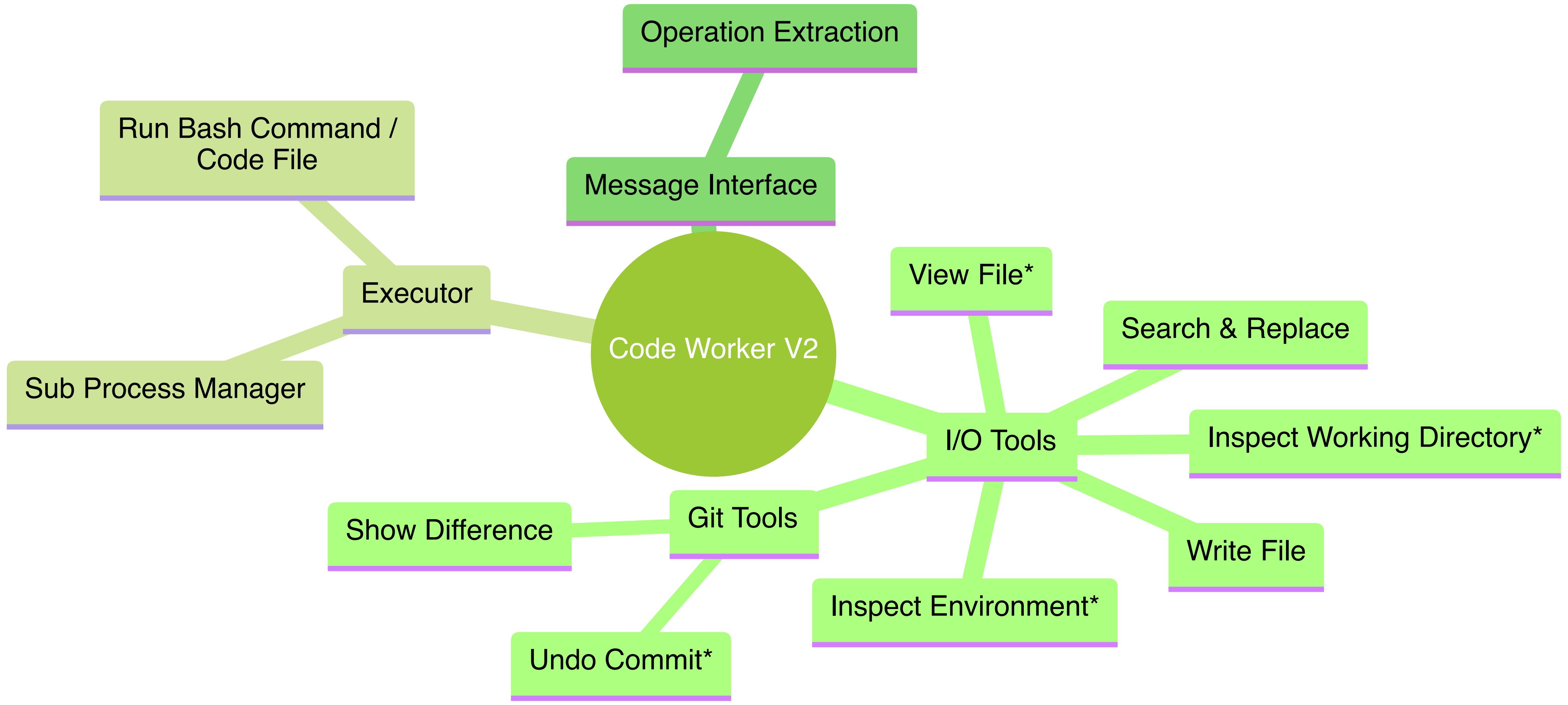
** Generated by AI*

✦ 10+ models
✦ 2 local tools
✦ 2600+ MCP tools
✦ In one server

# From answer to action: BESIII code worker

**Actuators**

**BESIII code worker**



- Code Worker V2
  - Message Interface
    - Operation Extraction
  - Executor
    - Run Bash Command / Code File
    - Sub Process Manager
  - I/O Tools
    - View File*
    - Search & Replace
    - Inspect Working Directory*
    - Write File
    - Inspect Environment*
    - Git Tools
      - Show Difference
      - Undo Commit*

Run on HPC Cluster

⇄

**HaiDDF**

⇄

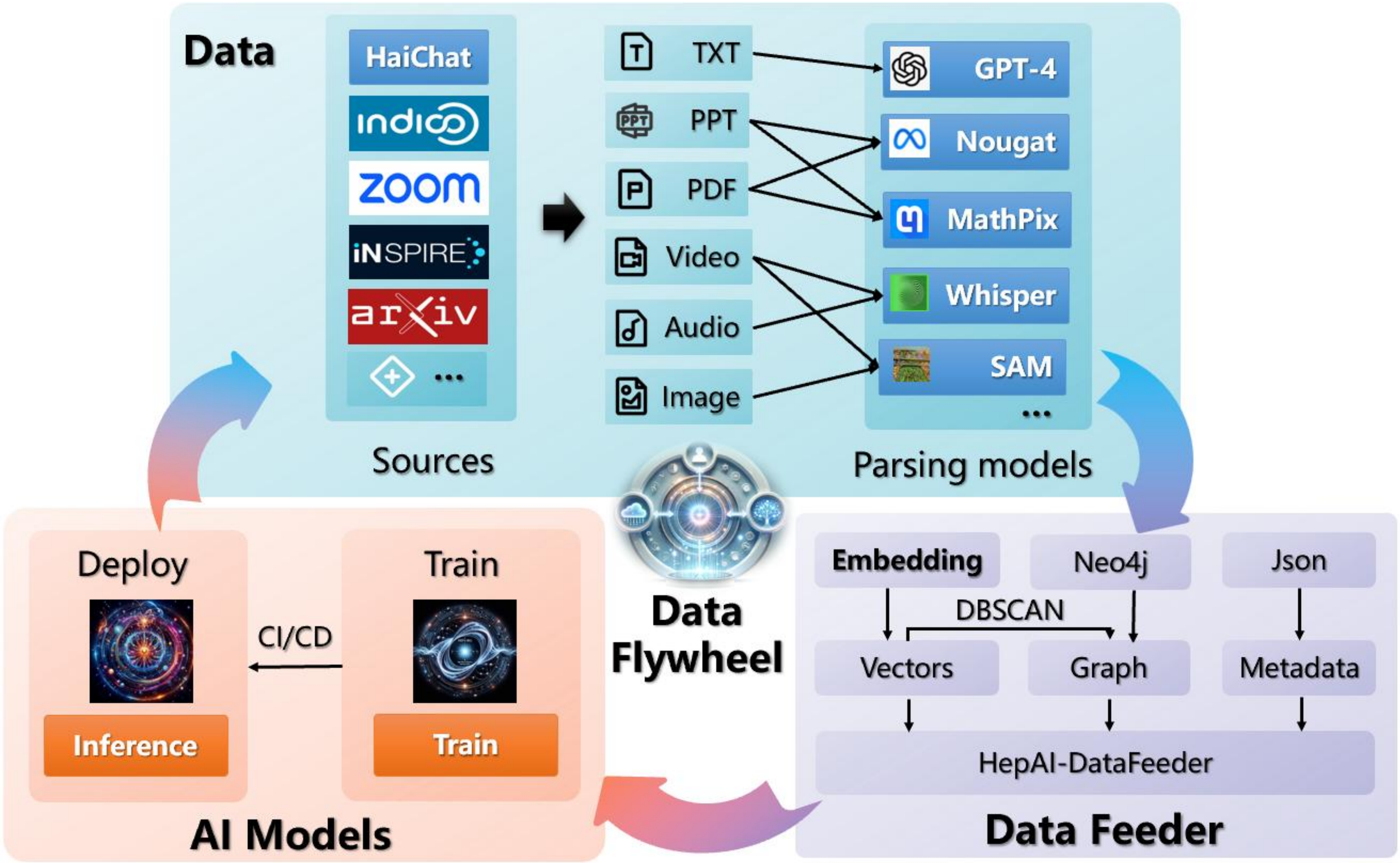
**Dr.Sai**

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

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

# Data flywheel system

**Sensors**

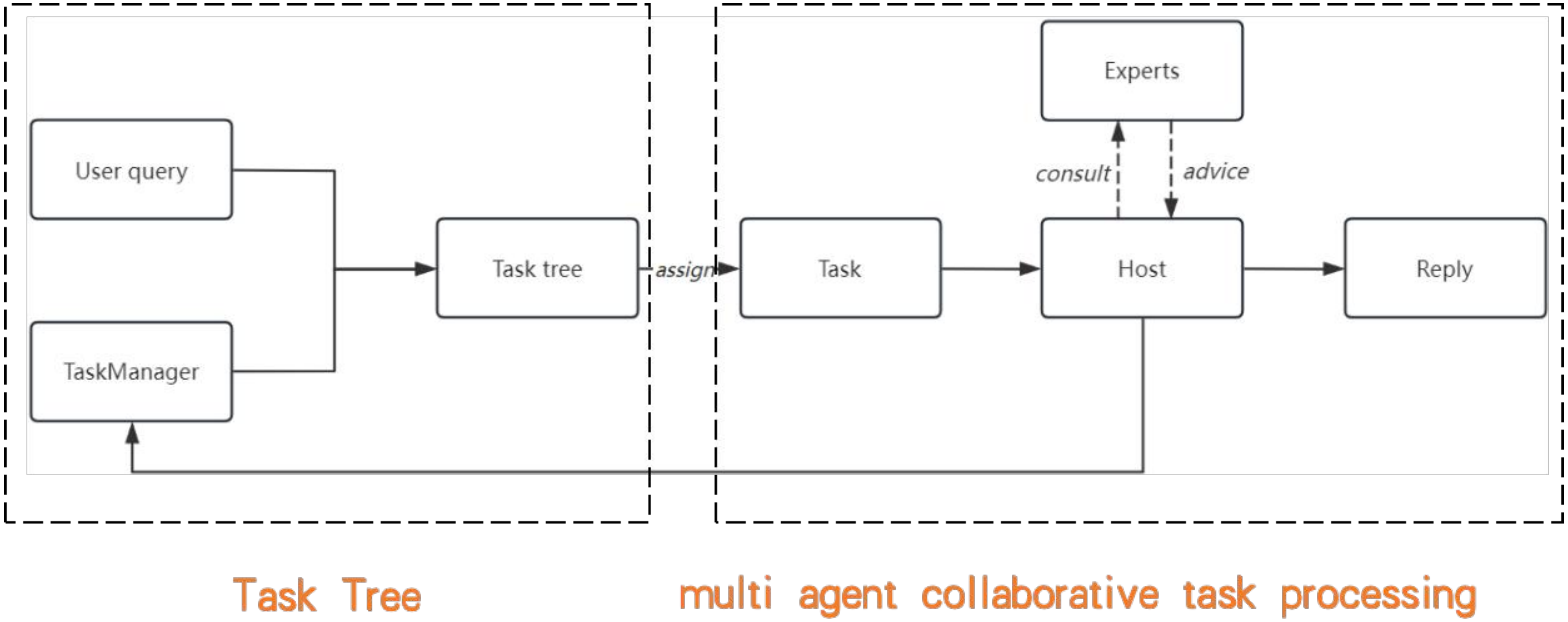


**The "data flywheel" enables continuous model iteration and evolution**

- Improves models by constructing circular data pathways
- With more collected data for cleaning and labeling, further enhance the models

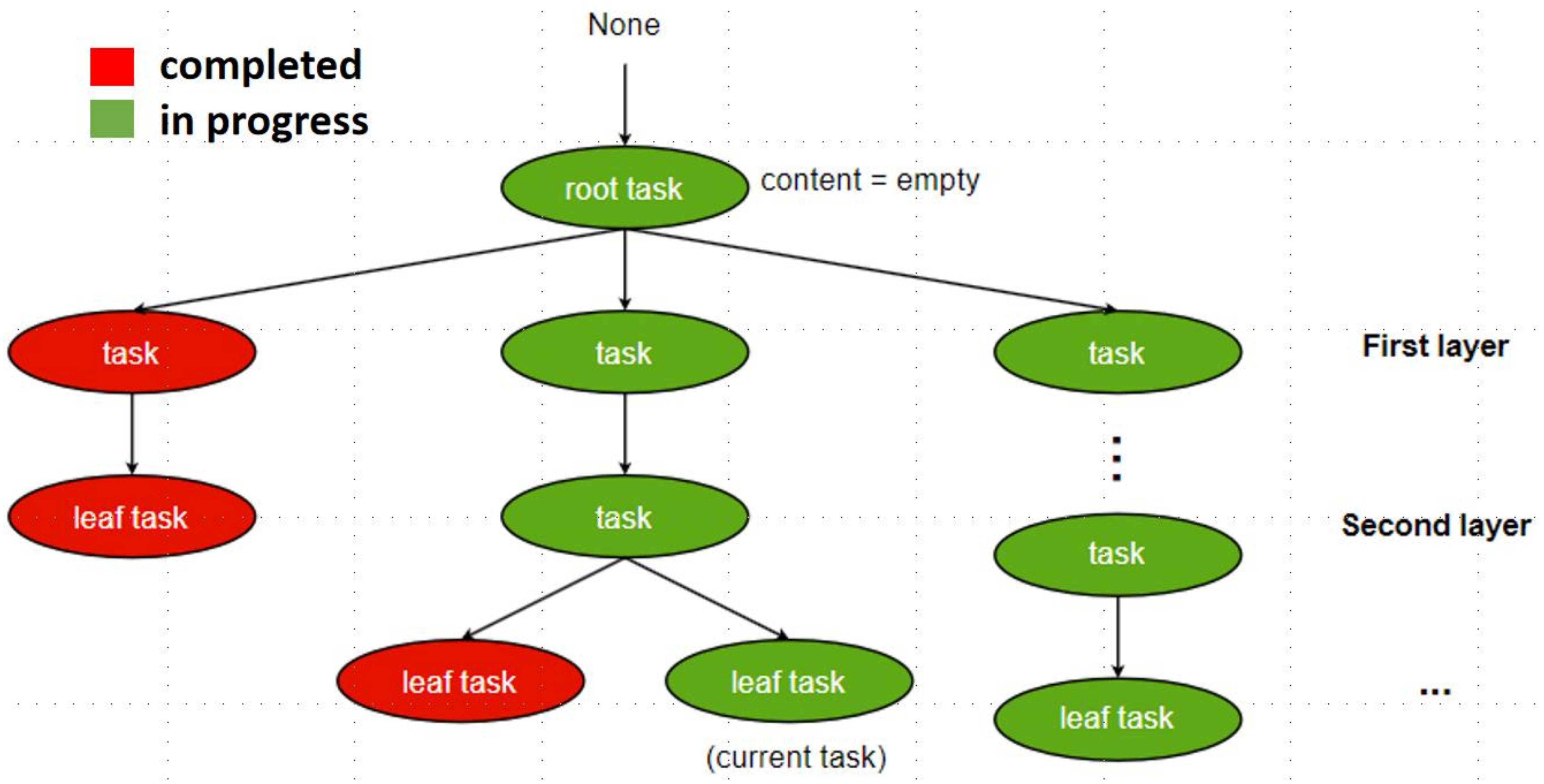
# MAS optimization: Collaboration mode

- Based **Talker-Reasoner** agent mode
- Multiple rounds of inquiry to achieve task execution results reflection
- Upgrade to **Host** led multi-agent collaboration architecture, with direct response to routine problems and intelligent allocation of complex tasks

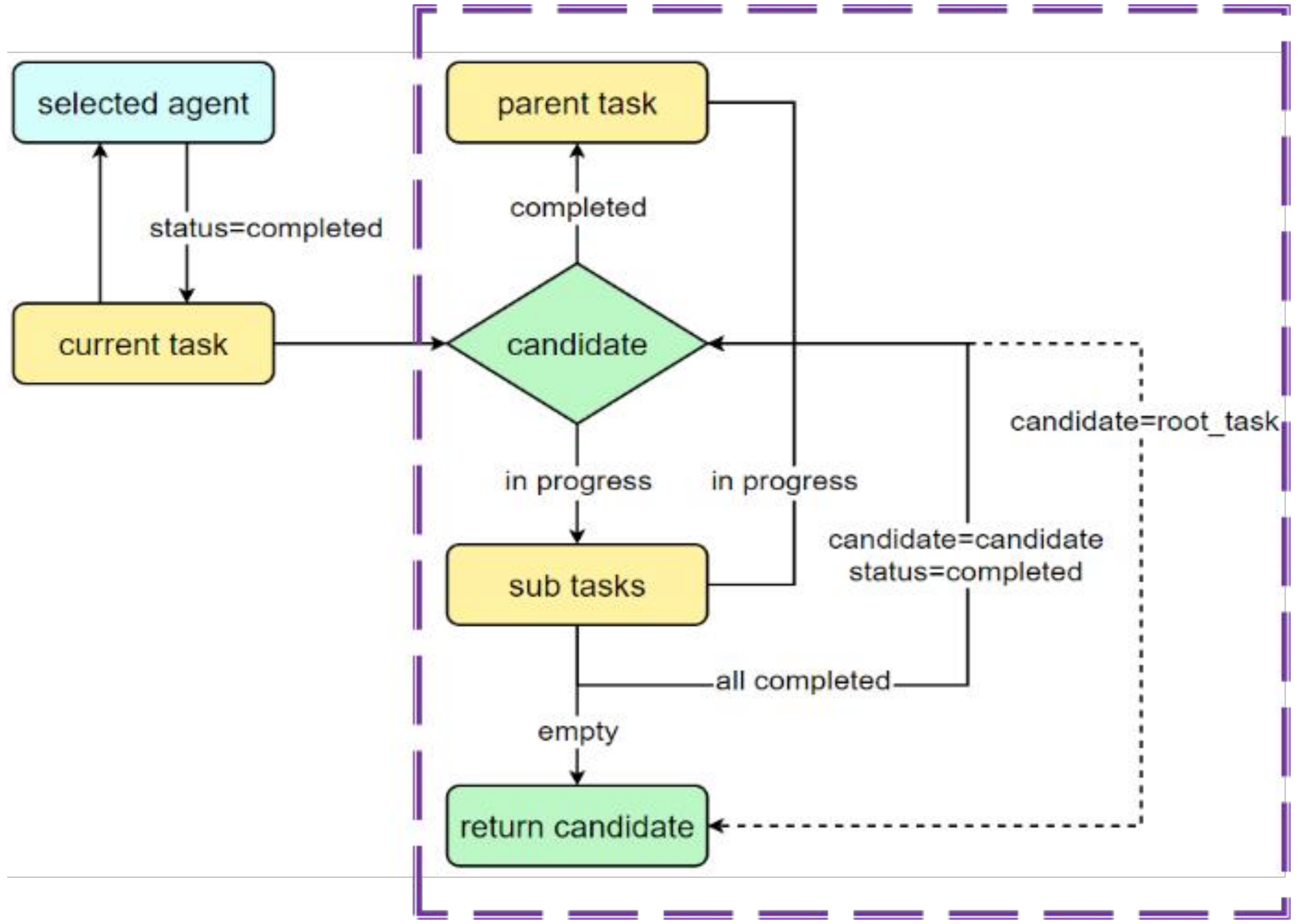


# MAS optimization: Task Tree

- Planner divide a user-query (parent task) to sub tasks, and use **Hierarchical** task model
- Adopt the **task tree** to manage complex physics analysis tasks
- Fractal design ensures a closed-loop logic for task backtracking
- Implement task tree addition, deletion, modification, and query



**Task Tree structure**

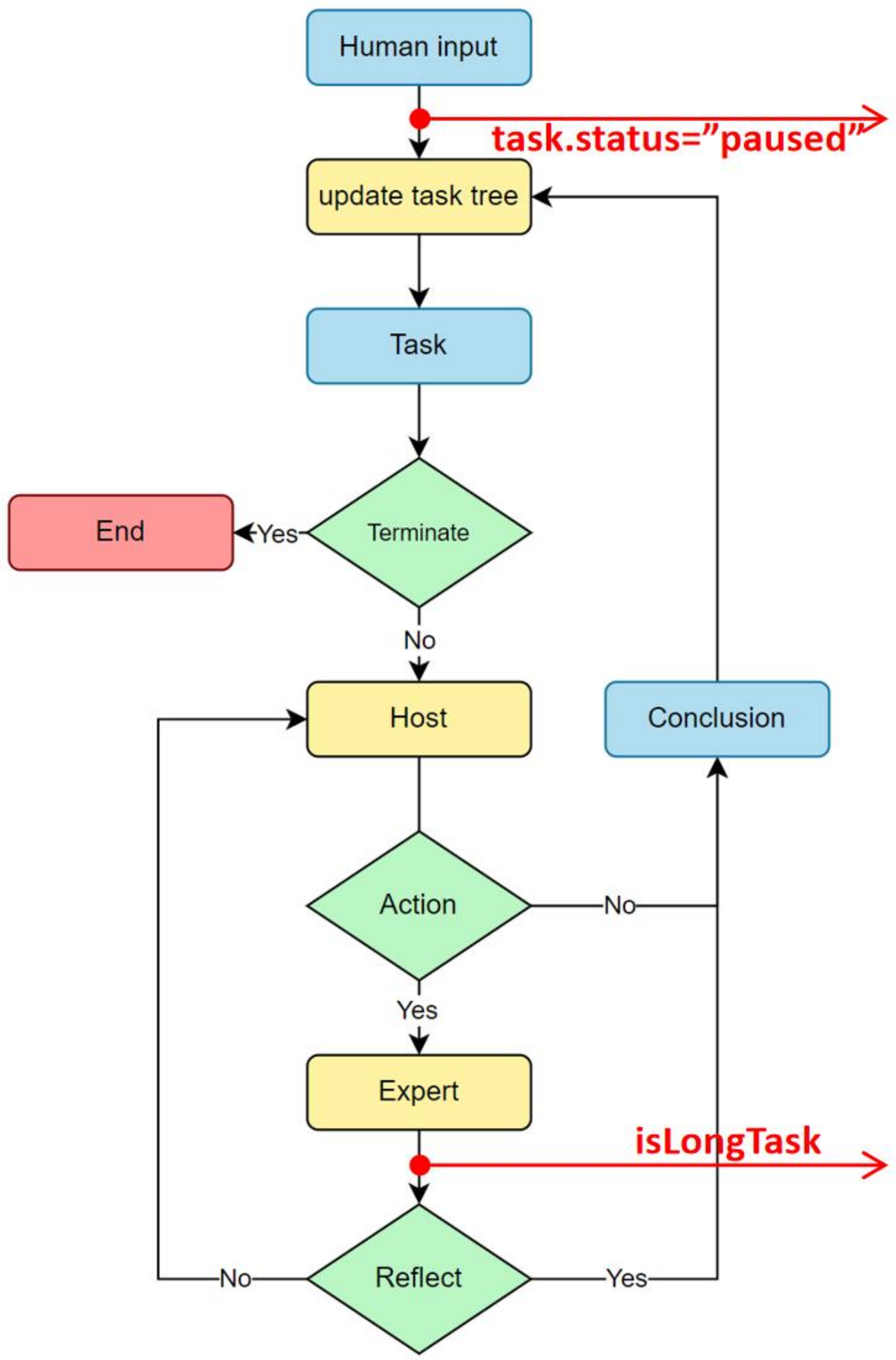


**Task retrieval logic**

# MAS optimization: Task Tree

- **Also support long-task execution**

*Long-task: tasks with long execution time*



**recover**

根据用户输入进行行动

- "status"：调用Tester查询任务状态
- "continue"：更新任务状态为"completed"，顺序执行旧任务系统中的下一个任务
- other：提示存在未完成长任务，退出groupchat

**pending**

1. 设置task.status="paused"
2. 文本处理：提取jobkey，延展reply内容
3. 更新groupchat
4. 结束groupchat

**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)

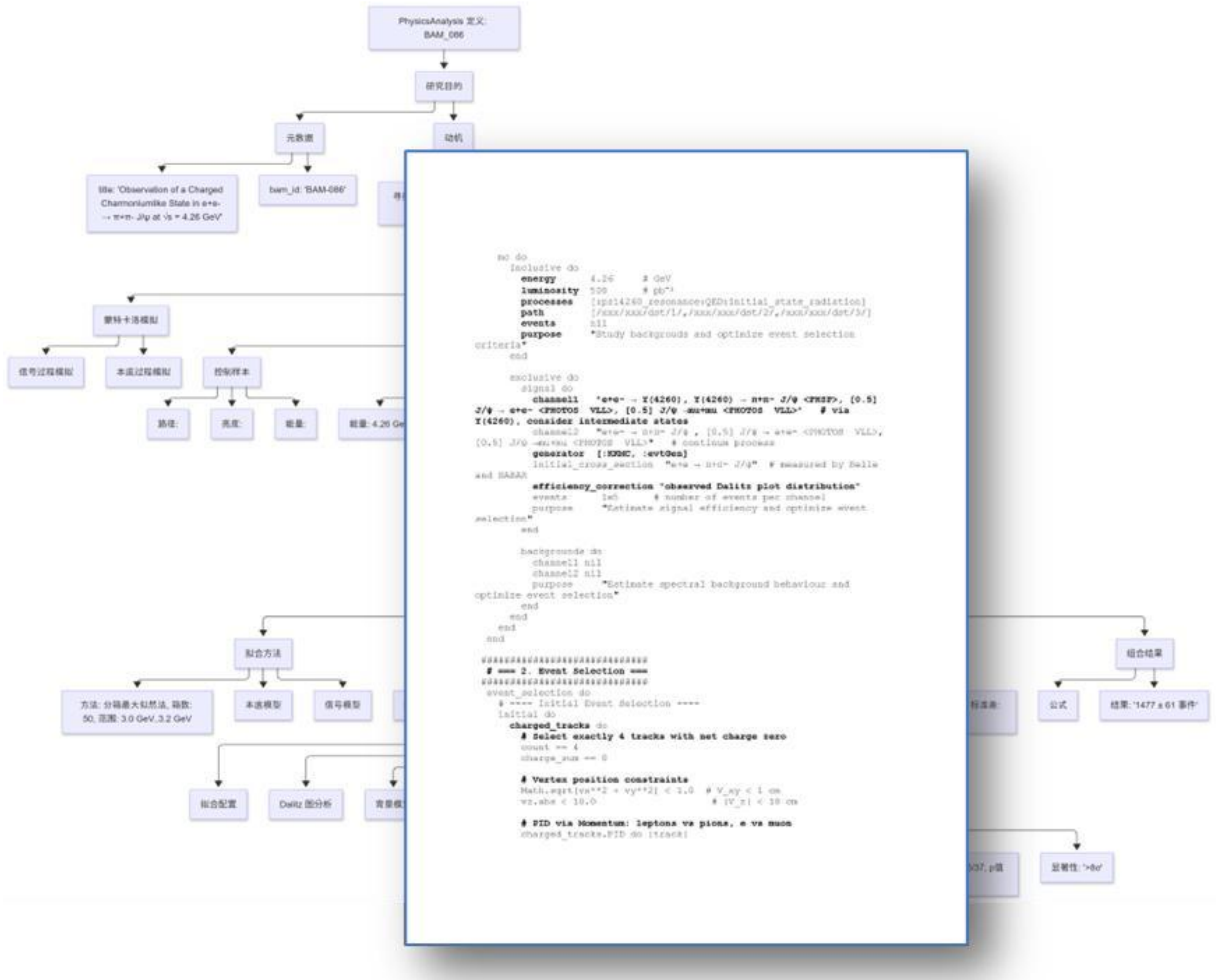
Q: <other query>

A: You have unfinished long tasks!

>> exit groupchat

# MAS optimization: Domain-specific language

- Current LLM do not know the HEP data analysis procedures and do not understand the logics
- We can interpret the analysis to a **Domain-specific language (DSL)**
  - Define each step of analysis in sequence, so LLM can “understand” the procedure
  - 600+ BESIII published paper → transfer to DSL manually
- **DSL is served as a guide to Dr.Sai-BESIII**
  - Stored into RAG; find the DSL for the analysis similar to user's target analysis and take it as reference

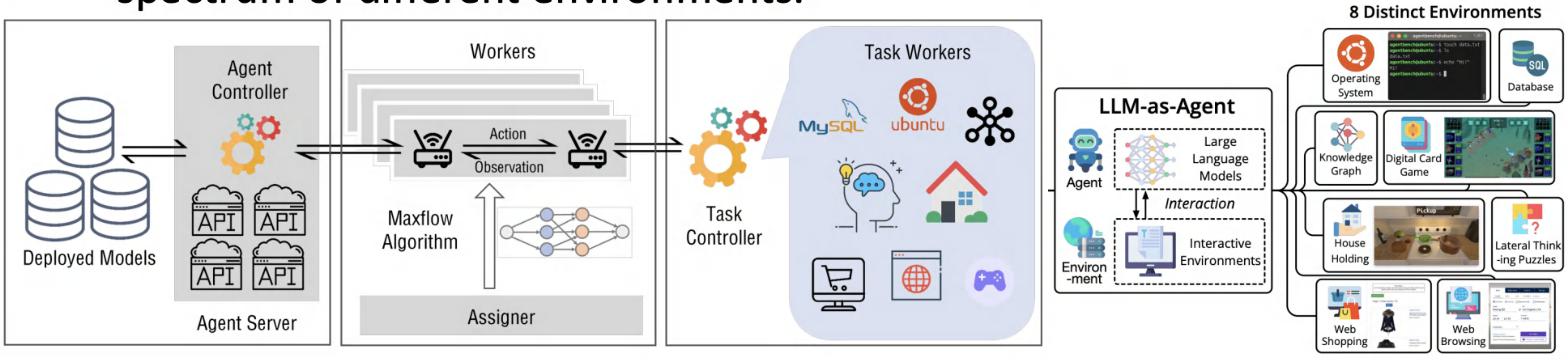


```
    mc do
      inclusive do
        energy      4.26     # GeV
        luminosity  500      # pb^-1
        processes   [:psi4260_resonance:QED:initial_state_radiation]
        path        [/xxx/xxx/dst/1/,/xxx/xxx/dst/2/,/xxx/xxx/dst/3/]
        events      nil
        purpose     "Study backgrouds and optimize event selection
criteria"
      end

      exclusive do
        signal do
          channel1   'e+e- → Y(4260), Y(4260) → π+π- J/ψ <PHSP>, [0.5]
J/ψ → e+e- <PHOTOS  VLL>, [0.5] J/ψ →mu+mu <PHOTOS  VLL>'   # via
Y(4260), consider intermediate states
          channel2   "e+e- → π+π- J/ψ , [0.5] J/ψ → e+e- <PHOTOS  VLL>,
[0.5] J/ψ →mu+mu <PHOTOS  VLL>"   # continue process
          generator  [:KKMC, :evtGen]
          initial_cross_section  "e+e → π+π- J/ψ"  # measured by Belle
and BABAR
          efficiency_correction "observed Dalitz plot distribution"
          events      1e5      # number of events per channel
          purpose     "Estimate signal efficiency and optimize event
selection"
        end

        backgrounds do
          channel1 nil
          channel2 nil
          purpose     "Estimate spectral background behaviour and
optimize event selection"
        end
      end
    end
  end

 ##############################
 # === 2. Event Selection ===
 ##############################
  event_selection do
    # ---- Initial Event Selection ----
    initial do
      charged_tracks do
        # Select exactly 4 tracks with net charge zero
        count == 4
        charge_sum == 0

        # Vertex position constraints
        Math.sqrt(vx**2 + vy**2) < 1.0   # V_xy < 1 cm
        vz.abs < 10.0                    # |V_z| < 10 cm

        # PID via Momentum: leptons vs pions, e vs muon
        charged_tracks.PID do |track|
```

# Performance of Dr.Sai-BESIII

## Introduction to AGENTBEENCH

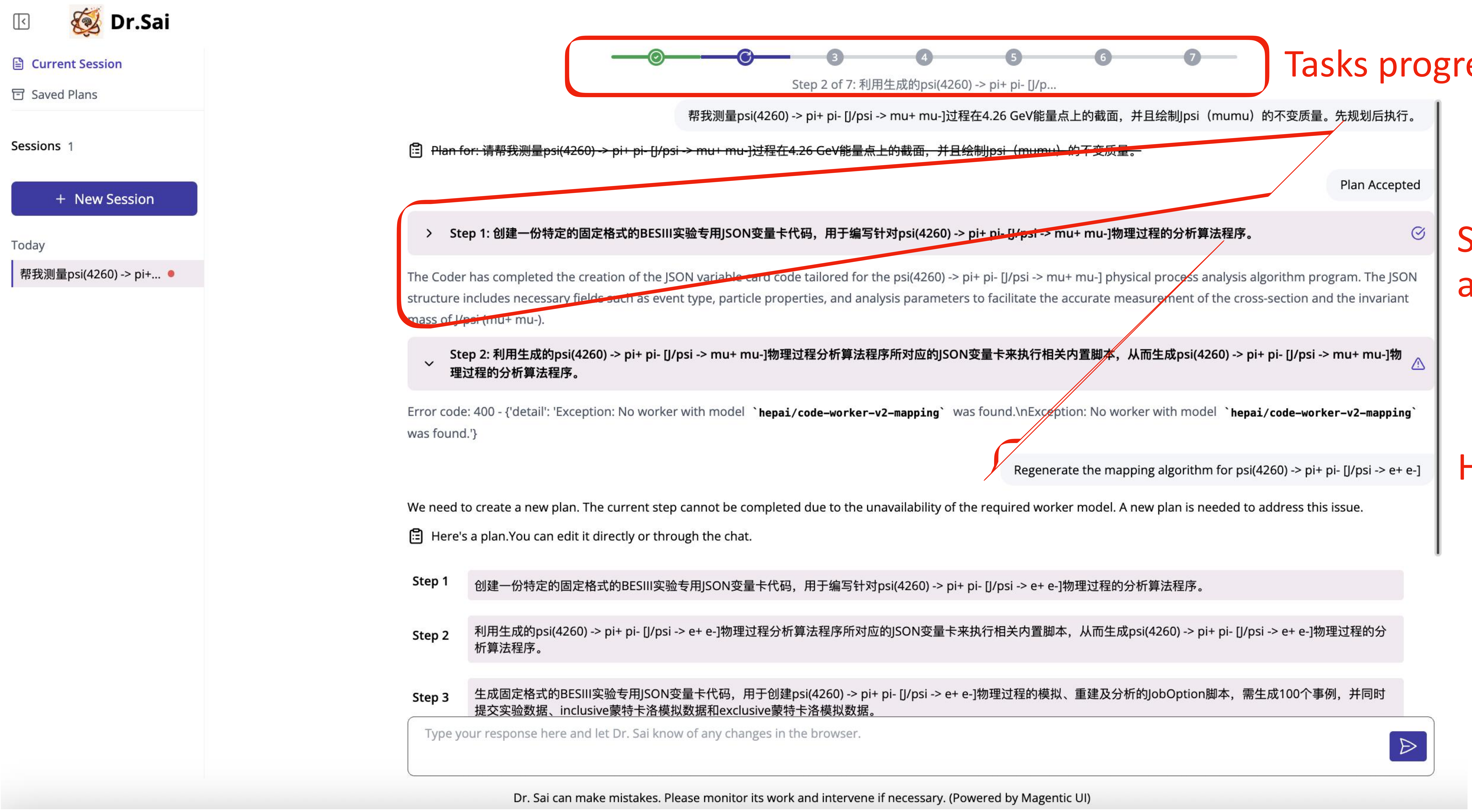
- a first benchmark designed to evaluate LLM-as-Agent across a diverse spectrum of different environments.



- Decouple framework design:
  - Agent controller: 4 agents in Dr.Sai.
  - Worker controller: run plenty of workers at the same time.
  - Task controller: run different task at the same time.
- Supports parallel testing for Multi-Agent and Multi-Task scenarios
- Could be used to Dr.Sai evaluation.

# Update: Magentic-UI

Human can access and interrupt at any time based on Magentic-UI



Step 2 of 7: 利用生成的psi(4260) -> pi+ pi- [J/p...

帮我测量psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]过程在4.26 GeV能量点上的截面，并且绘制Jpsi（mumu）的不变质量。先规划后执行。

~~Plan for: 请帮我测量psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]过程在4.26 GeV能量点上的截面，并且绘制Jpsi（mumu）的不变质量。~~

Plan Accepted

**Step 1: 创建一份特定的固定格式的BESIII实验专用JSON变量卡代码，用于编写针对psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程的分析算法程序。**

The Coder has completed the creation of the JSON variable card code tailored for the psi(4260) -> pi+ pi- [J/psi -> mu+ mu-] physical process analysis algorithm program. The JSON structure includes necessary fields such as event type, particle properties, and analysis parameters to facilitate the accurate measurement of the cross-section and the invariant mass of J/psi (mu+ mu-).

**Step 2: 利用生成的psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程分析算法程序所对应的JSON变量卡来执行相关内置脚本，从而生成psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程的分析算法程序。**

Error code: 400 - {'detail': 'Exception: No worker with model `hepai/code-worker-v2-mapping` was found.\nException: No worker with model `hepai/code-worker-v2-mapping` was found.'}

Regenerate the mapping algorithm for psi(4260) -> pi+ pi- [J/psi -> e+ e-]

We need to create a new plan. The current step cannot be completed due to the unavailability of the required worker model. A new plan is needed to address this issue.

Here's a plan.You can edit it directly or through the chat.

**Step 1** 创建一份特定的固定格式的BESIII实验专用JSON变量卡代码，用于编写针对psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程的分析算法程序。

**Step 2** 利用生成的psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程分析算法程序所对应的JSON变量卡来执行相关内置脚本，从而生成psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程的分析算法程序。

**Step 3** 生成固定格式的BESIII实验专用JSON变量卡代码，用于创建psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程的模拟、重建及分析的JobOption脚本，需生成100个事例，并同时提交实验数据、inclusive蒙特卡洛模拟数据和exclusive蒙特卡洛模拟数据。

Type your response here and let Dr. Sai know of any changes in the browser.

Dr. Sai can make mistakes. Please monitor its work and intervene if necessary. (Powered by Magentic UI)

Tasks progress

Single task's status and output/error

Human interruption