



中國科學院高能物理研究所  
Institute of High Energy Physics  
Chinese Academy of Sciences

# LLM-based human-AI collaboration for HEP - '**Dr. Sai**'

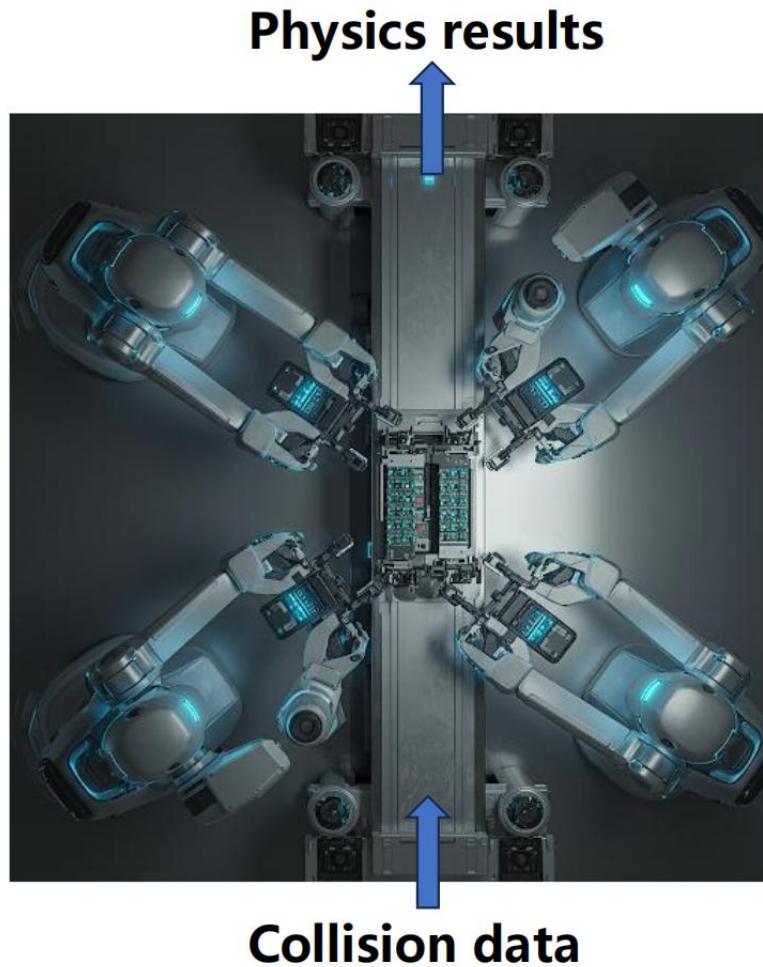
Ke Li ([like@ihep.ac.cn](mailto:like@ihep.ac.cn))

on behalf of Dr.Sai group

Institute of High Energy Physics, China

# Outline

- Motivation
- Introduction of BESIII
- Dr.Sai project
- Methodology
- Status
- Prospects



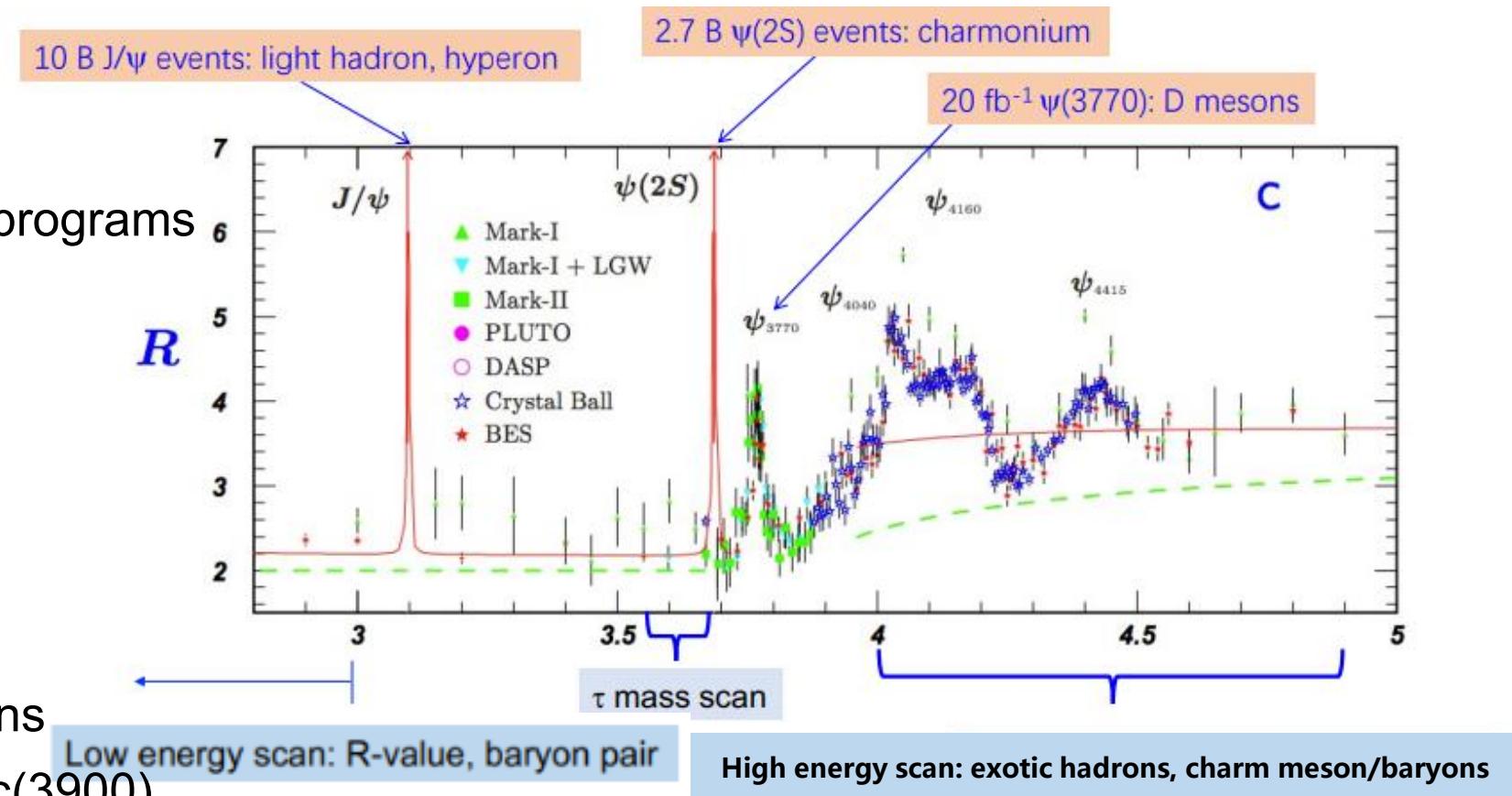
**Goal:**  
**A virtual**  
**“robots” to**  
**work on HEP**  
**data analysis**

# Motivation

- Physics analysis at HEP experiment become more and more complex
  - Big data (normally PB-EB), lots of data processing and checks ...
- Lots of **human-computer interactions**
  - Many tasks can be regarded as text/code generation
  - **LLM is good at text/code generation**
- We need an AI system which "understand" HEP knowledge (how to do physics analysis, how to deal with the tools/codes, etc. )
  - The key is **how to model the HEP knowledge, such as physics analysis**
  - **Start from lepton collider experiment (BESIII) where the analysis is relatively simpler**

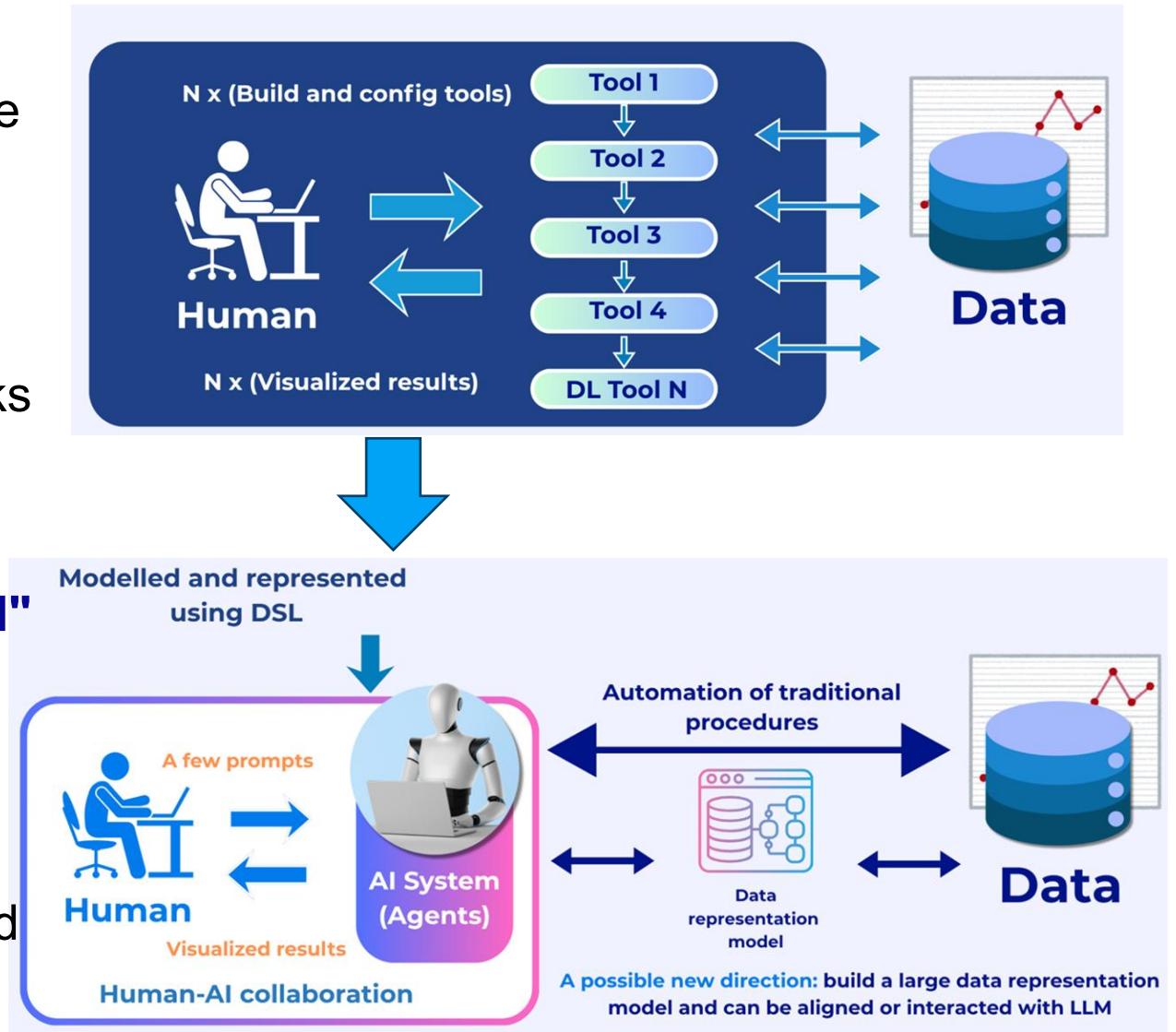
# Introduction of BESIII - physics program

- >700 scientists and engineers
- Tau-charm factory, rich physics programs
  - Light hadrons
  - Charm meson/baryons
  - Charmonium
  - Precise test of SM
  - Search for new physics
- Hundreds of physics results
  - Discovered >30 new hadrons
    - First tetraquark state: Zc(3900)
  - **Good for analysis modelling**



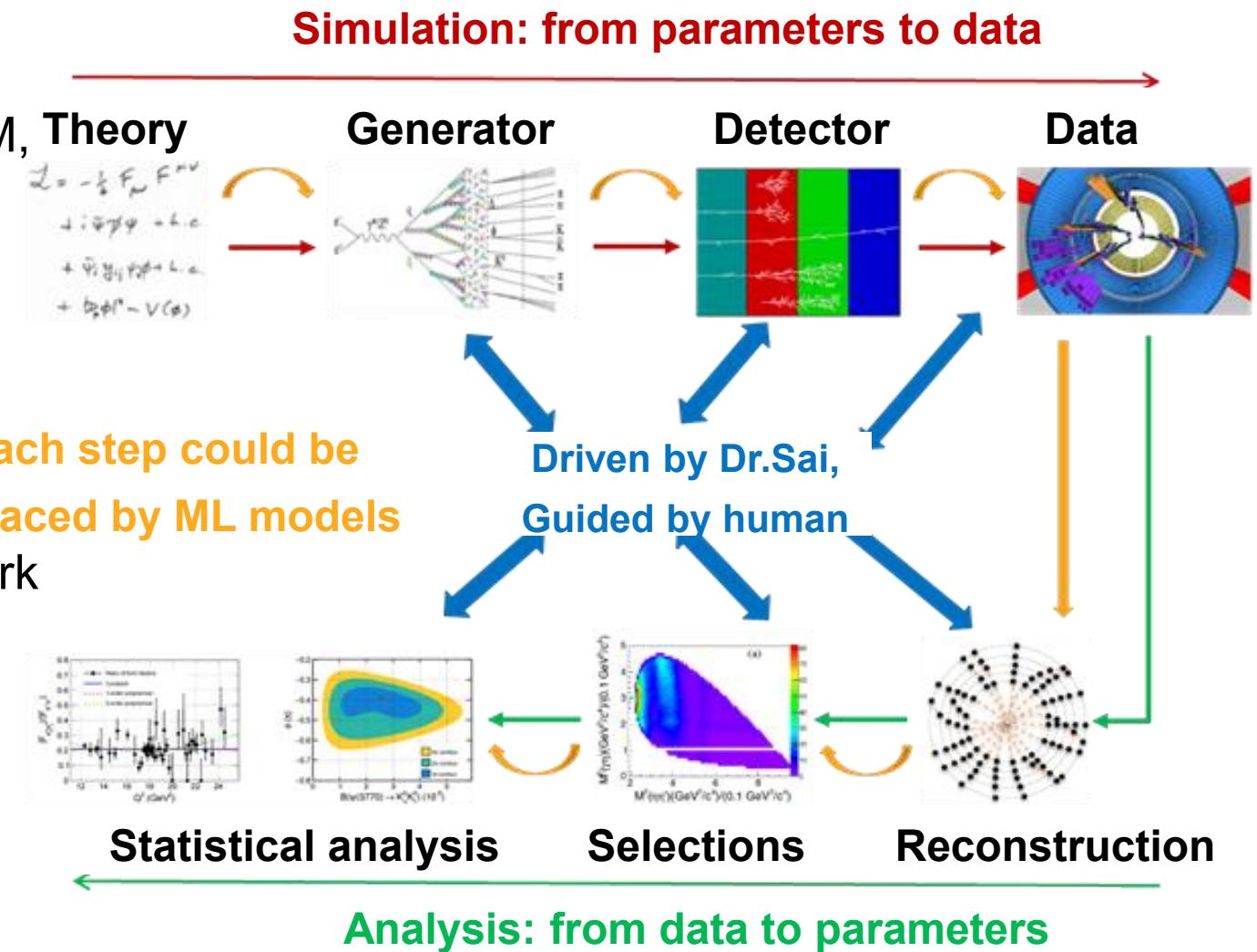
# How LLM can help

- LLM is good at **text/code generation**, but the rules in natural languages is different from HEP data, LLM can not touch data directly
- A new paradigm: AI assist scientists in reasoning, planning, and executing HEP tasks under human supervision
- Key challenge: LLM is not magic, it is just statistics, **how we can let LLM "understand" our "knowledge" ?**
- Possible solution:
  - Develop a Domain-Specific-Language (DSL) to represent and store HEP-related knowledge, i.e. how to do analysis



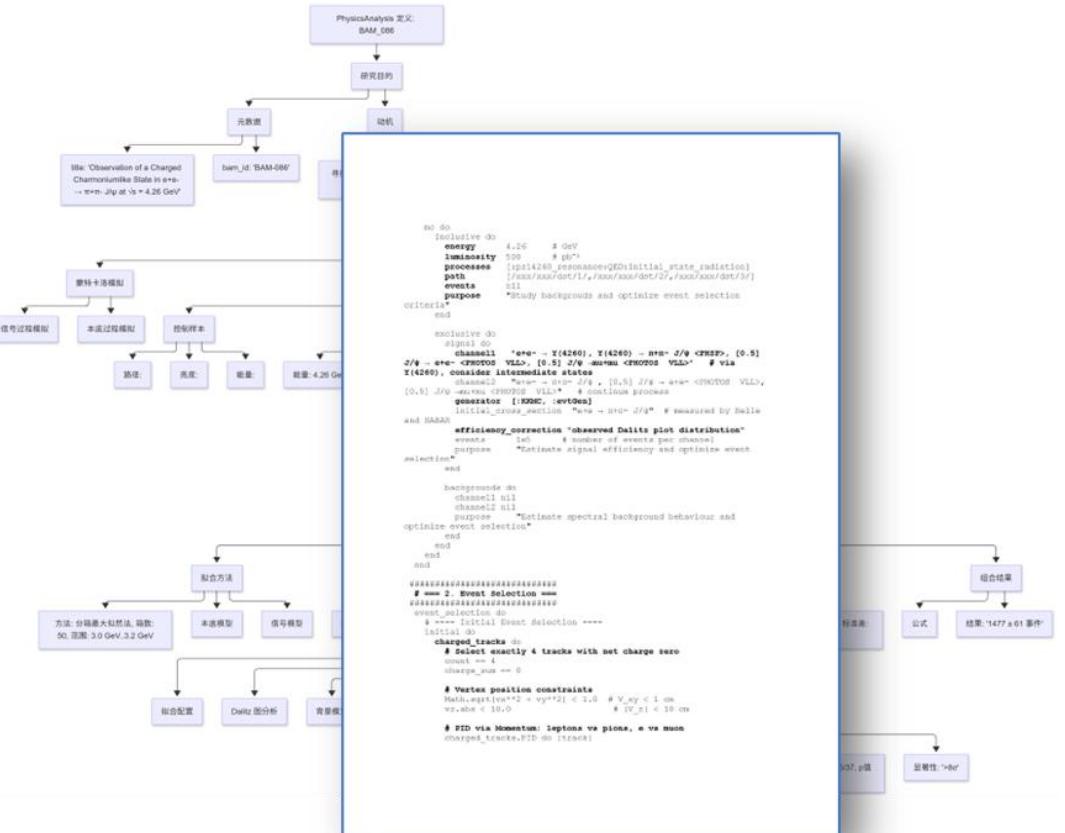
# Human-AI collaboration: Dr. Sai Project

- A multi-agents system based on LLM, Theory aim to **automate the HEP data analysis**
  - AI agent  $\approx$  LLM + Tools + knowledge
  - Developed on [autogen](#) framework
- LLM is switchable:
  - Now using DeepSeek V3/R1
- Analysis tools are integrated



# First attempts of analysis modelling

- 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 the LLM can "understand" the procedure
  - BESIII has published >600 physics results
  - We have to translate them to DSL manually now
- DSL is served as a guide to Dr. Sai
  - Dr.Sai will find the DSL for the analysis similar to the user's target analysis and take it as reference



# Preparation of DSLs

- The DSL can be converted from BESIII papers by VLM efficiently
- We plan to freeze the DSL version soon, then launch bulk production

**III. EVENT SELECTION AND DATA ANALYSIS**

Each  $\psi(3686) \rightarrow \omega K^+ K^-$ ,  $\omega \rightarrow \pi^+ \pi^- \pi^0$  candidate has four good charged tracks with zero net charge and at least two good photon candidates. A good charged track is required to satisfy track fitting and pass within 10 cm of the interaction point in the beam direction and within 1 cm in the plane perpendicular to the beam. Furthermore, it is required to lie within the angular coverage of the MDC, i.e.  $|\cos \theta| < 0.93$  in the laboratory frame, where  $\theta$  is the polar angle.

For photon candidates, the shower energy should be greater than 25 MeV in the barrel region and 50 MeV in the end-cap regions, where the barrel is defined as  $|\cos \theta| < 0.8$  and the end-cap regions as  $0.86 < |\cos \theta| < 0.92$ . Also the average time of the hit EMC crystals with respect to the event start time should be between 0 and 700 ns to suppress electronic noise and background hits. The angle between the direction of a photon candidate and any charged track is required to be greater than  $20^\circ$  to avoid showers caused by charged tracks.

The TOF and  $dE/dx$  information are combined for each charged track to calculate the particle identification probability ( $P_i$  with  $i = \pi, K$ ) of each particle-type hypothesis. For a pion candidate,  $P_\pi > 0.001$  and  $P_\pi > P_K$  are required, while for a kaon candidate  $P_K > 0.001$  and  $P_K > P_\pi$  are required.

A vertex fit is performed assuming all charged tracks are from the IP. A four-constraint (4C) energy-momentum-conserving kinematic fit is performed. If there are more than two photon candidates, we loop over all possible combinations, and the combination with the minimum 4C  $\chi^2$  is kept for further analysis. The invariant mass of the photon pair is required to be in the range  $0.11 < M_{\gamma\gamma} < 0.15$  GeV/ $c^2$ . Then a 5C kinematic fit is performed with the invariant mass of the two photons constrained to the mass of  $\pi^0$ , and  $\chi^2 < 90$  is required, which is based on the optimization of the figure of merit (FOM),  $FOM = N_{sg}/\sqrt{N_{sg} + N_{bg}}$ , where  $N_{sg}$  and  $N_{bg}$  are the numbers of signal and background events estimated by the inclusive MC, respectively.

Converted  
by LLM

DSL

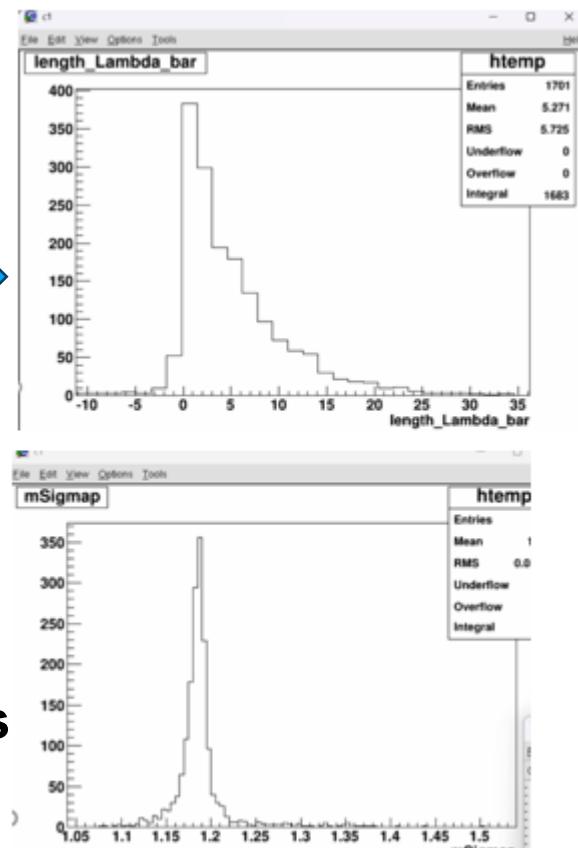
```
event_selection = Selection.new
event_selection.select_track {
  cos_theta 0.93 # |cosθ| < 0.93
  Vz 10.0 # cm (beam direction)
  Vr 1.0 # cm (transverse plane)
  nChp "==" # exactly 2 positive tracks
  nChn "==" # exactly 2 negative tracks
}
.event_selection.select_photon {
  angle_to_track 20.0 # degrees
  energyThreshold_b 0.025 # GeV (barrel: |cosθ| < 0.8)
  energyThreshold_e 0.050 # GeV (endcap: 0.86 < |cosθ| < 0.92)
  tdc_emc_start 0 # ns
  tdc_emc_end 700 # ns
  nGam "==" # at least 2 photons
}
.event_selection.pid(has_lepton = false) {
  prob_cut 0.001
  identify :kp, "from_pion" # K+ ID: P(K) > P(n) & P(K) > 0.001
  identify :km, "from_pion" # K- ID: P(K) > P(n) & P(K) > 0.001
  identify :kp, "from_kaon" # π+ ID: P(π) > P(K) & P(π) > 0.001
  identify :km, "from_kaon" # π- ID: P(π) > P(K) & P(π) > 0.001
  nkp "=" 1 # exactly 1 K+
  nkm "=" 1 # exactly 1 K-
  npip "=" 1 # exactly 1 π+
  npim "=" 1 # exactly 1 π-
}
```

BESIII paper

Source code

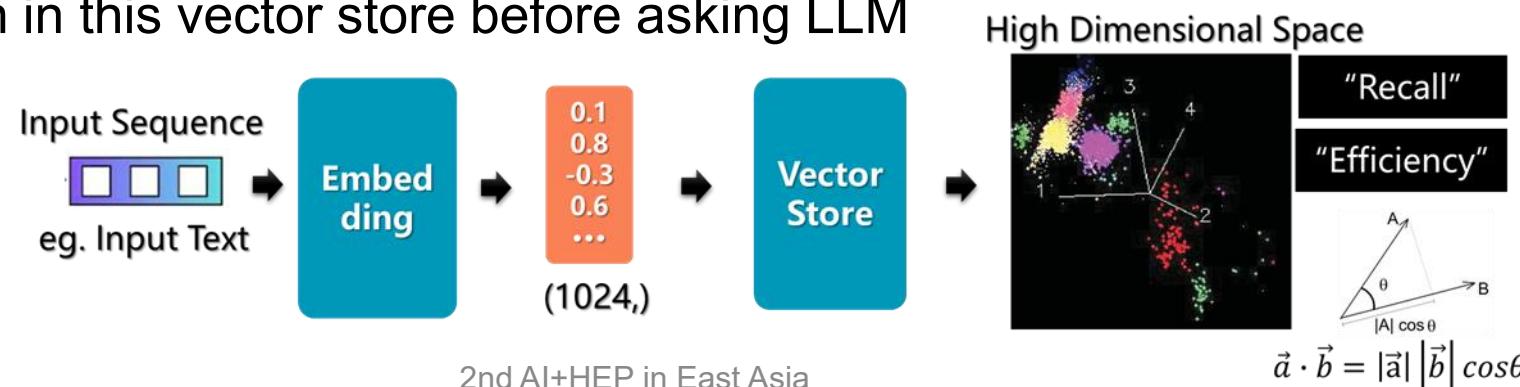
```
event_selection = Selection.new
KinematicFit* kfitt = &KinematicFit::instance();
int ipip_i = 1, ipip_j = 1, ipip_k = 1, ipip_l = 1;
for (size_t i = 0; i < index_ipip.size(); i++) {
  RecMcKallTrack* pip_iTrk = (*evdMcKallTrack.begin() + index_ipip[i]) - mcKallTrack();
  WTrackParameter* wip_iTrk = WTrackParameter::wip_ipip[2], pip_iTrk->getHelloX(), pip_iTrk->getZErrorX());
  for (size_t j = 0; j < index_ipip.size(); j++) {
    RecMcKallTrack* pip_jTrk = (*evdMcKallTrack.begin() + index_ipip[j]) - mcKallTrack();
    WTrackParameter* wip_jTrk = WTrackParameter::wip_ipip[2], pip_jTrk->getHelloX(), pip_jTrk->getZErrorX());
    for (size_t k = 0; k < index_ipip.size(); k++) {
      RecMcShower* g_kTrk = (*evdMcKallTrack.begin() + 16*i) - mcShower();
      for (size_t l = 0; l < index_ipip.size(); l++) {
        if (i == k) continue;
        double mass_of_gamma_gamma = (pdg_iTrk + pdg_kTrk).m();
        if ((0.11 < mass_of_gamma_gamma && mass_of_gamma_gamma < 0.15) && (0.11 < mass_of_gamma_gamma && mass_of_gamma_gamma < 0.15)) continue;
        kmfitt->init();
        kmfitt->AddTrack(0, wip_ipip[iTrk]);
        kmfitt->AddTrack(1, wip_ipip[jTrk]);
        kmfitt->AddTrack(2, 0.0, g_kTrk);
        kmfitt->AddTrack(3, 0.0, g_lTrk);
        kmfitt->AddFourMomentum(0, pdg_cm);
        bool oksq = kmfitt->Fit();
        double chisq = kmfitt->chiSq();
        if (oksq) {
          if (chisq < chisq_4C_kmfitt) {
            chisq_4C_kmfitt = chisq;
            ipip_i = index_ipip[i];
          }
        }
      }
    }
  }
}
```

Plots



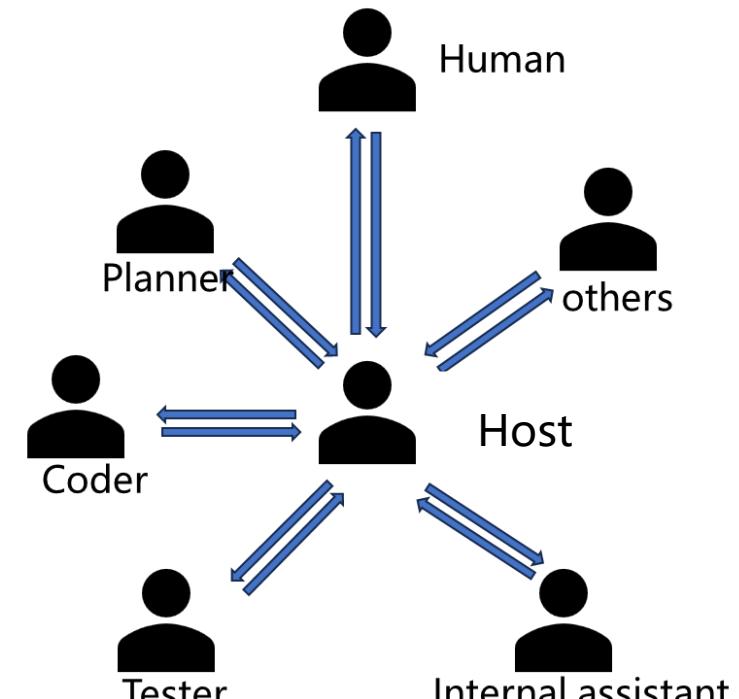
# Memory of Dr. Sai - RAG

- Retrieval-Augmented Generation (RAG)
  - Most-promising solution to suppress hallucinations
- Usage: store BESIII internal data from twiki, webpage, internal docs and reviews of analyses, and DSL
- Current approach: **vector store** (will move to knowledge graph)
  - Embedding models: **BGE-M3** and PhysBert
    - Convert input data into vectors in a multidimensional space
- Dr.Sai will search in this vector store before asking LLM



# Multi-Agents system

- The HEP data analysis is too complex for LLM now
- We can devide the complex task to small and simple task, and develop a **dedicated agent** for each kind of task
- Multi-Agents (foundation model is switchable):
  - **Host**: select correct agent
  - **Planner**: task decomposition
  - **Coder**: code generation
  - **Tester**: testing/execution
  - **Internal assistant**
- Human can chat with Host, then Host chat with other agents
- Each agent could have different LLM and RAG collection
- Support distributed deployment



Preliminary

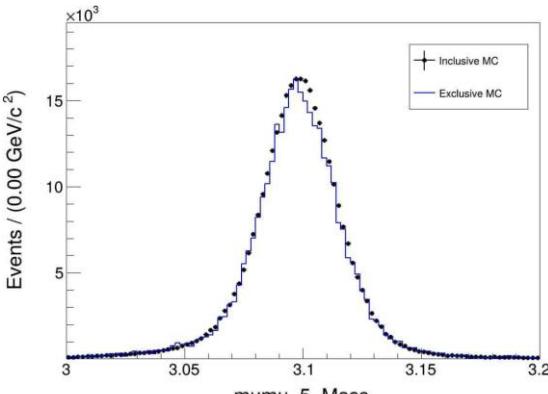
# What can it do now

- Simple analyses, e.g. BR measurement for some channels
  - A preliminary DSL is developed to represent the analysis
  - Some examples are provided
    - So Dr.Sai “know” some info of analysis procedures
  - Analysis workflow is re-organized
    - Predefined multiple modules, configured using json (easily converted from DSL)
    - More than x10 reduction on the amount of code to be generated
    - So Dr.Sai can easily generate the necessary code/configs
  - The analysis tools are integrated
    - So Dr.Sai can do tasks automatically, e.g. job submitting, plotting, simple fitting

# Current pipeline

## User's prompt

Please measure the cross-section of the  $\psi(3770) \rightarrow \pi^+ \pi^-$  process at the 3.686 GeV energy point and plot the invariant mass of  $\psi(3770)(\mu^+ \mu^-)$ . If Monte Carlo generation is required, simulate 10,000 events. Plan first, then execute.



Target distribution  
2026/01/19

## Task decomposition

Planner发言：  
检测到BE-SIM实验室物理分析任务，以下是对代码“psi3770TopBarP”的物理分析流程的默认任务规划：

任务规划：任务描述

- 1 读取物理分析JSon配置卡：指读一整条特定的本地或远端的BE-SIM实验任务配置卡。通过编写脚本psi3770 > pbar p物理分析程序读取分析的JSon配置卡并读取其内的配置参数。从人生地网psi3770 > pbar p物理分析程序读取参数。
- 2 读取物理分析JSon配置卡并分析：根据生成的psi3770 > pbar p物理分析程序读取的JSon配置卡并读取其内的配置参数。从人生地网psi3770 > pbar p物理分析程序读取参数。
- 3 生成JobOptionCard：读取物理分析JSon配置卡：指读一整条特定的本地或远端的BE-SIM实验任务配置卡。通过编写脚本psi3770 > pbar p物理分析程序生成的JSon配置卡并读取其内的配置参数。从人生地网psi3770 > pbar p物理分析程序生成参数。
- 4 执行物理分析JSon配置卡并分析JobOptionCard：通过编写脚本psi3770 > pbar p物理分析程序，重新读取分析JobOptionCard并读取其对应的JSon配置卡并读取其内的配置参数。从人生地网psi3770 > pbar p物理分析程序重新读取JobOptionCard并读取其内的配置参数。
- 5 打印配置参数：选择后台生成配置的一个名称或包含data/ems/lnmc的配置OOT文件（或配置Jedec），执行内部的程序并从屏幕上打印出所有的配置名称。
- 6 生成JobOptionCard：根据物理分析JSon配置卡，通过编写脚本psi3770 > pbar p物理分析程序，仅读取JobOptionCard并读取其内的配置参数的一致性。
- 7 执行物理分析JSon配置卡并分析：利用物理分析的JSon配置卡并读取其内的配置参数。从人生地网部分。
- 8 生成JobOptionCard：根据物理分析JSon配置卡，通过编写脚本psi3770 > pbar p物理分析程序，生成的配置将当成JobOptionCard文件。
- 9 读取物理分析JSon配置卡并分析：通过编写脚本psi3770 > pbar p物理分析程序读取并读取其内的配置参数。从人生地网部分。
- 10 分配CPU资源：自动地将物理分析的JSon配置卡并读取其内的配置参数并进行分配给可以运行的计算机psi3770 > pbar p物理分析程序的Jpsi3770分支党。能地地不要直接引用POD文件，专写物理分析的配置参数的配置文件。
- 11 调用python代码：执行刚刚生成的用于计算分支比的Python代码。用计算Jpsi3770分支比。输出结果。
- 12 在任务完成：执行物理分析任务后psi3770TopBarP的进度。进度也包括每日的进度情况。已执行的工作及成果。面向的对手及资源情况。并建议使用资源。并给予评价。

keywords: psi3770 > pbar p -l 3773 GeV

键入命令以下参数：

- 能量: 3.773 GeV
- 变量: psi3770TopBarP >obar p

## Finish tasks in sequence

# Error monitoring & Reflection

## IEP in East Asia

## Code generation

```

DECLARE_CORPORATE_PSA24060P1P135010MuNsAI

PsA24060P1P135010MuNsAI::PsA24060P1P135010MuNsAI(const std::string name, ISVlocator* pSVlocator) :
    AlgorithmName(name), pSVlocator(pSVlocator)
{
    declareProperty("IsSignalMC", m_IsSignalMC = 1);
    declareProperty("Ecm", m_Ecm = 4.26);
}

40 Status PsA24060P1P135010MuNsAI::initialize()
{
    AlgorithmName(m_IsSignalMC, m_Ecm);
    log << MSG::INFO << "***** in initialize()" << endl;
    StatusCode status;

    StatusCode status;

    NtupleList* nt(ntupleList("FILE/truth"));
    if (nt) m_ntuple_truth = nt;
    else {
        m_ntuple_truth = ntupleList(>>Book("FILE/truth", CLID_ColumnListtuple, "truth information"));
        if (m_ntuple_truth) {
            status = m_ntuple_truth->add("run0", m_run0_truth);
            status = m_ntuple_truth->add("evtn0", m_evt0_truth);
            status = m_ntuple_truth->add("run0", m_run0_truth);
            status = m_ntuple_truth->add("evtn0", m_evt0_truth);

            status = m_ntuple_truth->add("p1_plus_1_p4_truth", 4, m_p1_plus_1_p4_truth);
            status = m_ntuple_truth->add("p1_minus_2_p4_truth", 4, m_p1_minus_2_p4_truth);
            status = m_ntuple_truth->add("mu_3_plus_3_p4_truth", 4, m_mu3_plus_3_p4_truth);
            status = m_ntuple_truth->add("mu_minus_4_p4_truth", 4, m_mu_minus_4_p4_truth);
        }
    }
    else {
        log << MSG::ERROR << "Can not book N-tuple:" << long(m_ntuple_truth) << endl;
        return StatusCode::FAILURE;
    }
}

```

# High performance computing system

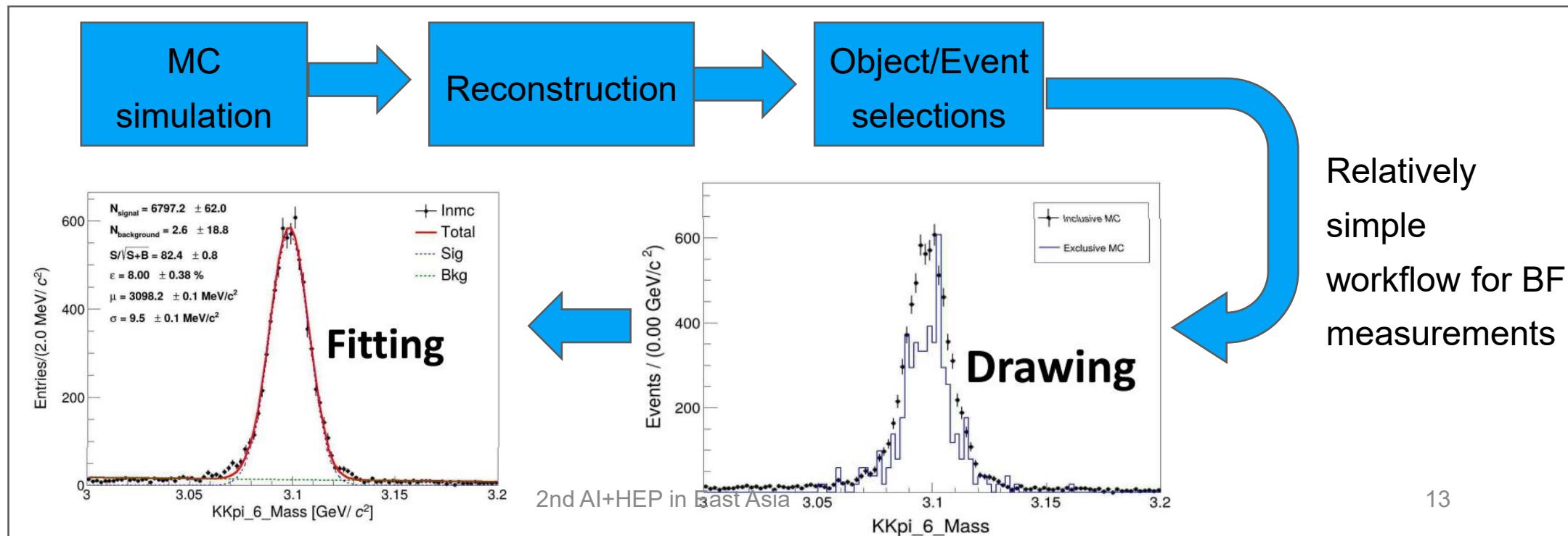
## Execution using htcondor

# One example

- Now Dr.Sai can config/call BESIII analysis tools to do simple analysis automatically
- One test, use the  $\psi(3686) \rightarrow \pi^+ \pi^- J/\psi$  to measure multiple branching fractions of  $J/\psi$  decays
  - Prompt: “please measure xxx using xxx dataset, fit xxx and estimate the BFs

Paper under  
preparation

2026/01/19



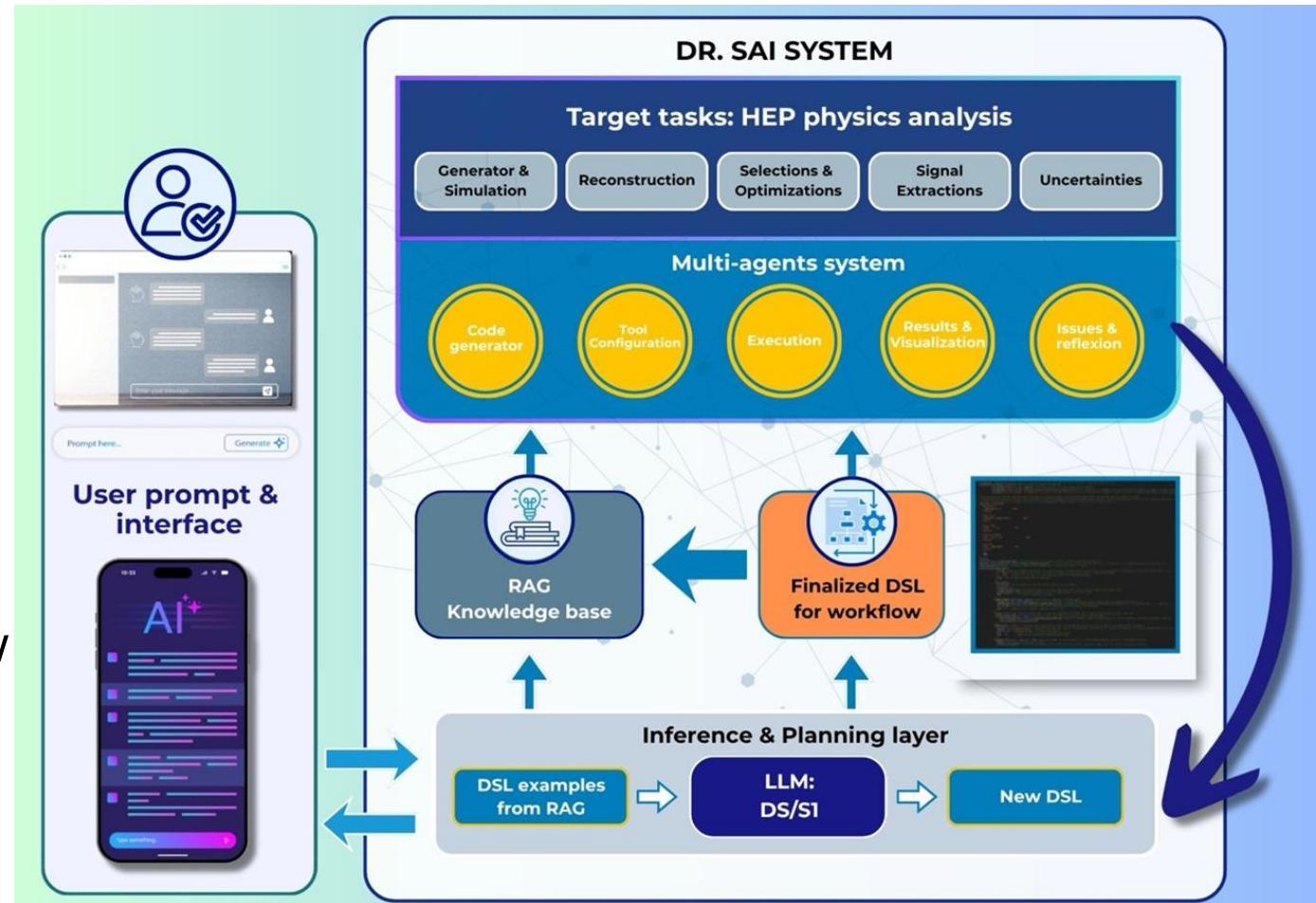
13

# Experience and plan

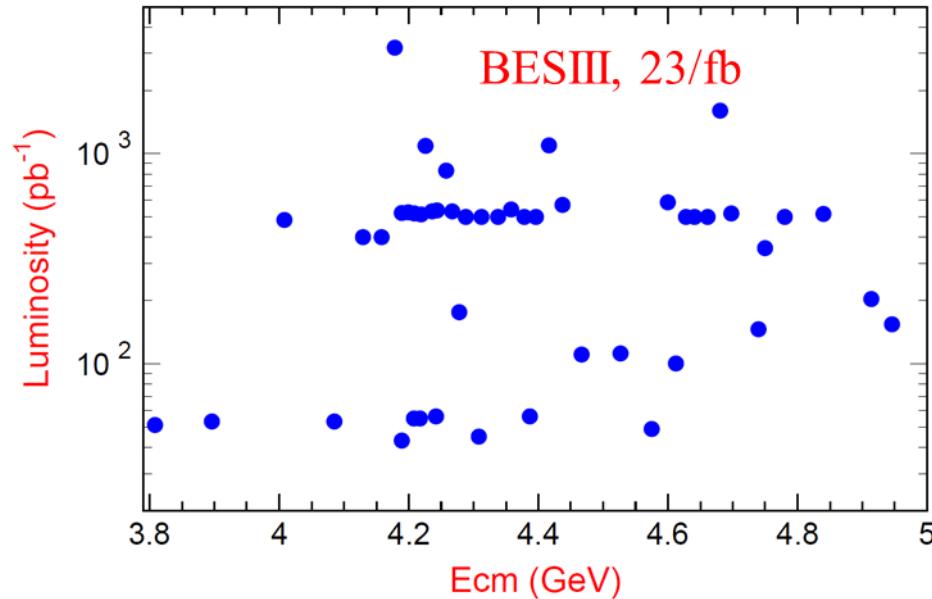
- The key is **HEP knowledge representation and embedding** !
  - Knowledge means how to do physics analysis
- Current solution: interpret analysis procedure into DSL and store in RAG
- Next (target: June 2026):
  - Use DSL to re-organize the whole workflow of Dr.Sai
  - **Human-AI collaborate to work on real analysis**
  - Assist expert to work on real physics analysis at BESIII
- Lots of works on-going, stay tuned

# The Next Dr.Sai

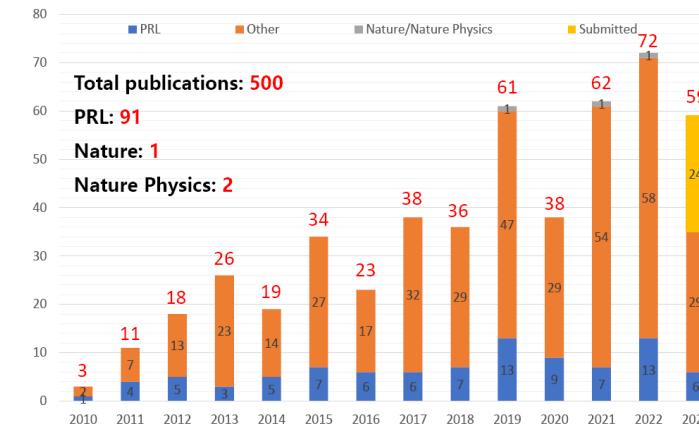
- We need to make it useful
  - Current prototype looks promising
- Next will be a **human-AI collaboration**
  - User raise ideas
  - Dr.Sai generate draft-DSL
  - User review and correct the DSL
  - Dr.Sai execute following the DSL
  - User monitor & control the workflow
- Can be easily migrated to other HEP experiments, e.g. ATLAS, CMS, LHCb



# Prospects



**BESIII publications**  
(May 9, 2023)



- >700 physics results from ~700 people in the past 14 years
  - More than 30 new hadrons are discovered from hundreds of decay channels
- More data will be collected after BEPCII-upgrade
- We can use Dr. Sai to **go through all the channels quickly** once new data were collected
- Or we can use natural language to guide Dr. Sai to do new analysis

# Summary

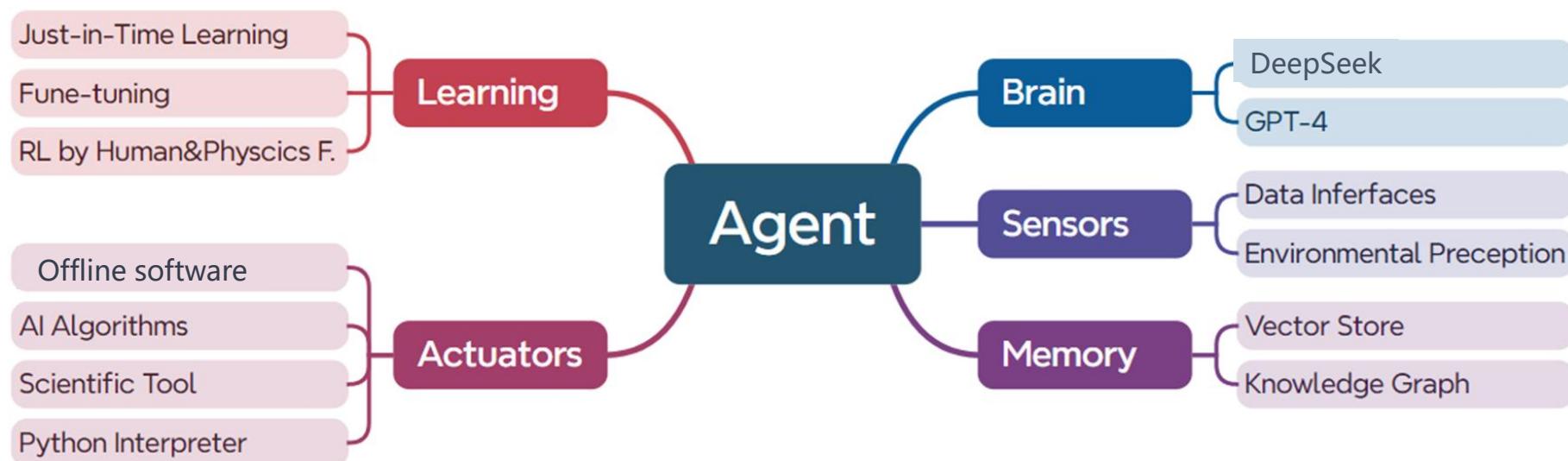
- LLM could be very helpful for HEP
  - Not just generate draft code/text, but also can be used to **automate the analysis**
- A prototype of Dr.Sai is built
  - It can be used to **automate the workflow of simple analysis at BESIII** from user's query to preliminary BFs correctly, more validations are on-going
  - More advanced usage of LLM **need new ideas**, e.g. knowledge representation and embedding
  - We plan to build Dr.Sai
- Next will be a more intelligent Dr.Sai
  - We are migrating to other experiments now
  - Welcome to discuss/collaborate !

# backup

# Dr. Sai (赛博士) project

Short for Dr. Science and Dr. Cyber in Chinese

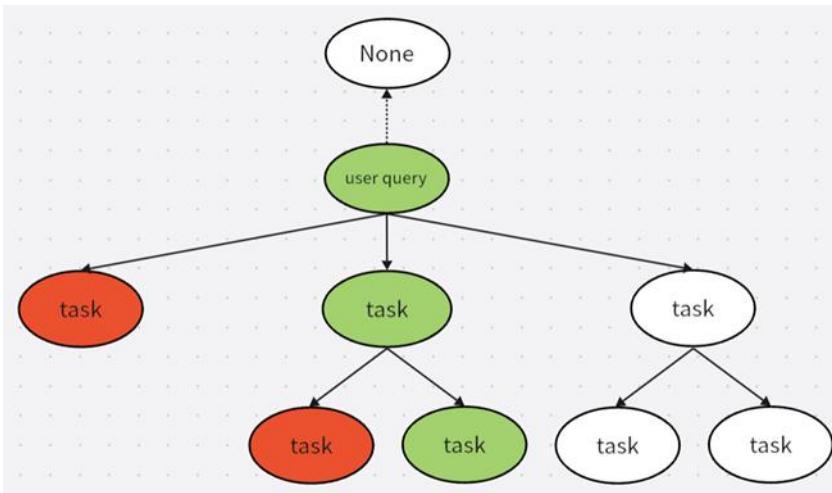
- LLM is switchable: GPT/LLaMA/DeepSeek
  - Default: DeepSeek V3/R1
  - Investigating the approaches to build better domain LLM



Developed based on [autogen](#) framework.

# One example of Dr. Sai V1

- Task decomposition
  - Compose complex task into multiple small, simple, and well-defined sub-tasks
  - Sub-tasks will be executed in sequence



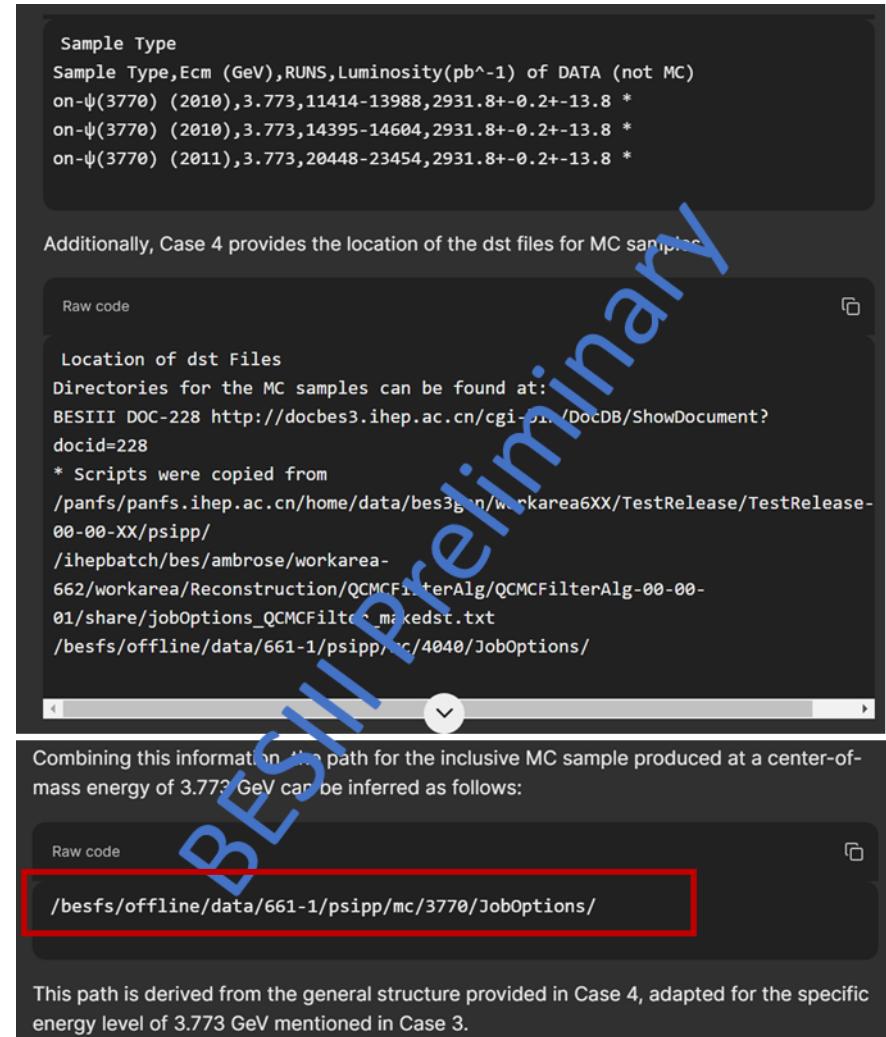
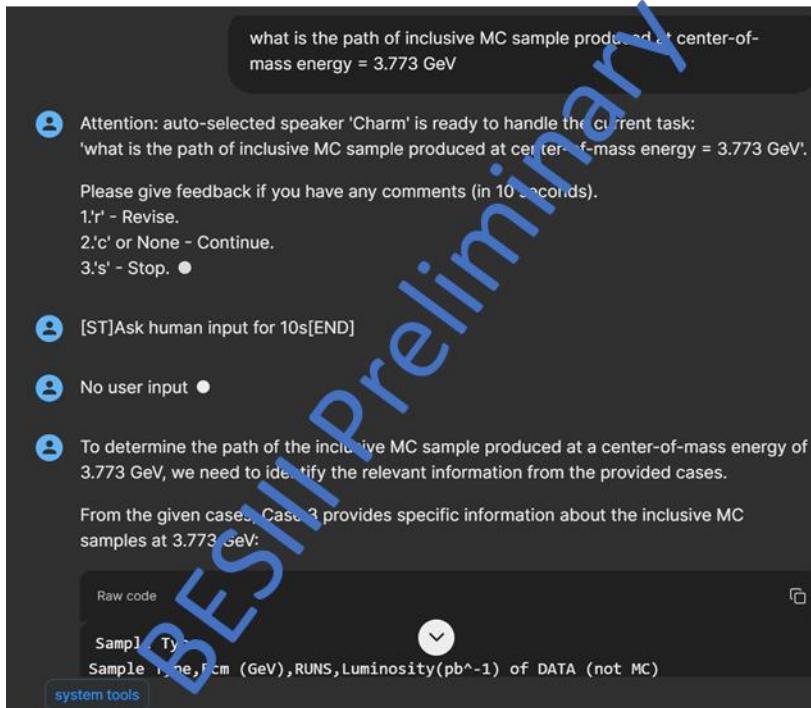
## How to search for Zc(3900)

"Observation": "The user requests to decompose the task of searching for Zc(3900) from electron-positron collision data at a center-of-mass energy of 4.26 GeV.",  
"Thoughts": "To decompose the task, we need to break down the process into manageable steps, including data collection, event selection, background suppression, signal extraction, and analysis.",  
"Status": "CONTINUE",  
"Plan": [  
 "(1) Data Collection: Gather the electron-positron collision data at a center-of-mass energy of 4.26 GeV from the BESIII experiment.",  
 "(2) Event Selection: Identify and select events that are likely to contain the Zc(3900) signal. This typically involves selecting events with the final state particles of interest, such as  $\pi^\pm J/\psi$ .",  
 "(3) Background Suppression: Apply criteria to reduce background events that can mimic the signal. This may involve using kinematic cuts, particle identification, and other techniques.",  
 "(4) Signal Extraction: Use invariant mass calculations and other methods to extract the Zc(3900) signal from the selected events. This involves plotting the invariant mass spectrum of the  $\pi^\pm J/\psi$  system and identifying peaks corresponding to Zc(3900).",  
 "(5) Statistical Analysis: Perform a statistical analysis to determine the significance of the observed signal. This includes fitting the invariant mass spectrum and estimating the signal yield and background.",  
 "(6) Systematic Uncertainties: Evaluate and account for systematic uncertainties in the measurement. This may involve varying selection criteria, fitting methods, and other parameters.",  
 "(7) Results and Interpretation: Interpret the results in the context of theoretical models and previous measurements. Compare the observed properties of Zc(3900) with theoretical predictions and other experimental results.",  
 "(8) Documentation and Reporting: Document the analysis process, results, and conclusions. Prepare a report or publication to share the findings with the scientific community."  
,  
"Comment": "The task involves a detailed and systematic approach to search for the Zc(3900) particle in electron-positron collision data. Each step is crucial to ensure the accuracy and reliability of the results."  
} ●

# One example of Dr. Sai V1

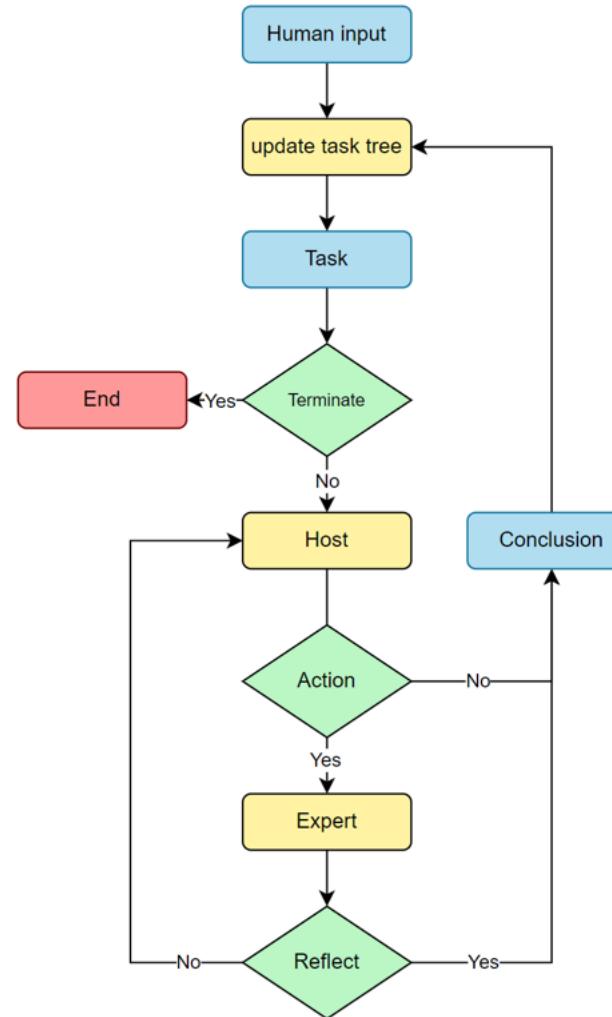
- BESIII internal assistant

- Prompt: where is the xxx MC sample
- Then it search in RAG collections
- LLM read the RAG outputs and conclude correctly



# Multi-Agents communication logics

1. Human pass task to Dr.Sai
2. It will think if this task is simple or complex and if all tasks in task tree are finished
3. The Host need to think to select the next agent
  1. Planner, coder, tester, or others
  2. Planner will make/update task tree
  3. Coder will write corresponding code
  4. Tester will launch a worker in a specific computing environment and do execution
4. We are testing a better definition of agents and logic



# Evaluation system

- Constructed our own benchmark in [AgentBench](#) framework
- RAG evaluation
  - Signal: correct 100 Q-A pairs
  - Background: incorrect 1200 Q-A pairs (random combinations)
  - Tested different embedding models
- Agent-level evaluation
  - Task decompositions: check the similarity between agent output and reference
  - Next agent selection: compare the name of next agent from host to reference
- Dr. Sai evaluation
  - Comparisons on Dr. Sai output, e.g. histograms

# Evaluation system

## 1. RAG recall (recall: 100%, precision: 89%(Bge-M3) & 93%(Physicsbert))

ID	top_k	Question	paragraph	Answer	Bge-M3		Physicsbert	
					context_recall	context_pricision	context_recall	context_pricision
1	1	How you calculate the decay length of and (Lambda)	Considering that $(\Sigma^0)$ has long vertex fit for $(n^+ n^-)$	We performed a secondary	1	0.9	1	1
1	5	How you calculate the decay length of and (Lambda)	Considering that $(\Sigma^0)$ has long vertex fit for $(n^+ n^-)$	We performed a secondary	1	1	1	1
1	10	How you calculate the decay length of and (Lambda)	Considering that $(\Sigma^0)$ has long vertex fit for $(p^+ \pi^-)$	We performed a secondary	1	0.9	1	1
1	30	How you calculate the decay length of and (Lambda)	Considering that $(\Sigma^0)$ has long vertex fit for $(n^+ n^-)$	We performed a secondary	1	1	1	1



## 2. Task decomposition

## 3. Agent selection

Final results:  
Number of task: 100%  
Completeness: 91%  
Logic: 81%

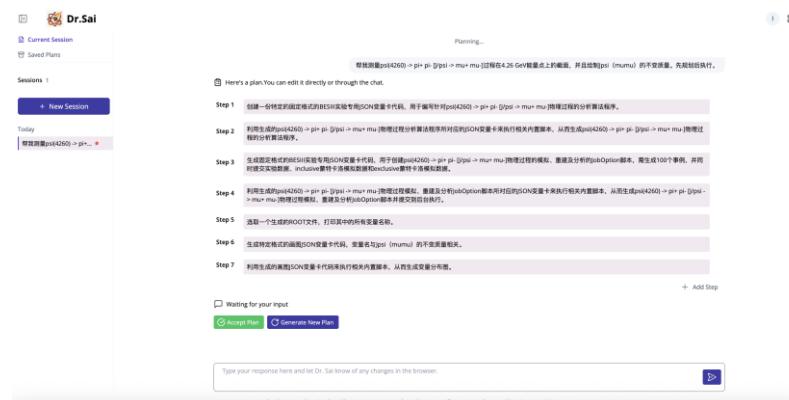
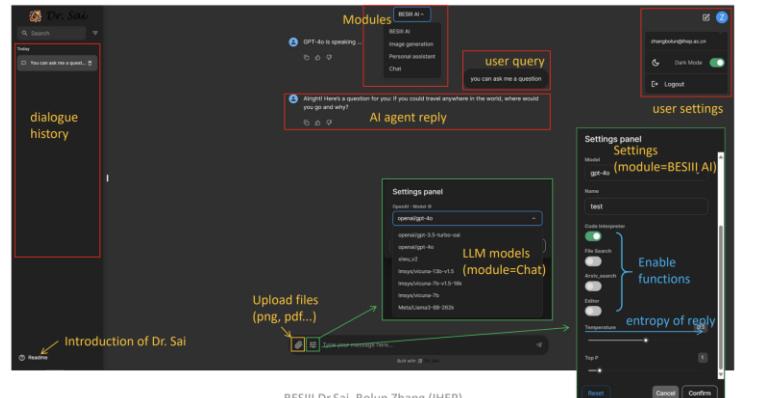
Final results without Anchor words:  
Agent score 44%  
Tool score: 44%



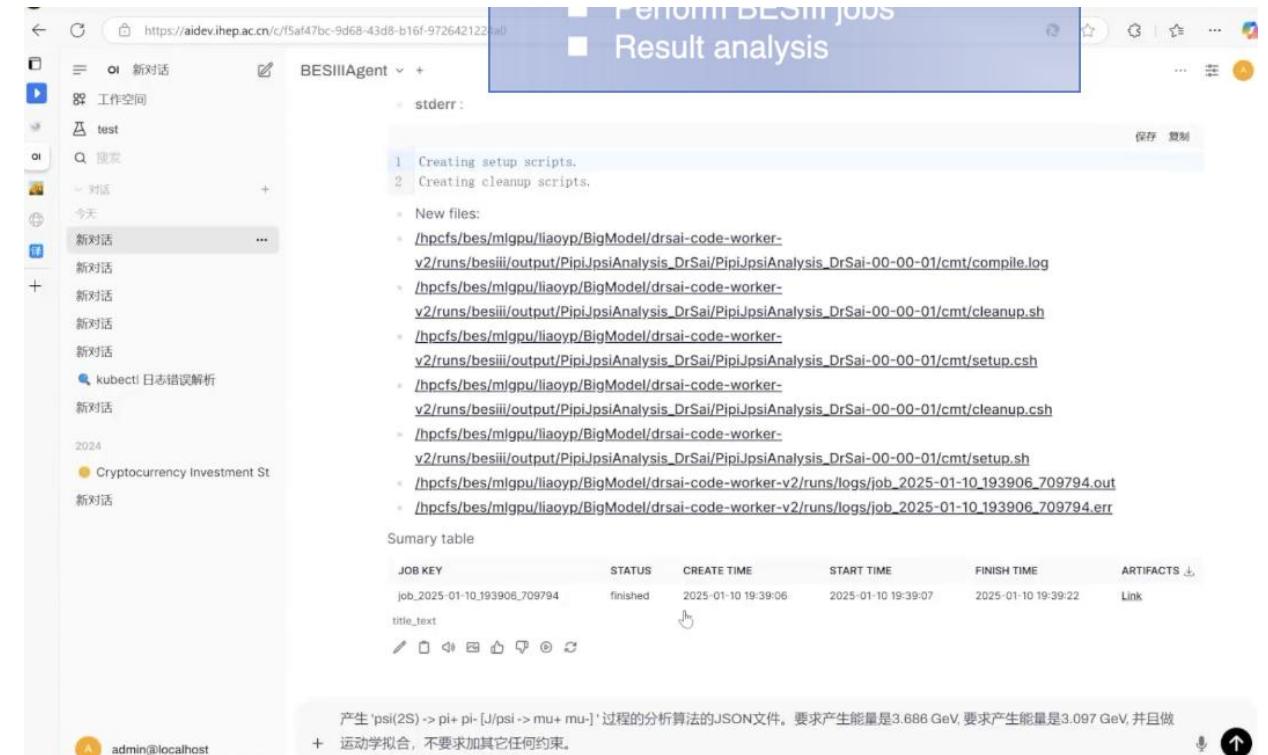
Final results with Anchor words:  
Agent score : 98%  
Tool score: 98%

# New interface of Dr.Sai

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



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



On-going

Plan to move to Magentic-UI

More friendly for interactive

2nd AI+HEP in East Asia

25