中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

LLM-based physics analysis assistant at BESIII - 'Dr. Sai'

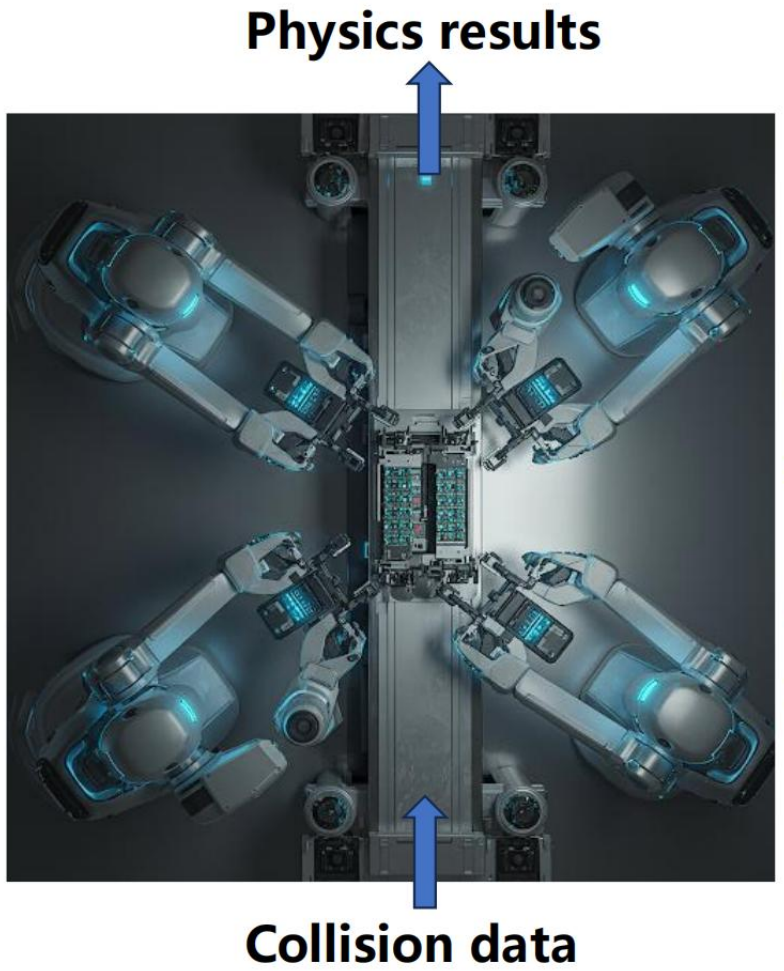
Ke Li (like@ihep.ac.cn)

on behalf of Dr. Sai working group

Institute of High Energy Physics, China

Outline

- Motivation
- Introduction of BESIII
- Dr.Sai project
- Methodology
- Status and prospects

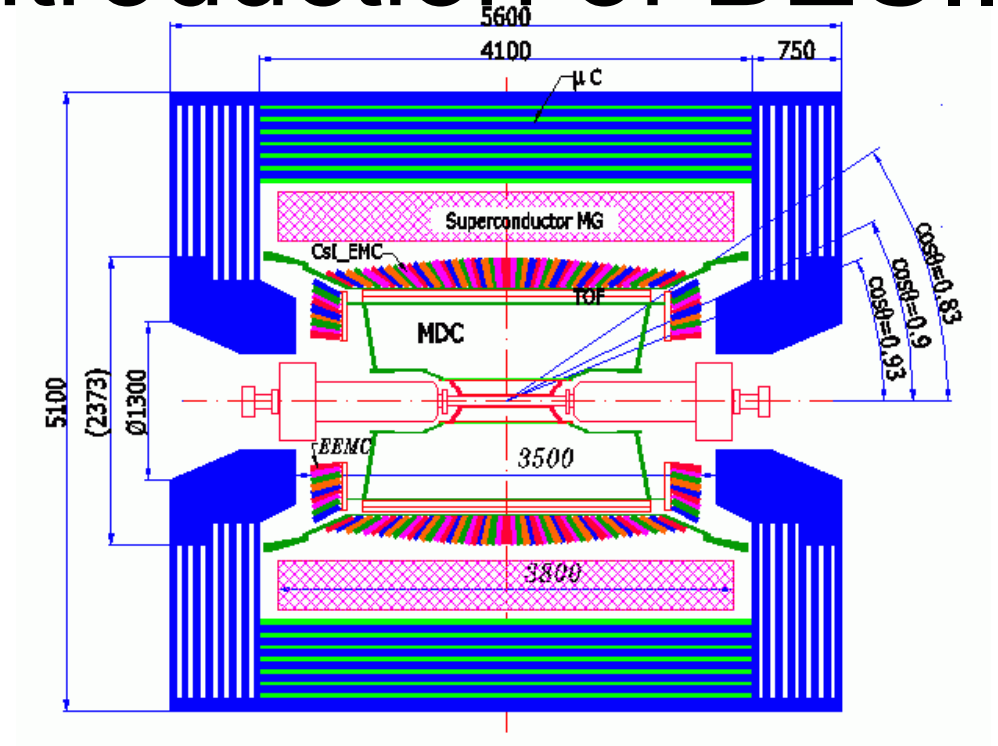


Goal:
A virtual "robots" to work on HEP data analysis

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

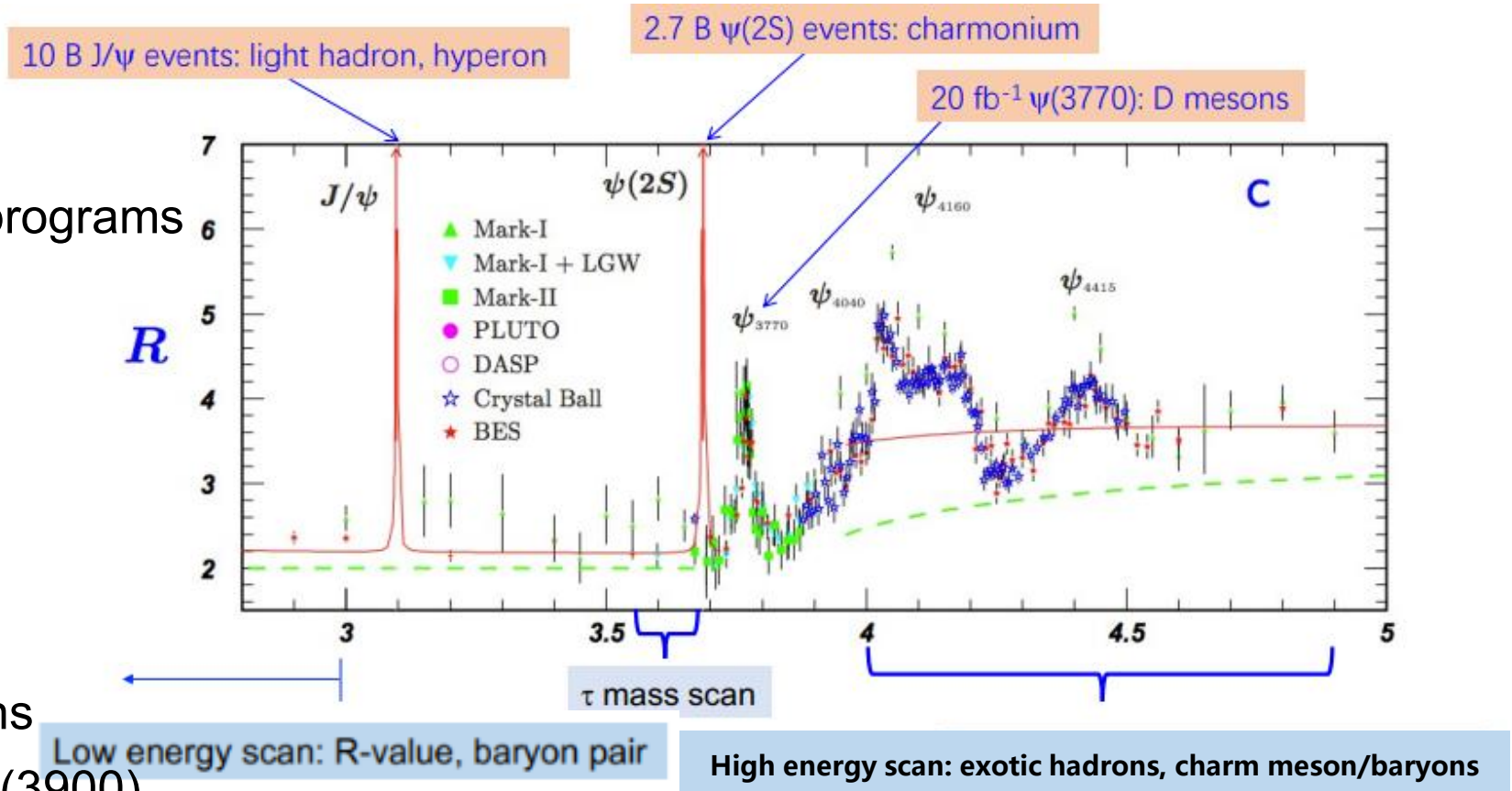


Start running from 2009, biggest electron-positron collision data around 3-4 GeV over the world

- Beijing Electron Positron Collider (BEPCII)
  - Design luminosity L_D=1 x 10^{33} $cm^{-2}s^{-1}$ @ 3.773 GeV (2016 achieved), x3 is expected after upgrade
  - Continuous injection (top-up mode)
- BEijing Spectrometer (BESIII), almost a 4pi detector
  - Good spatial resolution (130um) and energy resolution (2.5%)

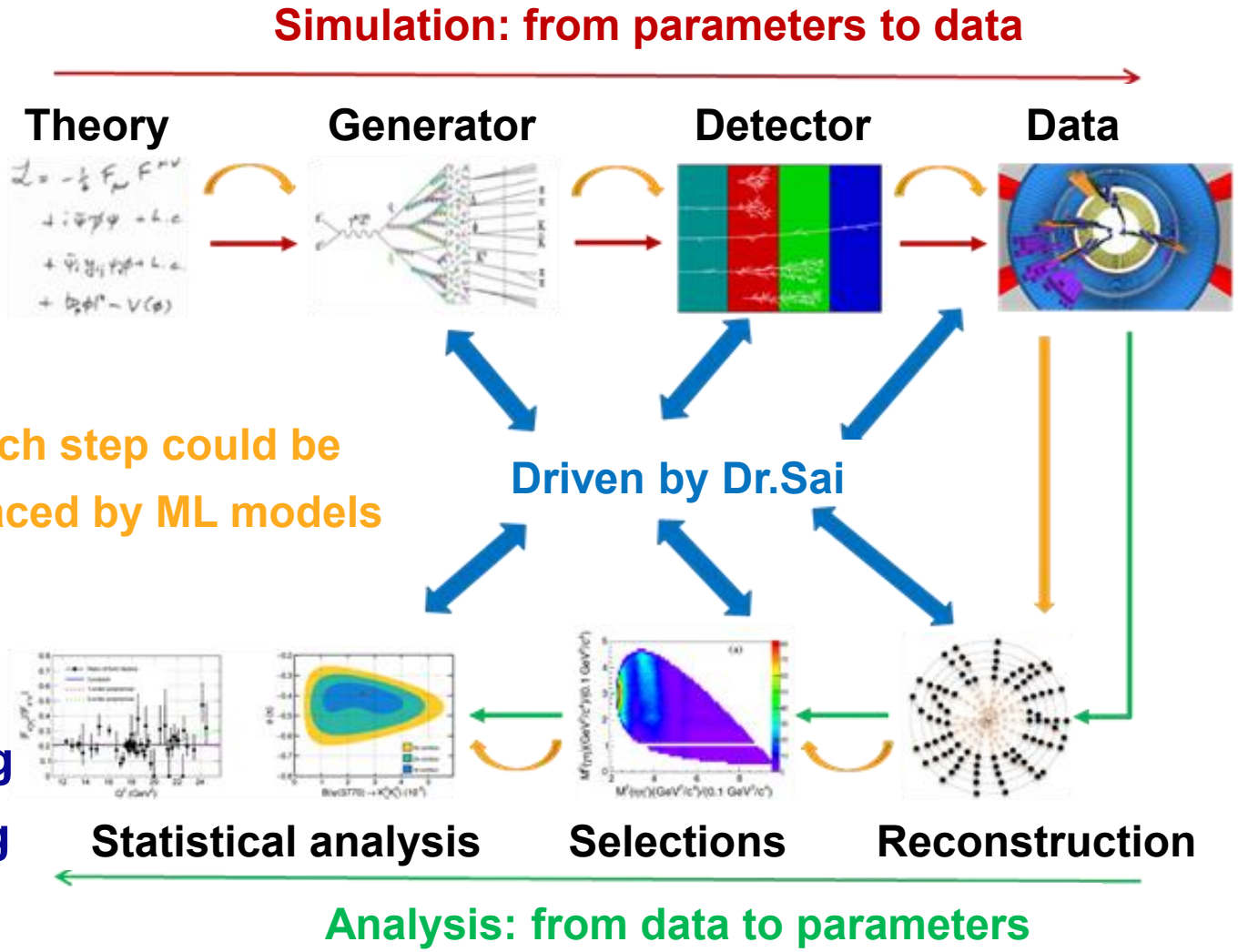
Introduction of BESIII - physics program

- >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 hardrons
    - First tetraquark state: Zc(3900)
  - **Good for analysis modelling**



How LLM can help

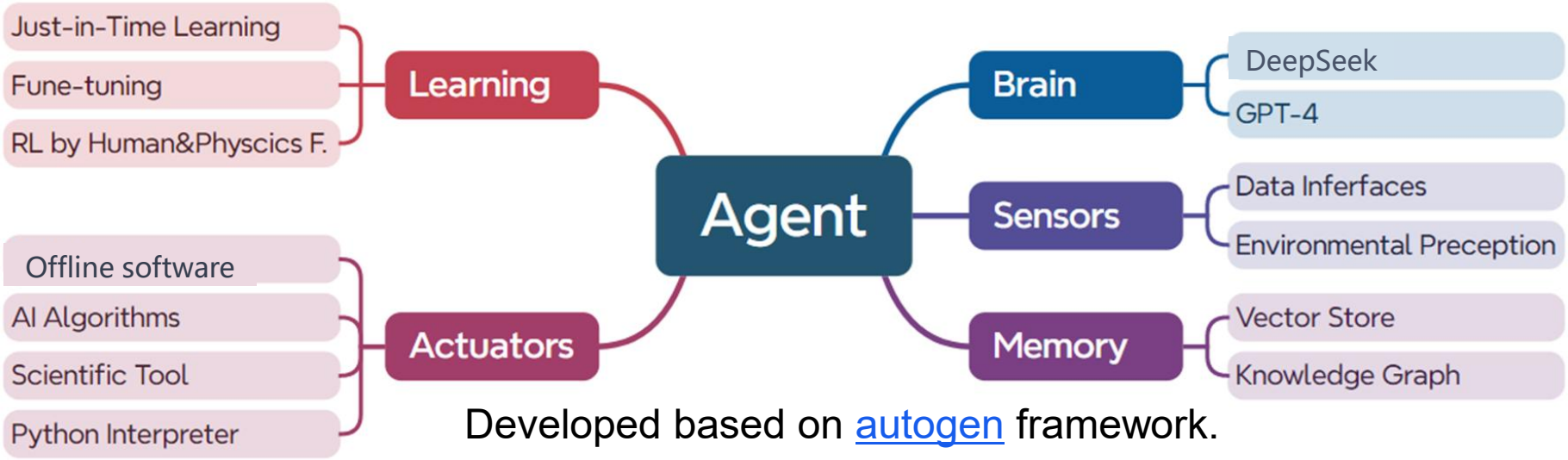
- LLM is good at text/code generation
- But rules in natural languages is different from HEP data
- One possible approach
  - **Use LLM to automate the data analysis workflow**
  - Similar to self-driving
  - It is possible given the LLM is rapidly developing
  - The missing part is the **modelling of the workflow and embedding to LLM**



Dr. Sai (赛博士）project

Short for Dr. Science and Dr. Cyber in Chinese

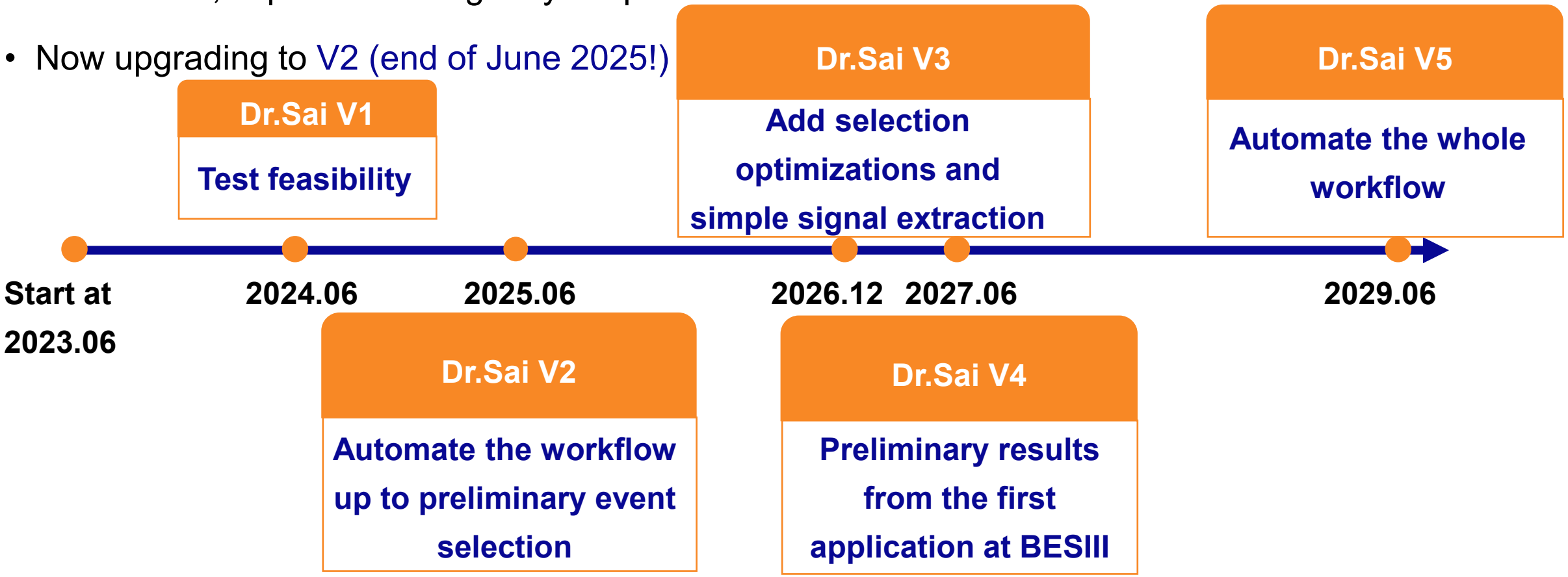
- A multi-agents system based on LLM, aim to **automate the HEP data analysis**
  - LLM = brain, AI agent = human
- LLM is switchable: GPT/LLaMA/DeepSeek
  - A demo of domain LLM: Xiwu V2 (fine-tuned on LLaMA3)
  - Investigating the approaches to build better domain LLM



Developed based on autogen framework.

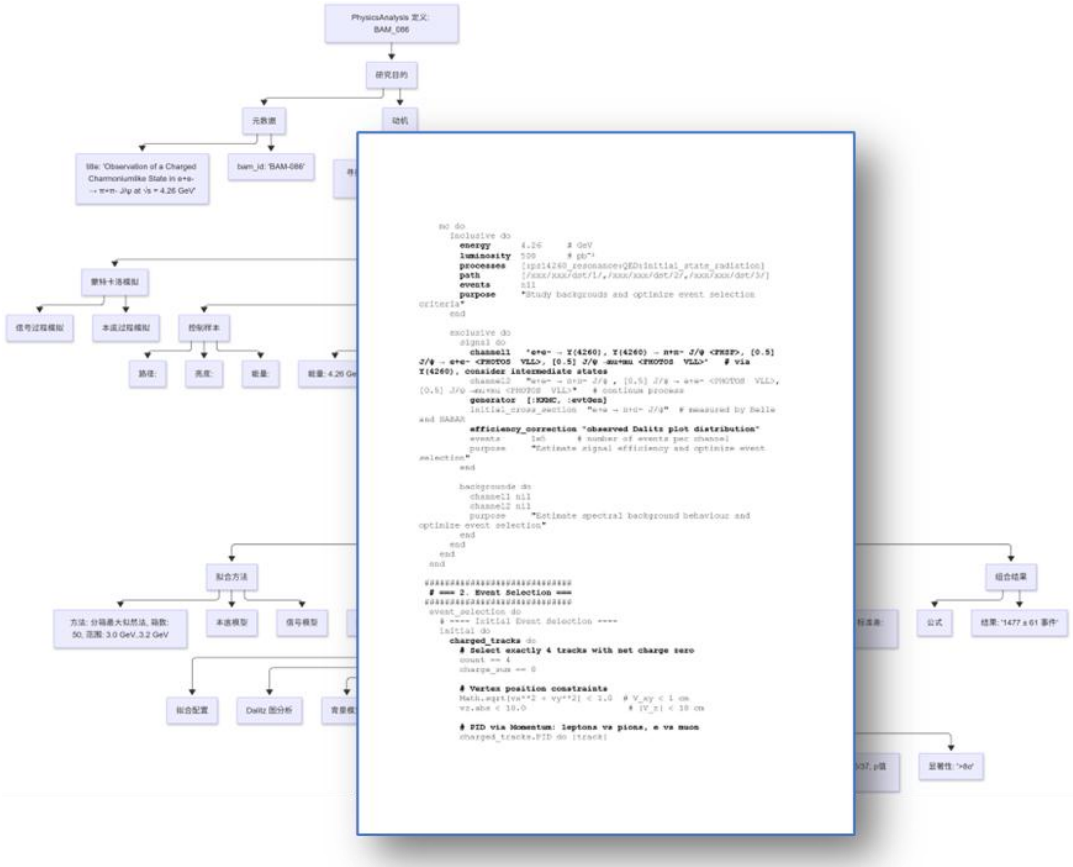
Dr. Sai (赛博士) project - timeline

- AI assistant for BESIII: Dr. Sai V1
  - chatbot, capable of doing very simple tasks
- Now upgrading to V2 (end of June 2025!)



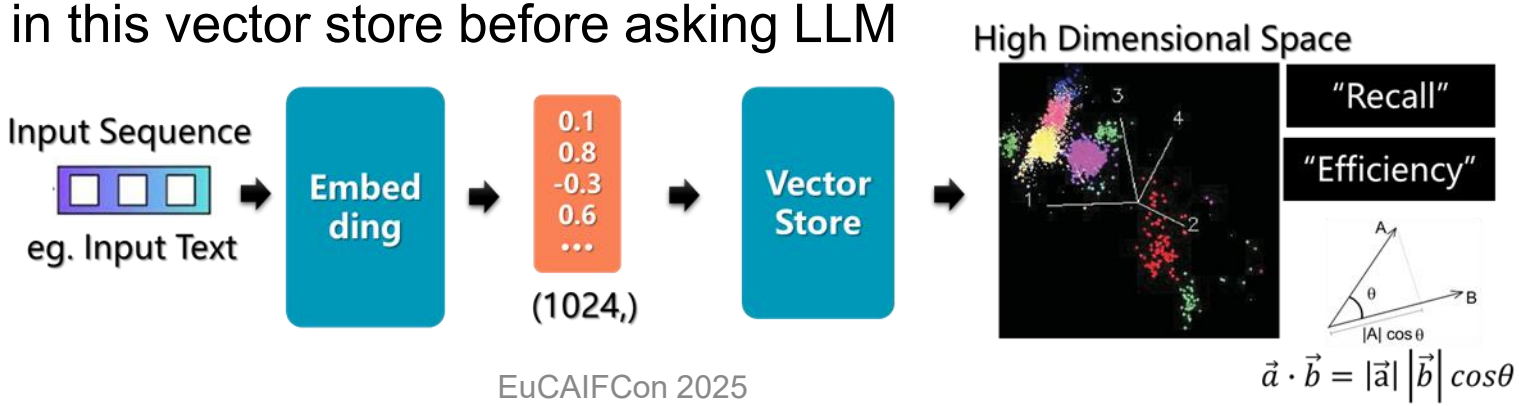
First attempts of analysis modelling

- 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



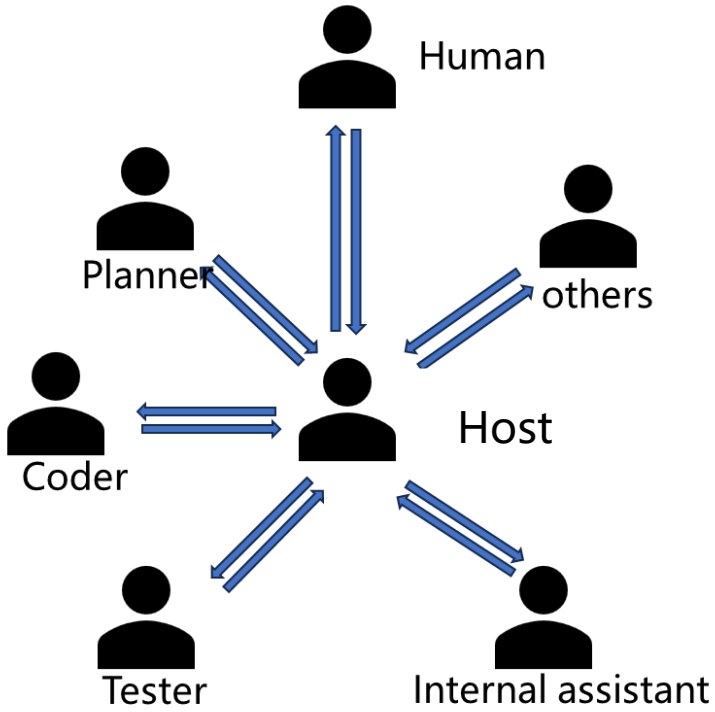
Memory of Dr. Sai - RAG

- 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

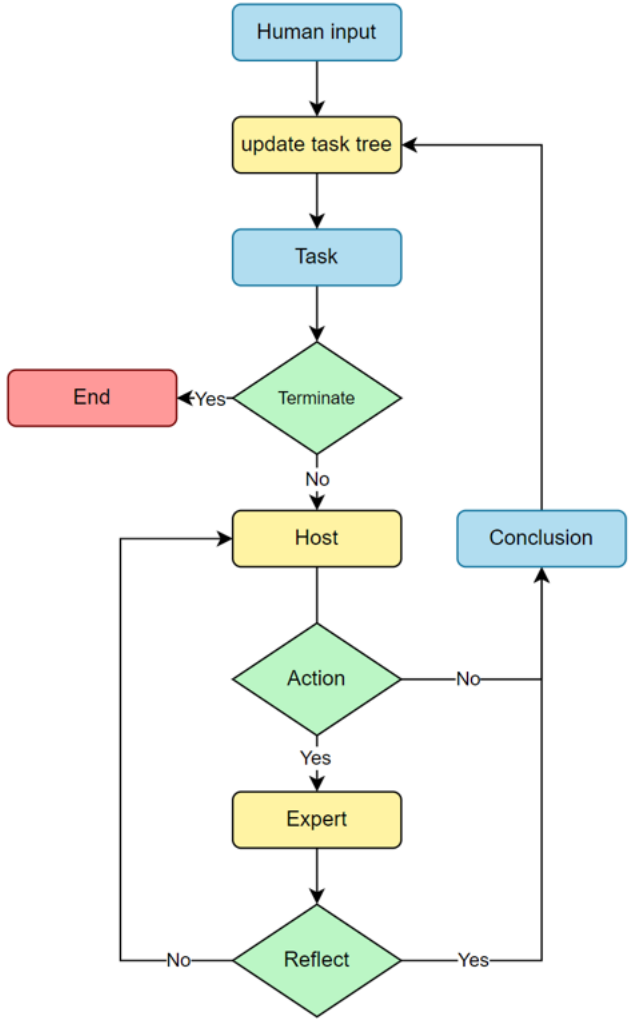
- The HEP data analysis is too complex for LLM now
- We can devided the complex task to small and simple task, and develop a dedicated agent for each kind of task
- Multi-Agents (foundation model is switchable):
  - Host: select correct agent
  - Planner: task decomposition
  - Coder: code generation
  - 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



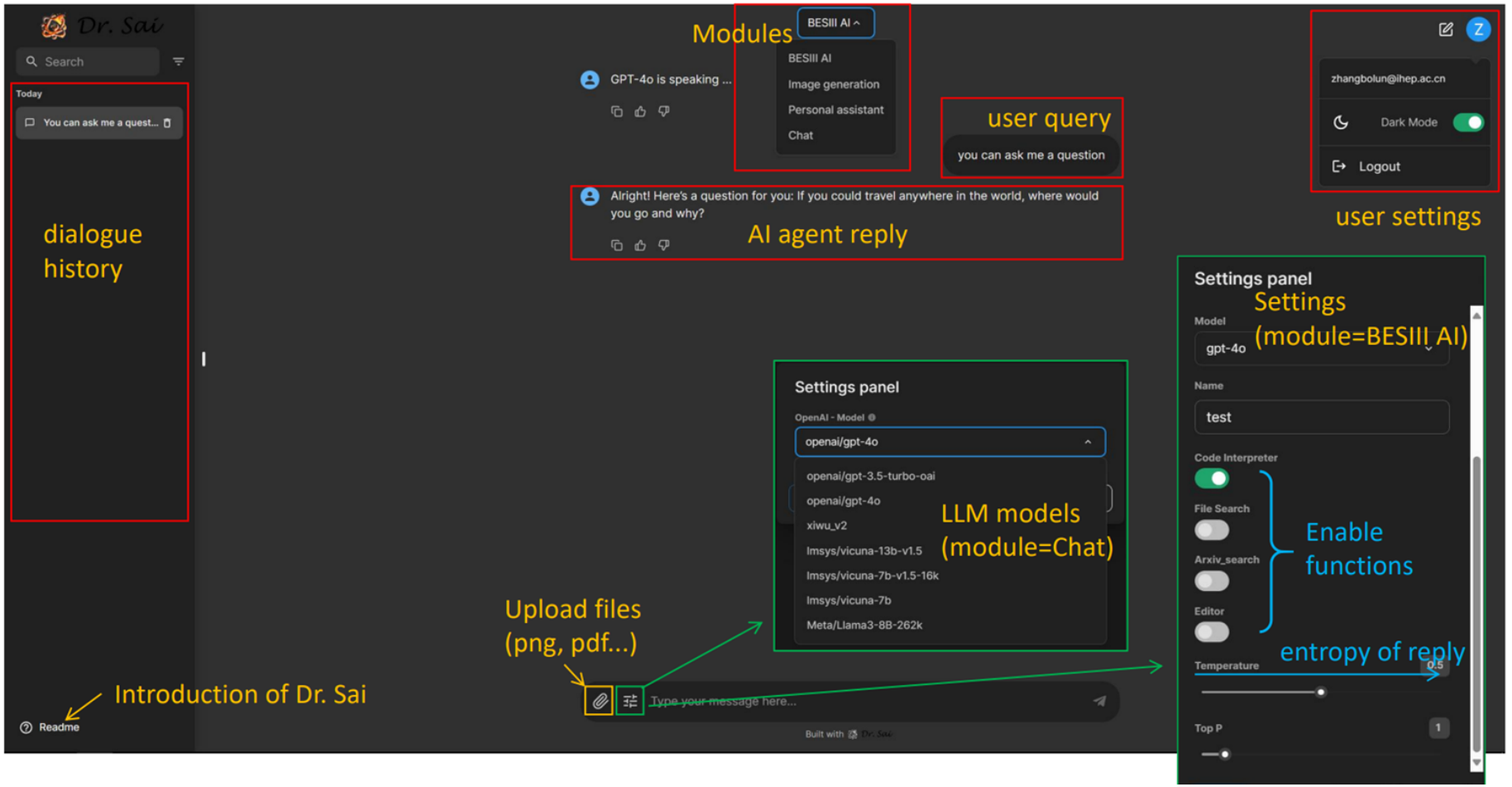
Preliminary

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



WebUI



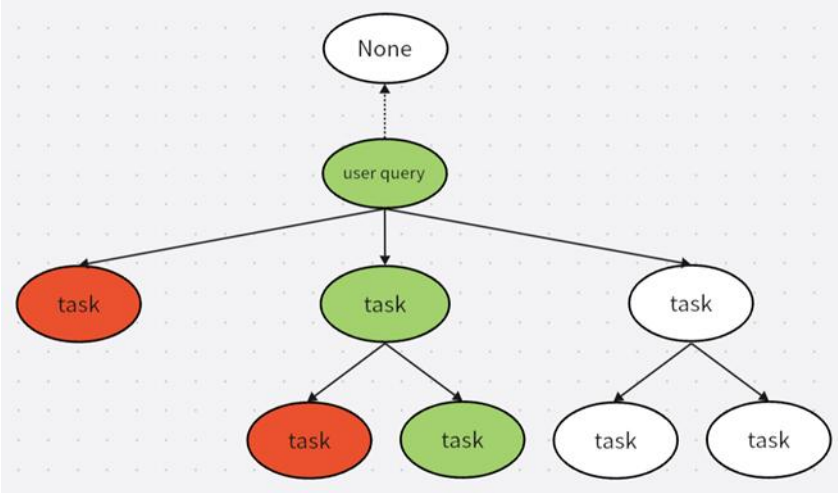
Upgrading, please stay tuned

Status of Dr. Sai

- A demo (Dr. Sai V1) is released at June 2024
  - Internal assistant, similar to chATLAS
  - Capable of small and simple task, e.g. write a simple MC generator configuration
  - The automation looks feasible
- Now upgrading all components, expected to have a new version at the **end of June !**
  - Better multi-agents communication logics
  - DSL for multiple analyses
  - Re-organize the workflow of BESIII data analysis
  - A comprehensive evaluation system and benchmarks
  - Support MCP, etc.

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

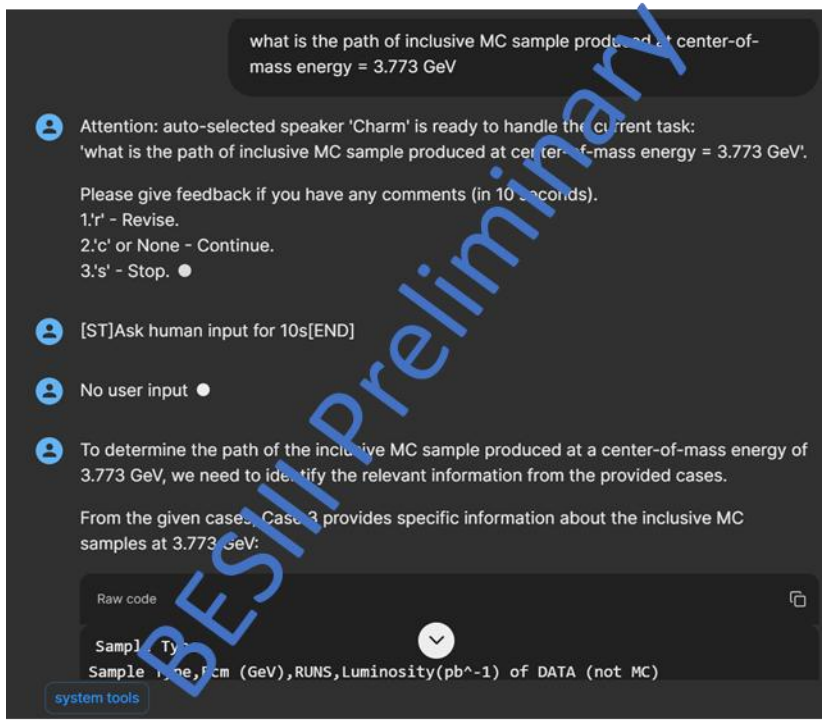


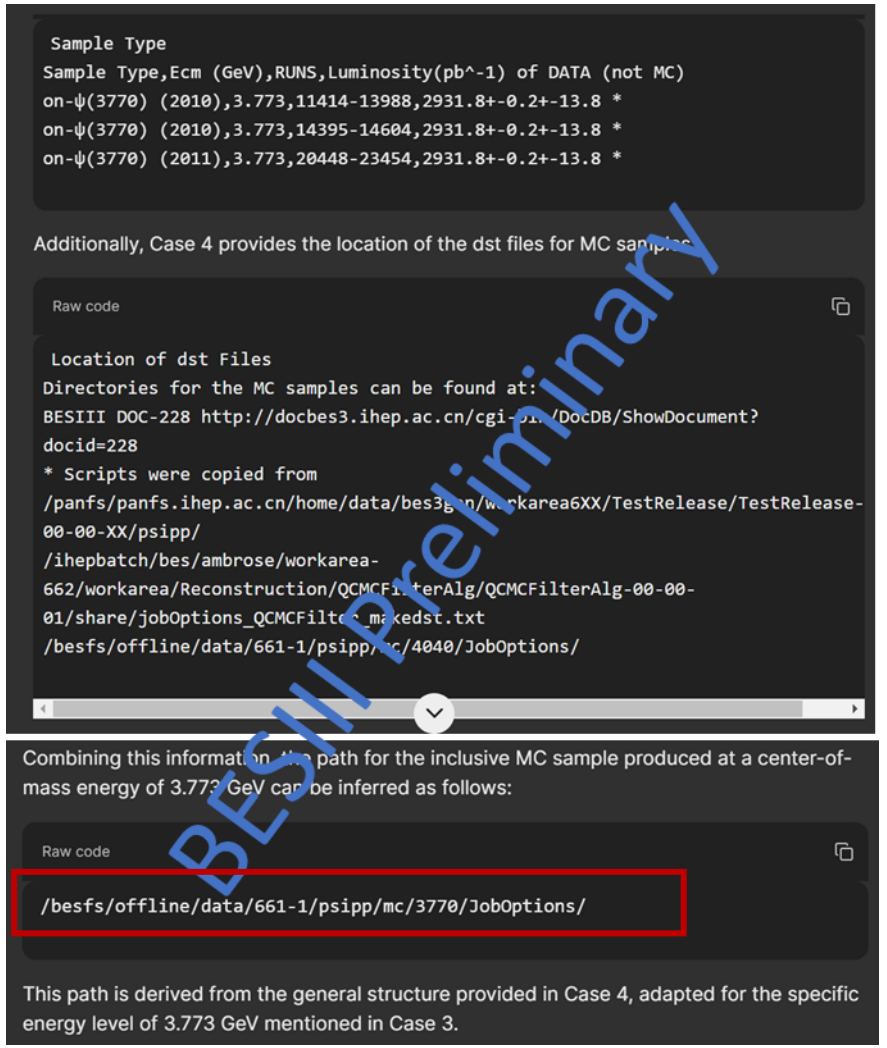
How to search for Zc(3900)

"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





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 manually and store in RAG
- Next:
  - Align the DSL and scientific tools/codes
  - **Interpret analysis as Markov chain and use reinforcement learning to build a new LLM**
  - Investigating other approaches
- Lots of works on-going, stay tuned

Summary

- LLM could be very helpful for HEP
  - Not just generate draft code/text, but also can be used to automate the analysis
- A demo of AI assistant is built to test the feasibility
  - Chatbot with BESIII internal textual data
- A new version of Dr. Sai will be ready soon
  - It is expected to automate the workflow of analysis at BESIII from user's query to histogram after preliminary selections
- More advanced usage of LLM need new ideas, e.g. knowledge representation and embedding
  - Should be similar to all HEP experiments
- There are lots of on-going AI/ML activities at IHEP and BESIII to push "AI for HEP"
  - Welcome to discuss/collaborate !

back-up