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'

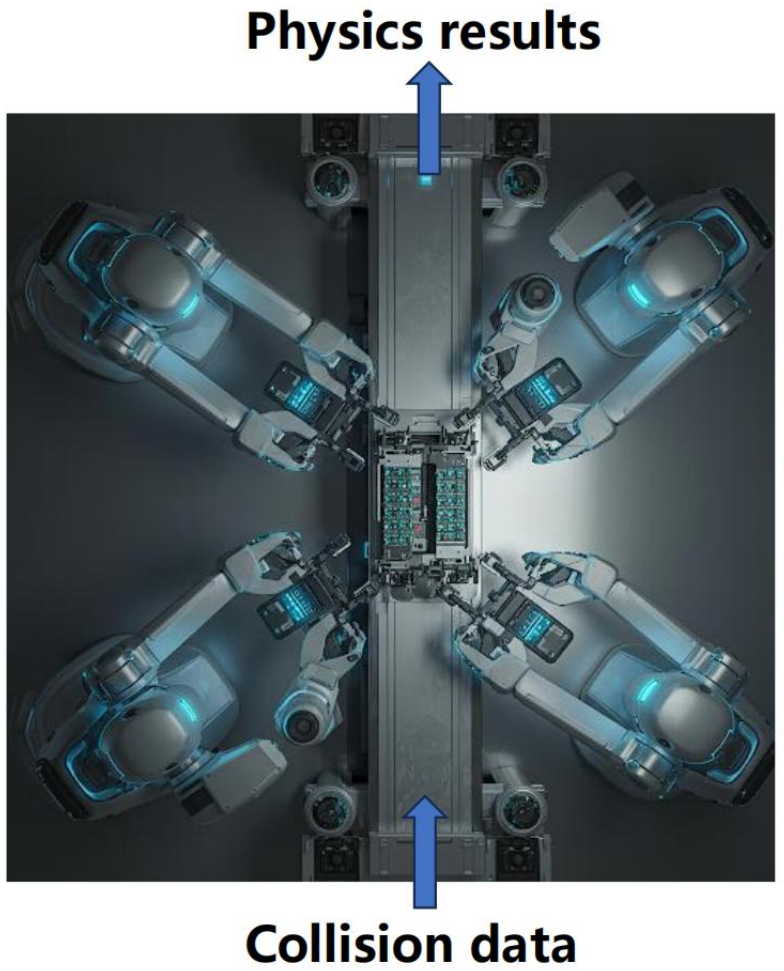
Ke Li (like@ihep.ac.cn)

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

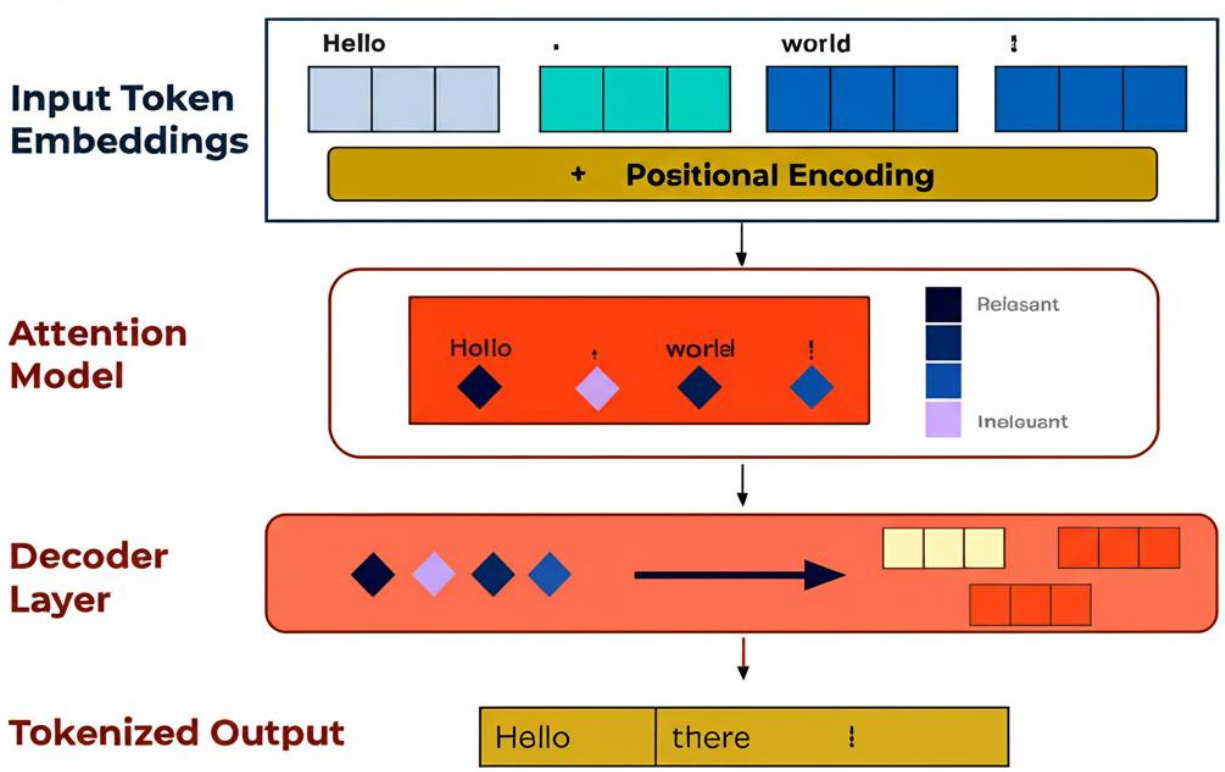


Goal:
A virtual "robots" to work on HEP data analysis

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

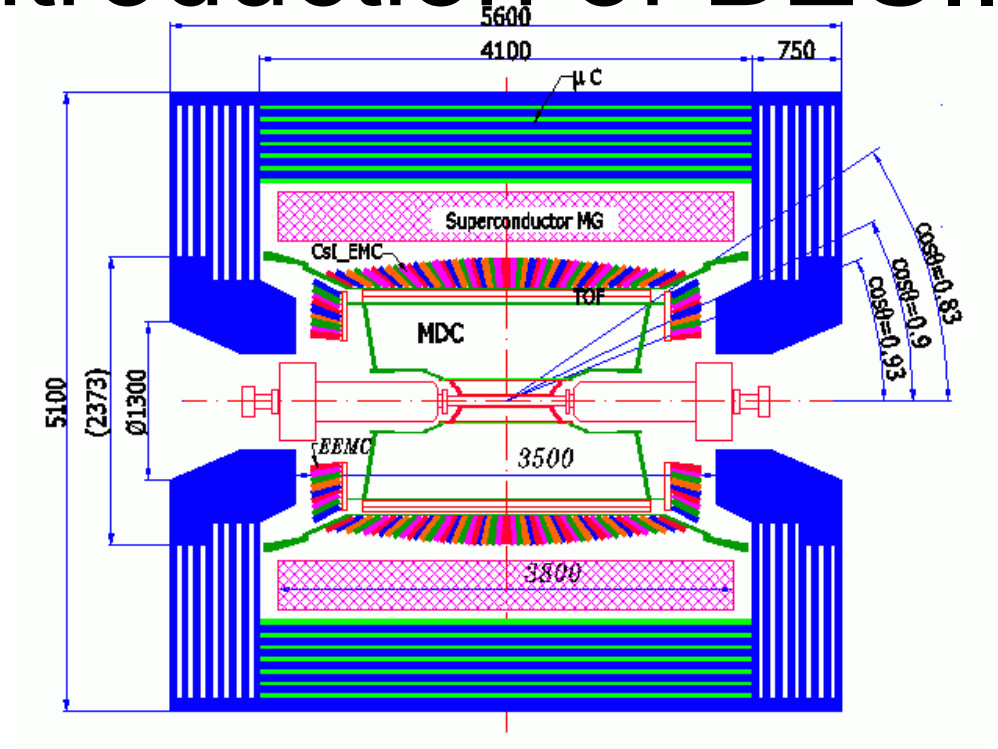
Introduction of Large Language Model (LLM)



- **Trillions of parameters** in the model
- Trained on massive amount of text data
- Predict the next most **statistically** probable token
- This training allow the model **"understand" the patterns** in the training data

We should embed our knowledge into these patterns

Introduction of BESIII

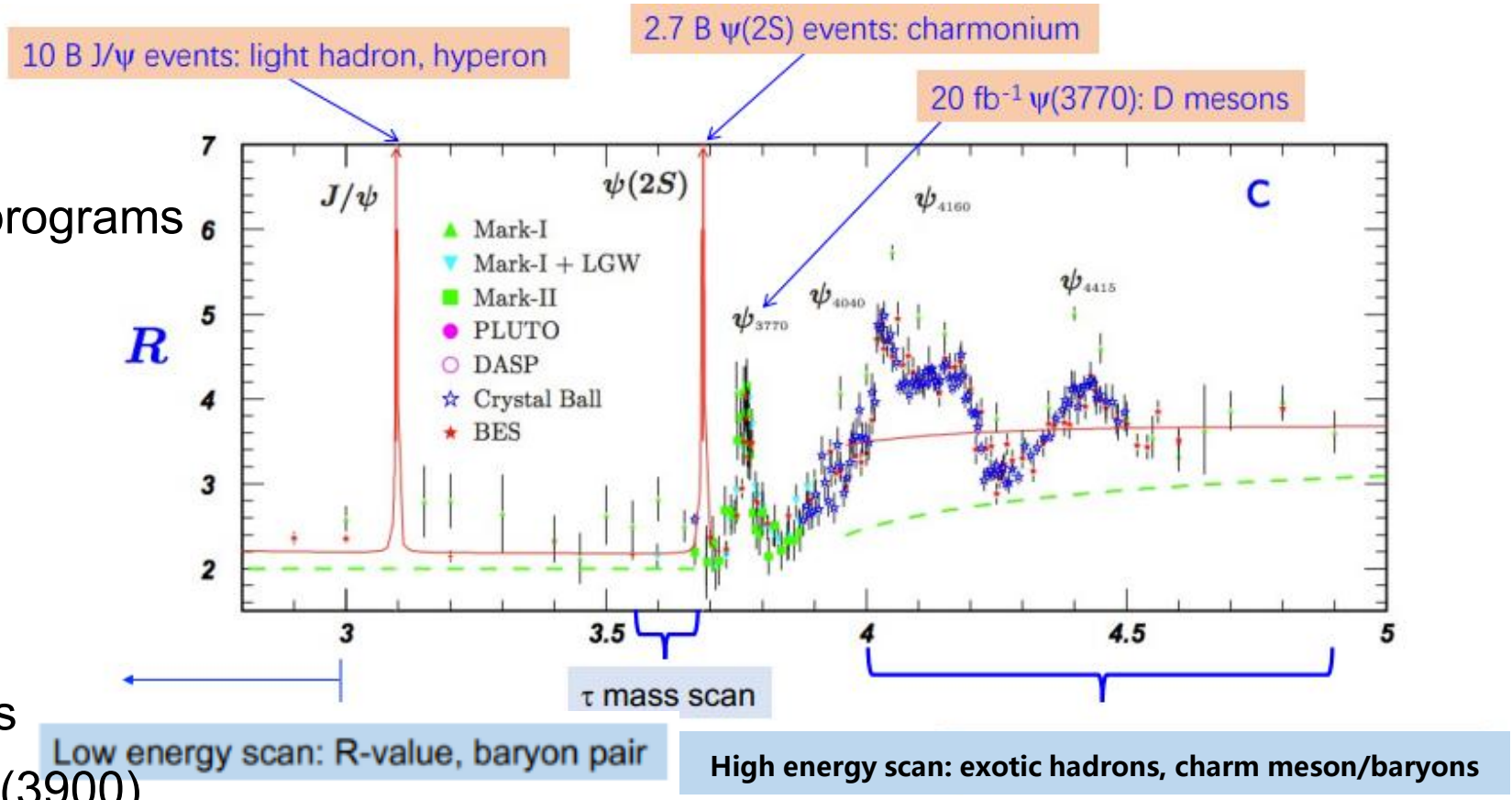


Start running from 2009, biggest electron-positron collision data around 3-4 GeV over the world

- 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%)

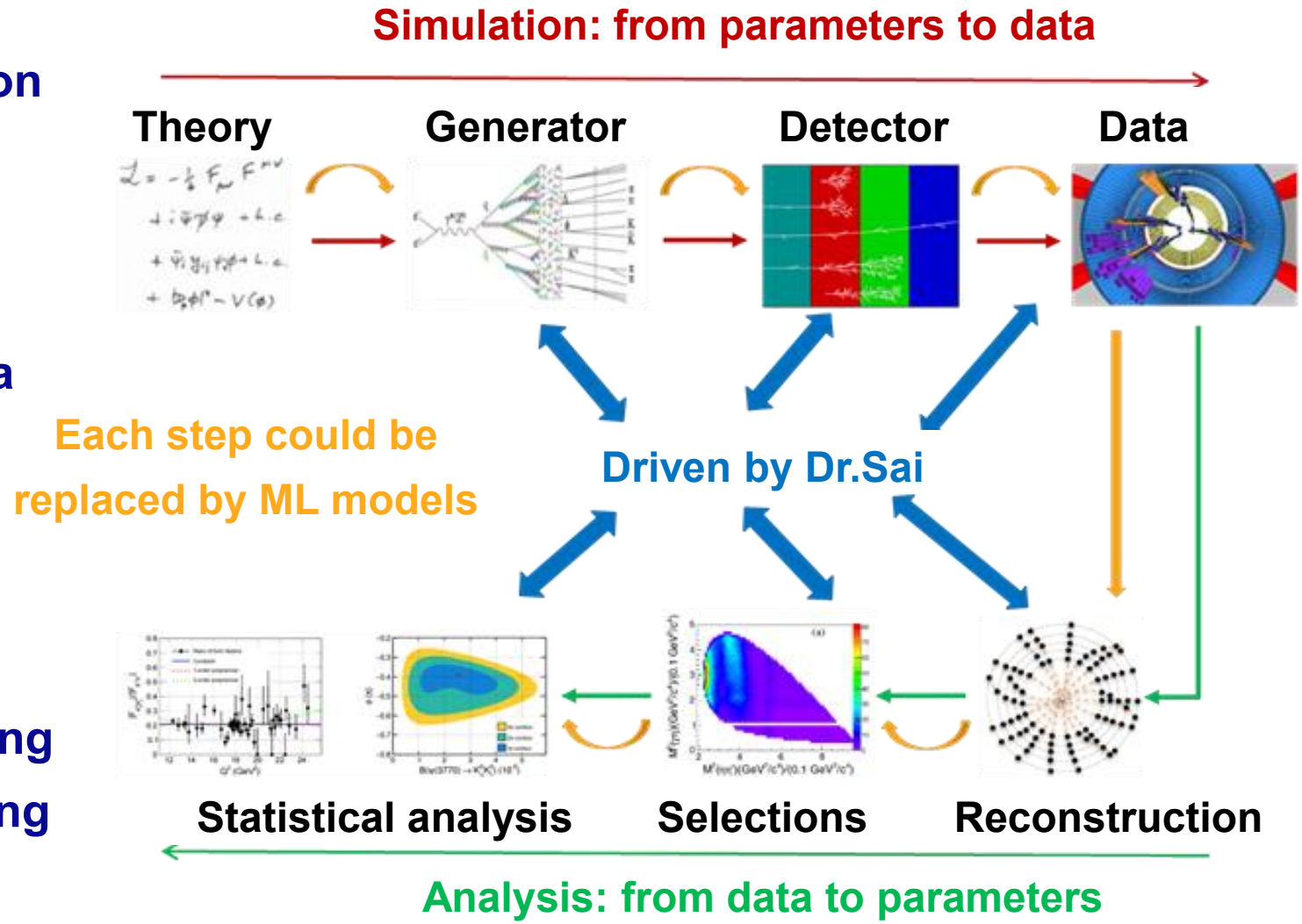
Introduction of BESIII - physics program

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



How LLM can help

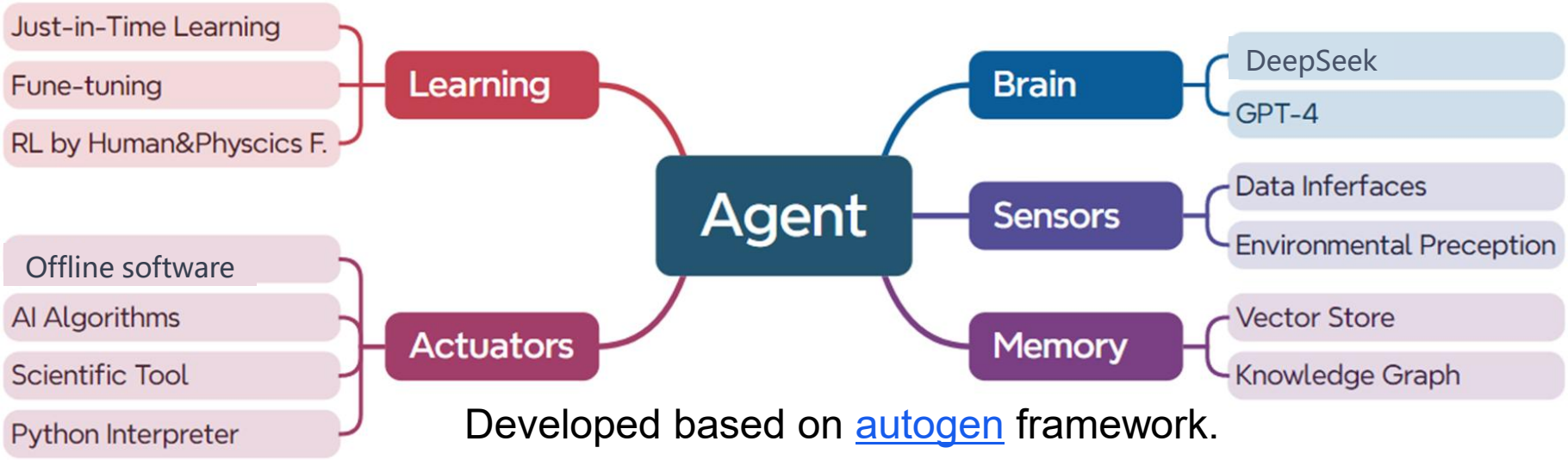
- LLM is good at **text/code generation**
- But rules in natural languages is different from HEP data
- One possible approach
  - **Use LLM to automate the data analysis workflow**
  - Similar to self-driving
  - It is possible given the LLM is rapidly developing
  - The missing part is the **modelling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



Dr. Sai (赛博士）project

Short for Dr. Science and Dr. Cyber in Chinese

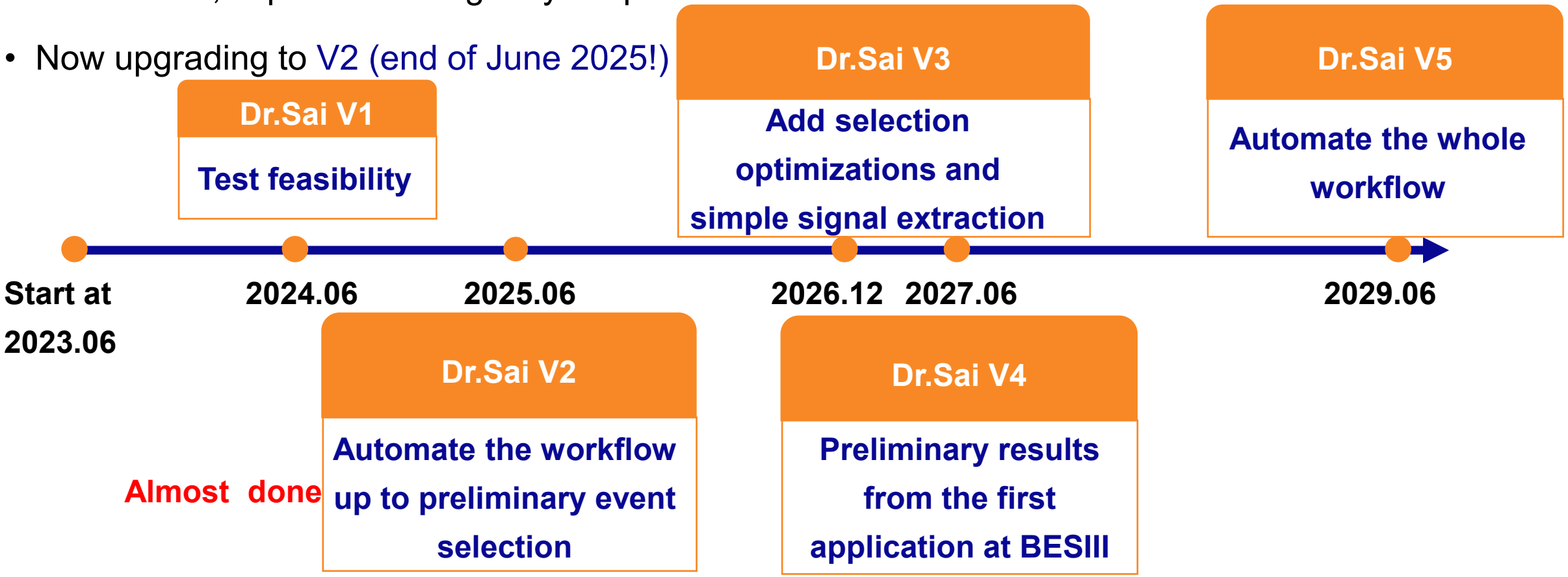
- A multi-agents system based on LLM, aim to **automate the HEP data analysis**
  - LLM = brain, AI agent = human
- LLM is switchable: GPT/LLaMA/DeepSeek
  - Default is DeepSeek V3/R1 deployed at IHEP
  - Investigating the approaches to build better domain LLM



Developed based on autogen framework.

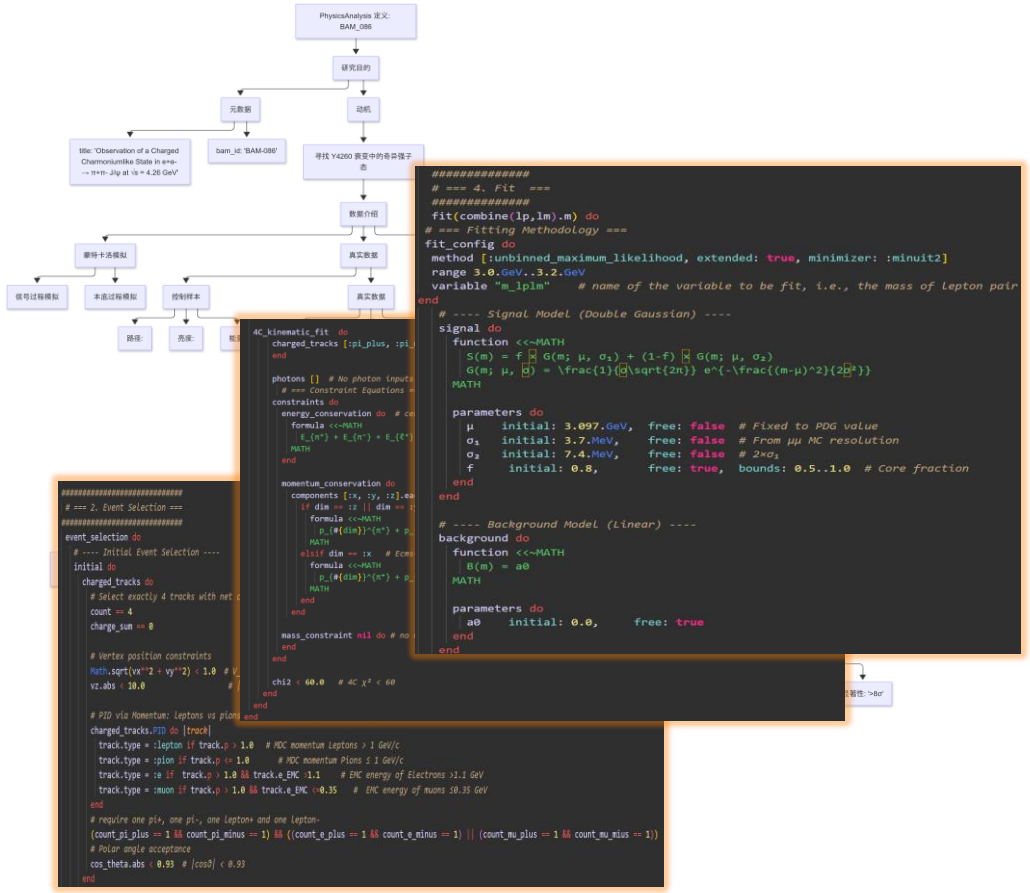
Dr. Sai (赛博士) project - timeline

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



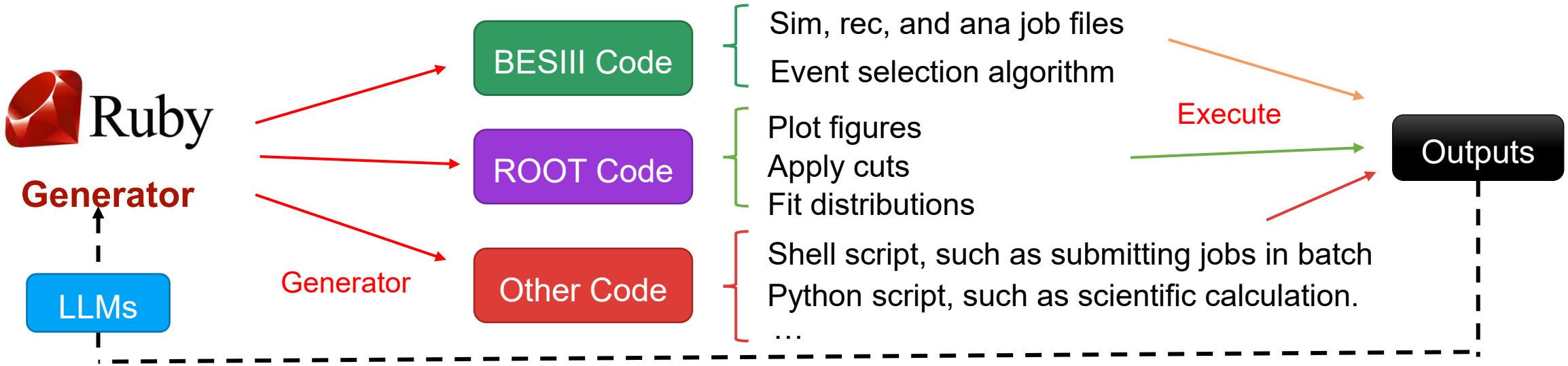
First attempt of analysis modelling

- 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 >700 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



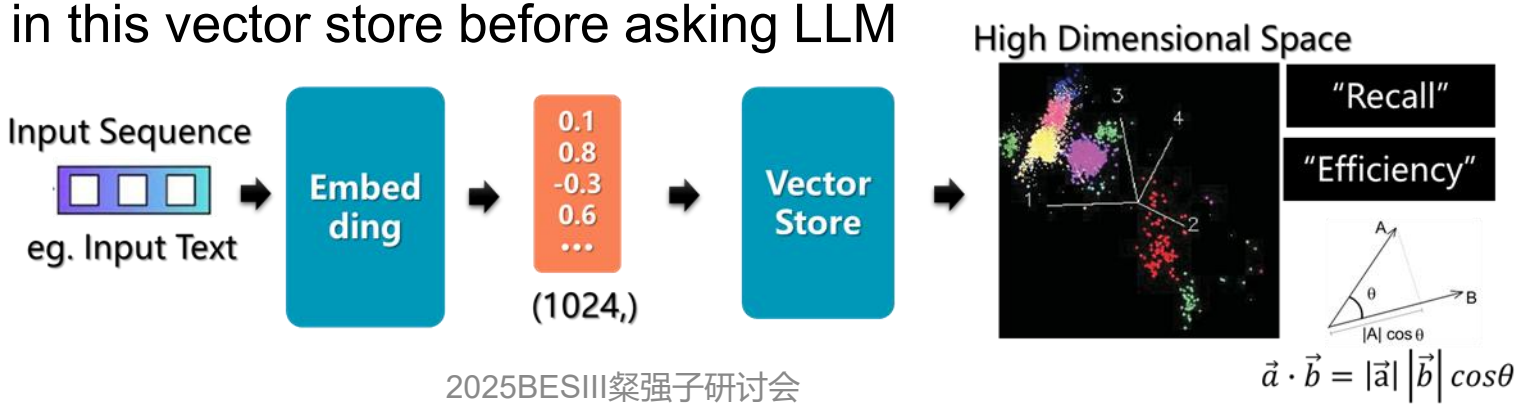
DSL for BESIII analysis

- Currently **ALL LLMs** are not smart enough to understand and reason properly, much less for BESIII analysis
- DSL V1 is a **“guide” to teach LLM** how do work
- DSL V2 will align with BESIII/ROOT code and will be more flexible



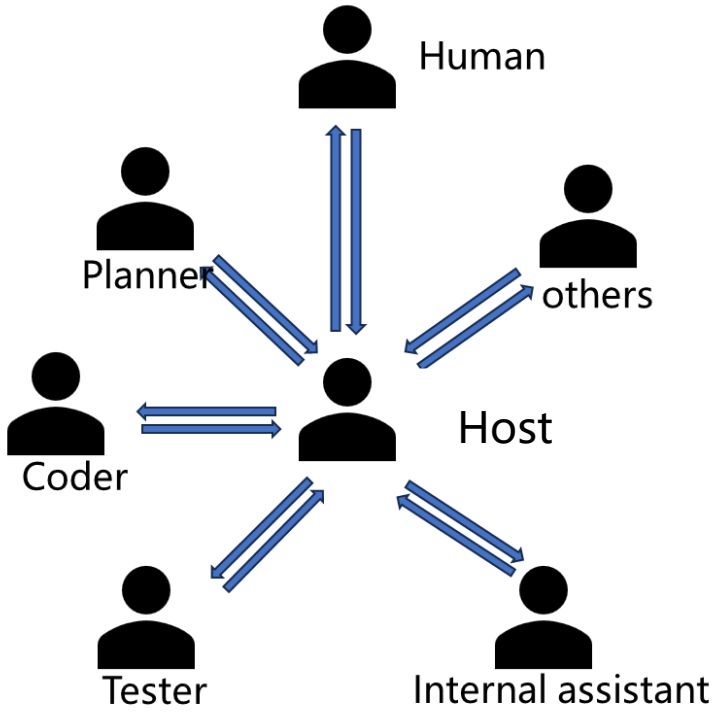
Memory of Dr. Sai - RAG

- 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

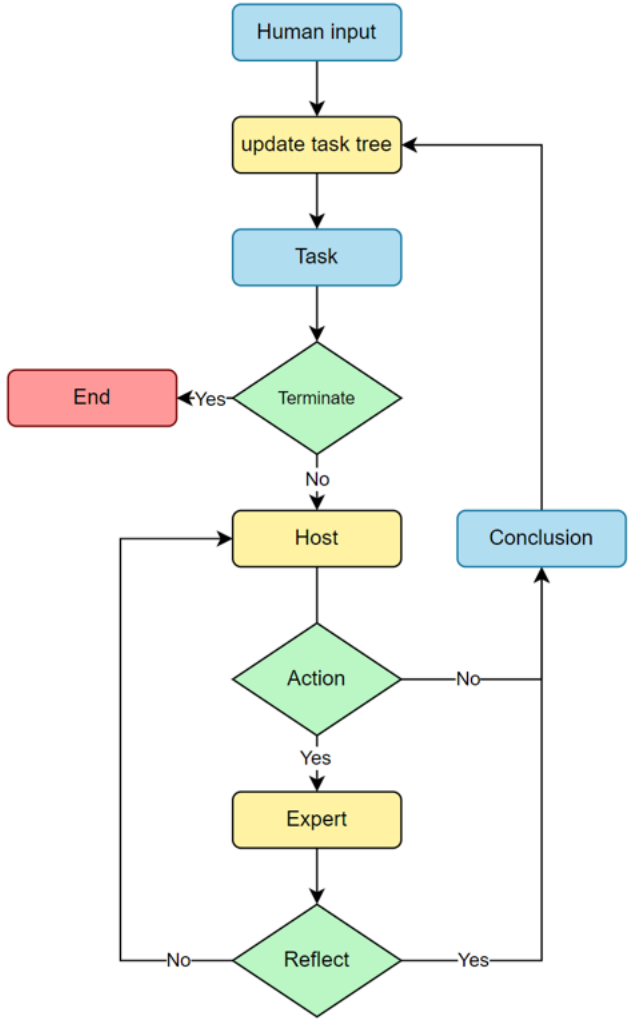
- The HEP data analysis is too complex for LLM now
- We can decompose the complex task to small and simple task, and develop a dedicated agent for each kind of task
- Multi-Agents (LLM is switchable):
  - Host: select correct agent
  - Planner: task decomposition
  - Coder: code generation
  - Tester: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



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))

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

2. Task decomposition

Final results:

Number of task: 100%

Completeness: 91%

Logic: 81%

3. Agent selection

Final results without Anchor words:

Agent score 44%

Tool score: 44%

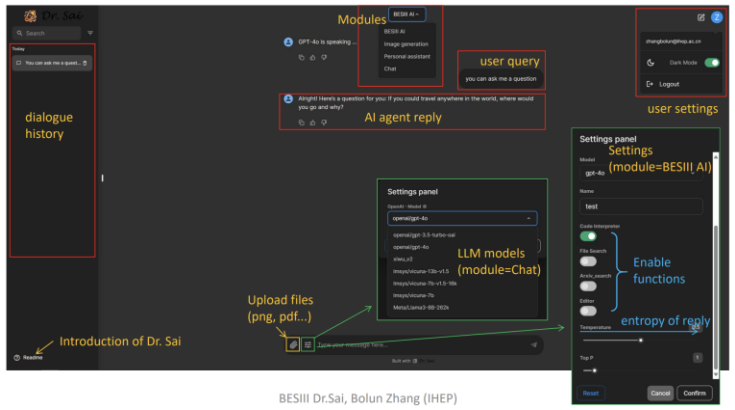
Final results with Anchor words:

Agent score : 98%

Tool score: 98%

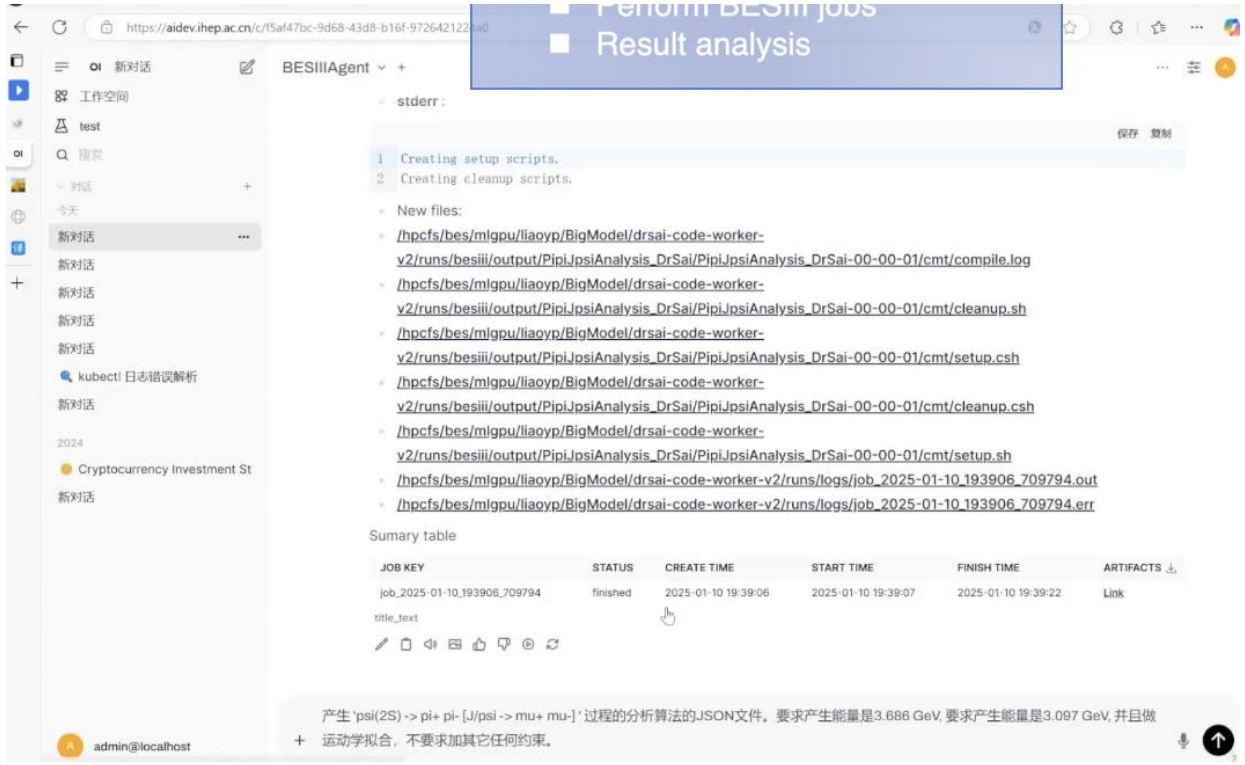
New interface of Dr.Sai

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

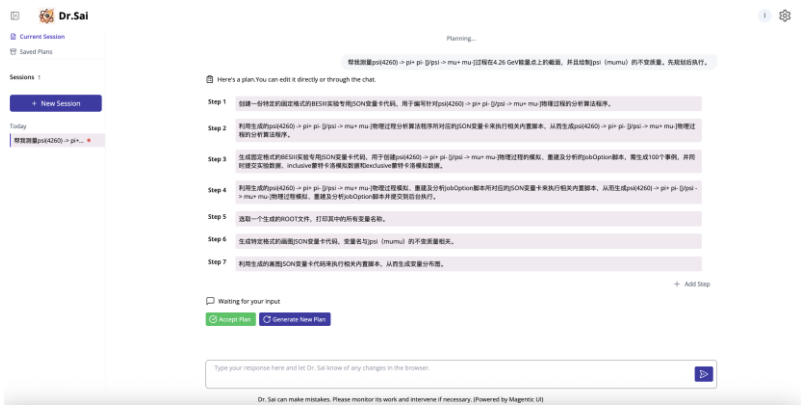


BESIII Dr.Sai, Bolun Zhang (IHEP)

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



On-going



Plan to move to Magentic-UI

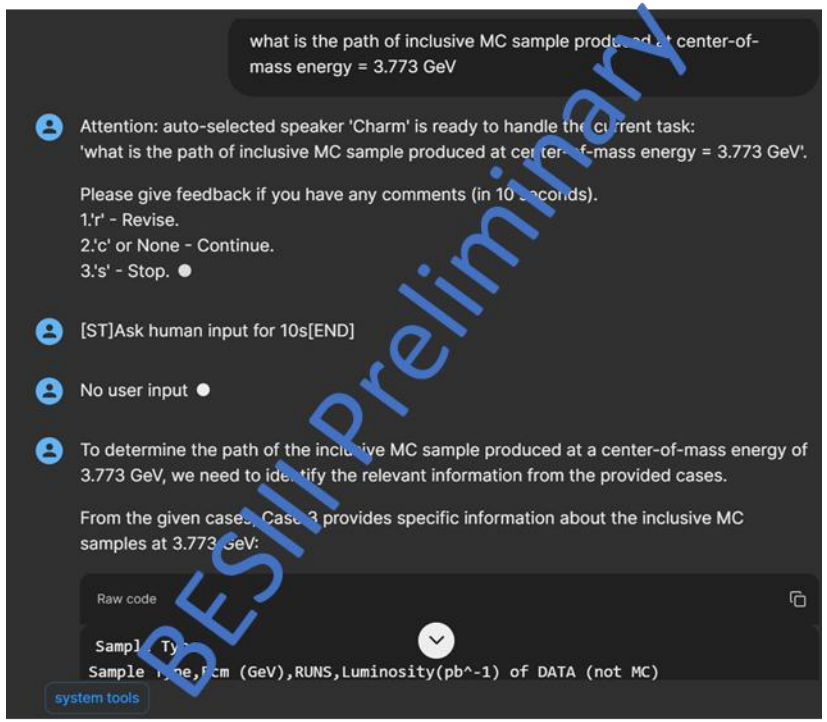
More friendly for interactive

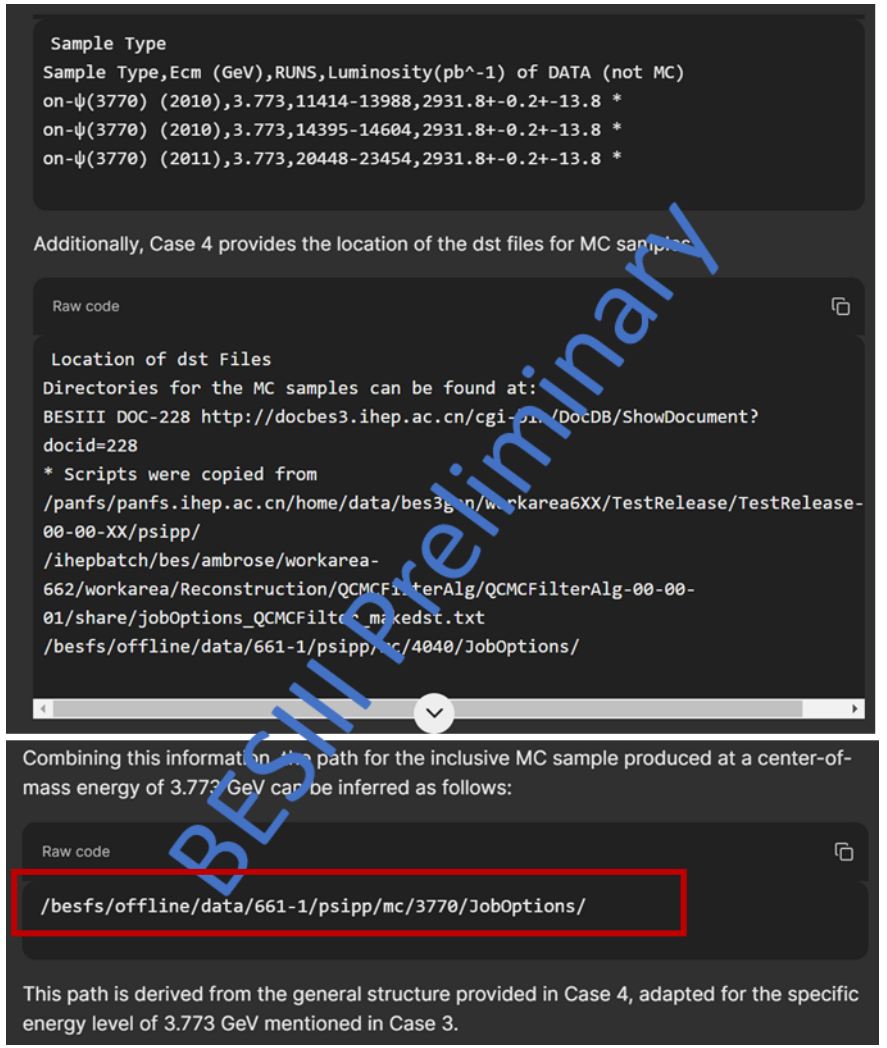
Status of Dr. Sai

- 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 mid-2025 **(almost done)** **!**
  - 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

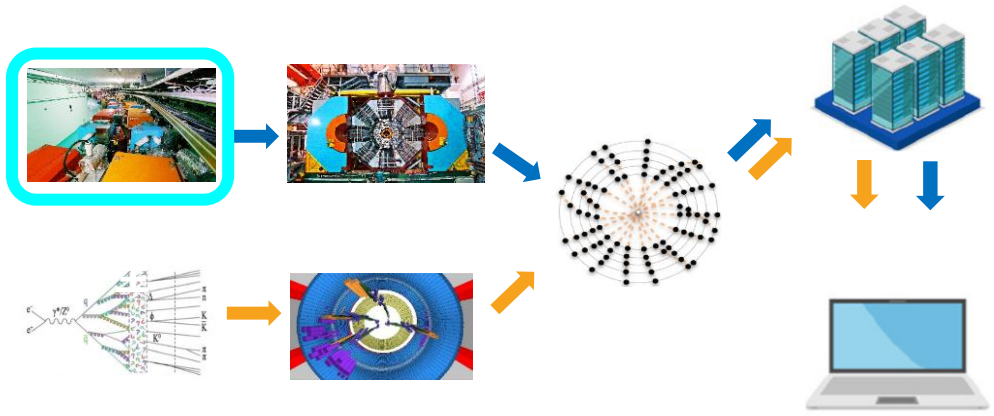
- BESIII internal assistant
  - Prompt: where is the xxx MC sample
  - Then it search in RAG collections
  - LLM read the RAG outputs and conclude correctly

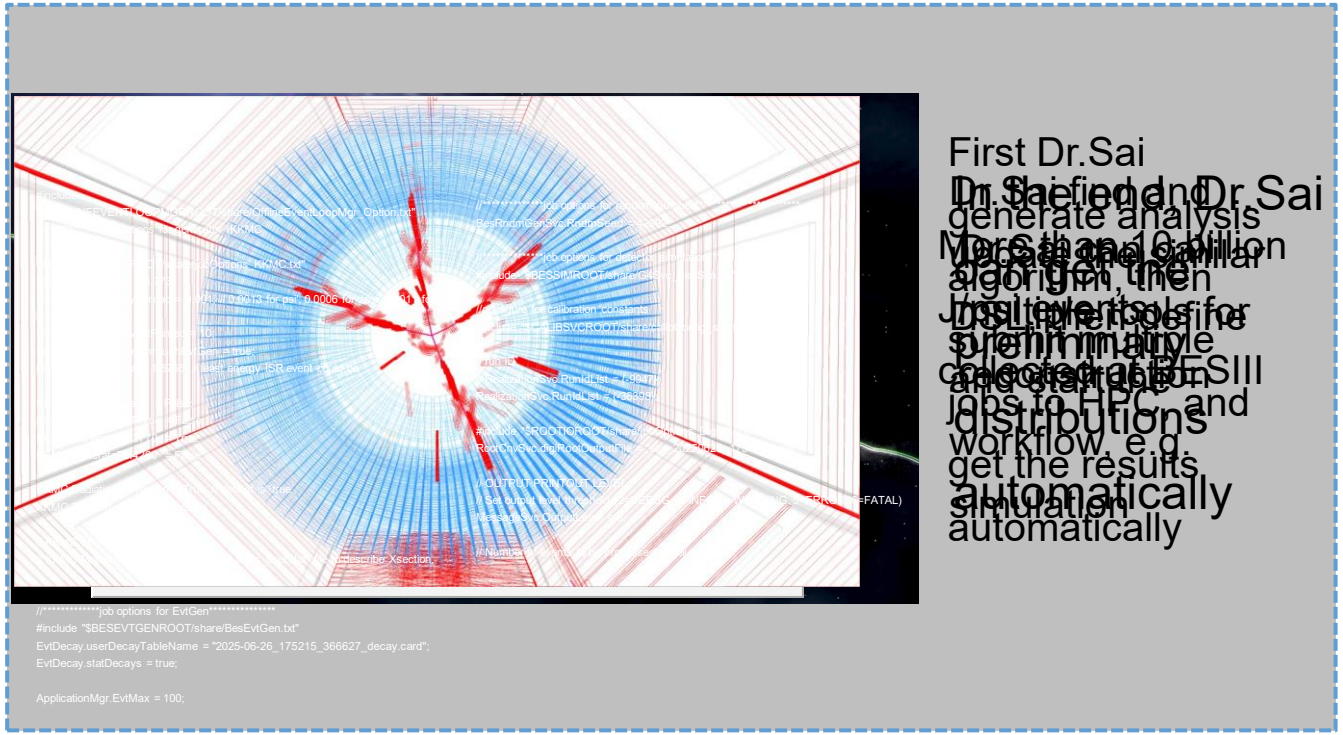




Preliminary results of Dr. Sai V2

User prompt: “please measure the cross section of $e^+e^- \to \pi^+\pi^- J/\psi$ at 4.26 GeV”

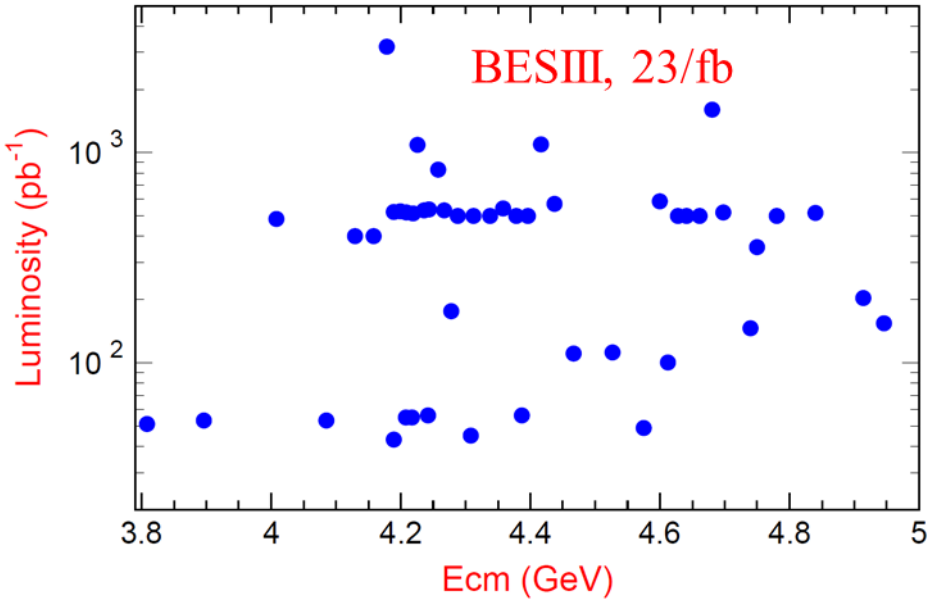


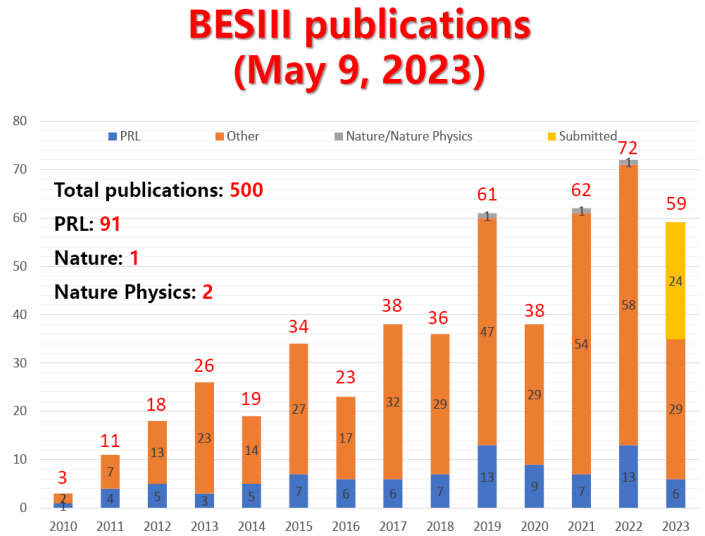


Experience and plan

- 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
- **We plan to use Dr. Sai to remeasure tens of cross sections by the end of 2026** !

Prospects





- >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 code/text draft, but also can be used to automate the analysis
- A demo of AI assistant is built to test the feasibility
- A new version of Dr. Sai will be ready soon
  - It can **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
  - Easy to integrate to other HEP experiments
- There are lots of on-going AI/ML activities at IHEP and BESIII to push "AI for HEP"
  - Welcome to discuss/collaborate !

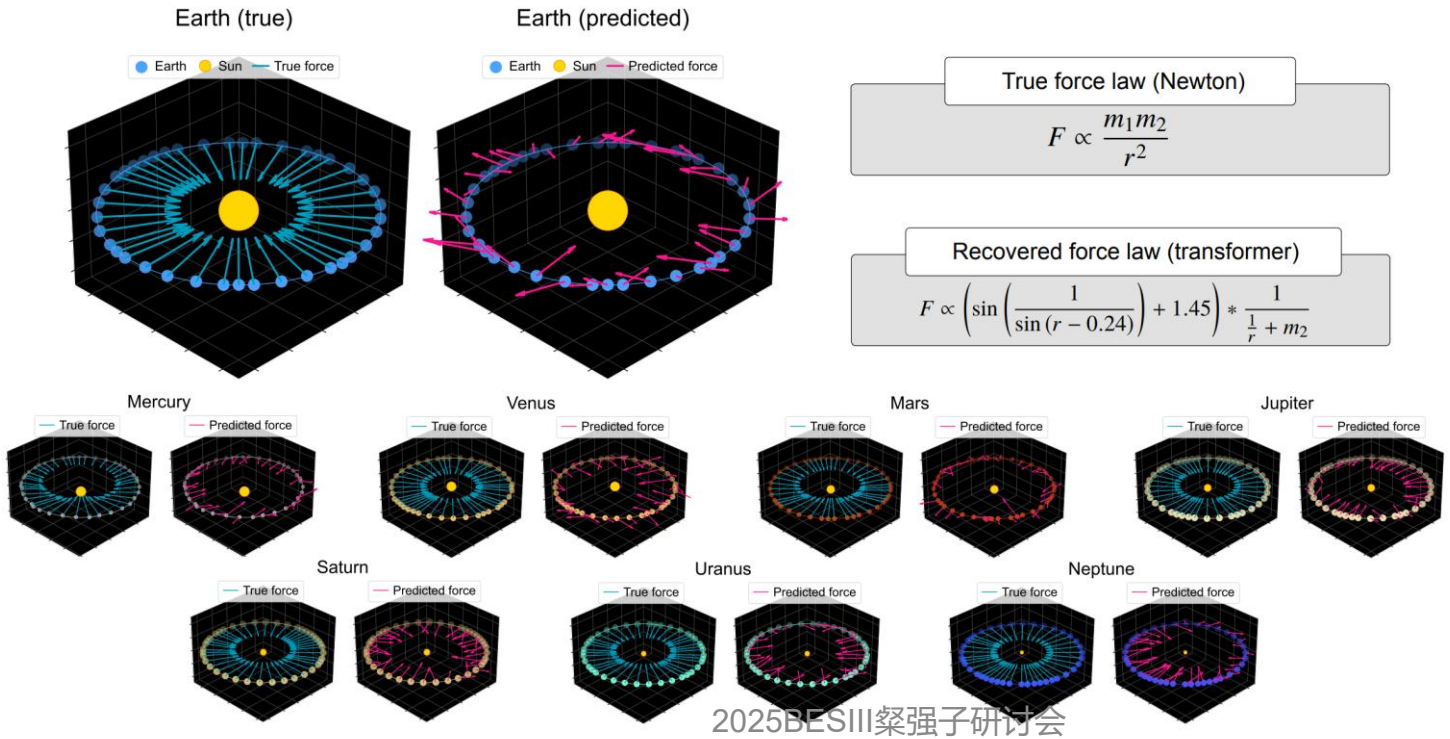
Next: AI+HEP

- Now almost all studies at HEP is “ML”, can we make it **“intelligent”?**
- “AI+” will be the priority task for most of disciplines, could be the next evolution, HEP should not be absent
- Please join us to discuss and design the “AI+HEP” roadmap
  - Quantum computing and machine learning workshop 2025 !
  - https://indico.ihep.ac.cn/event/25857/
  - Save the date, **2025/08/19 – 2025/08/23**
  - See you at Qingdao

back-up

推理与计算能力 ➔ 物理学家？

- 开普勒行星轨道预测模型→牛顿万有引力定律
- 学习预测序列能否揭示更深层次的规律？[arXiv: 2507.06952]
  - 哈佛 MIT研究团队在 1,000 万条仿真太阳系数据上训练的 Transformer 模型能够精准地预测行星轨道，但是并不能得到牛顿定律



推理与计算能力 ➔ 物理学家？

- PHYBench:大规模物理场景下的复杂推理能力评估基准 [arXiv:2504.16074]
  - 由北京大学物理学院和人工智能研究所等机构共同创建
  - 覆盖力、电、热、光、现代物理等多个领域
  - 难度从高中习题到本科习题和物理奥赛

排名	模型	组织机构	得分
1	**Human Expert**	**Peking University**	**71.40**
2	Grok 4	xAI	53.03
3	Gemini 2.5 pro	Google	49.46
4	o3 (high)	OpenAI	46.37
5	Doubao-Seed 1.6-Thinking	Bytedance	45.53
6	DeepSeek-R1 0528	DeepSeek	44.25

- 最强的推理模型仍落后于人类专家
- 大语言模型物理推理能力有限