LLM-based physics analysis agent at BESIII —— Dr. Sai (赛博士)

Yipu Liao (廖一朴)

Institute of High Energy Physics, CAS, Beijing

on behalf of **Dr. Sai** working group

Joint-efforts from IHEP-UCAS-LZU-JLU

Erice, Italy

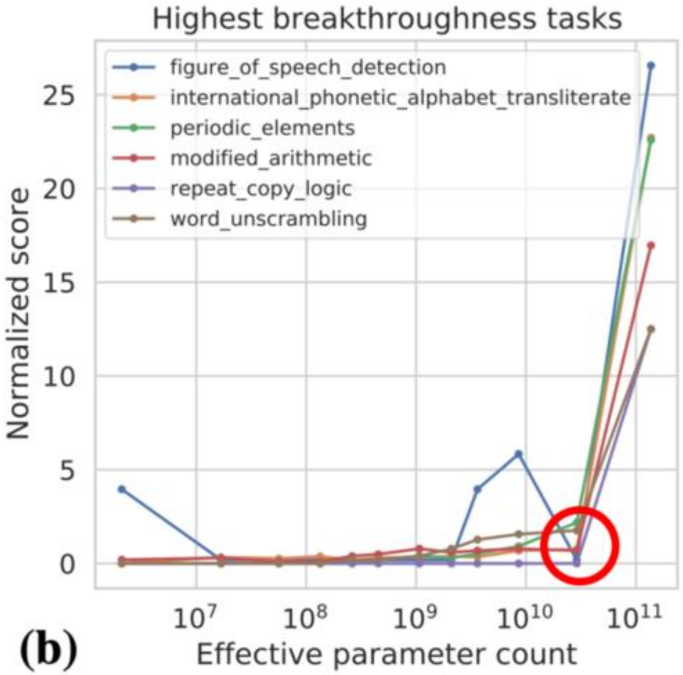
2024.06.20

What is Large Language Model (LLM) ?

- **Large language models (LLMs), normally build on Transformer architecture (Deep Learning), 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)

- **For us, open-source foundation model + higher level model for HEP + fine tuning for BESIII**



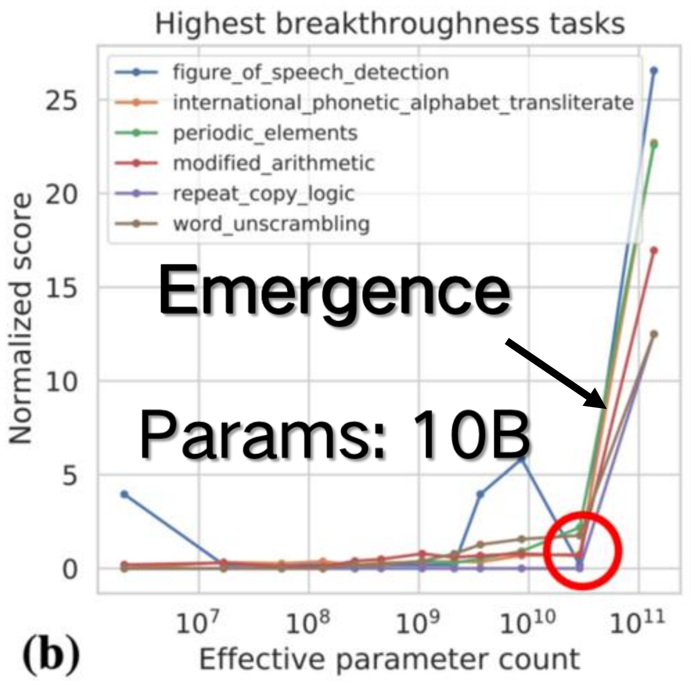
Scientific Reports volume13, Article number: 18562 (2023)

What is Large Language Model (LLM) ?

- **Large language models (LLMs), normally build on Transformer architecture (Deep Learning), has demonstrated impressive performance in text / code generation**
  - GPT4o, Gemini, LLaMa3 ...
  - Could be used for HEP studies
  - Game changer

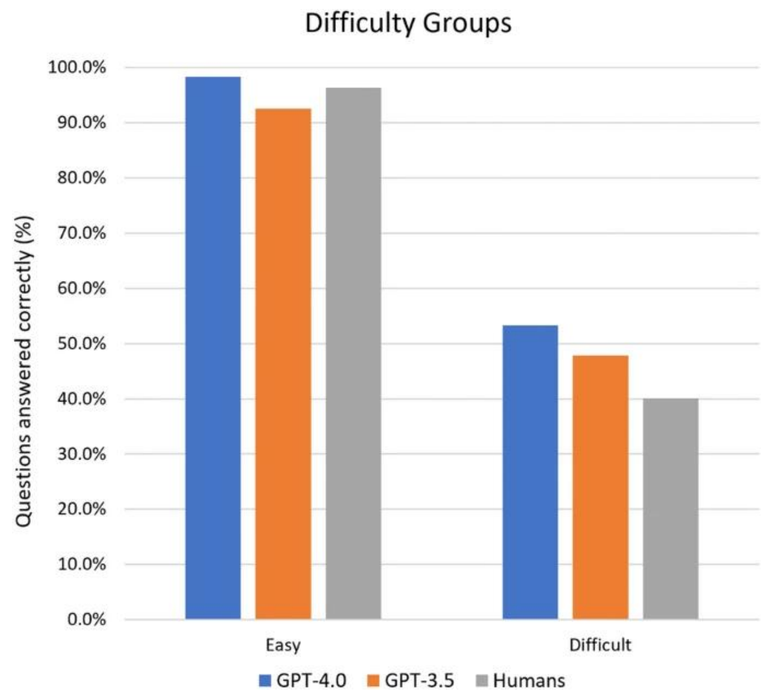
- A foundation model (large, computing intensive) + fine tuning for each task individually (smaller data set)

- **For us, open-source foundation model + higher level model for HEP + fine tuning for BESIII**



Emergence

Params: 10B



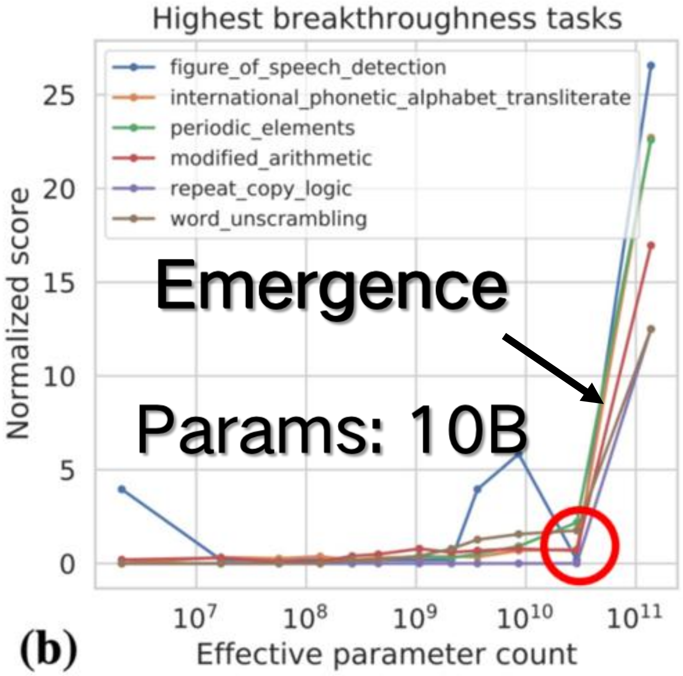
Scientific Reports volume13, Article number: 18562 (2023)

What is Large Language Model (LLM) ?

- **Large language models (LLMs), normally build on Transformer architecture (Deep Learning), has demonstrated impressive performance in text / code generation**
  - GPT4o, Gemini, LLaMa3 ...
  - Could be used for HEP studies
  - Game changer

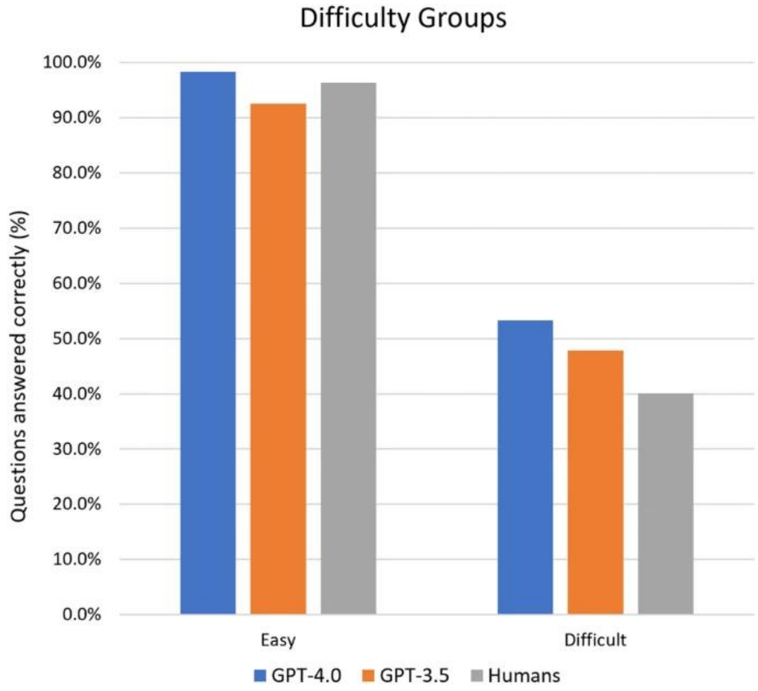
- A foundation model (large, computing intensive) + fine tuning for each task individually (smaller data set)

- **For us, open-source foundation model + higher level model for HEP + fine tuning for BESIII**



Emergence

Params: 10B



Scientific Reports volume13, Article number: 18562 (2023)

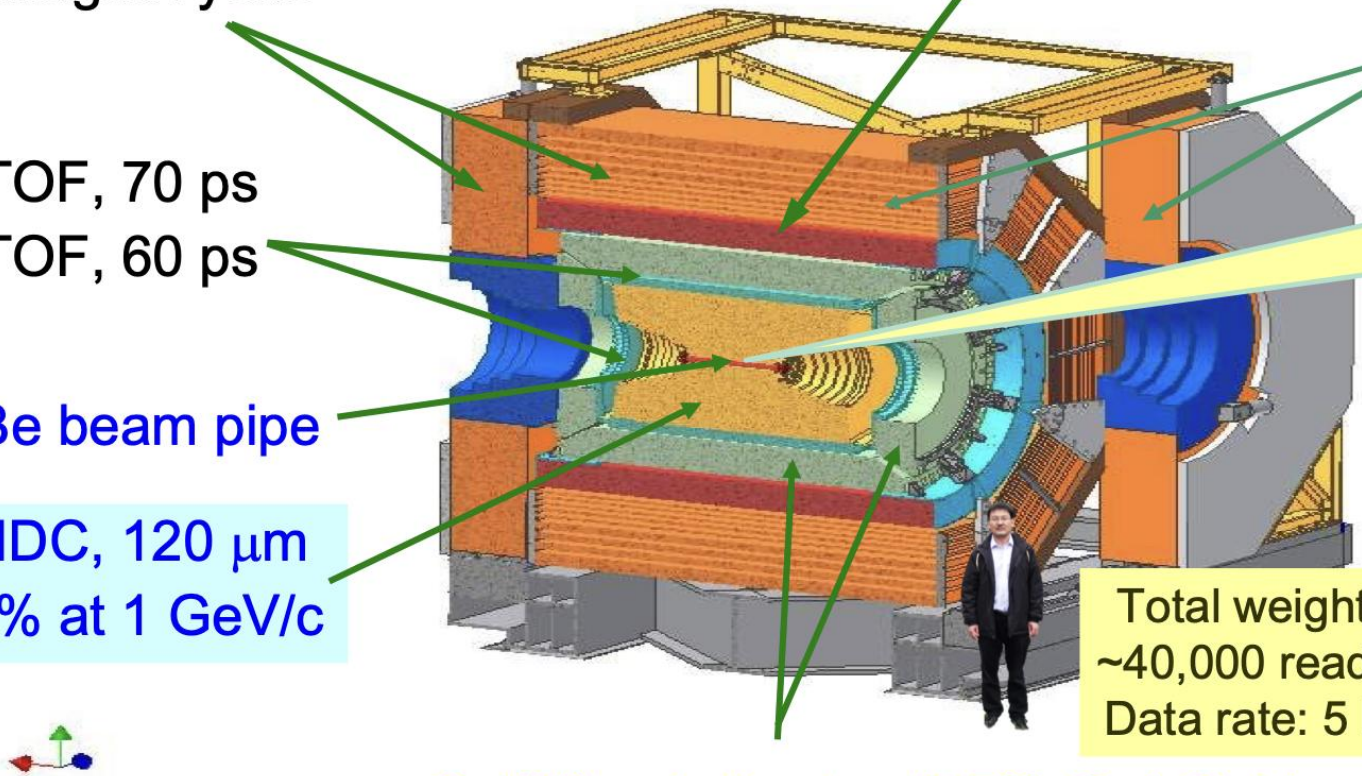
Beijing Electron Positron Collider II (BEPCII)

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

BESIII experiment



SC magnet, 1 T

Magnet yoke

RPC

CGEM-IT is being built & placed here in 2024

BTOF, 70 ps
ETOF, 60 ps

Be beam pipe

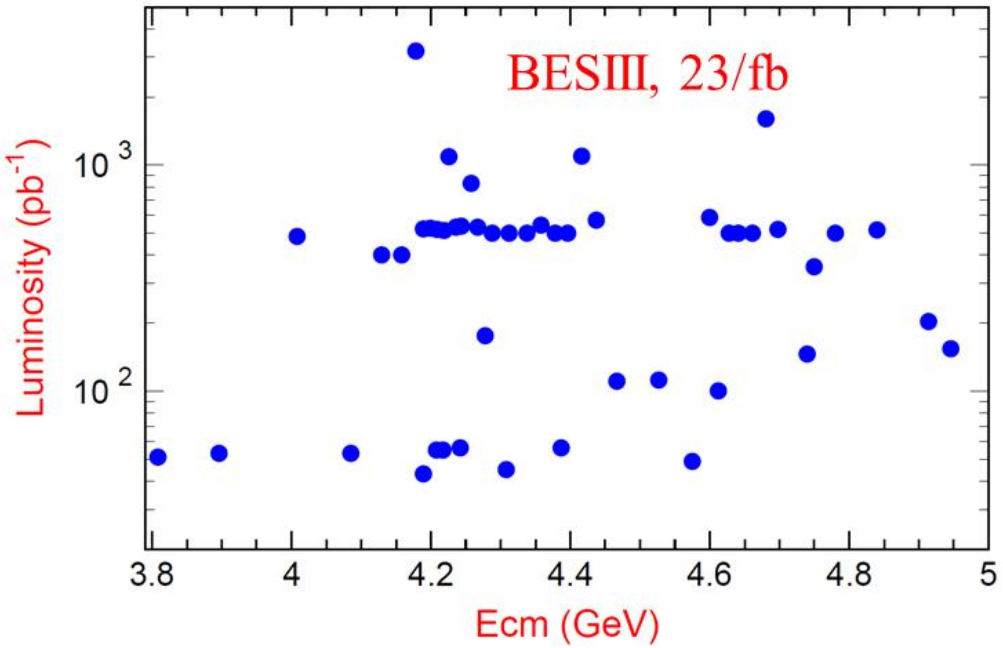
MDC, 120 μm
0.5% at 1 GeV/c

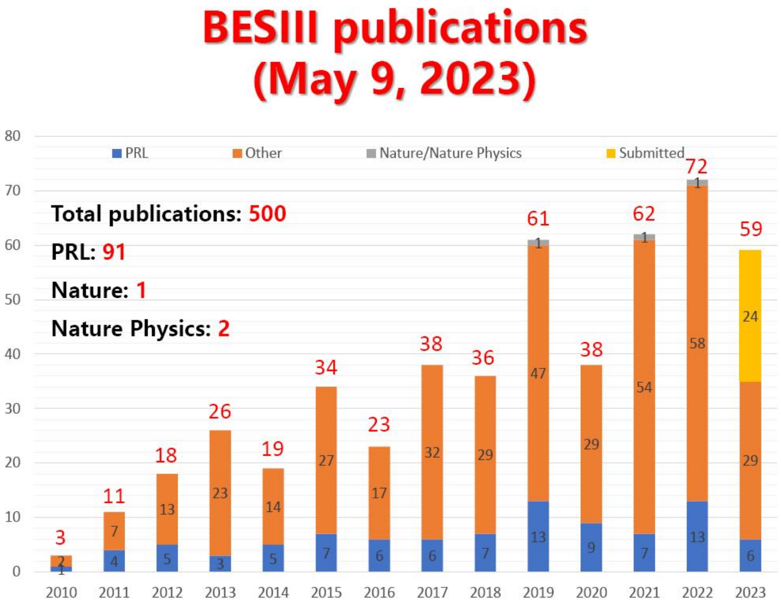
Total weight 750 tonnes,
~40,000 readout channels,
Data rate: 5 kHz, 50 MB/s

CsI(Tl) calorimeter, 2.5% @ 1 GeV

Has been in full operation since 2008,
all subdetectors are in very good status!

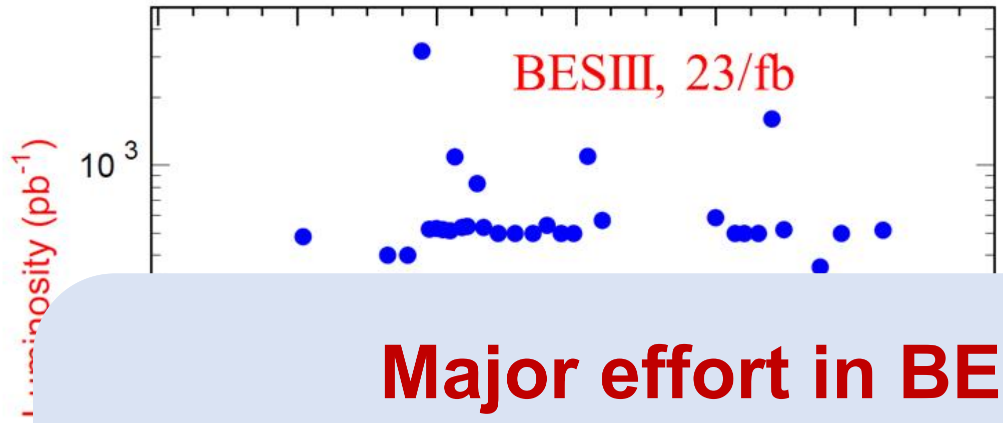
Why we need LLM ?

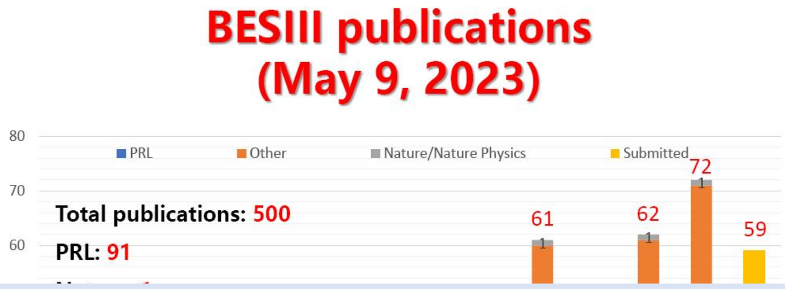




- More data will be collected after BEPCII-upgrade
- > 500 physics results from ~500 people in the past 14 years
  - One result normally took **~3 years**
- We need **a more efficient workflow** in order to achieve the goals in BESIII white paper

Why we need LLM ?





Major effort in BESIII analysis is spent in writing / coding / testing

AND LLM is good at code / text generation !

- More data will be collected after BEPCII-upgrade
- > 500 physics results from ~500 people in the past 14 years
  - One result normally took **~3 years**
- We need **a more efficient workflow** in order to achieve the goals in BESIII white paper

Dr. Sai (赛博士) project for BESIII / HEP

- **AI Agent: AI tools capable of autonomously performing complex tasks**
  - LLM = brain → AI agent = human
- AI agent based on **Xiwu** model (LLM for HEP)
  - based on Llama 2/3, will train with BESIII internal data, e.g. memo/drafts, BOSS source code, Q-A in HyperNews (BESIII internal contact page)
- One milestone: **AI assistant**, It can help scientist on data analysis, e.g. MC generation, signal extraction, and a navigator inside BESIII
  - Internal version release, target at **End of June 2024** !
- Goal: **AI scientist**, it can analyze the data automatically like a real person who have Ph.D degree

Dr. Sai (赛博士) project for BESIII / HEP

- **AI Agent: AI tools capable of autonomously performing complex tasks**
  - LLM = brain → AI agent = human
- AI agent based on **Xiwu** model (LLM for HEP)
  - based on Llama 2/3, will train with BESIII internal data, e.g. memo/drafts, BOSS source code, Q-A in HyperNews (BESIII internal contact page)
- One milestone: **AI assistant**, It can help scientist on data analysis, e.g. MC generation, signal extraction, and a navigator inside BESIII
  - Internal version release, target at **End of June 2024** !
- Goal: **AI scientist**, it can analyze the data automatically like a real person who have Ph.D degree

~20 people from IHEP, UCAS, LZU and JLU, lots of fun stuffs, **welcome to contact and join us !**

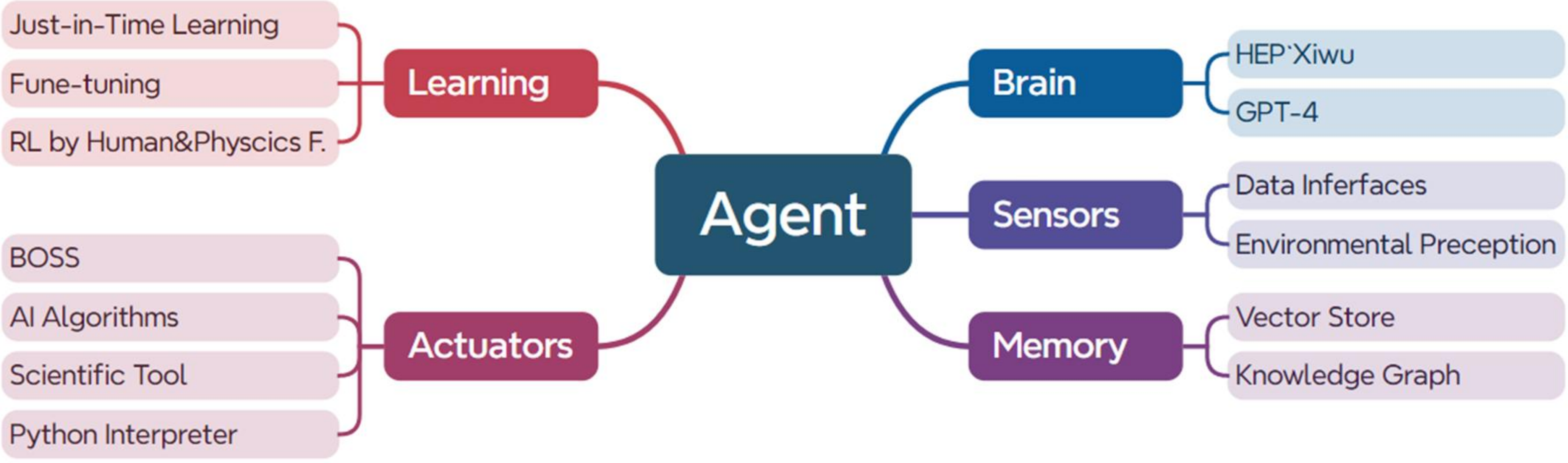
Dr. Sai (赛博士) project for BESIII / HEP

- **AI Agent: AI tools capable of autonomously performing complex tasks**
  - LLM = brain → AI agent = human
- AI agent based on **Xiwu** model (LLM for HEP)
  - based on Llama 2/3, will train with BESIII internal data, e.g. memo/drafts, BOSS source code, Q-A in HyperNews (BESIII internal contact page)
- One milestone: **AI assistant**, It can help scientist on data analysis, e.g. MC generation, signal extraction, and a navigator inside BESIII
  - Internal version release, target at **End of June 2024** !
- Goal: **AI scientist**, it can analyze the data automatically like a real person who have Ph.D degree

~20 people from IHEP, UCAS, LZU and JLU, lots of fun stuffs, **welcome to contact and join us !**

Dr. Sai (赛博士) research agent

Multi-agents framework is developed based on AutoGen



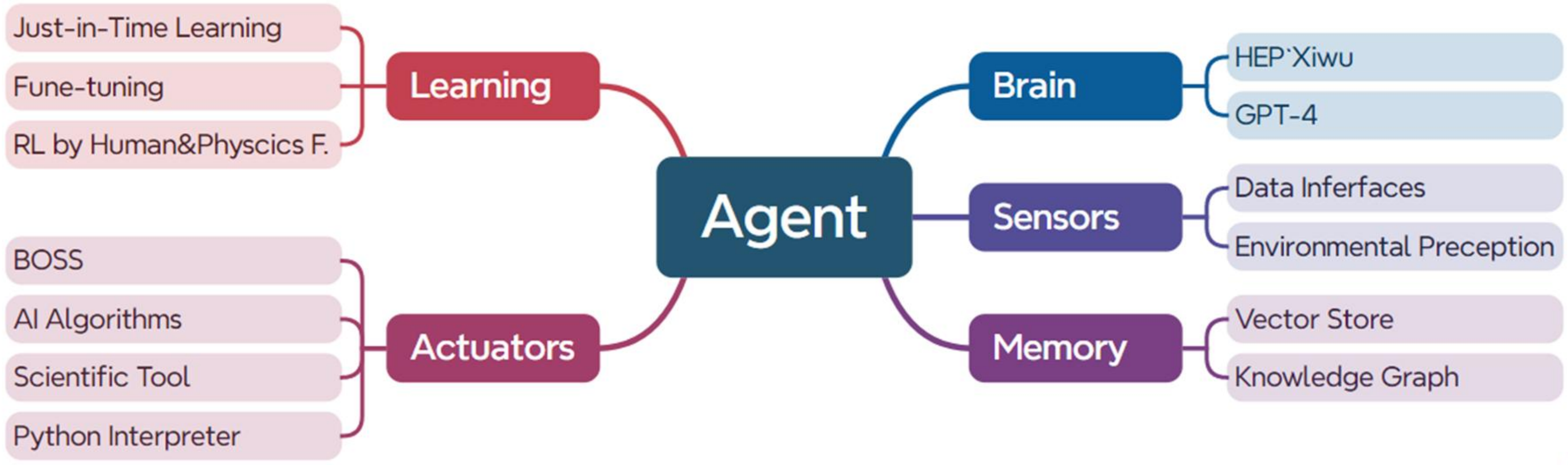
- Agent
  - Learning
    - Just-in-Time Learning
    - Fune-tuning
    - RL by Human&Physcics F.
  - Actuators
    - BOSS
    - AI Algorithms
    - Scientific Tool
    - Python Interpreter
  - Brain
    - HEP`Xiwu
    - GPT-4
  - Sensors
    - Data Inferfaces
    - Environmental Preception
  - Memory
    - Vector Store
    - Knowledge Graph

Key of this project: **make the results from AI more reliable**

- New architecture
- Good quality data
- In-the-fly validation and test **(For next generation!)**

Dr. Sai (赛博士) research agent

Multi-agents framework is developed based on AutoGen



Agent

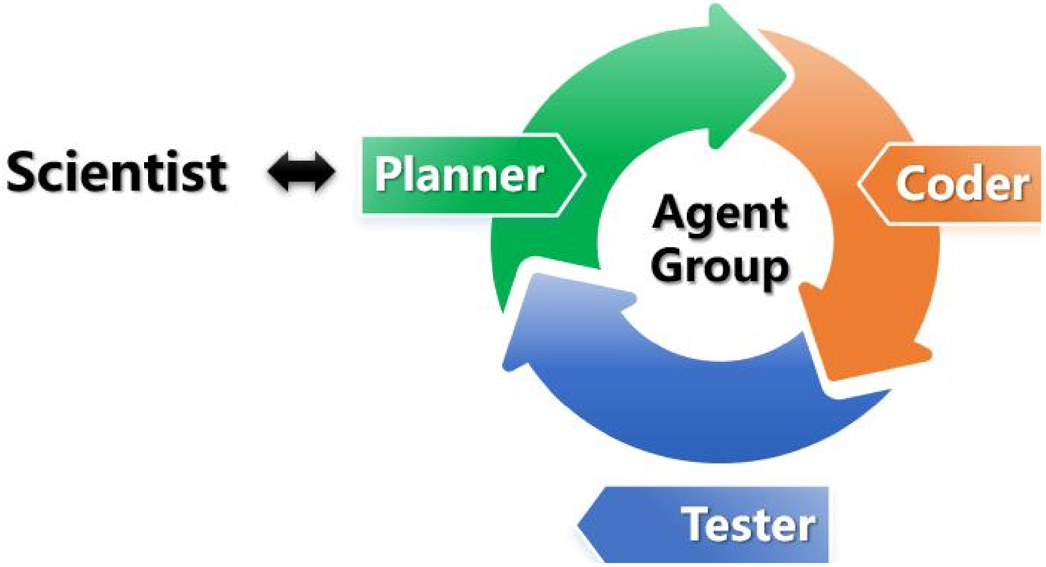
Learning
- Just-in-Time Learning
- Fune-tuning
- RL by Human&Physcics F.

Actuators
- BOSS
- AI Algorithms
- Scientific Tool
- Python Interpreter

Brain
- HEP˙Xiwu
- GPT-4

Sensors
- Data Inferfaces
- Environmental Preception

Memory
- Vector Store
- Knowledge Graph



Scientist ⬌ Planner

Agent Group

Coder

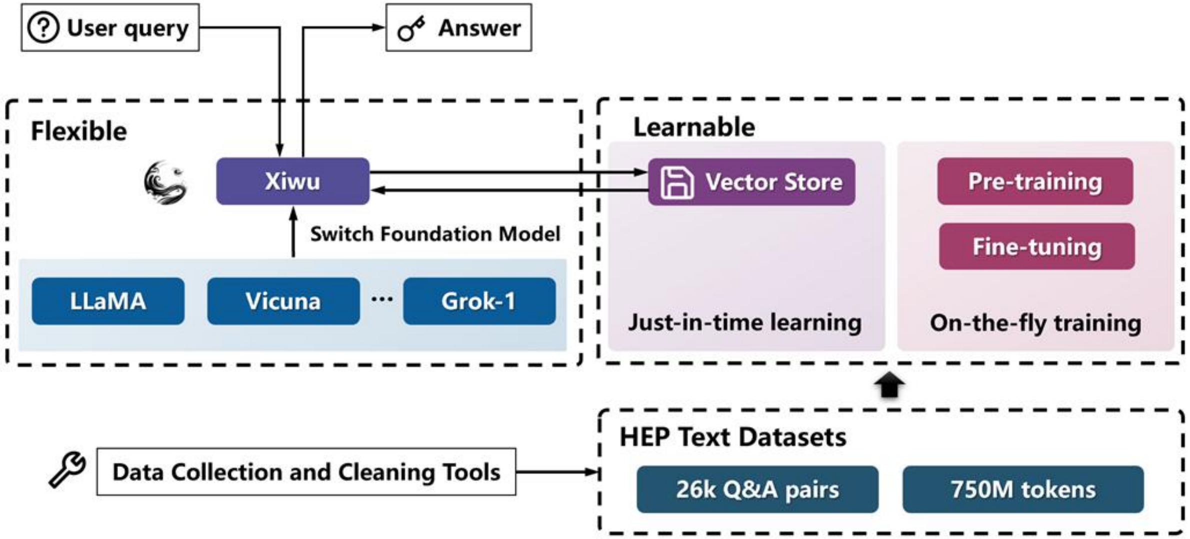
Tester

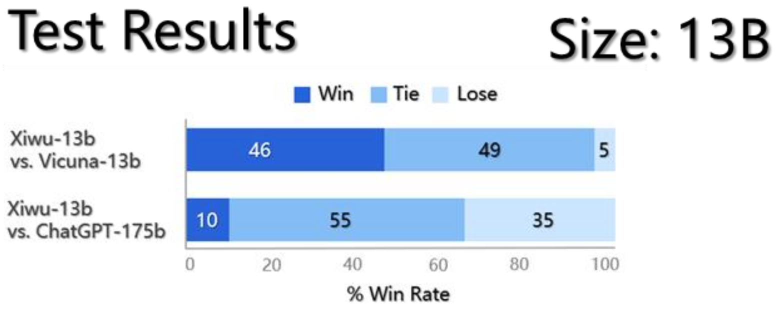
Agents:
- Planner: Planning and tasks decomposition
- Coder: Write BESIII-related codes
- Tester: Using scientific tools for testing

Human can interact via HumanProxy

The brain of Dr. Sai - Xiwu (溪悟) model

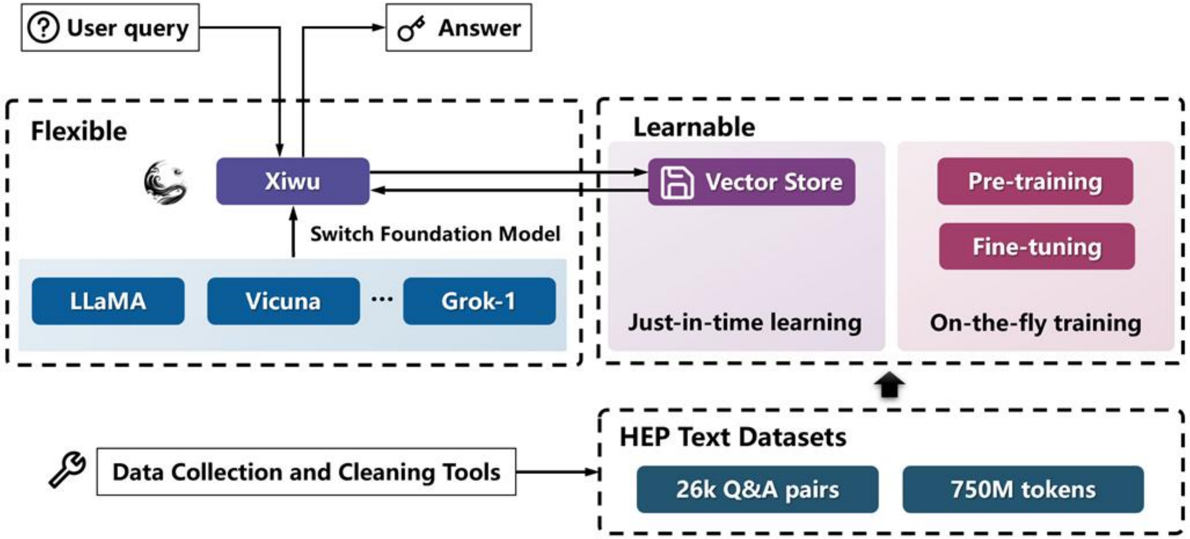
- Xiwu: a basis flexible and learnable LLM for HEP
  Xi(溪)：Streamlet → Drops of water
  Wu(悟)：Understand and inferring
- First version release at April (refer to arXiv:2404.08001 for more details)
  - high level model based on open-source foundational LLM, e.g. LLaMa
  - **First LLM for HEP, version 2 based on LLaMA-3-70B is on-going**

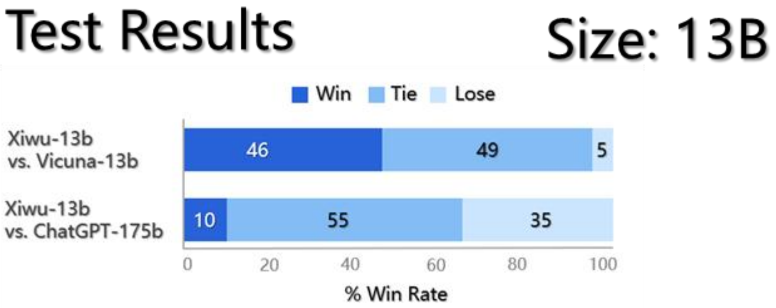




The brain of Dr. Sai - Xiwu (溪悟) model

- Xiwu: a basis flexible and learnable LLM for HEP
  Xi(溪)：Streamlet → Drops of water
  Wu(悟)：Understand and inferring
- First version release at April (refer to arXiv:2404.08001 for more details)
  - high level model based on open-source foundational LLM, e.g. LLaMa
  - **First LLM for HEP, version 2 based on LLaMA-3-70B is on-going**





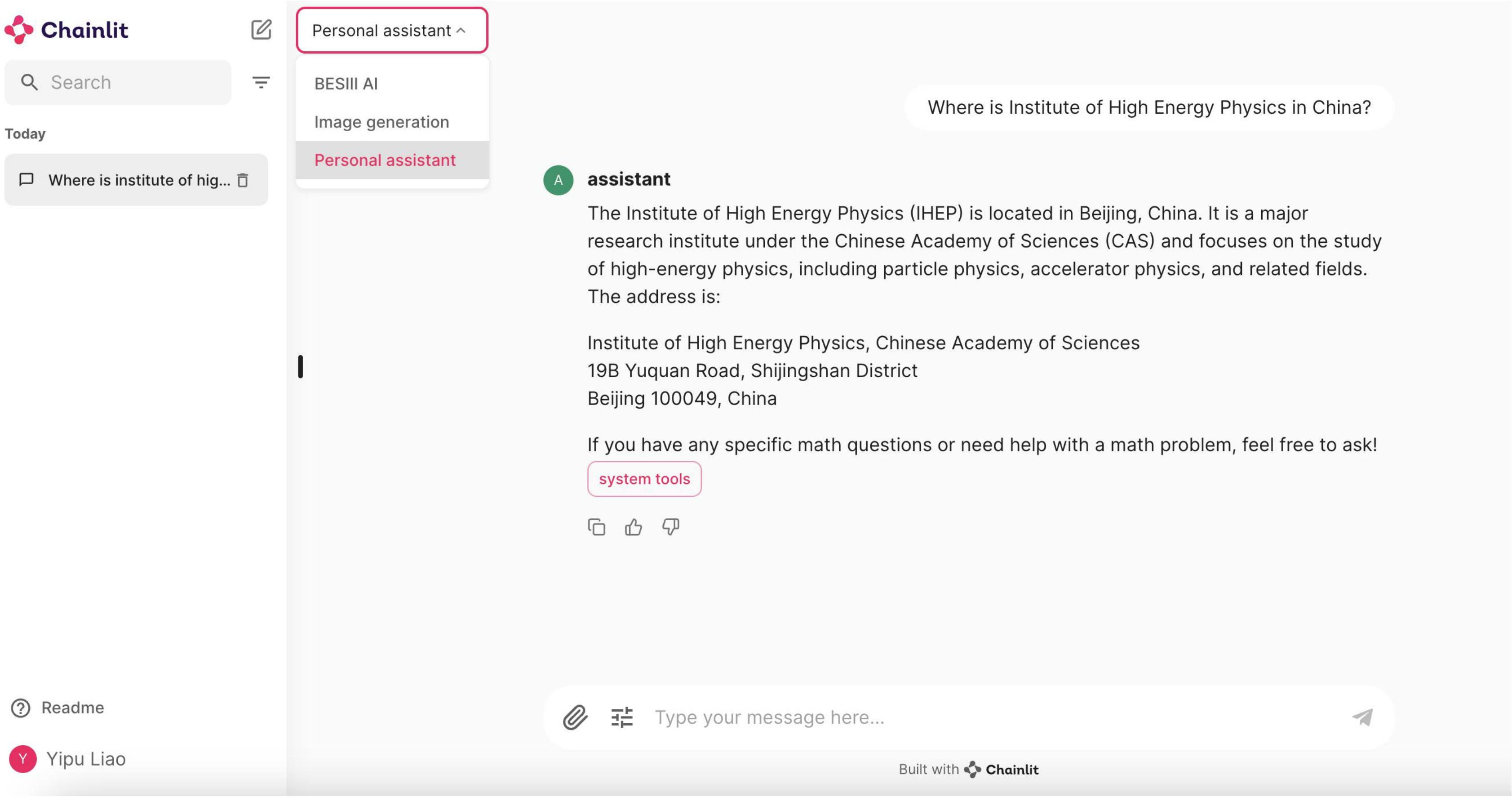
Training data

- Recent papers on arXiv
  - PDF files parser: Hai-Nougat, advanced iteration of the Nougat model
- Good quality chat history from IHEP-AI platform
  - The data is cleaned by human or AI (GPT4)
  - 180k Question-Answer pairs in 3 months
- **BESIII internal data**
  - internal memo, parsed by Hai-Nougat
  - Question-Answer pairs from HyperNews during internal paper review
  - BESIII Offline Software System (BOSS) source code
  - BESIII public webpages and internal webpages
  - The data on indico will be used later
- All the BESIII internal data sets are stored in Retrieval Augmented Generation (RAG) or used in training and fine-tuning

Training data

- Recent papers on arXiv
  - PDF files parser: Hai-Nougat, advanced iteration of the Nougat model
- Good quality chat history from IHEP-AI platform
  - The data is cleaned by human or AI (GPT4)
  - 180k Question-Answer pairs in 3 months
- **BESIII internal data**
  - internal memo, parsed by Hai-Nougat
  - Question-Answer pairs from HyperNews during internal paper review
  - BESIII Offline Software System (BOSS) source code
  - BESIII public webpages and internal webpages
  - The data on indico will be used later
- All the BESIII internal data sets are stored in Retrieval Augmented Generation (RAG) or used in training and fine-tuning

Performance: General LLM for Q&A



Where is Institute of High Energy Physics in China?

assistant

The Institute of High Energy Physics (IHEP) is located in Beijing, China. It is a major research institute under the Chinese Academy of Sciences (CAS) and focuses on the study of high-energy physics, including particle physics, accelerator physics, and related fields. The address is:

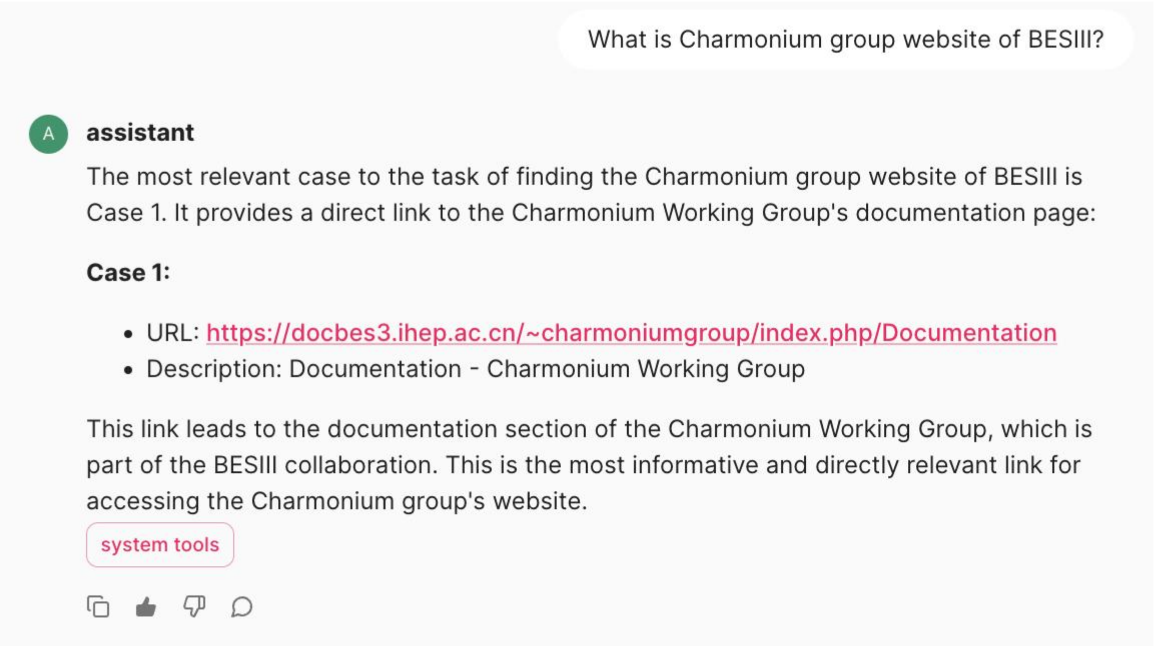
Institute of High Energy Physics, Chinese Academy of Sciences
19B Yuquan Road, Shijingshan District
Beijing 100049, China

If you have any specific math questions or need help with a math problem, feel free to ask!

- WebUI based on Chainlit, rewriting with JaveScript now
- Same to general GPT
- Transfer IHEP AI platform from https://ai.ihep.ac.cn/ to here

Performance: Internal navigator at BESIII

- Same with the chATLAS project at ATLAS
  - Navigator and assistant to replace the simple ‘search’
  - BESIII internal data at websites (bes3.ihep.ac.cn) and HyperNews
    - Not public yet
  - In general, better performance than I expected
    - E.g. Question ‘where is the XXXX MC sample’,
    - Answer ‘The path of the sample is in XXXXXX’



What is Charmonium group website of BESIII?

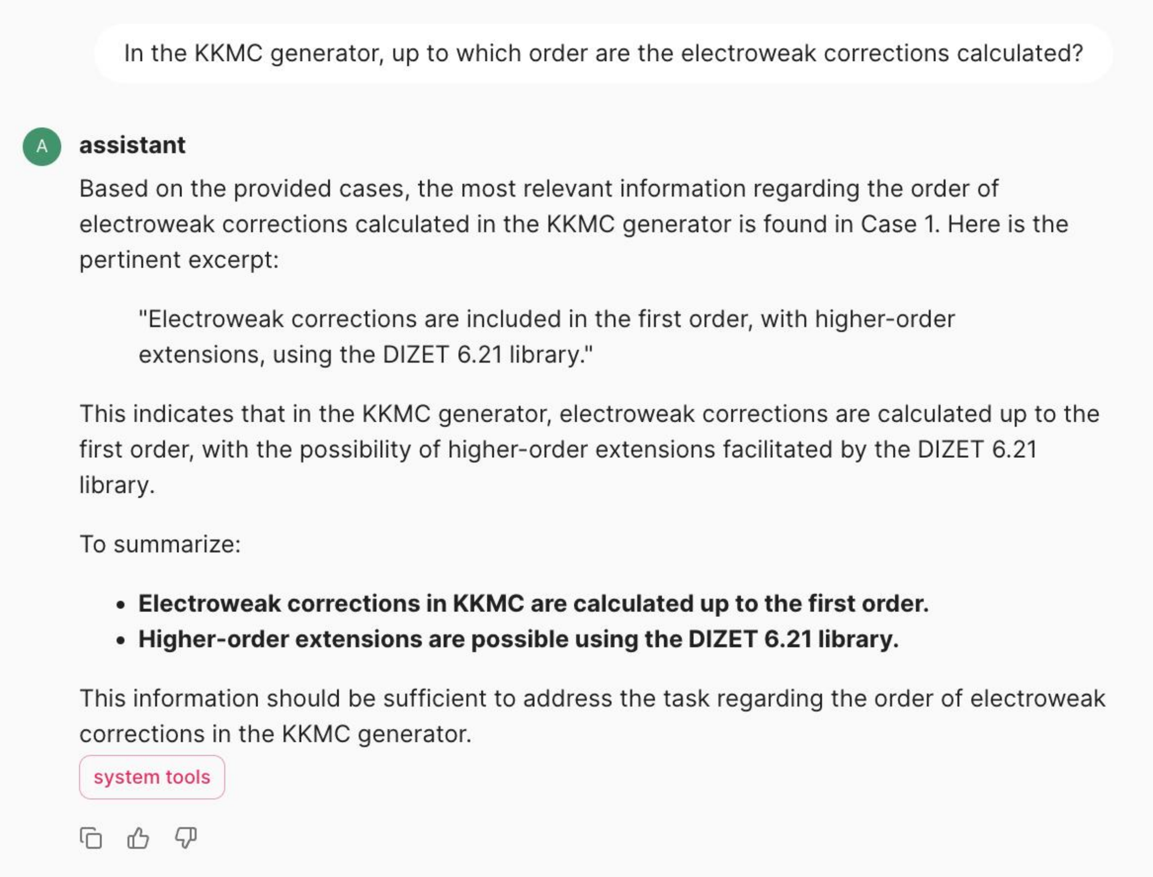
assistant

The most relevant case to the task of finding the Charmonium group website of BESIII is Case 1. It provides a direct link to the Charmonium Working Group's documentation page:

Case 1:

- URL: https://docbes3.ihep.ac.cn/~charmoniumgroup/index.php/Documentation
- Description: Documentation - Charmonium Working Group

This link leads to the documentation section of the Charmonium Working Group, which is part of the BESIII collaboration. This is the most informative and directly relevant link for accessing the Charmonium group's website.



In the KKMC generator, up to which order are the electroweak corrections calculated?

assistant

Based on the provided cases, the most relevant information regarding the order of electroweak corrections calculated in the KKMC generator is found in Case 1. Here is the pertinent excerpt:

> "Electroweak corrections are included in the first order, with higher-order extensions, using the DIZET 6.21 library."

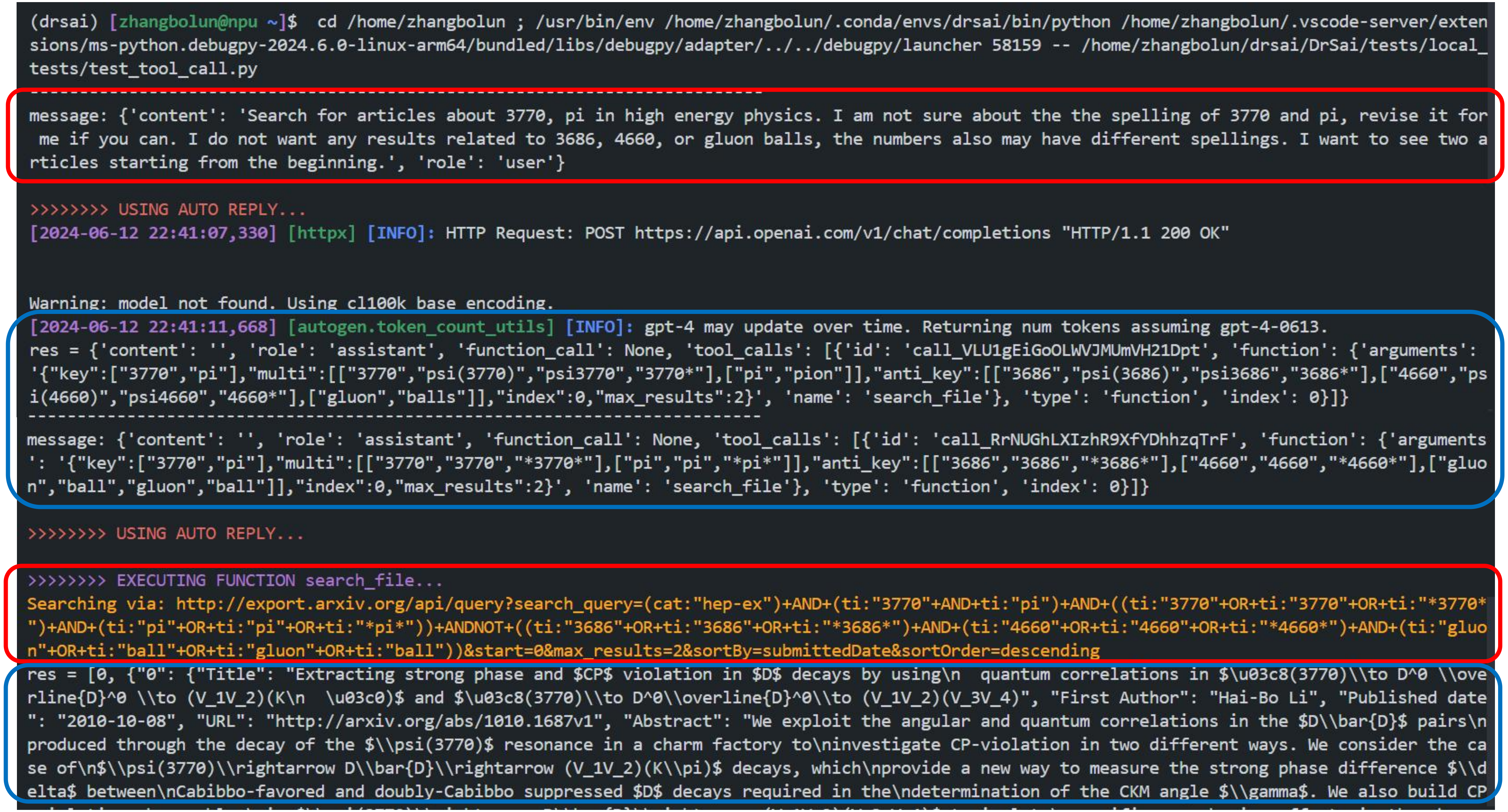
This indicates that in the KKMC generator, electroweak corrections are calculated up to the first order, with the possibility of higher-order extensions facilitated by the DIZET 6.21 library.

To summarize:

- **Electroweak corrections in KKMC are calculated up to the first order.**
- **Higher-order extensions are possible using the DIZET 6.21 library.**

This information should be sufficient to address the task regarding the order of electroweak corrections in the KKMC generator.

Performance: arXiv paper searching



```
(drsai) [zhangbolun@npu ~]$  cd /home/zhangbolun ; /usr/bin/env /home/zhangbolun/.conda/envs/drsai/bin/python /home/zhangbolun/.vscode-server/exten
sions/ms-python.debugpy-2024.6.0-linux-arm64/bundled/libs/debugpy/adapter/../../debugpy/launcher 58159 -- /home/zhangbolun/drsai/DrSai/tests/local_
tests/test_tool_call.py
--------------------------------------------------------------------------
```

User request

```
message: {'content': 'Search for articles about 3770, pi in high energy physics. I am not sure about the the spelling of 3770 and pi, revise it for
 me if you can. I do not want any results related to 3686, 4660, or gluon balls, the numbers also may have different spellings. I want to see two a
rticles starting from the beginning.', 'role': 'user'}

>>>>>>>> USING AUTO REPLY...
[2024-06-12 22:41:07,330] [httpx] [INFO]: HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"


Warning: model not found. Using cl100k base encoding.
```

Parameters extracted by LLM

```
[2024-06-12 22:41:11,668] [autogen.token_count_utils] [INFO]: gpt-4 may update over time. Returning num tokens assuming gpt-4-0613.
res = {'content': '', 'role': 'assistant', 'function_call': None, 'tool_calls': [{'id': 'call_VLU1gEiGoOLWVJMUmVH21Dpt', 'function': {'arguments':
'{"key":["3770","pi"],"multi":[["3770","psi(3770)","psi3770","3770*"],["pi","pion"]],"anti_key":[["3686","psi(3686)","psi3686","3686*"],["4660","ps
i(4660)","psi4660","4660*"],["gluon","balls"]],"index":0,"max_results":2}', 'name': 'search_file'}, 'type': 'function', 'index': 0}]}
--------------------------------------------------------------------------
message: {'content': '', 'role': 'assistant', 'function_call': None, 'tool_calls': [{'id': 'call_RrNUGhLXIzhR9XfYDhhzqTrF', 'function': {'arguments
': '{"key":["3770","pi"],"multi":[["3770","3770","*3770*"],["pi","pi","*pi*"]],"anti_key":[["3686","3686","*3686*"],["4660","4660","*4660*"],["gluo
n","ball","gluon","ball"]],"index":0,"max_results":2}', 'name': 'search_file'}, 'type': 'function', 'index': 0}]}

>>>>>>>> USING AUTO REPLY...
```

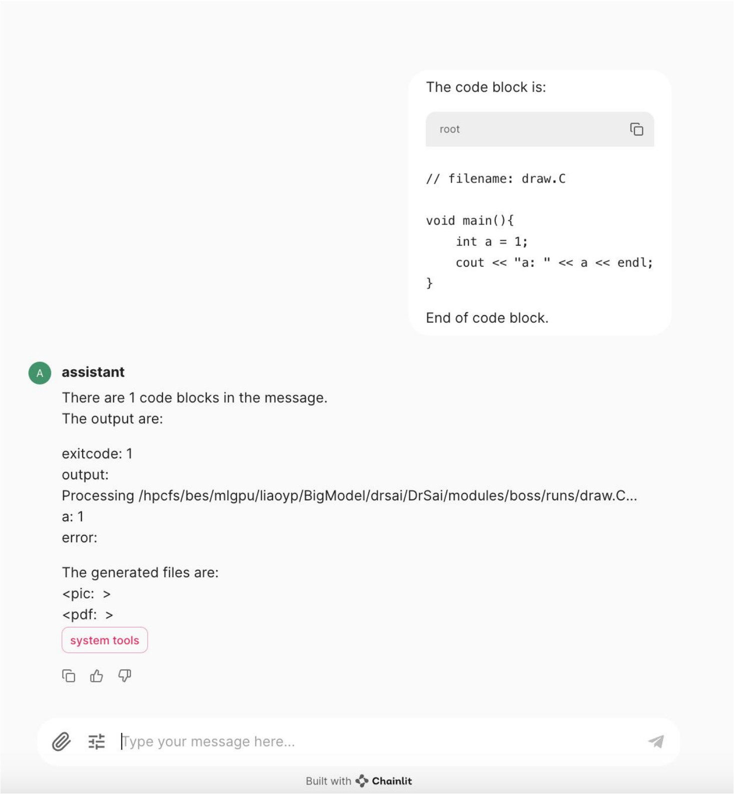
Quary URL
to arXiv

```
>>>>>>>> EXECUTING FUNCTION search_file...
Searching via: http://export.arxiv.org/api/query?search_query=(cat:"hep-ex")+AND+(ti:"3770"+AND+ti:"pi")+AND+((ti:"3770"+OR+ti:"3770"+OR+ti:"*3770*
")+AND+(ti:"pi"+OR+ti:"pi"+OR+ti:"*pi*"))+ANDNOT+((ti:"3686"+OR+ti:"3686"+OR+ti:"*3686*")+AND+(ti:"4660"+OR+ti:"4660"+OR+ti:"*4660*")+AND+(ti:"gluo
n"+OR+ti:"ball"+OR+ti:"gluon"+OR+ti:"ball"))&start=0&max_results=2&sortBy=submittedDate&sortOrder=descending
```

Quary results
(title, abstract,
authors ...)

```
res = [0, {"0": {"Title": "Extracting strong phase and $CP$ violation in $D$ decays by using\n  quantum correlations in $\u03c8(3770)\\to D^0 \\ove
rline{D}^0 \\to (V_1V_2)(K\n  \u03c0)$ and $\u03c8(3770)\\to D^0\\overline{D}^0\\to (V_1V_2)(V_3V_4)", "First Author": "Hai-Bo Li", "Published date
": "2010-10-08", "URL": "http://arxiv.org/abs/1010.1687v1", "Abstract": "We exploit the angular and quantum correlations in the $D\\bar{D}$ pairs\n
produced through the decay of the $\\psi(3770)$ resonance in a charm factory to\ninvestigate CP-violation in two different ways. We consider the ca
se of\n$\\psi(3770)\\rightarrow D\\bar{D}\\rightarrow (V_1V_2)(K\\pi)$ decays, which\nprovide a new way to measure the strong phase difference $\\d
elta$ between\nCabibbo-favored and doubly-Cabibbo suppressed $D$ decays required in the\ndetermination of the CKM angle $\\gamma$. We also build CP
```

Performance: coding & execution



The code block is:

```
root
```

```
// filename: draw.C

void main(){
    int a = 1;
    cout << "a: " << a << endl;
}
```

End of code block.

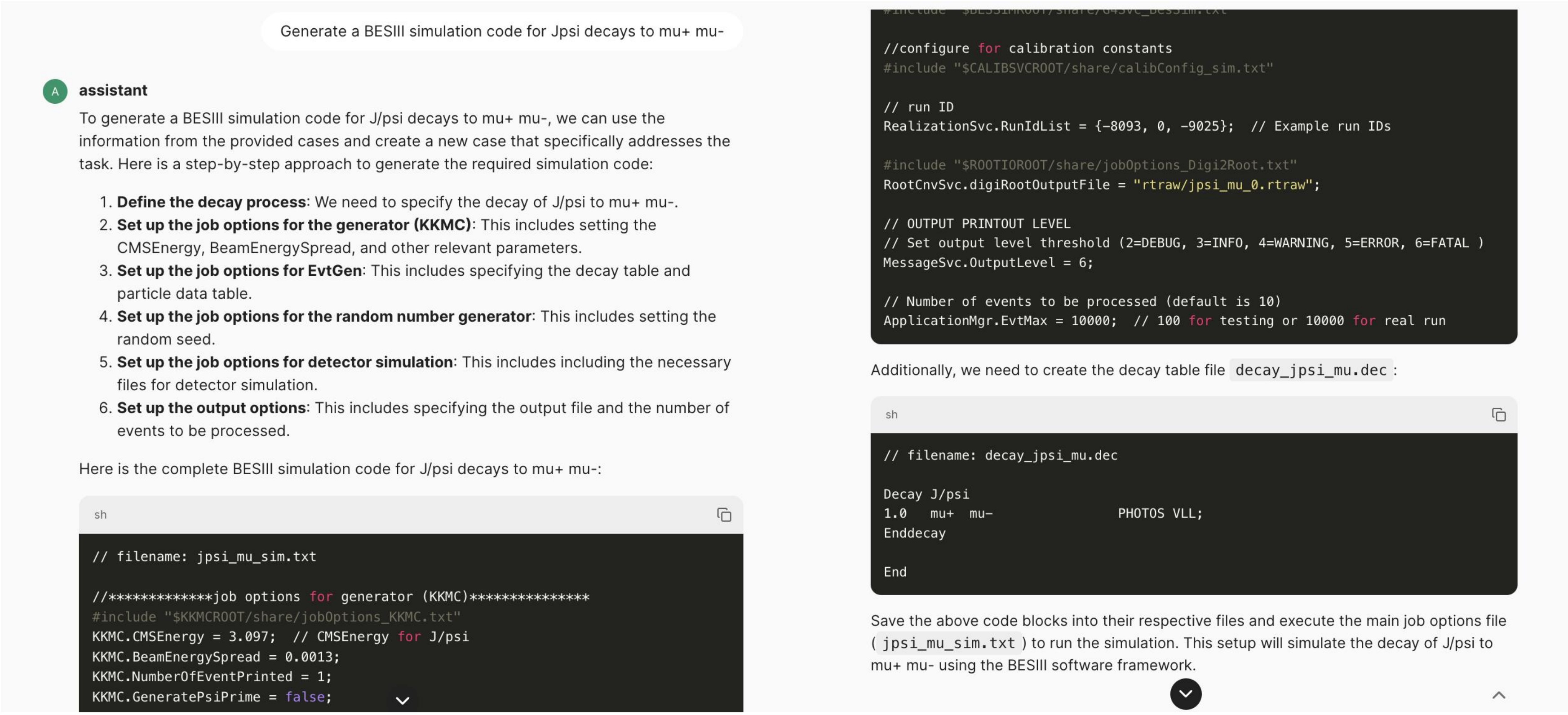
assistant

There are 1 code blocks in the message.
The output are:

```
exitcode: 1
output:
Processing /hpcfs/bes/mlgpu/liaoyp/BigModel/drsai/DrSai/modules/boss/runs/draw.C...
a: 1
error:
```

The generated files are:
<pic: >
<pdf: >

CERN ROOT code execution



Generate a BESIII simulation code for Jpsi decays to mu+ mu-

assistant

To generate a BESIII simulation code for J/psi decays to mu+ mu-, we can use the information from the provided cases and create a new case that specifically addresses the task. Here is a step-by-step approach to generate the required simulation code:

1. **Define the decay process**: We need to specify the decay of J/psi to mu+ mu-.
2. **Set up the job options for the generator (KKMC)**: This includes setting the CMSEnergy, BeamEnergySpread, and other relevant parameters.
3. **Set up the job options for EvtGen**: This includes specifying the decay table and particle data table.
4. **Set up the job options for the random number generator**: This includes setting the random seed.
5. **Set up the job options for detector simulation**: This includes including the necessary files for detector simulation.
6. **Set up the output options**: This includes specifying the output file and the number of events to be processed.

Here is the complete BESIII simulation code for J/psi decays to mu+ mu-:

```sh
// filename: jpsi_mu_sim.txt

//*************job options for generator (KKMC)***************
#include "$KKMCROOT/share/jobOptions_KKMC.txt"
KKMC.CMSEnergy = 3.097;  // CMSEnergy for J/psi
KKMC.BeamEnergySpread = 0.0013;
KKMC.NumberOfEventPrinted = 1;
KKMC.GeneratePsiPrime = false;
```

```
//configure for calibration constants
#include "$CALIBSVCROOT/share/calibConfig_sim.txt"

// run ID
RealizationSvc.RunIdList = {-8093, 0, -9025};  // Example run IDs

#include "$ROOTIOROOT/share/jobOptions_Digi2Root.txt"
RootCnvSvc.digiRootOutputFile = "rtraw/jpsi_mu_0.rtraw";

// OUTPUT PRINTOUT LEVEL
// Set output level threshold (2=DEBUG, 3=INFO, 4=WARNING, 5=ERROR, 6=FATAL )
MessageSvc.OutputLevel = 6;

// Number of events to be processed (default is 10)
ApplicationMgr.EvtMax = 10000;  // 100 for testing or 10000 for real run
```

Additionally, we need to create the decay table file `decay_jpsi_mu.dec`:

```sh
// filename: decay_jpsi_mu.dec

Decay J/psi
1.0    mu+  mu-                PHOTOS VLL;
Enddecay

End
```

Save the above code blocks into their respective files and execute the main job options file (`jpsi_mu_sim.txt`) to run the simulation. This setup will simulate the decay of J/psi to mu+ mu- using the BESIII software framework.

BESIII Offline Sofware System (BOSS) coding

Summary

- **AI era is coming !**
  - It will not replace you but will help you to work more efficiently
- First AI agent for HEP - **Dr. Sai**
  - For now it can do simple tasks, e.g. write / update code
  - Timeline: beta version at end of June 2024, stable version at the end of 2024
- Next: more data, multi-model, e.g. slides on indico, experts' chat history at IHEP AI platform https://ai.ihep.ac.cn/ (or https://chat.ihep.ac.cn/)
  - **We want and need your experience and expertise** !
- Similar projects from other experiments:
  - LHC： AccGPT (LLaMa), AI assistant for accelerator control
  - ATLAS： chATLAS (GPT), AI assistant for internal navigator, same with one component of Dr. Sai
  - Outreach/education: outreach assistant (GPT), train people outside of HEP to analyze ATLAS open-data, same with one component of Dr. Sai

Yipu Liao (廖一朴)
Institute of High Energy Physics, CAS, Beijing
Email: liaoyp@ihep.ac.cn

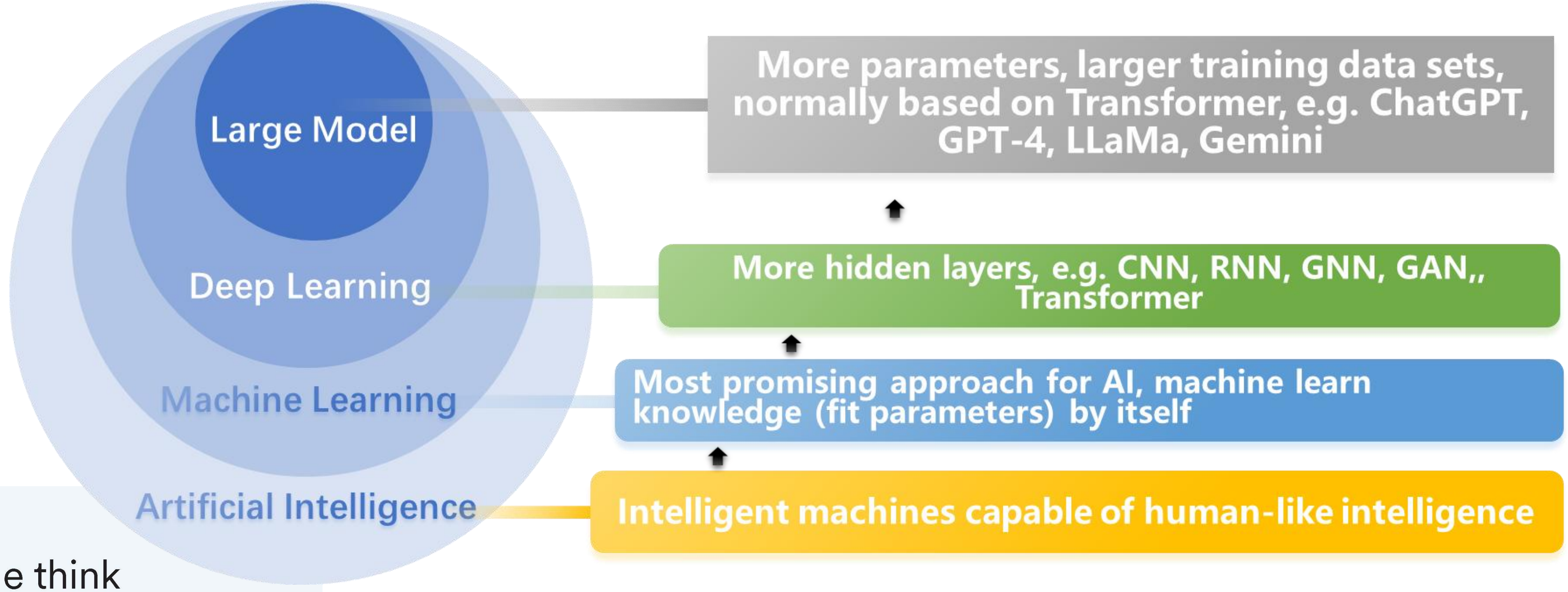
Thank you for listening!

Group members

Zheng-De Zhang, Yi-Yu Zhang, Jian-Fang Li, Dong-Bo Xiong, Siyang Chen, Qian-Ran Sun, Hao-Fan Wang, Fa-Zhi Qi, Chang-Zheng Yuan, Ke Li, Yi-Pu Liao, Bo-Lun Zhang, Ming-Run Li, Pan Huang, Jun-Kun Jiao (JLU), Zijie Shang (LZU), Jian-Wen Luo (UCAS) ...

Many thanks to them!

Machine learning and AI



Large Model

Deep Learning

Machine Learning

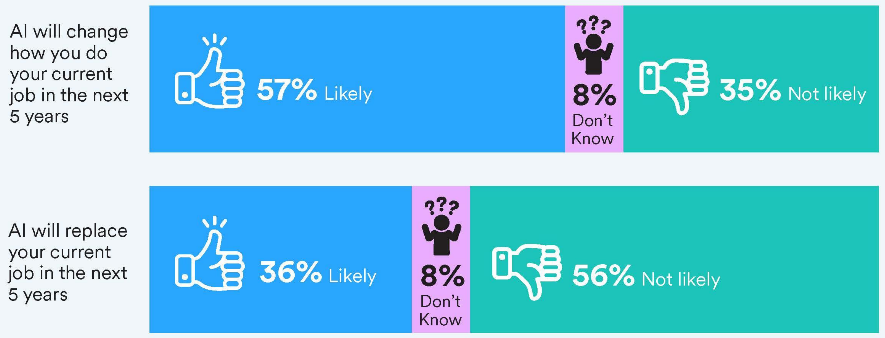
Artificial Intelligence

More parameters, larger training data sets, normally based on Transformer, e.g. ChatGPT, GPT-4, LLaMa, Gemini

More hidden layers, e.g. CNN, RNN, GNN, GAN,, Transformer

Most promising approach for AI, machine learn knowledge (fit parameters) by itself

Intelligent machines capable of human-like intelligence



Global opinions: How people think AI will affect jobs, 2023

Source: Ipsos, 2023 | Chart: 2024 AI Index report

2024 AI index report

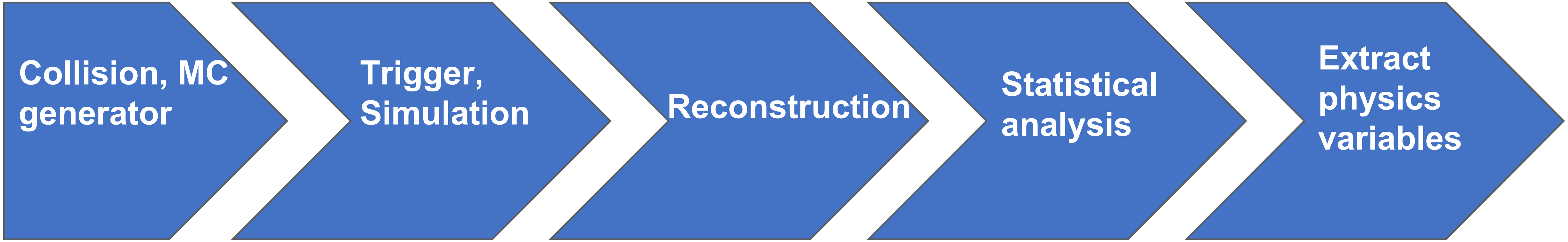
The next industrial revolution

Why we need LLM

- Major effort in BESIII analysis is spent in writing / testing / coding / text
    - **LLM is good at code/text generation !**

- Key problems for LLM at HEP
    - how to make sure the outputs are reliable?
    - how to avoid hallucinations ?
    - Current solutions:
        - **More accurate and good quality data for training**
        - **More tests and validations**
        - **More proper architecture**

Data process workflow at HEP experiment



Collision, MC generator

Accelerator control, initial-state-radiation, parton showering, hadronization, NP-correction, pileup , et. al.

Trigger, Simulation

Data acquisition, fast reconstruction, data input/output, online monitoring, detector geometry, detector noise, calibration, multi-scattering , et. al.

Reconstruction

Track and vertex finding and fitting, clusterization and reconstruction of jet, jet tagging, kinematic fit, detector calibration, et. al.

Statistical analysis

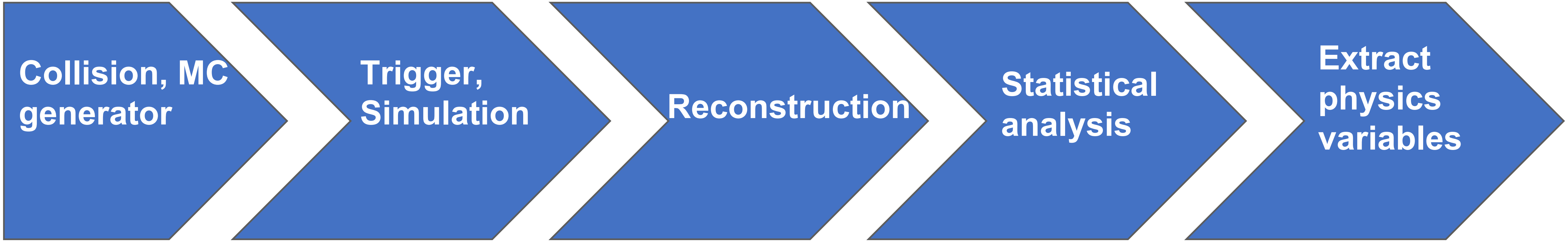
Event selection, optimizations, background analysis, injection test, reweighting, correlation corrections, et. al.

Extract physics variables

systematic uncertainty, fitting, uncertainty propagation, radiation and VP corrections, et. al.

Too complicated, similar lines of code as windows/macOS

Data process workflow at HEP experiment



Collision, MC generator

Trigger, Simulation

Reconstruction

Statistical analysis

Extract physics variables

Accelerator
control, ini
state-radi
parton sh
hadroniza
NP-correc
pileup , et

Data acquisition, fast

Track and vertex

Event selection,

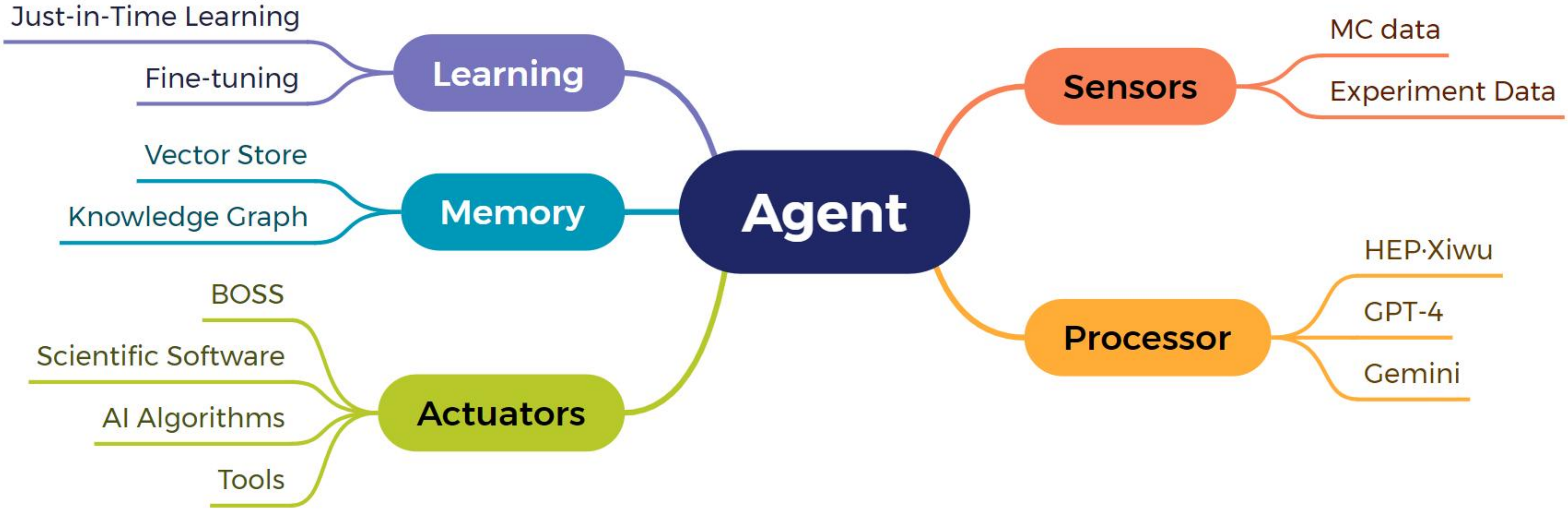
systematic
ng,
l.

One small task needs : several people + several years !
we have to make it more efficient !

Too complicated, similar lines of code as windows/macOS

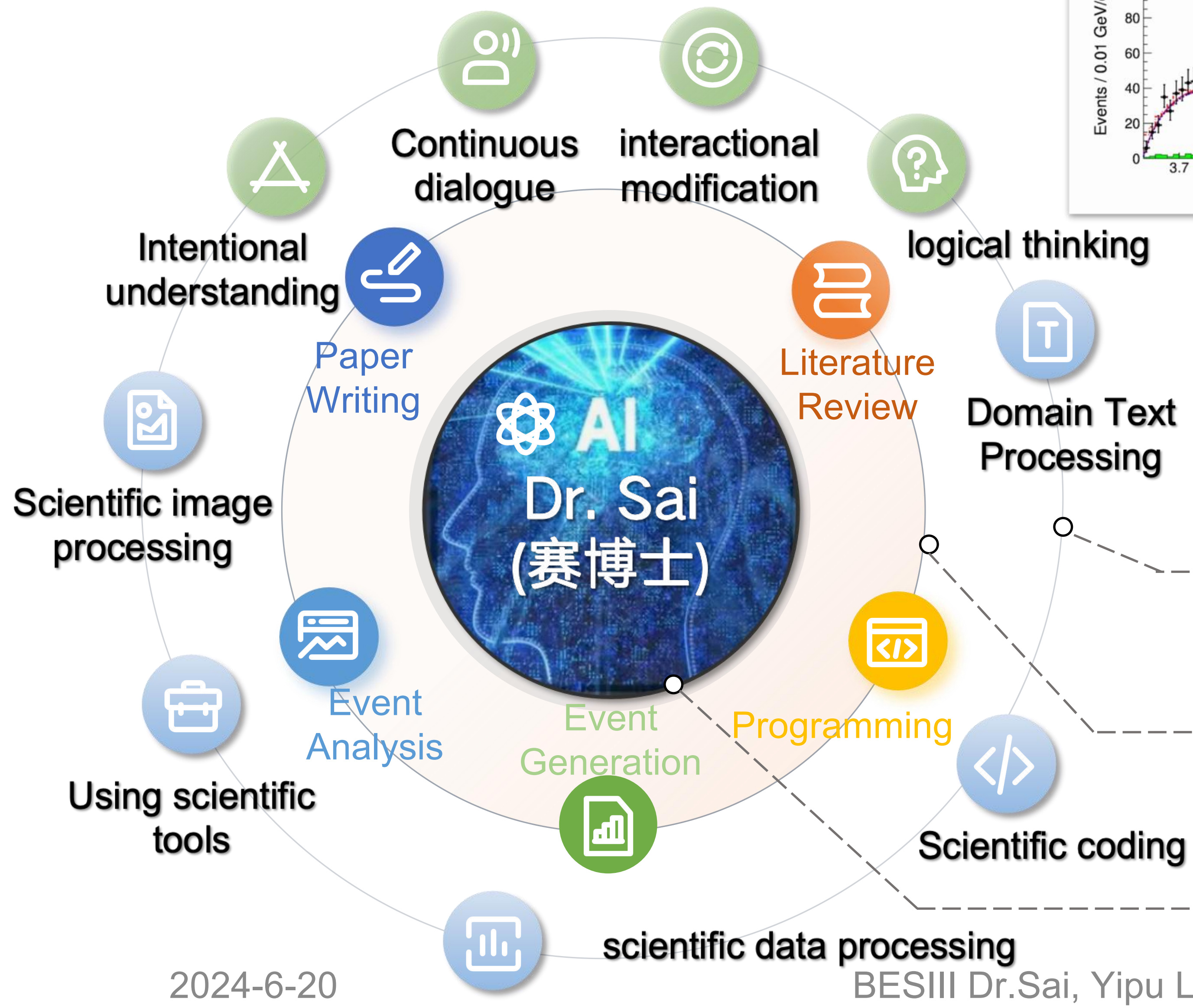
What is Agent

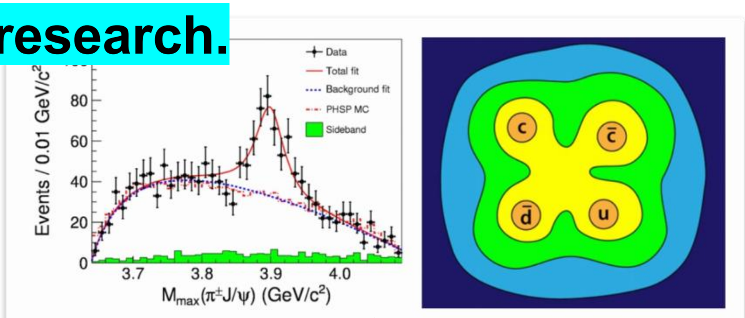
AI agent refers to a **system** or **software** that can make autonomous decisions or perform actions on behalf of its users or other systems based on its knowledge, programming, environment, and inputs.

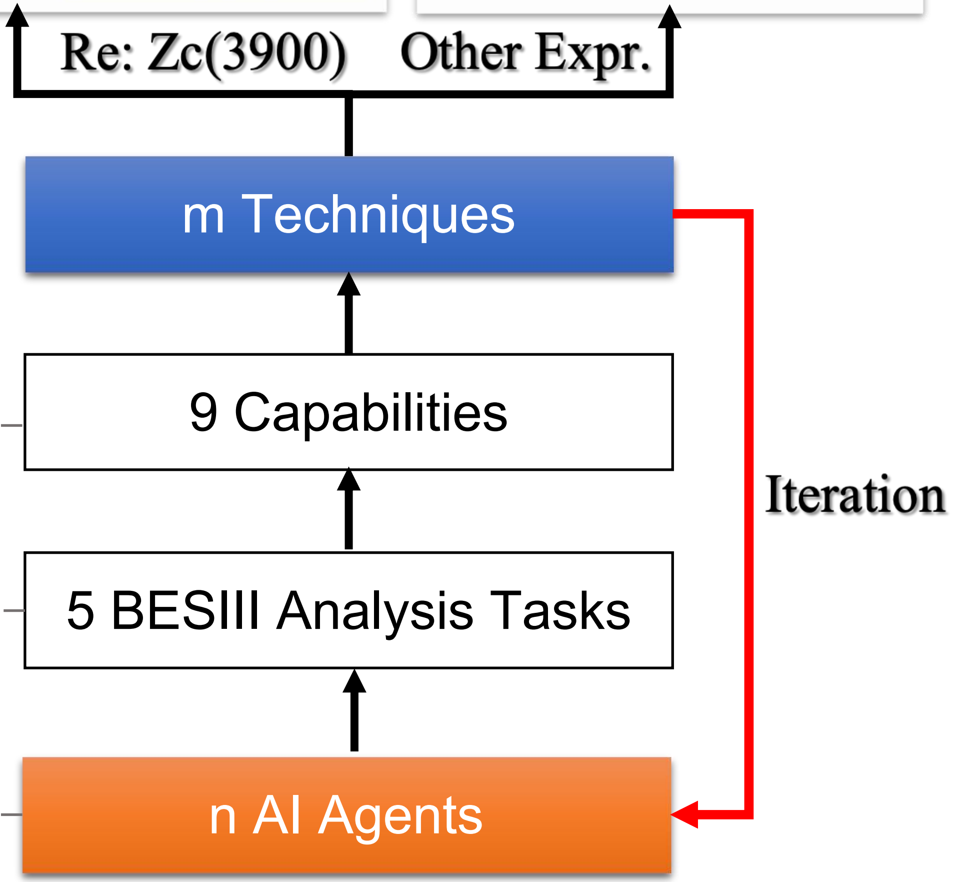


Dr. Sai (赛博士) research agent

Let the large model conduct particle physics research.







Re: Zc(3900)

Other Expr.

m Techniques

9 Capabilities

5 BESIII Analysis Tasks

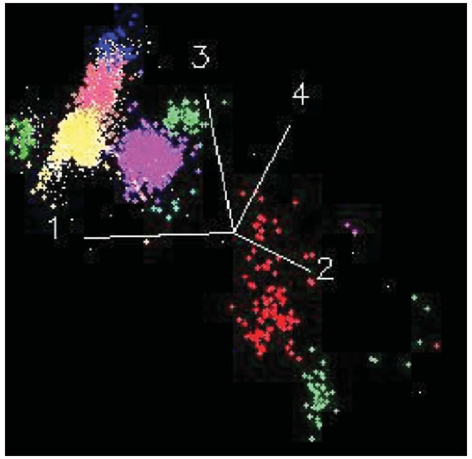
n AI Agents

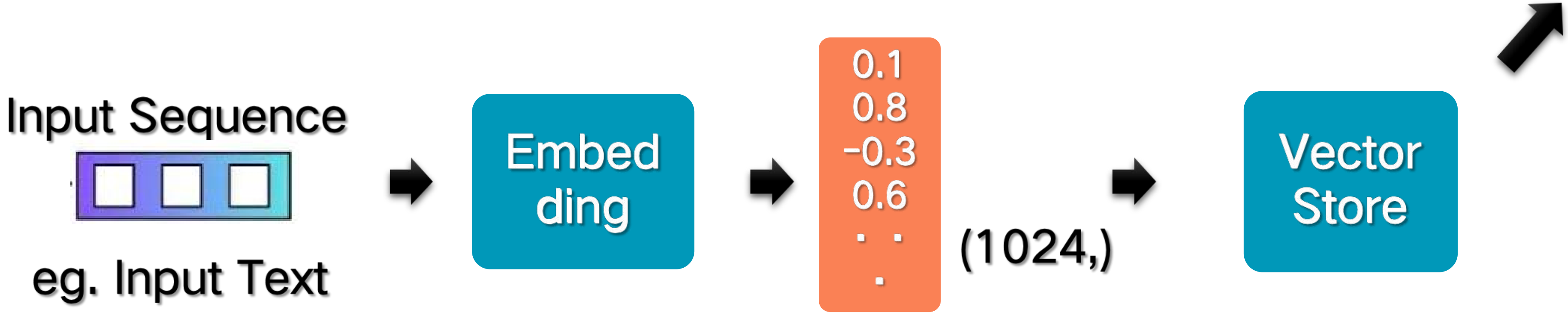
Iteration

The memory of Dr. Sai - RAG

- Retrieval-Augmented Generation (RAG)
  - Most promising solution to avoid hallucinations
  - Goal: store private data so no need for retraining
  - Current approach is based on LlamaIndex
  - Vector store (done, based on LangChain) and knowledge graph (in development) are also considered
    - Embeddings (BGE-M3 model), convert input data into vectors of a multi-dimensional space
- Usage: store BESIII internal data
  - User send BESIII related questions
  - RAG return question + BESIII internal data to LLM

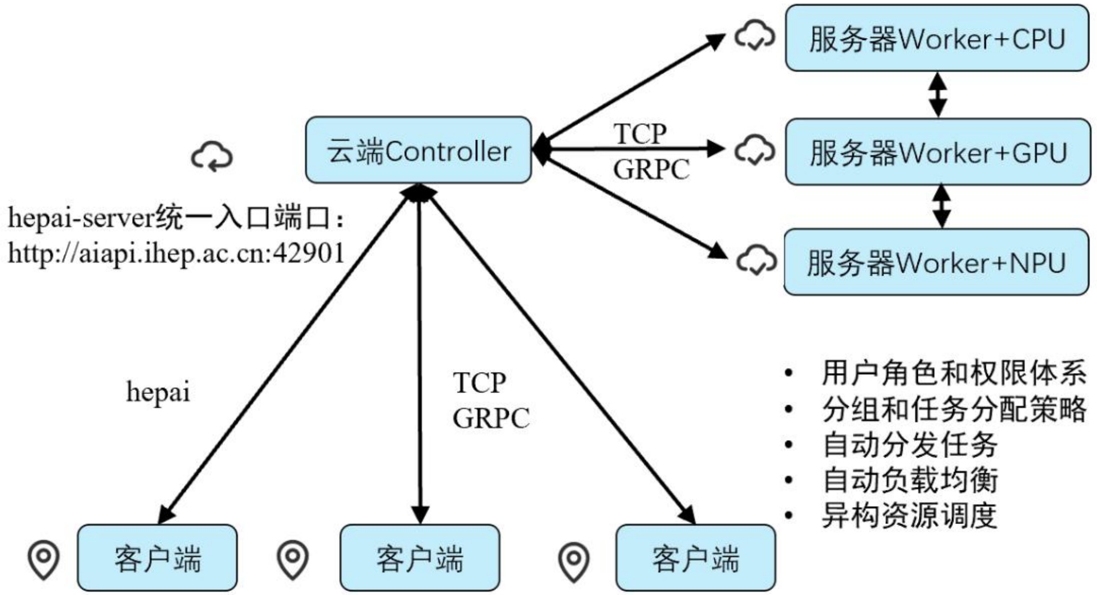
High Dimensional Space





The actuators of Dr. Sai

The Distributed Deployment Framework (HepAI-DDF)



- 用户角色和权限体系
- 分组和任务分配策略
- 自动分发任务
- 自动负载均衡
- 异构资源调度

Client

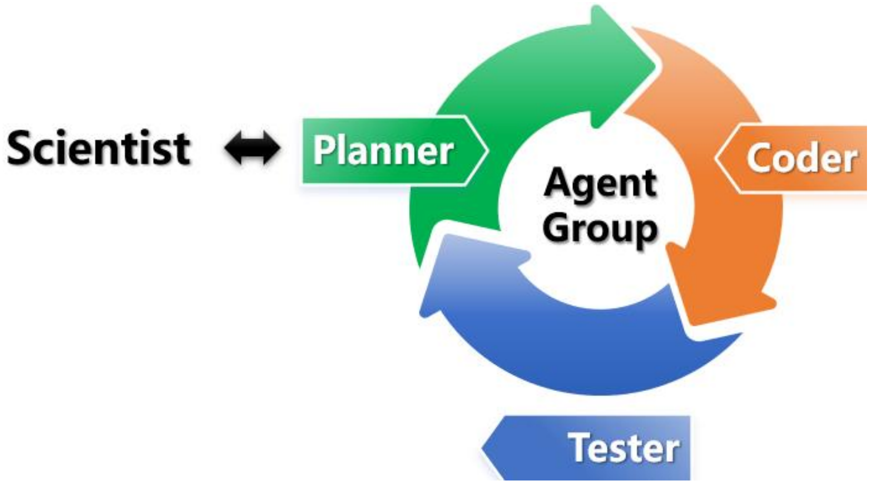
BOSS 7.1.0

Drawing tools

Scipy etc.

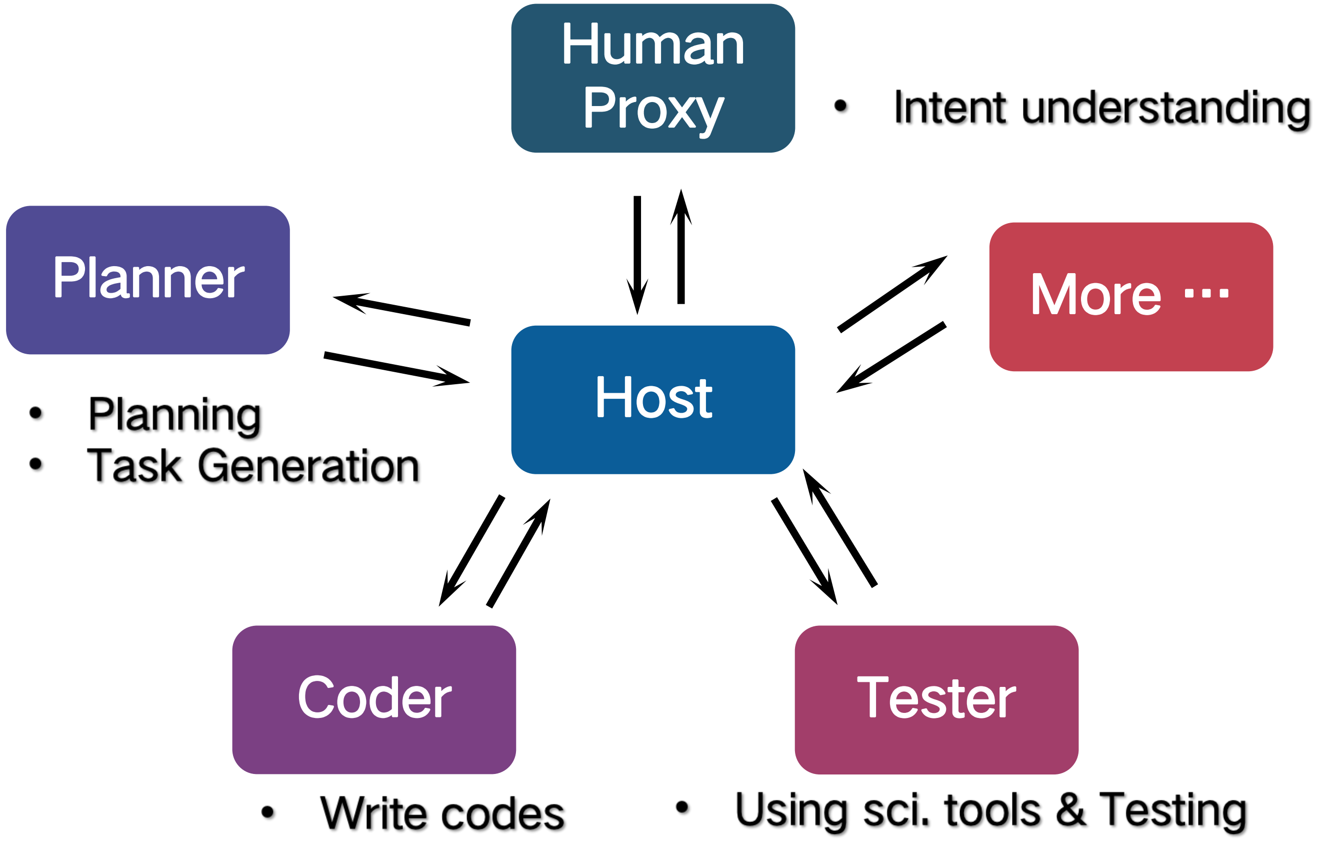
Multi-agents management system

- Developed based on AutoGen framework
- Normally one agent is dedicated for one task, HEP data processing is very complicated, impossible for one agent



Previous version
(in Nov. 2023)

upgrade


Human Proxy
- Intent understanding

Planner
- Planning
- Task Generation

Host

More …

Coder
- Write codes

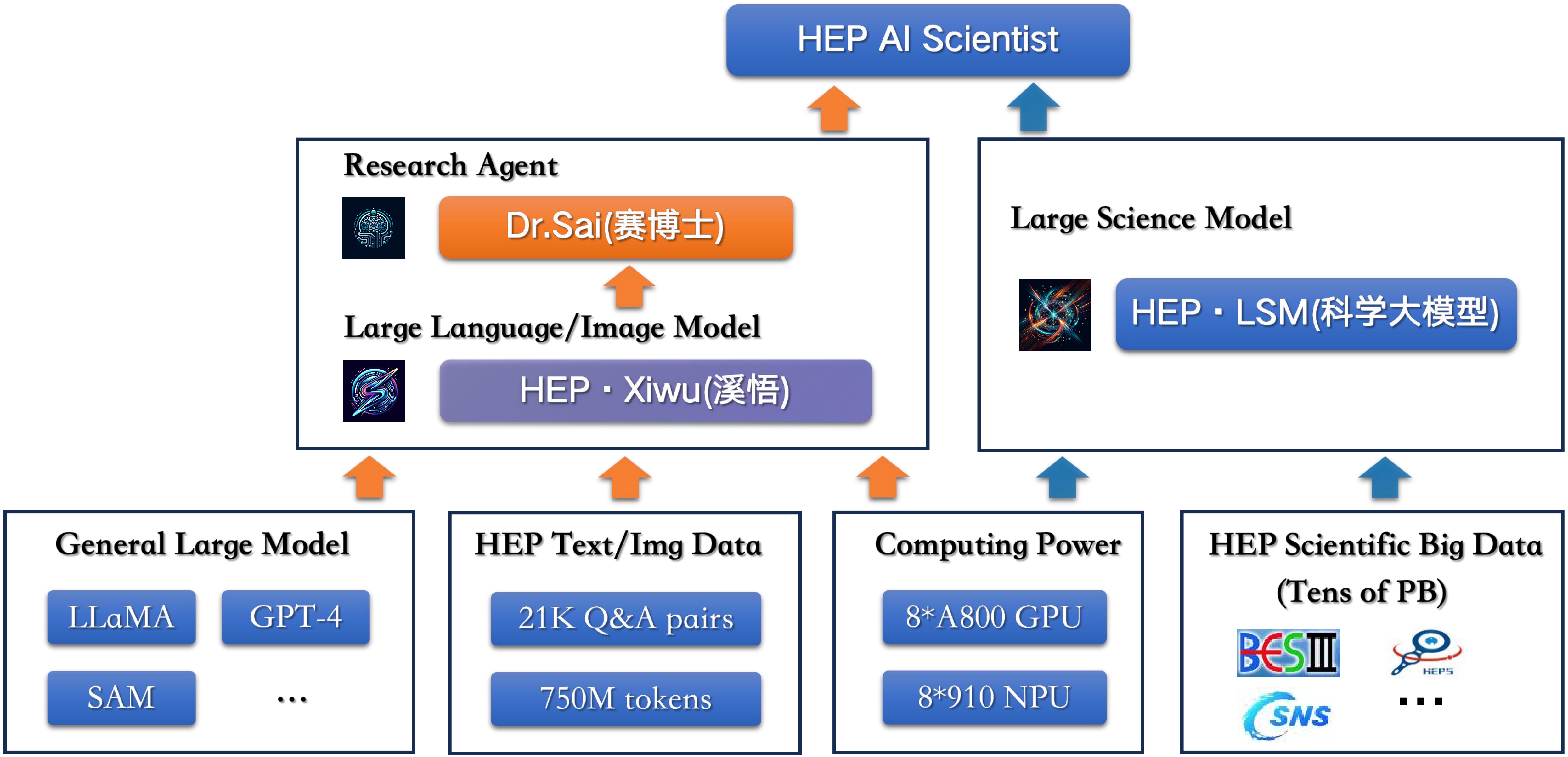
Tester
- Using sci. tools & Testing

Status and prospects for Dr. Sai

- **Under construction and testing, plan to release the first version (two AI agents) at June 2024**
  - one dedicated for BESIII and another for public, **stay tune**

- One application: software and training
  - BOSS (C++ code) upgrade
    - step 1: simple improvements using new C++ features, e.g. array to vector
    - step 2: re-structure the code for each file individually
    - step 3: AI-assisted update on algorithms
  - Outreach and training:
    - Train junior graduated students to understand BESIII and data analysis better

Roadmap of High Energy Physics AI Scientist



HEP AI Scientist

Research Agent

Dr.Sai(赛博士)

Large Language/Image Model

HEP · Xiwu(溪悟)

Large Science Model

HEP · LSM(科学大模型)

General Large Model

- LLaMA
- GPT-4
- SAM
- …

HEP Text/Img Data

- 21K Q&A pairs
- 750M tokens

Computing Power

- 8*A800 GPU
- 8*910 NPU

HEP Scientific Big Data (Tens of PB)