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

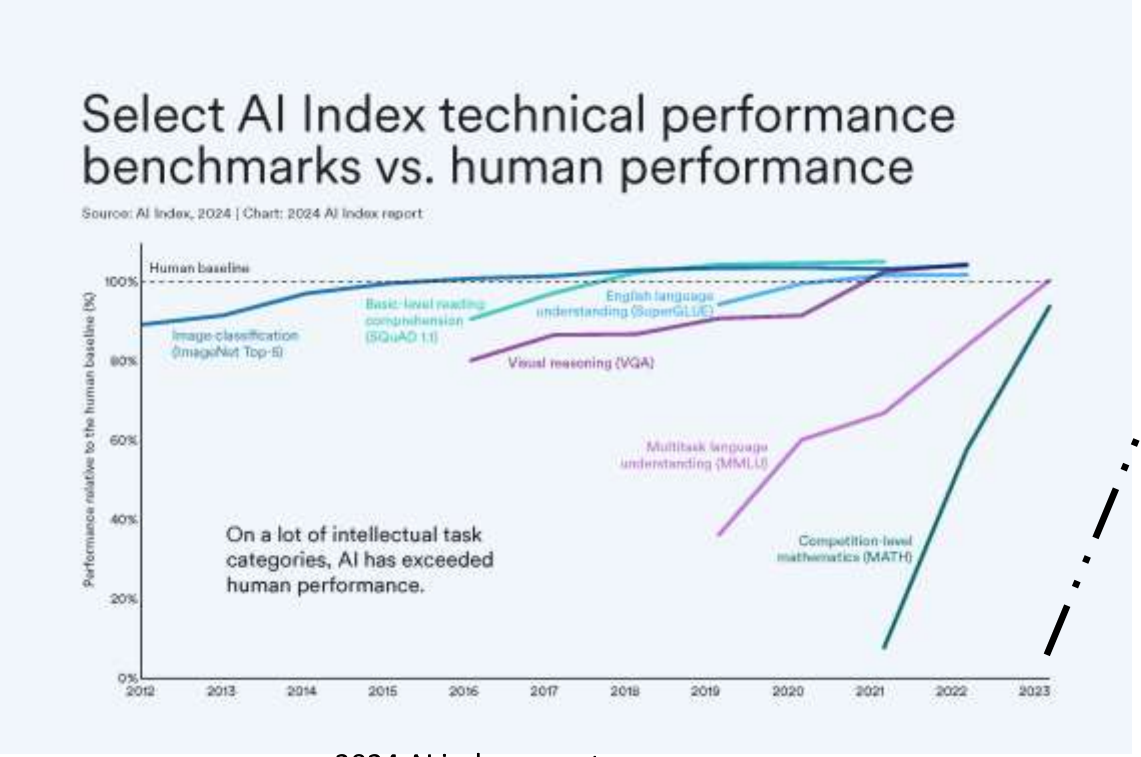
LLM-based physics analysis agent (Dr. Sai) at BESIII and exploration of future AI virtual scientist

Ke Li (like@ihep.ac.cn)

on behalf of Dr. Sai working group

Outline

- Motivation
  - what is LLM
  - why we need LLM
- Dr. Sai project
  - architecture
  - training
- Status and Prospects



Select AI Index technical performance benchmarks vs. human performance

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

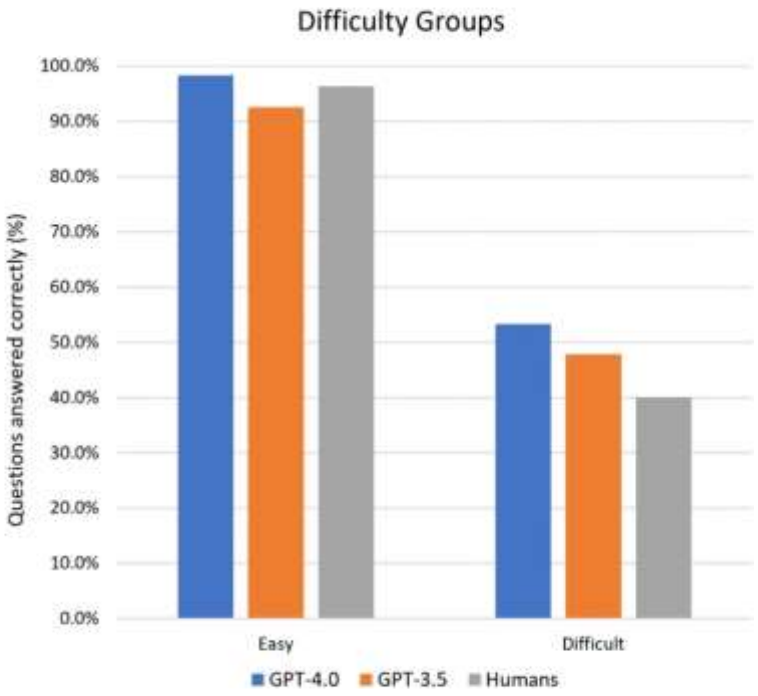
On a lot of intellectual task categories, AI has exceeded human performance.

2024 AI index report

Analysis at High energy physics ?

What is Large Language Model (LLM)

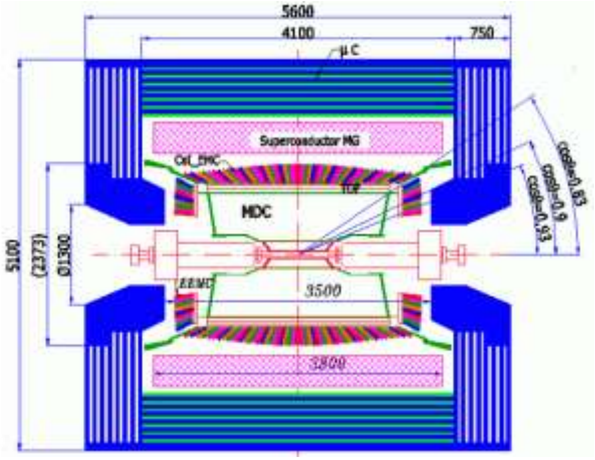
- Large language models (LLMs), normally build on transformer architecture, has demonstrated impressive performance in **text/code generation**
    - GPT4o, Gemini, LLaMa3…
    - Could be used for HEP studies
    - Game changer
- **For us, open-source foundation model + higher level model for HEP + fine tuning for BESIII**



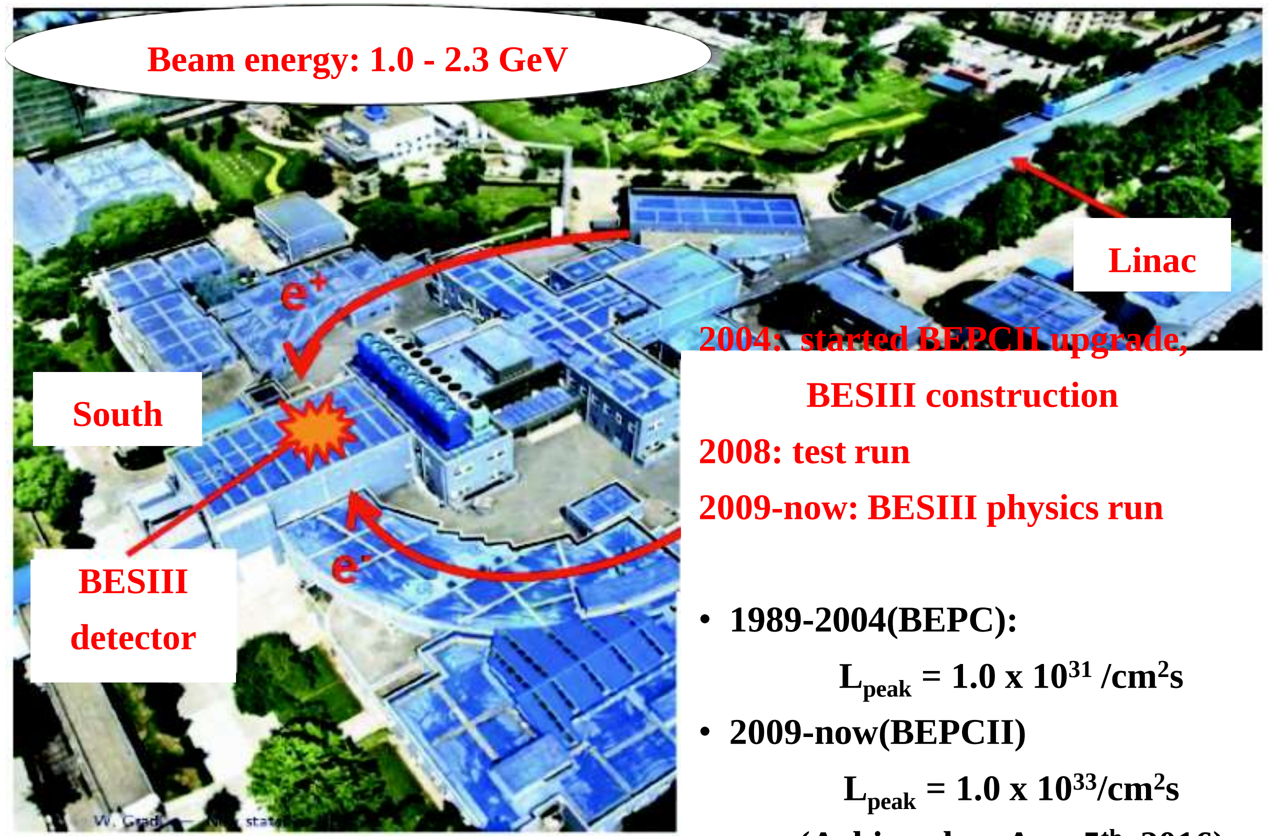
Scientific Reports volume13, Article number: 18562 (2023)

BESIII at Beijing Electron-Positron Collider II

BESIII detector



A double-ring collider with high luminosity

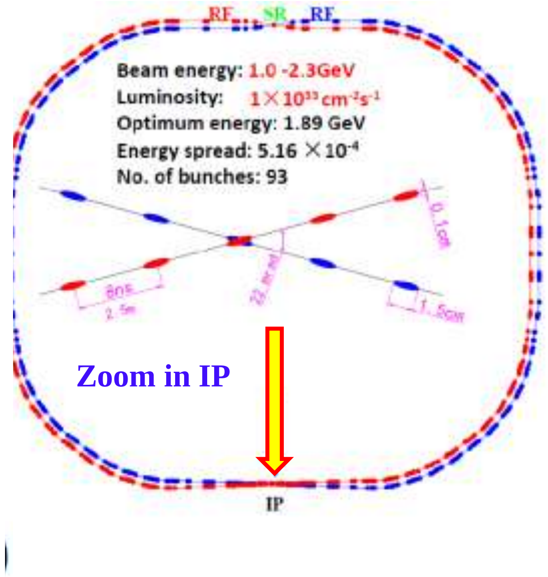


2004: started BEPCII upgrade, BESIII construction

2008: test run

2009-now: BESIII physics run

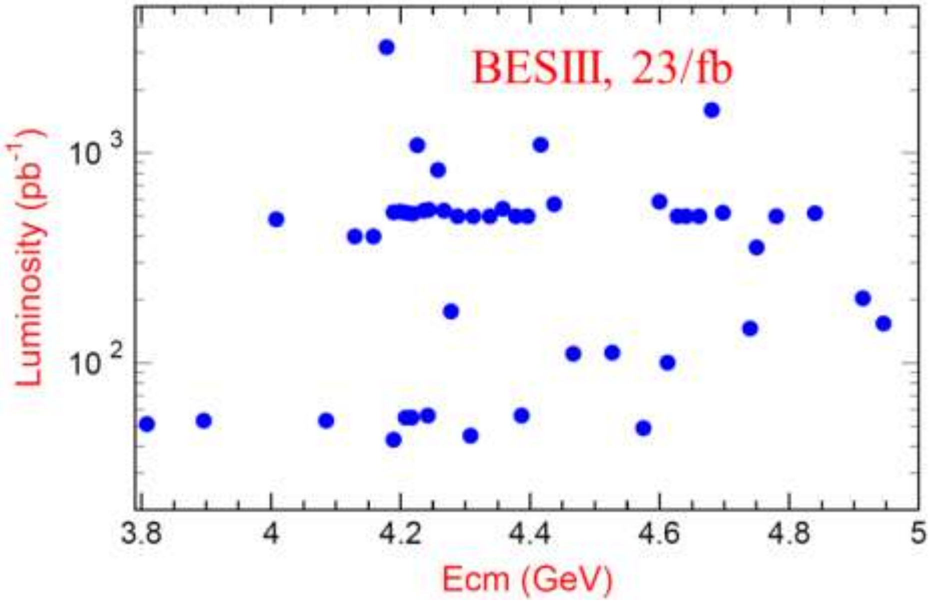
- 1989-2004(BEPC): $L_{peak} = 1.0 \times 10^{31}$ /cm²s
- 2009-now(BEPCII) $L_{peak} = 1.0 \times 10^{33}$/cm²s (Achieved on Apr. 5th, 2016)

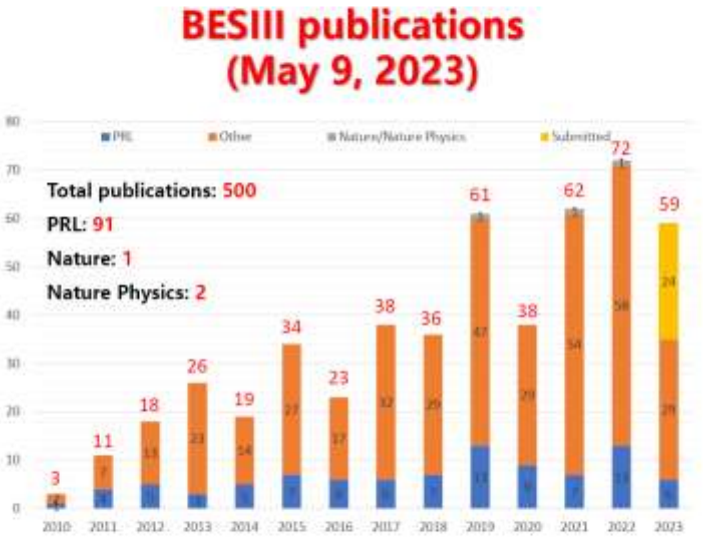


Beam energy: 1.0 -2.3GeV
Luminosity: 1×10^{33} cm^{-2}s^{-1}
Optimum energy: 1.89 GeV
Energy spread: 5.16×10^{-4}
No. of bunches: 93

- Preparing for upgrade (2024-2025)
- More data will be collected

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 !
  - E.g. Test Lattice-QCD, light/charm hadron spectroscopy, charm quark weak decay

Why we need LLM

- Major effort in BESIII analysis is spent in writting/testing/updating code/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**

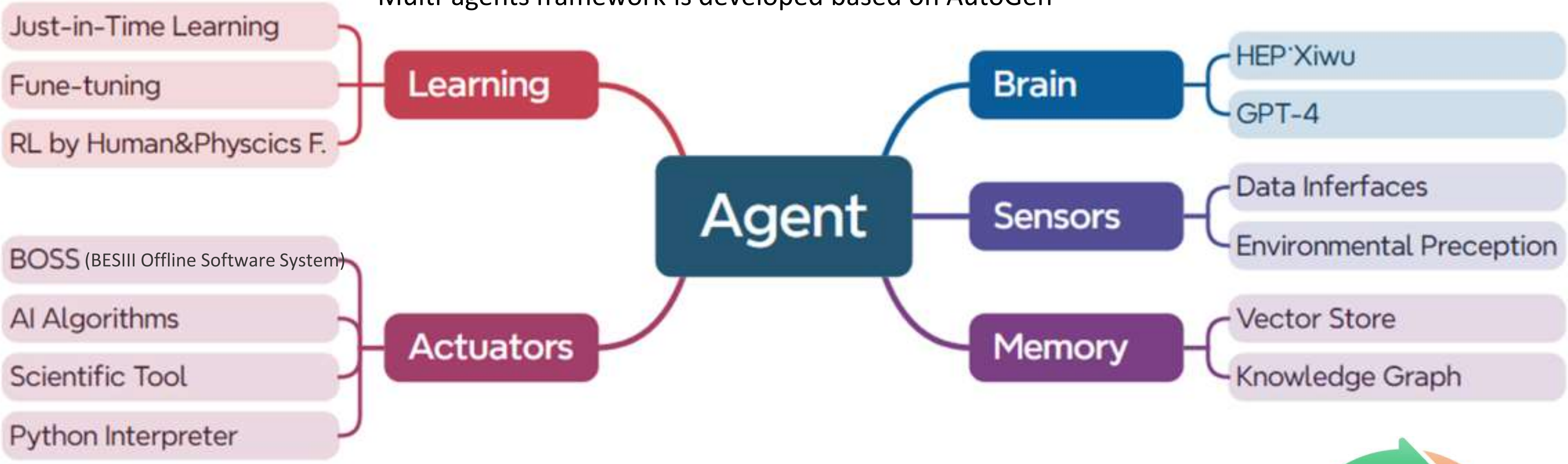
Dr. Sai project for BESIII/HEP

- 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, source code, Q-As from internal review
- One milestone: **AI assistant**, It can help scientist on data analysis, e.g. MC generation, signal extraction, and a navigator inside BESIII
  - target at **June 2024** !
- Goal: **AI scientist**, it can analyze the data automatically like an expert

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

Dr. Sai

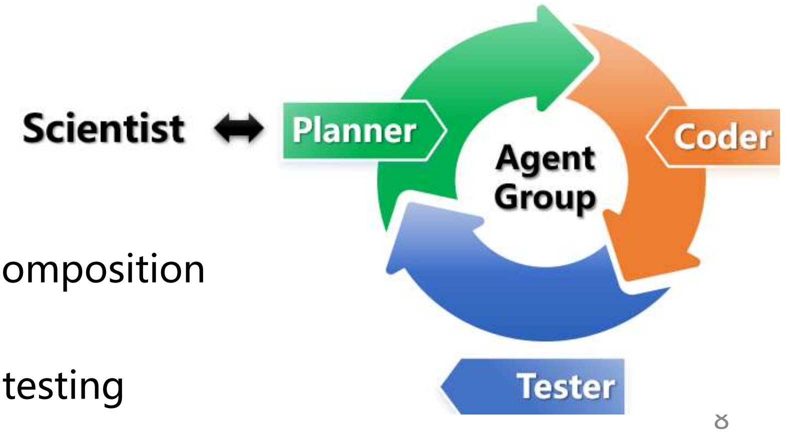
Multi-agents framework is developed based on AutoGen



Key of this project:

make the results from AI more reliable

- New architecture
- Good quality data
- In-the-fly validation and test



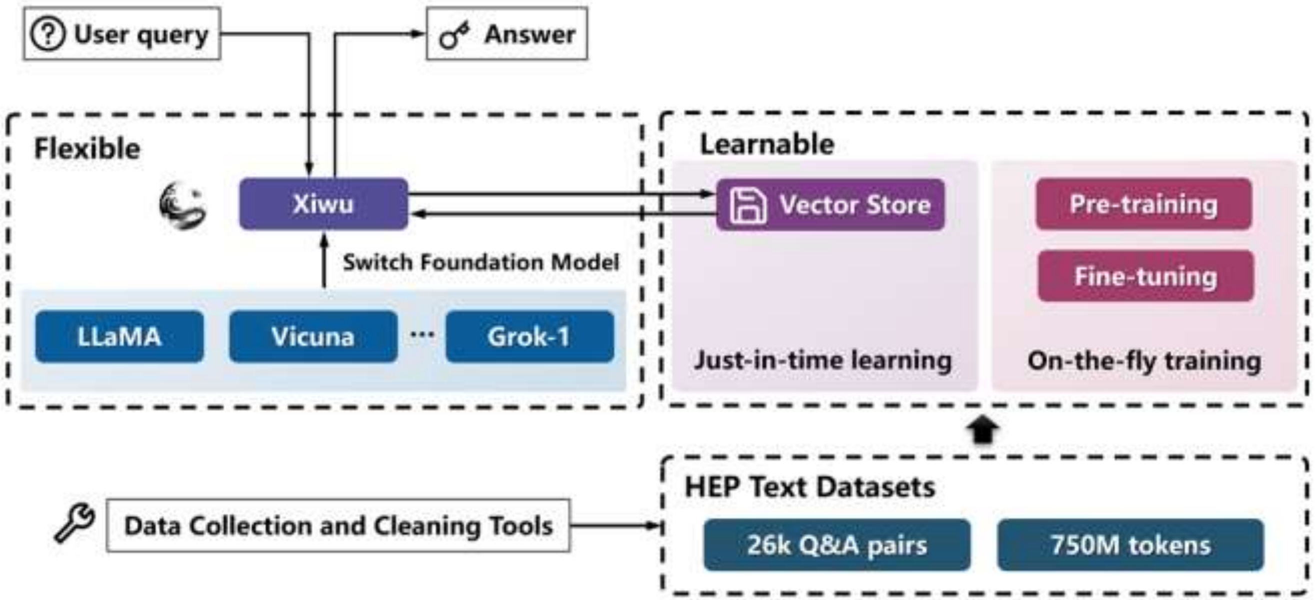
Agents:

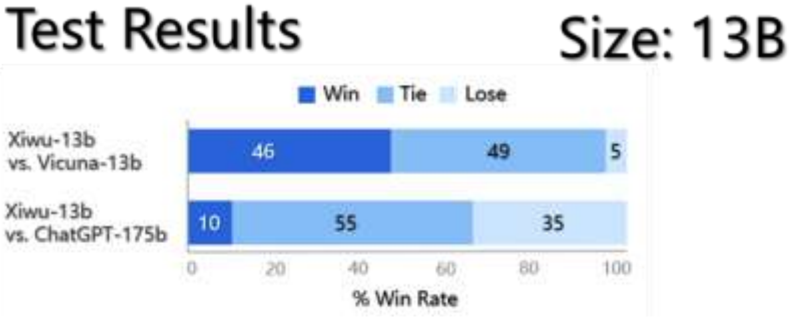
- Planner: Planning and tasks decomposition
- Coder: Write code/text
- Tester: Using scientific tools for testing

Human can interact via HumanProxy

The brain of Dr. Sai - Xiwu model

- Xiwu: a basis flexible and learnable LLM for HEP
- 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 is on-going**

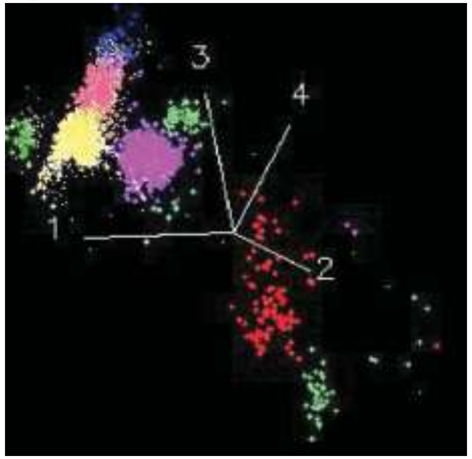


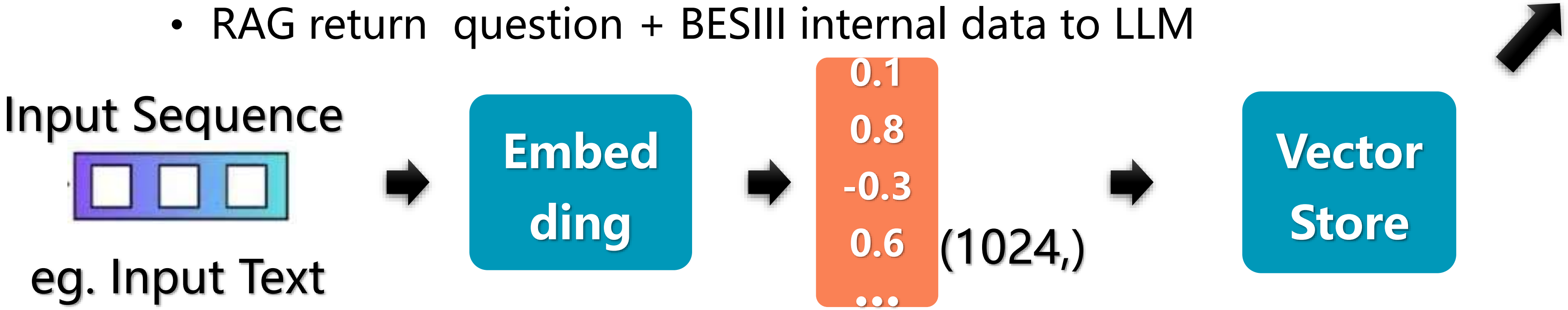


The Memory of Dr. Sai - RAG

- Retrieval-Augmented Generation (RAG)
  - Most promising solution to avoid hallucinations
  - Goal: store private data so no need for retraining
  - Current approach: vector store
    - 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




Input Sequence

eg. Input Text

Embedding

0.1
0.8
-0.3
0.6
...

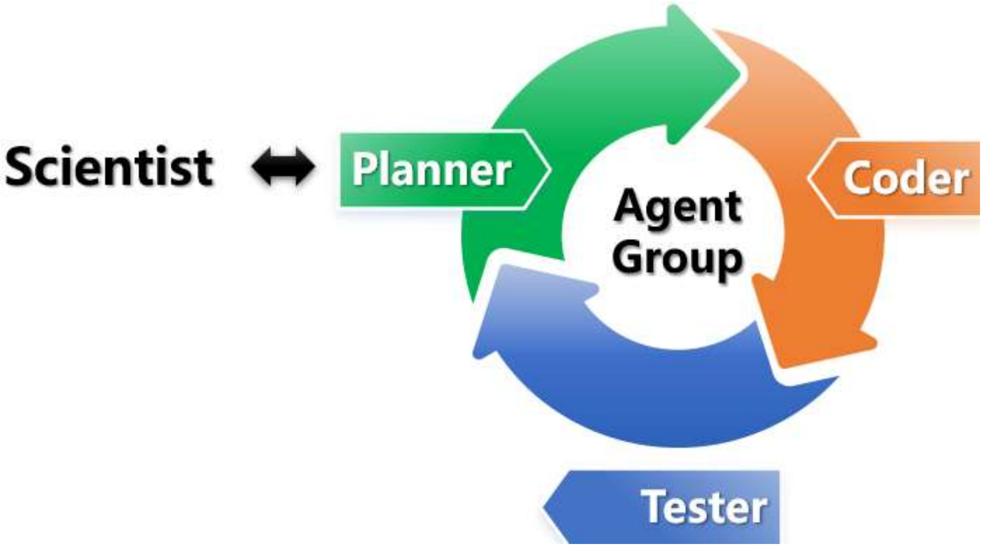
(1024,)

Vector Store

Multi-agents managment system

- Developed based on AutoGen framework
- Normally one agent is dedicated for one task, HEP data processing is very complicated, impossible to build one common agent
- Multi-Agents:
  - **Planner**: Planning and tasks decomposition
  - **Coder**: Write code/text
  - **Tester**: Using scientific tools for testing
  - Common tools: plotting et.al.
  - …

•Human can interact via HumanProxy



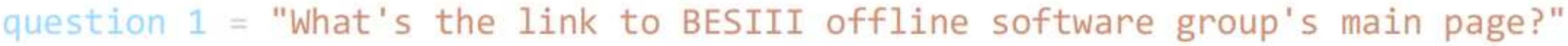
- If test failed, the feedback will be used to improve the prompts at next iteration.
- For each task, we have multi-unittests

Training data

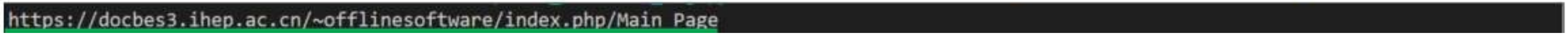
- Recent papers on arXiv
    - PDF files parser: HaiNougat, advanced iteration of the Nougat model
- Good quality chat history from IHEP-AI platform
    - The data is cleaned by human or AI (GPT4)
    - 180k Question-Answer pairs in 3 months
- **BESIII internal data**
    - internal memo, parsered by HaiNougat (>5000 PDF files)
    - Q-A pairs from internal paper review (>50k Q-A pairs)
    - BESIII Offline Software System (BOSS) source code
    - BESIII public webpages and internal webpages
    - The data (agenda, slides) on indico will be used later
- All the BESIII internal data sets are stored in RAG or used in training and fine-tuning

Simple test: internal navigator at BESIII

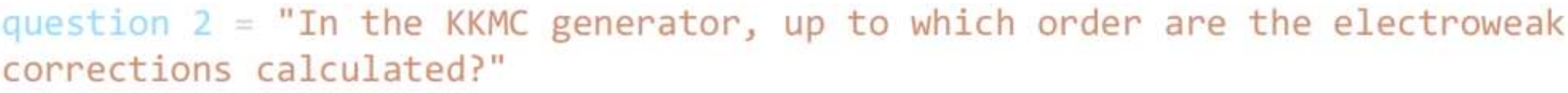
- Same with the chATLAS project at ATLAS
  - Navigator and assistant to replace the simple ‘search’
  - BESIII internal data at websites (bes3.ihep.ac.cn)
    - Not public yet
  - In general, better performance than I expected
    - E.g. Question ‘where is the XXXX MC sample’,
    - Answer ‘The path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



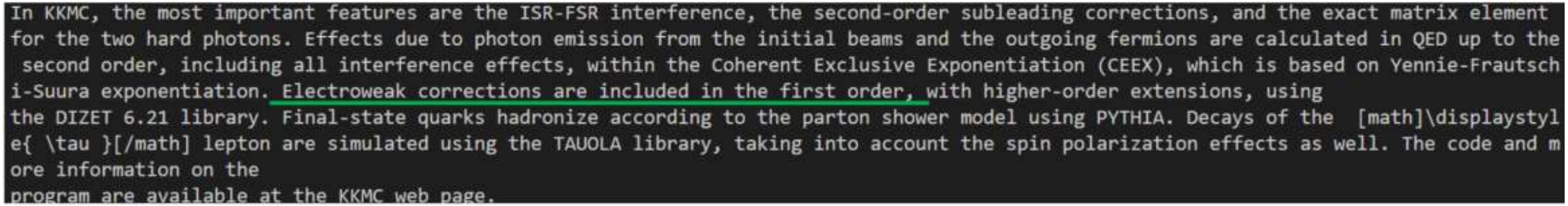
```
question 1 = "What's the link to BESIII offline software group's main page?"
```



```
https://docbes3.ihep.ac.cn/~offlinesoftware/index.php/Main_Page
```



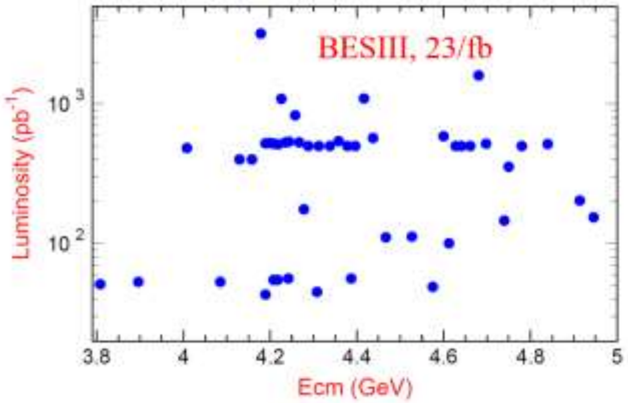
```
question 2 = "In the KKMC generator, up to which order are the electroweak
corrections calculated?"
```

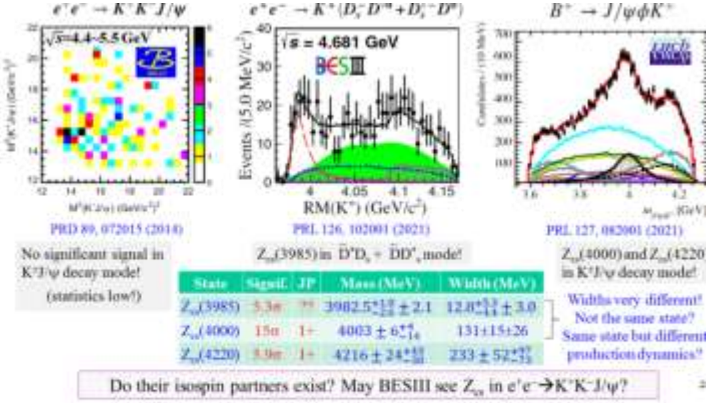


```
In KKMC, the most important features are the ISR-FSR interference, the second-order subleading corrections, and the exact matrix element
for the two hard photons. Effects due to photon emission from the initial beams and the outgoing fermions are calculated in QED up to the
 second order, including all interference effects, within the Coherent Exclusive Exponentiation (CEEX), which is based on Yennie-Frautsch
i-Suura exponentiation. Electroweak corrections are included in the first order, with higher-order extensions, using
the DIZET 6.21 library. Final-state quarks hadronize according to the parton shower model using PYTHIA. Decays of the  [math]\displaystyl
e{ \tau }[/math] lepton are simulated using the TAUOLA library, taking into account the spin polarization effects as well. The code and m
ore information on the
program are available at the KKMC web page.
```

Status and prospects for Dr. Sai

- XYZ studies at BESIII:
  - Many (intermediate) resonants, many decay channels -> interferences
  - Difficult to retrieve useful information correctly





PRD 89, 072015 (2014)

PRL 126, 102001 (2021)

PRL 127, 082001 (2021)

No significant signal in K^+J/ψ decay mode! (statistics low!)

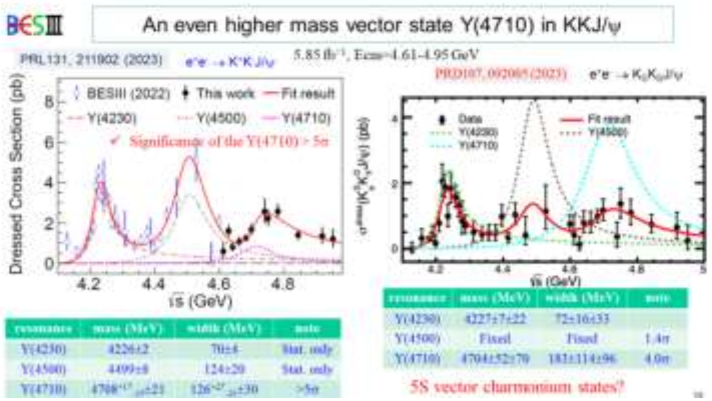
Z_{cs}(3985) in $\bar{D}^*D_s + \bar{D}D^*_s$ mode!

Z_{cs}(4000) and Z_{cs}(4220) in K^+J/ψ decay mode!

State	Signif.	JP	Mass (MeV)	Width (MeV)
Z_{cs}(3985)	5.3σ	??	$3982.5^{+1.8}_{-2.6} \pm 2.1$	$12.8^{+5.3}_{-4.4} \pm 3.0$
Z_{cs}(4000)	15σ	1+	$4003 \pm 6^{+4}_{-14}$	131±15±26
Z_{cs}(4220)	5.9σ	1+	$4216 \pm 24^{+43}_{-30}$	$233 \pm 52^{+97}_{-73}$

Widths very different! Not the same state? Same state but different production dynamics?

Do their isospin partners exist? May BESIII see Z_{cs} in $e^+e^- \to K^+K^-J/\psi$?



An even higher mass vector state Y(4710) in KKJ/ψ

PRL131, 211902 (2023) $e^+e^- \to K^+K^-J/\psi$ 5.85 fb^{-1}, Ecm=4.61-4.95 GeV

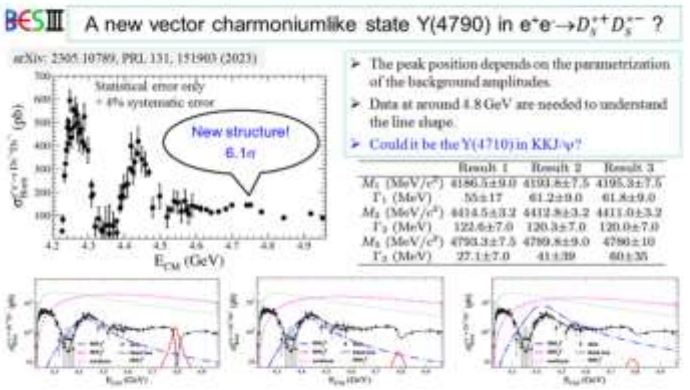
✓ Significance of the Y(4710) > 5σ

resonance	mass (MeV)	width (MeV)	note
Y(4230)	4226±2	70±4	Stat. only
Y(4500)	4499±8	124±20	Stat. only
Y(4710)	4708^{+17}_{-15}±21	126^{+27}_{-23}±30	>5σ

PRD107, 092005 (2023) $e^+e^- \to K_SK_SJ/\psi$

resonance	mass (MeV)	width (MeV)	note
Y(4230)	4227±7±22	72±16±33	
Y(4500)	Fixed	Fixed	1.4σ
Y(4710)	4704±52±70	183±114±96	4.0σ

5S vector charmonium states?



A new vector charmoniumlike state Y(4790) in $e^+e^- \to D_s^{*+}D_s^{*-}$?

arXiv: 2305.10789, PRL 131, 151903 (2023)

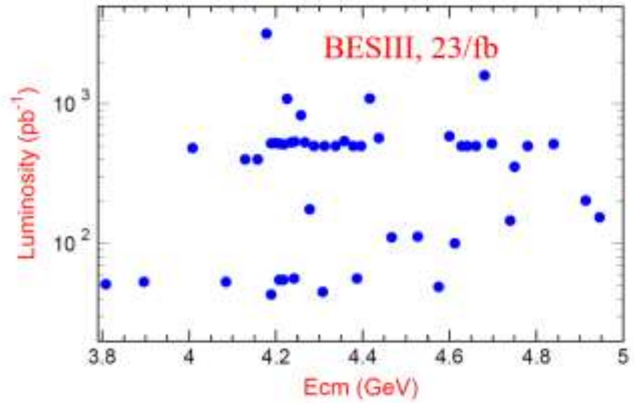
- The peak position depends on the parametrization of the background amplitudes.
- Data at around 4.8 GeV are needed to understand the line shape.
- Could it be the Y(4710) in KKJ/ψ?

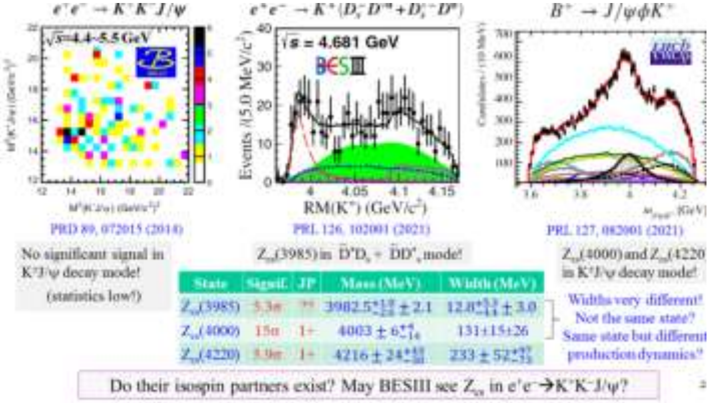
	Result 1	Result 2	Result 3
M_1 (MeV/c^2)	4186.5±9.0	4193.8±7.5	4195.3±7.5
Γ_1 (MeV)	55±17	61.2±9.0	61.8±9.0
M_2 (MeV/c^2)	4414.5±3.2	4412.8±3.2	4411.0±3.2
Γ_2 (MeV)	122.6±7.0	120.3±7.0	120.0±7.0
M_3 (MeV/c^2)	4793.3±7.5	4789.8±9.0	4790±10
Γ_3 (MeV)	27.1±7.0	41±39	60±35

How to understand them? what is the physics behind them ?

Status and prospects for Dr. Sai

- Exotic hadron (XYZ) studies at BESIII:
  - Many (intermediate) resonants, many decay channels -> interferences
  - Difficult to retrieve useful information correctly





PRD 89, 072015 (2014)

No significant signal in K^+J/ψ decay mode! (statistics low?)

PRL 126, 102001 (2021)

$Z_{cs}(3985)$ in $\bar{D}^*D_s + \bar{D}D^*_s$ mode!

State	Signif.	JP	Mass (MeV)	Width (MeV)
$Z_{cs}(3985)$	5.3σ	??	$3982.5^{+1.8}_{-2.6} \pm 2.1$	$12.8^{+5.3}_{-4.4} \pm 3.0$
$Z_{cs}(4000)$	15σ	1+	$4003 \pm 6^{+4}_{-14}$	131±15±26
$Z_{cs}(4220)$	5.9σ	1+	$4216 \pm 24^{+43}_{-30}$	$233 \pm 52^{+97}_{-73}$

PRL 127, 082001 (2021)

$Z_{cs}(4000)$ and $Z_{cs}(4220)$ in K^+J/ψ decay mode!

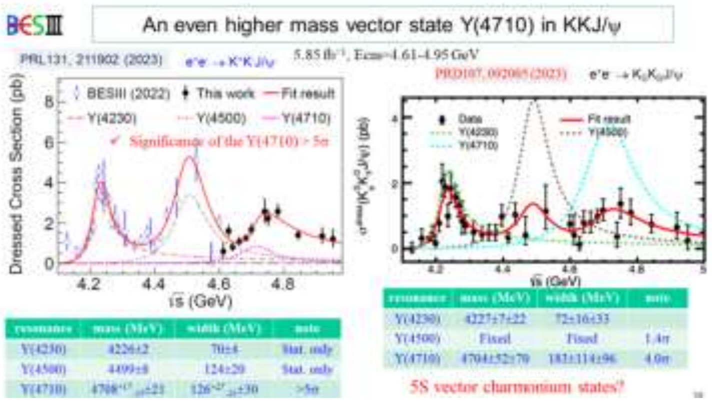
Widths very different! Not the same state? Same state but different production dynamics?

Do their isospin partners exist? May BESIII see Z_{cs} in $e^+e^- \to K^+K^-J/\psi$?

How to understand them? what is the physics behind them ?

As a experimental people, I don't know.

But the cross section measurements of ALL channels should be one necessary condition.



An even higher mass vector state Y(4710) in KKJ/ψ

PRL131, 211902 (2023) $e^+e^- \to K^+K^-J/\psi$

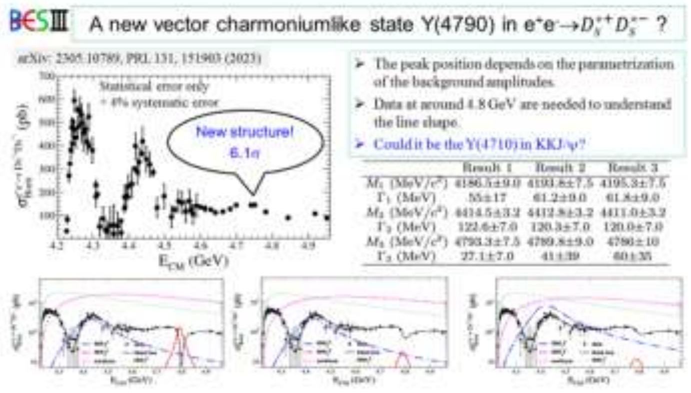
5.85 fb^{-1}, Ecm=4.61-4.95 GeV

resonance	mass (MeV)	width (MeV)	note
Y(4230)	4226±2	70±4	Stat. only
Y(4500)	4499±8	124±20	Stat. only
Y(4710)	4708^{+17}_{-15}±21	126^{+27}_{-23}±30	>5σ

PRD107, 092005 (2023) $e^+e^- \to K_SK_SJ/\psi$

resonance	mass (MeV)	width (MeV)	note
Y(4230)	4227±7±22	72±16±33	
Y(4500)	Fixed	Fixed	1.4σ
Y(4710)	4704±52±70	183±114±96	4.0σ

5S vector charmonium states?



A new vector charmoniumlike state Y(4790) in $e^+e^- \to D_s^{*+}D_s^{*-}$?

arXiv: 2305.10789, PRL 131, 151903 (2023)

- The peak position depends on the parametrization of the background amplitudes.
- Data at around 4.8 GeV are needed to understand the line shape.
- Could it be the Y(4710) in KKJ/ψ?

	Result 1	Result 2	Result 3
M_1 (MeV/c^2)	4186.5±9.0	4193.8±7.5	4195.3±7.5
Γ_1 (MeV)	55±17	61.2±9.0	61.8±9.0
M_2 (MeV/c^2)	4414.5±3.2	4412.8±3.2	4411.0±3.2
Γ_2 (MeV)	122.6±7.0	120.3±7.0	120.0±7.0
M_3 (MeV/c^2)	4793.3±7.5	4789.8±9.0	4786±10
Γ_3 (MeV)	27.1±7.0	41±39	60±35

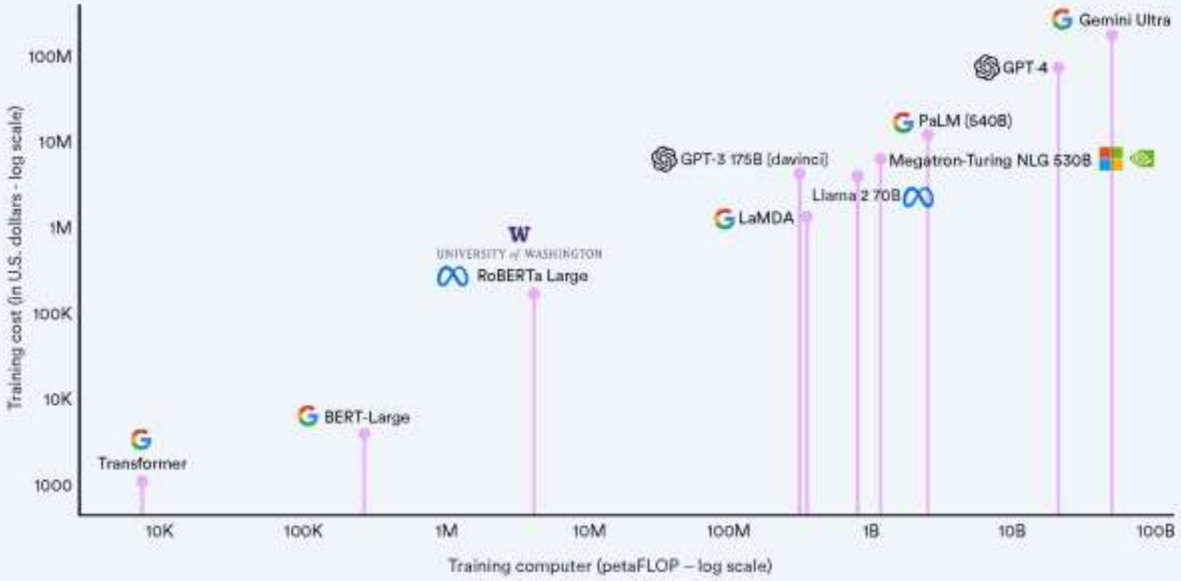
Summary

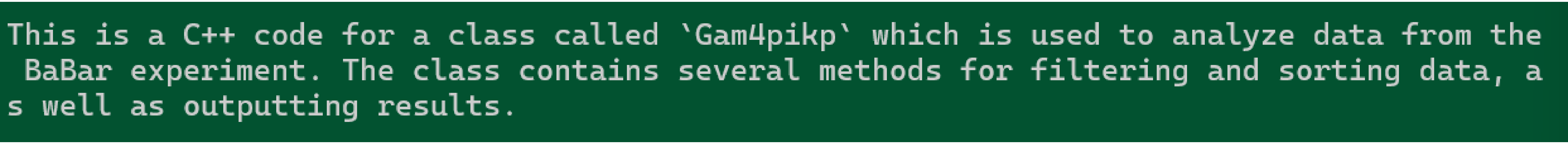
- AI agents for HEP - **Dr. Sai** is build for BESIII experiment
  - Assist people on multiple tasks, e.g. data analysis
  - Internal testing and debugging
  - Beta version at June 2024, stable version at the end of 2024
  - Personpowers are more than welcome
- Next:
  - Xiwu2: train with more data, multi-modal, e.g. slides on indico, experts' chat history at IHEP AI platform https://ai.ihep.ac.cn/
  - AI virtual scientist: automate full chain of physics analysis

back-up

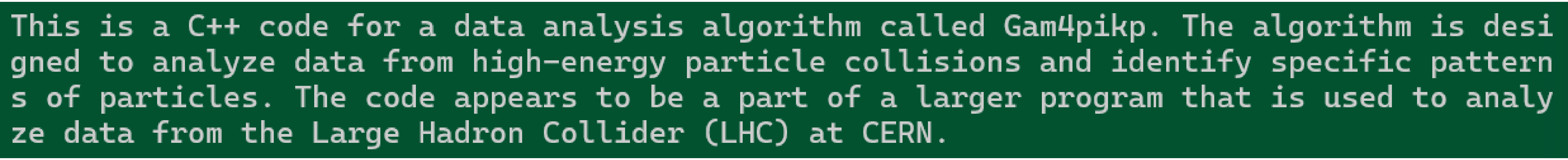
Estimated training cost and compute of select AI models

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

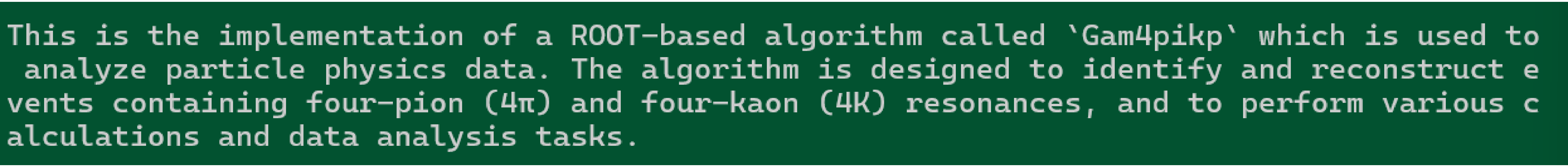




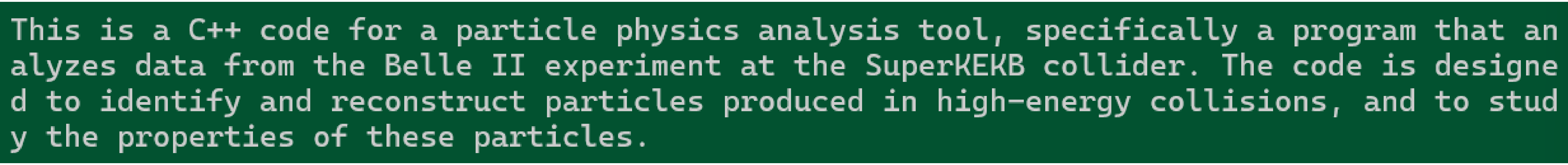
```
This is a C++ code for a class called `Gam4pikp` which is used to analyze data from the
 BaBar experiment. The class contains several methods for filtering and sorting data, a
s well as outputting results.
```



```
This is a C++ code for a data analysis algorithm called Gam4pikp. The algorithm is desi
gned to analyze data from high-energy particle collisions and identify specific pattern
s of particles. The code appears to be a part of a larger program that is used to analy
ze data from the Large Hadron Collider (LHC) at CERN.
```



```
This is the implementation of a ROOT-based algorithm called `Gam4pikp` which is used to
 analyze particle physics data. The algorithm is designed to identify and reconstruct e
vents containing four-pion (4π) and four-kaon (4K) resonances, and to perform various c
alculations and data analysis tasks.
```



```
This is a C++ code for a particle physics analysis tool, specifically a program that an
alyzes data from the Belle II experiment at the SuperKEKB collider. The code is designe
d to identify and reconstruct particles produced in high-energy collisions, and to stud
y the properties of these particles.
```

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