



Dr. Sai

BESIII Internal Data cleaning and RAG for LLM

Junkun Jiao¹, Zijie Shang²

¹Jilin University, ²Lanzhou University
on behalf of **Dr. Sai** working group

Joint-efforts from IHEP-UCAS-LZU-JLU

Quantum Computing and Machine Learning Workshop
2024.08.04, Changchun

Outline

➤ Introduction

- Xiwu model for high energy physics (HEP)
- Retrieval-Augmented Generation (RAG)

➤ BESIII internal data cleaning

- Scraping and document collection

➤ Evaluation

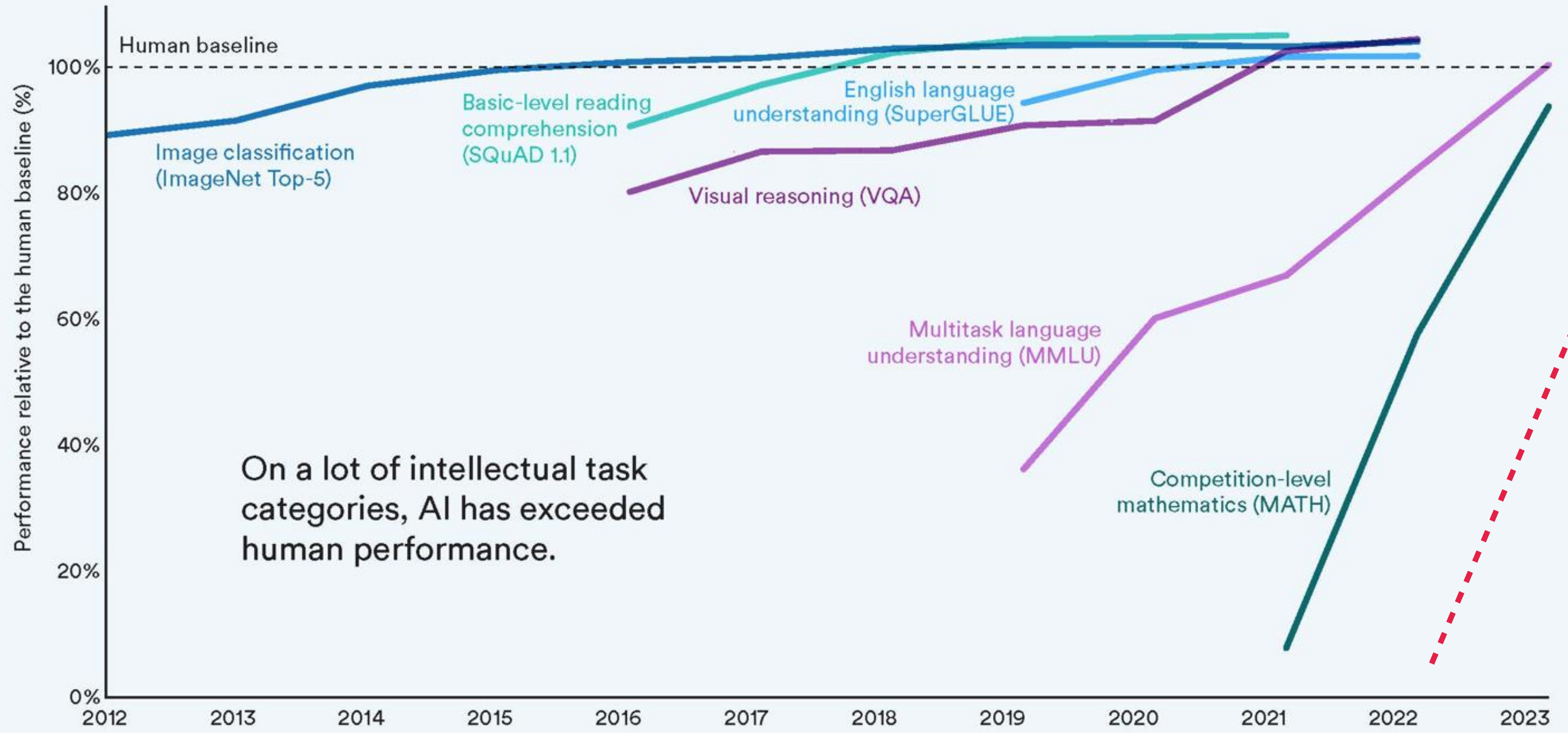
➤ Summary

Xiwu: A LLM for HEP research

The AI era is approaching. AI for science has the potential to dramatically accelerate research progresses and transform the research paradigm in many fields.

Select AI Index technical performance benchmarks vs. human performance

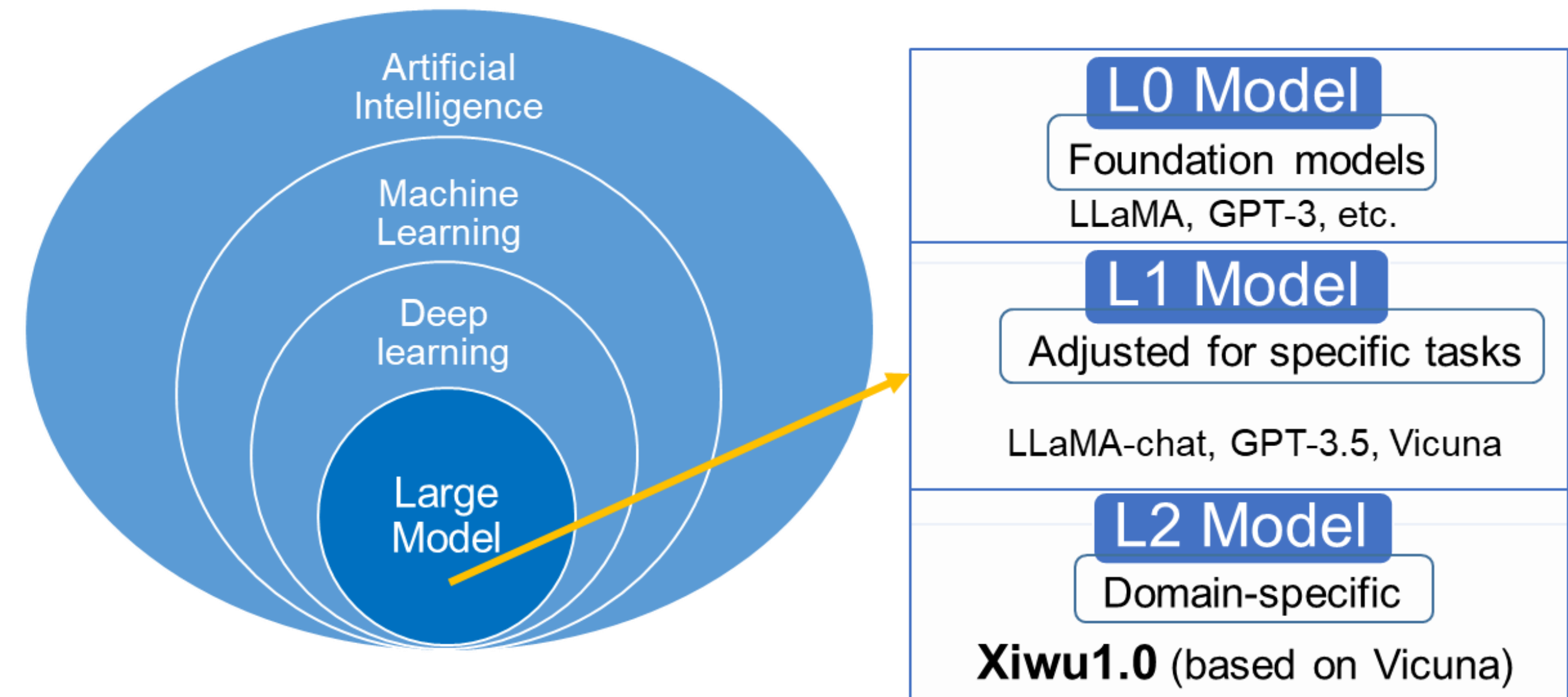
Source: AI Index, 2024 | Chart: 2024 AI Index report



High energy physics?

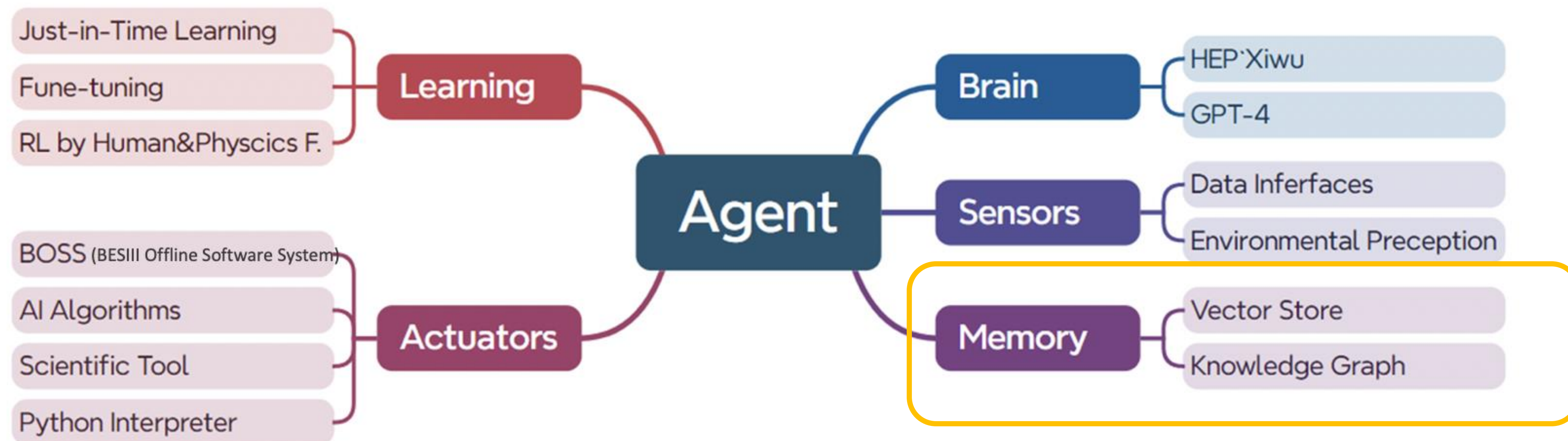
XiWu: A LLM for HEP research

- Large language models, on the base of the transformer architecture, demonstrate impressive performance in text/code generation.
 - GPT4o, Gemini, LLaMa3 (LLaMa3.1), etc
 - Could be trained for HEP research
- For us, open source foundation model + higher level model for HEP + fine tuning for BESIII.
 - **XiWu 1.0**, released in April, 2024 (refer to [arXiv: 2404.08001](https://arxiv.org/abs/2404.08001) for more details).
 - **XiWu 2.0**, based on LLaMa3-8B-Instruct, still in training.



Dr. Sai

- AI agent: AI tools capable of autonomously performing complex tasks
 - Based on Xiwu model, trained with BESIII internal data, e.g. memo, source code, Q-A pairs from internal review.
 - Current: **AI assistant**, helps scientist on data analysis.
 - Ultimate Goal: **AI virtual scientist**, analyzes the data automatically like a human expert.



- BESIII internal data is collected in Vector Store (Database for RAG)
- Knowledge Graph is under developing.

Retrieval-Augmented Generation (RAG) model

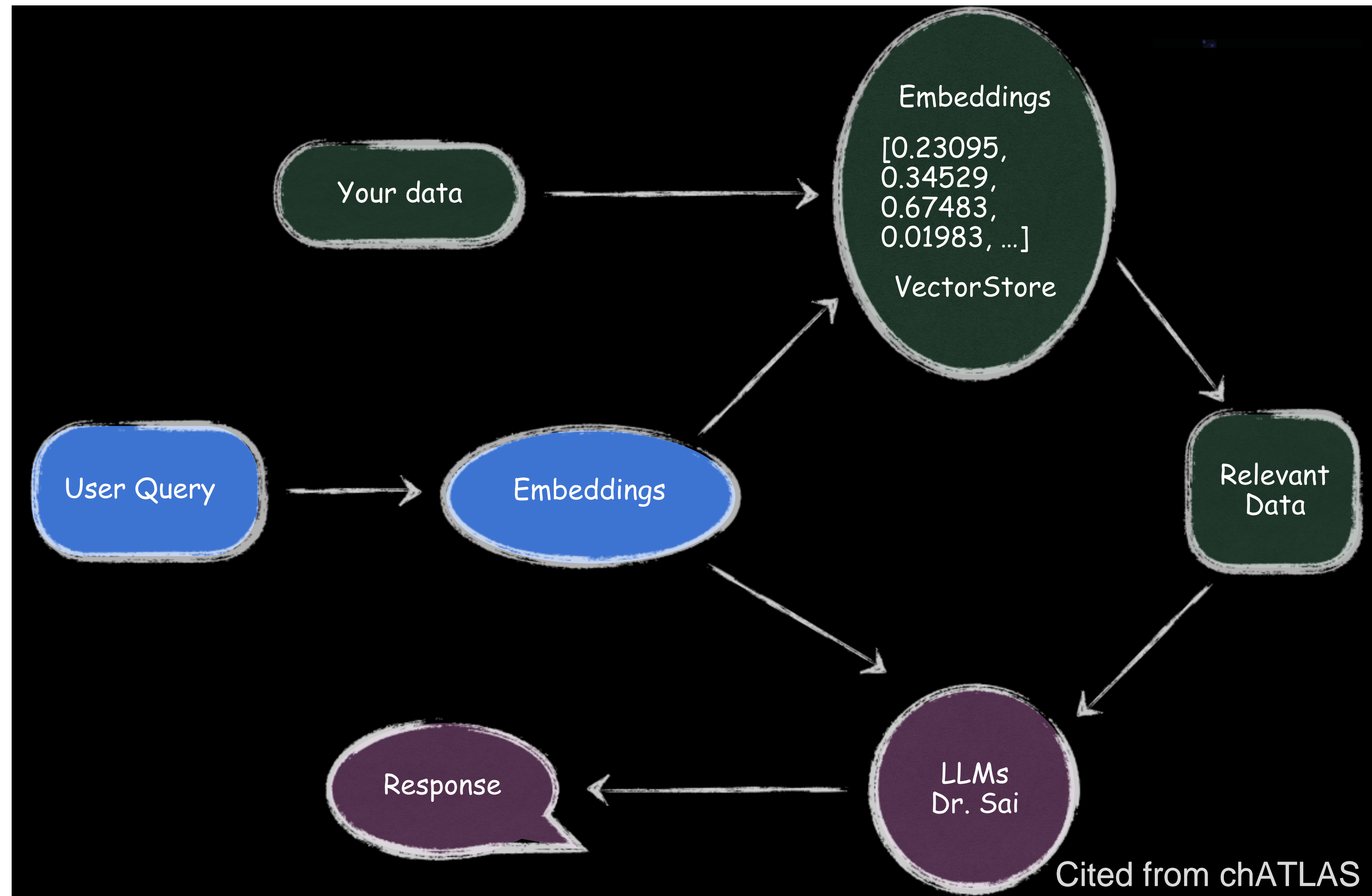
Most effective method to avoid hallucinations!

Goal: Ask questions to a LLM about private information without retraining

Approach: VectorStore: Convert input data into vectors in a multi-dimensional space. (BEG-M3)

Usage:

1. Query vector space
2. Send Query + relevant data to LLM
3. LLM generates response



Outline

➤ Introduction

- Xiwu model for high energy physics (HEP)
- Retrieval-Augmented Generation (RAG)

➤ **BESIII internal data cleaning**

- **Scraping and document collection**

➤ Evaluation

➤ Summary

BESIII data cleaning

To achieve better performance of RAG and Dr. Sai, more accurate and high-quality datasets are needed. BESIII internal data (BESIII website, code examples, memo, draft, etc.).

BESIII Official Website:

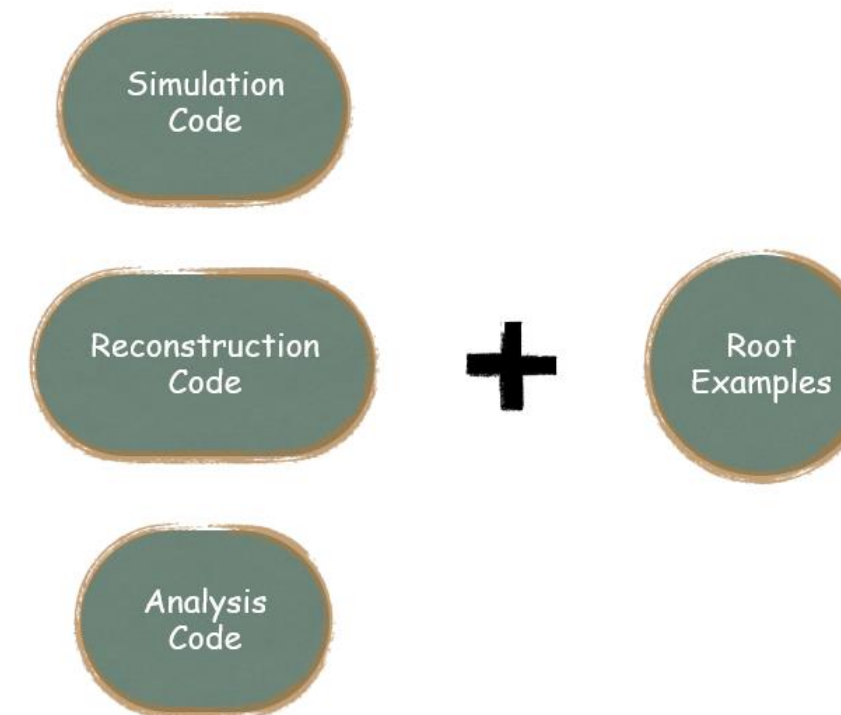
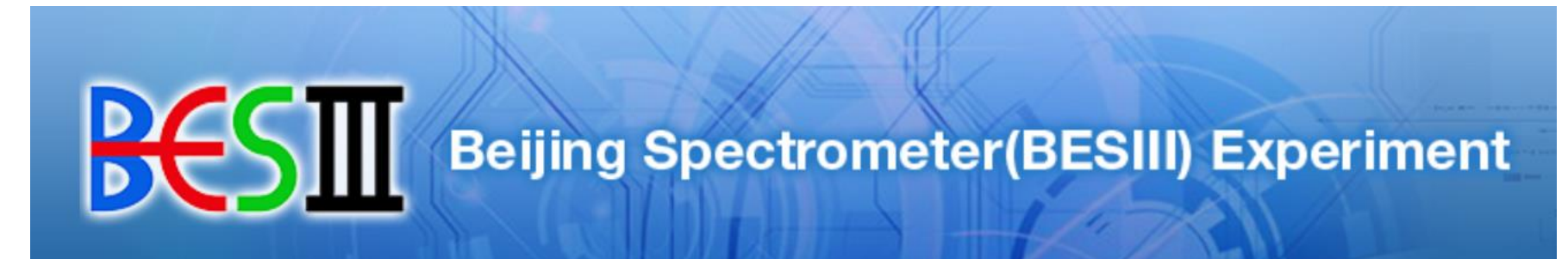
- Physics working groups
- ...

Code examples: Decay channels:

- $J/\psi \rightarrow \rho\pi$
- $\psi(3686) \rightarrow \pi^+\pi^-J/\psi$
- $e^+e^- \rightarrow Z_c(3900)\pi^\pm, Z_c(3900) \rightarrow J/\psi\pi^\pm$

DocDB: Latest memos (only published works)

Hyper-news: Question and answer pairs between referees and the author during memo review stage



Document Database

[\[DocDB Home \]](#) [\[New \]](#) [\[Search \]](#) [\[Last 20 Days \]](#) [\[List Authors \]](#) [\[List Topics \]](#) [\[List Events \]](#) [\[Help \]](#)

HyperNews at BESIII Experiment. Forum List by Category				
	Forums by Category	Recent Postings	Member Info	Overview
	Forums by Time Order	Search in Forums	Members List	Contact Admin
	Request a New Forum	Subscribe to Forums	New Member	

Document collection: Websites

Write python scripts to fetch text contents on websites

- Using **Selenium** (simulate login) and **BeautifulSoup** (parse websites).
- Start from the main page of the website and recursively visit included links.
- Check redirect. A list to record visited links, new links and depth of current page.
- Find all pages without repetition (or reaches maximum depth) and save text contents on each page into documents.
- Append metadata to each document for RAG.

□ WebSite

'metadata': {"author": author, "doc_title": doc_title, "origin": origin, "version": "0.1"}

-BESIII official website (<http://english.ihep.cas.cn/bes/>)

-charm group (https://docbes3.ihep.ac.cn/~charmgroup/index.php/Main_Page)

-charmonium group (https://docbes3.ihep.ac.cn/~charmoniumgroup/index.php/Main_Page)

-light hadron group (https://docbes3.ihep.ac.cn/~lightheadron/index.php/Main_Page)

-tau QCD group (https://docbes3.ihep.ac.cn/~tauqcdgroup/index.php/Main_Page)

-new physics group (https://docbes3.ihep.ac.cn/~newphysgroup/index.php/Main_Page)

-publicatoin committee (https://docbes3.ihep.ac.cn/~pubcom/index.php/Main_Page)

-offline software group (https://docbes3.ihep.ac.cn/~offlinesoftware/index.php/Main_Page)

Information on BESIII websites

Why selenium? Slow but stable, suit for small website.

Document collection: Websites

An example of the document stored in RAG

For conditional search

```
{
  "id_": "0033705d-1d86-4524-b3a2-7d4b23690d14",
  "embedding": null,
  "metadata": {
    "author": "charmonium working group",
    "doc_title": "4260 data details - Charmonium Working Group",
    "origin": "charmonium group page",
    "version": "0.1",
    "hash": "cc704dbcbe39b285a85ef21a9bccadb4fc4face456f7f5c96e831bc9245e4803"
  },
  "excluded_embed_metadata_keys": ["hash"],
  "excluded_llm_metadata_keys": ["hash"],
  "relationships": {
    "1": {
      "node_id": "34fa38e2-ba4b-4047-9de6-35880f05f444",
      "node_type": "4",
      "metadata": {
        "author": "charmonium working group",
        "doc_title": "4260 data details - Charmonium Working Group",
        "origin": "charmonium group page",
        "version": "0.1",
        "hash": "cc704dbcbe39b285a85ef21a9bccadb4fc4face456f7f5c96e831bc9245e4803"
      },
      "hash": "fe54dd3ae530f782461672b7690beb269fda9bdaf03a138268d57c67381a020f",
      "class_name": "RelatedNodeInfo"
    },
    "3": {
      "node_id": "c512d4c4-a898-4702-ac0a-a677a2fe2f96",
      "node_type": "1",
      "metadata": {},
      "hash": "615044724d8252b7dc879db6139389c8bf89a4603dd5a869eb5c28426ebb032d",
      "class_name": "RelatedNodeInfo"
    }
  },
  "text": "https://docbes3.ihep.ac.cn/~charmoniumgroup/index.php/4260_data_details\n4260 data details - Charmonium Working Group\n\n4260 data detailsFrom Charmonium Working GroupData around 4260\nTaken 2012-12-14 to 2013-06-03\nEc.m.= 4090 MeV: 50 pb-1, taken from 2013-06-01 to 2013-06-03.the runno: 33659-- 33719 , read Elog, text format,\nEc.m.= 3900 MeV: 50 pb-1, taken from 2013-05-29 to 2013-06-01.the runno: 33571-- 33657 , read Elog , text format\nEc.m.= 3810 MeV: 48 pb-1, taken from 2013-05-25 to 2013-05-29.the runno: 33490-- 33556 , read Elog, text format\nu00a0;\nEc.m.= 4230 MeV: 1011 pb-1, taken from 2013-04-01 to 2013-05-25.the runno: 32239-- 33484 , read Elog_Apir/ text format, Elog_May/ text format,\nEc.m.= 4245 MeV: 53 pb-1, taken from 2013-03-29 to 2013-04-01.the runno: 32141-- 32226 , read Elog, text format\nEc.m.= 4220 MeV: 52 pb-1, taken from 2013-03-25 to 2013-03-29.the runno: 32046-- 32140 , read Elog, text format\nEc.m.= 4210 MeV: 52 pb-1,",
  "start_char_idx": 0,
  "end_char_idx": 967,
  "text_template": "{metadata_str}\n\n{content}",
  "metadata_template": "{key}: {value}",
  "metadata_seperator": "\n",
  "class_name": "TextNode"
}
```

Link of the page

content in the page

Document collection: Code example

BESIII Code for signal MC

Simulation
Code

Reconstruction
Code

+

Root
Examples

Analysis
Code

BESIII Official code

$$\left\{ \begin{array}{l} - J/\psi \rightarrow \rho\pi \\ - \psi(3686) \rightarrow \pi^+\pi^-J/\psi \\ - e^+e^- \rightarrow Z_c(3900)\pi^\pm, Z_c(3900) \rightarrow J/\psi\pi^\pm \end{array} \right.$$

Stored for RAG:

```
{
  "id": "fb73df0c-1372-451f-bf9f-40372f1105ee",
  "embedding": null,
  "metadata": {
    "author": "huangpan",
    "doc_title": "rhopi_reconstruction",
    "hash": "9aee946fcc61f345281adfeabde3539b7863dd27ae390f35c044a1e162940a5a"
  },
  "excluded_embed_metadata_keys": ["hash"],
  "excluded_llm_metadata_keys": ["hash"],
  "relationships": {
    "1": {
      "node_id": "ddb8da37-53ca-4be7-9ab2-d337a2e4e99e",
      "node_type": "4",
      "metadata": {
        "author": "huangpan",
        "doc_title": "rhopi_reconstruction",
        "hash": "9aee946fcc61f345281adfeabde3539b7863dd27ae390f35c044a1e162940a5a"
      },
      "hash": "44f5ff88565812fc92662b5172974fceac945351865814efc5d9064eb3604efa",
      "class_name": "RelatedNodeInfo"
    },
    "3": {
      "node_id": "71b3e568-5760-45ee-92aa-2e2d255df7d8",
      "node_type": "1",
      "metadata": {},
      "hash": "5e49e6dcb27922c9da4d64db8659c84d12b35b77e6ae7932533729e0064b8616",
      "class_name": "RelatedNodeInfo"
    }
  },
  "text": "//input ROOT MC data\n#include \"$ROOTIOROOT/share/jobOptions_ReadRoot.txt\"\n#include \"$OFFLINEEVENTLOOPMGRROOT/share/OfflineEventLoopMgr_Option.txt\"\n\n// background mixing\n#include \"$BESEVENTMIXERROOT/share/jobOptions_EventMixer_rec.txt\"\n\n#include \"$CALIBSVCROOT/share/job-CalibData.txt\"\n#include \"$MAGNETICFIELDROOT/share/MagneticField.txt\"\n#include \"$ESTIMEALGROOT/share/job_EsTimeAlg.txt\"\n\n// PAT+TSF method for MDC reconstruction\n#include \"$MDCXRECOROOT/share/jobOptions_MdcPatTsfRec.txt\"\n#include \"$KALFITALGROOT/share/job_kalfit_numf_data.txt\"\n#include \"$MDCDEDXALGROOT/share/job_dedx_all.txt\"\n#include \"$TRKEXTALGROOT/share/TrkExtAlgOption.txt\"\n#include \"$TOFRECRROOT/share/jobOptions_TofRec.txt\"\n#include \"$TOFENERGYRECRROOT/share/TofEnergyRecOptions_MC.txt\"\n#include \"$EMCRECRROOT/share/EmcRecOptions.txt\"\n#include \"$MUCRECALGROOT/share/jobOptions_MucRec.txt\"\n#include \"$EVENTASSEMBLYROOT/share/EventAssembly.txt\"\n#include \"$PRIMARYVERTEXALGROOT/share/jobOptions_kalman.txt\"\n#include \"$VEEVERTEXALGROOT/share/jobOptions_veeVertex.txt\"\n#include \"$HLTMAKERALGROOT/share/jobOptions_HltMakerAlg.txt\"\n#include \"$EVENTNAVIGATORROOT/share/EventNavigator.txt\"\n\n//output ROOT REC data\n#include \"$ROOTIOROOT/share/jobOptions_Dst2Root.txt\"\n\n//configure of calibration constants for MC\n#include \"$CALIBSVCROOT/share/calibConfig_rec_mc.txt\"",
  "start_char_idx": 0,
  "end_char_idx": 1333,
  "text_template": "{metadata_str}\n\n{content}",
  "metadata_template": "{key}: {value}",
  "metadata_seperator": "\n",
  "class_name": "TextNode"
}
```


Document Collection: File in PDF format

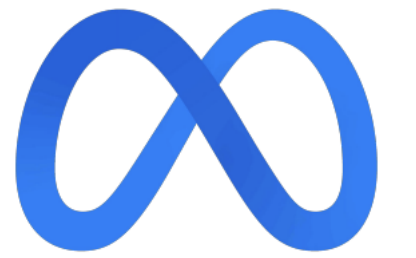
HaiNougat 



Special technology for document parse: **HaiNougat**

- Develop a model for HEP (HaiNougat), the advanced iteration of the Meta Nougat model, facilitating the extraction of complex information from HEP documents.
- A demo of HaiNougat accessible via: <https://ai.ihep.ac.cn/m/hai-nougat>
- More details in <https://github.com/ai4hep/hai-nougat>

Document Collection: File in PDF format



How to extract correct information from PDF files?

HaiNougat

HaiNougat, the advanced iteration of the Nougat model, specializes in transforming PDFs into Mathpix Markdown, facilitating the extraction of complex information from high-energy physics documents. On this platform, you can experience HaiNougat's capabilities, upload PDFs, obtain Mathpix Markdown versions, visualize the results in an editor and preview area, and download the processed output.

You can drop a PDF here:

File processed successfully

```
\begin{table}
\begin{tabular}{l c c c c} \hline \hline Mass (MeV/(c^{2})) & \Gamma_{\rm tot} (MeV) & \Gamma_{ee} (keV) & \mathcal{B}(D^0\bar{D}^0) & \mathcal{B}(D^+D^-) \\
(MeV) & \Gamma_{ee} (keV) & \mathcal{B}(D^0\bar{D}^0) & \mathcal{B}(D^+D^-) \\
(D^{+,+}D^{,-,-}) \\ \hline
3769.9(\pm)2.5 & 23.6(\pm)2.7 & 0.26(\pm)0.04 & 58\% & 42\% \\ \hline
\hline \end{tabular}
\end{table}
Table 1: Properties of the \psi(3770).
```

Mass (MeV/c ²)	Γ _{tot} (MeV)	Γ _{ee} (keV)	B(D ⁰ \bar{D} ⁰)	B(D ⁺ D ⁻)
3769.9 ± 2.5	23.6 ± 2.7	0.26 ± 0.04	58%	42%

Table 1: Properties of the $\psi' = \psi(3770)$.

```
\langle \rho \pi | \psi' \rangle = \langle \rho \pi | 2^3 S_1 \rangle \cos \phi - \langle \rho \pi | 1^3 D_1 \rangle \sin \phi = 0,
\langle \rho \pi | \psi' \rangle = \langle \rho \pi | 2^3 S_1 \rangle \sin \phi + \langle \rho \pi | 1^3 D_1 \rangle \cos \phi
\langle \rho \pi | \psi' \rangle = \langle \rho \pi | 2^3 S_1 \rangle / \sin \phi,
```

$$\begin{aligned} \langle \rho \pi | \psi' \rangle &= \langle \rho \pi | 2^3 S_1 \rangle \cos \phi - \langle \rho \pi | 1^3 D_1 \rangle \sin \phi = 0, \\ \langle \rho \pi | \psi' \rangle &= \langle \rho \pi | 2^3 S_1 \rangle \sin \phi + \langle \rho \pi | 1^3 D_1 \rangle \cos \phi \\ &= \langle \rho \pi | 2^3 S_1 \rangle / \sin \phi, \end{aligned} \tag{16}$$

An example of using HaiNougat to parse HEP domain paper.

Document Collection: File in PDF format

- **BESIII Memo**

500+ memos (only published) from DocDB (using **HaiNougat** converts PDFs to markdown

preserving math symbols and formulas)

Another data cleaning using LLM (GPT3.5-turbo/GPT-4.0)

Used for training for XiWu-2.0

- **Question-Answer pairs (QAs)**

Using Selenium and BeautifulSoup, cracked Hyper-news.

Cleaning **Q-As (~2K)** from posts using HaiNougat and LLM (GPT3.5-turbo/GPT-4.0)

Append into RAG

Outline

➤ Introduction

- Xiwu model for high energy physics (HEP)
- Retrieval-Augmented Generation (RAG)

➤ BESIII internal data cleaning

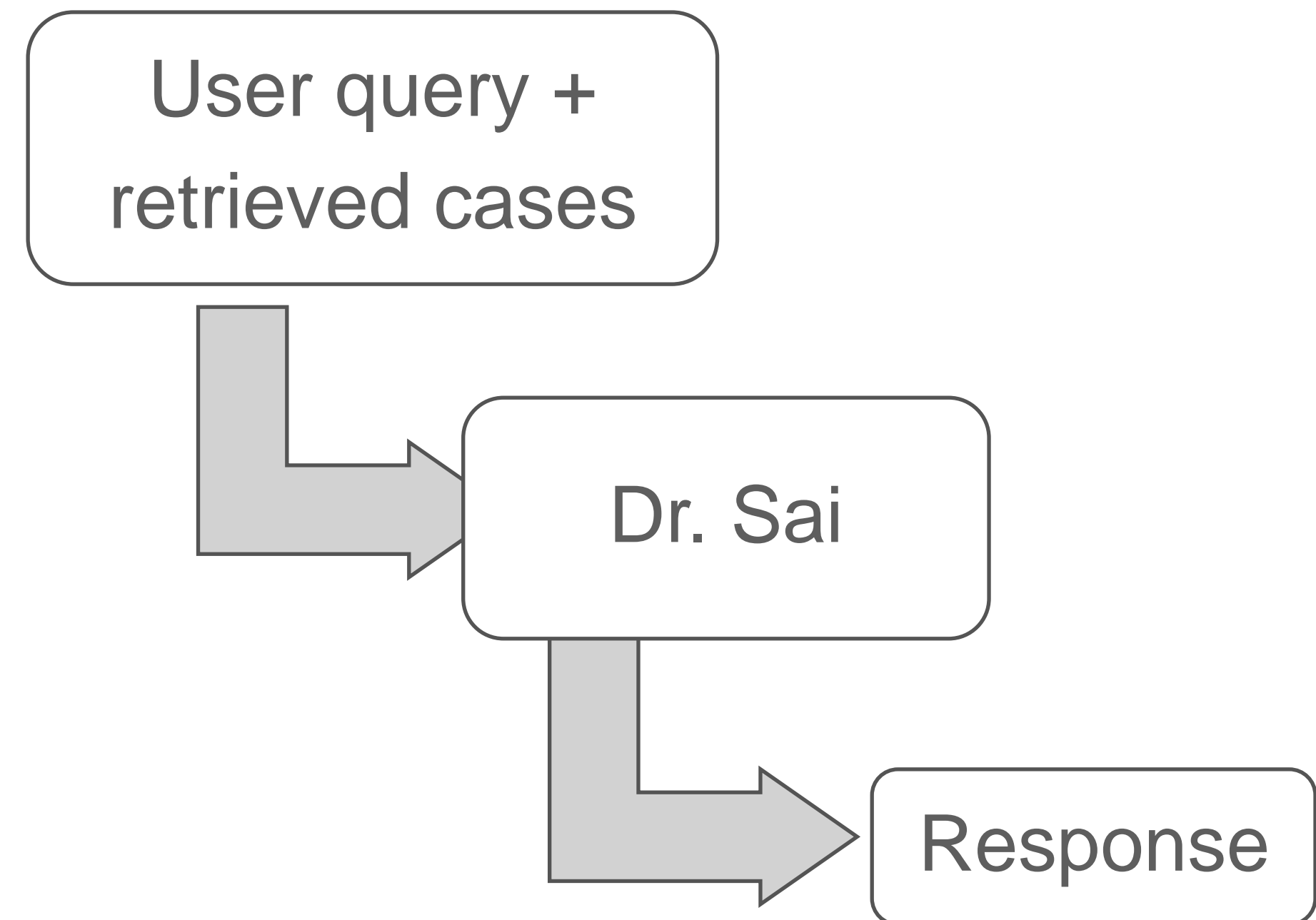
- Scraping and Document Collection

➤ **Evaluation**

➤ Summary

Simple evaluation

- Internal information search
- RAG based query
- Code generation



Evaluation: Internal information search

- Q1: What's link to BESIII offline software group's main page?

The link to the BESIII offline software group's main page is:

[\[https://docbes3.ihep.ac.cn/~offlinesoftware/index.php/Main_Page\]](https://docbes3.ihep.ac.cn/~offlinesoftware/index.php/Main_Page)(https://docbes3.ihep.ac.cn/~offlinesoftware/index.php/Main_Page)

- Q2: Considering BESIII detector, what intensity of magnetic field can the superconducting solenoidal magnet provide?

To determine the intensity of the magnetic field provided by the superconducting solenoidal magnet in the BESIII detector, we can analyze the provided cases for relevant information.

Relevant Cases:

Case 1:

- The superconducting solenoidal magnet in the BESIII detector provides a magnetic field of 1.0 T (0.9 T in 2012).

Case 2:

- Similar to Case 1, it states that the superconducting solenoidal magnet provides a magnetic field of 1.0 T (0.9 T in 2012).

Irrelevant Cases:

- **Case 3 to Case 10**: These cases do not provide information about the magnetic field intensity of the superconducting solenoidal magnet in the BESIII detector. They focus on other aspects of the BESIII experiment, such as data acquisition, particle studies, and theoretical predictions.

Conclusion:

The superconducting solenoidal magnet in the BESIII detector provides a magnetic field intensity of 1.0 Tesla (T). This information is consistently mentioned in both Case 1 and Case 2.

Evaluation: Internal information search

- Q3: Who's on duty of the spokesperson of BESIII collaboration?

To determine who is on duty as the spokesperson of the BESIII collaboration, we can analyze the provided cases for relevant information.

Relevant Information from Cases:

1. **Case 2**:

- Prof. Yuan Changzheng from the Institute of High Energy Physics (IHEP) was elected the new spokesperson of the BESIII Collaboration during the BESIII Physics and Software Workshop held from March 15 to 18, 2018.

2. **Case 3**:

- Reiterates that Prof. Yuan Changzheng was elected as the new spokesperson during the BESIII Physics and Software Workshop in March 2018.

3. **Case 6**:

- Xiaorui Lyu from UCAS was elected co-spokesperson of the collaboration, replacing Haiping Peng from USTC.

4. **Case 10**:

- Provides the current spokespersons:

- **Spokesperson**: Hai-Bo Li, Prof. of IHEP

- **Co-spokespersons**:

- Wolfgang Gradl, Prof. of Johannes Gutenberg University of Mainz

- Xiaorui Lyu, Prof. of UCAS

Conclusion:

Based on the most recent and relevant information from Case 10, the current spokesperson of the BESIII collaboration is **Hai-Bo Li**, with co-spokespersons **Wolfgang Gradl** and **Xiaorui Lyu**.

Evaluation: RAG based query

- Q: How to consider the uncertainty of gamma conversion rate?

Without using
RAG

The uncertainty in the gamma conversion rate can be considered by combining the statistical and systematic uncertainties. The formula for the total uncertainty (ΔR) in the gamma conversion rate (R) can be expressed as: $\Delta R = \sqrt{(\Delta R_{\text{stat}})^2 + (\Delta R_{\text{sys}})^2}$ where ΔR_{stat} is the statistical uncertainty and ΔR_{sys} is the systematic uncertainty. Statistical uncertainty typically arises from the finite number of observed events, while systematic uncertainty can come from various sources such as detector efficiency, background estimation, and calibration errors. To minimize the total uncertainty, both statistical and systematic uncertainties should be carefully evaluated and reduced as much as possible.

With RAG

The uncertainty of gamma conversion rate can be calculated using the formula: $(\sigma_{\{R\}}/R) = n * (\sigma_{\{r\}}/(1-r))$, where r is the conversion rate of 1 photon, n is the multiplicity of photons, and R is the probability of an event not having any gamma conversion. The conversion rate r is typically around -1% for BES, determined by the photon mass attenuation length of the front-end material.

Evaluation: RAG based query

- Q: What is the predicted decay rate of η' to 4π according to the broken-SU(6)xO(3) quark model?

Without using
RAG

The broken-SU(6)xO(3) quark model is a theoretical framework used to predict various particle decay rates. The specific prediction for η' to 4π might be detailed in specialized research papers or reviews.

Using RAG

The decay rate of η' to 4π has been studied in various theoretical and experimental contexts. According to the broken-SU(6)xO(3) quark model, the decay rate of η' to 4π is predicted to be $\Gamma(\eta' \rightarrow 4\pi) = 2.4 \times 10^{-4} \text{ MeV}$, corresponding to a branching fraction of approximately 1.2×10^{-3} . Experimental data provide 90% confidence level upper limits on the branching fractions: $B(\eta' \rightarrow \pi^+\pi^-\pi^+\pi^-) < 2.4 \times 10^{-4}$ and $B(\eta' \rightarrow \pi^+\pi^-\pi^0\pi^0) < 2.6 \times 10^{-3}$. The measured branching fractions are $B(\eta' \rightarrow \pi^+\pi^-\pi^+\pi^-) = (8.40 \pm 0.68(\text{stat}) \pm 0.63(\text{syst})) \times 10^{-5}$ and $B(\eta' \rightarrow \pi^+\pi^-\pi^0\pi^0) = (1.90 \pm 0.36(\text{stat}) \pm 0.17(\text{syst})) \times 10^{-4}$.

Without RAG, the response is general, whereas the answer with RAG is more specific and tailored for BESIII.

Evaluation: Code generation

Separate user requirement into a workflow consisting of subtasks.

Example: $Z_c(3900)$

- Q "I want to analyze $Z_c(3900)$ on BESIII. Please generate signal MC, reconstruct it, then analyze the signal MC and data. At the end, please draw the feature distribution plots
- Planer:
 - Task(1) Generate the signal MC for $Z_c(3900)$
 - Task(2) Reconstruct the signal MC using BOSS.
 - Task(3) Analyze the reconstructed signal MC.
 - Task(4) Analyze the real data.
 - Task(5) Draw the feature distribution plots using ROOT.

One step of task (1):

J

I want to generate signal MC sample of $Z_c(3900)$. Please write the decay card



To generate a signal Monte Carlo (MC) sample for the ($Z_c(3900)$) using BESIII's BOSS software, you need to create a decay card that specifies the decay channels and properties of the ($Z_c(3900)$) particle. Here is an example of such a decay card where ($Z_c(3900)$) decays to ($J/\psi \pi$) and (J/ψ) further decays to ($\mu^+ \mu^-$):

plaintext

```
# filename: Zc3900.dec

# Particle and decay definitions for Zc(3900) and J/psi
Particle Zc(3900)+ 3.900 0.001

Decay Zc(3900)+
  1.0 J/psi pi+ PHSP
Enddecay

Decay J/psi
  1.0 mu+ mu- VLL
Enddecay

End
```

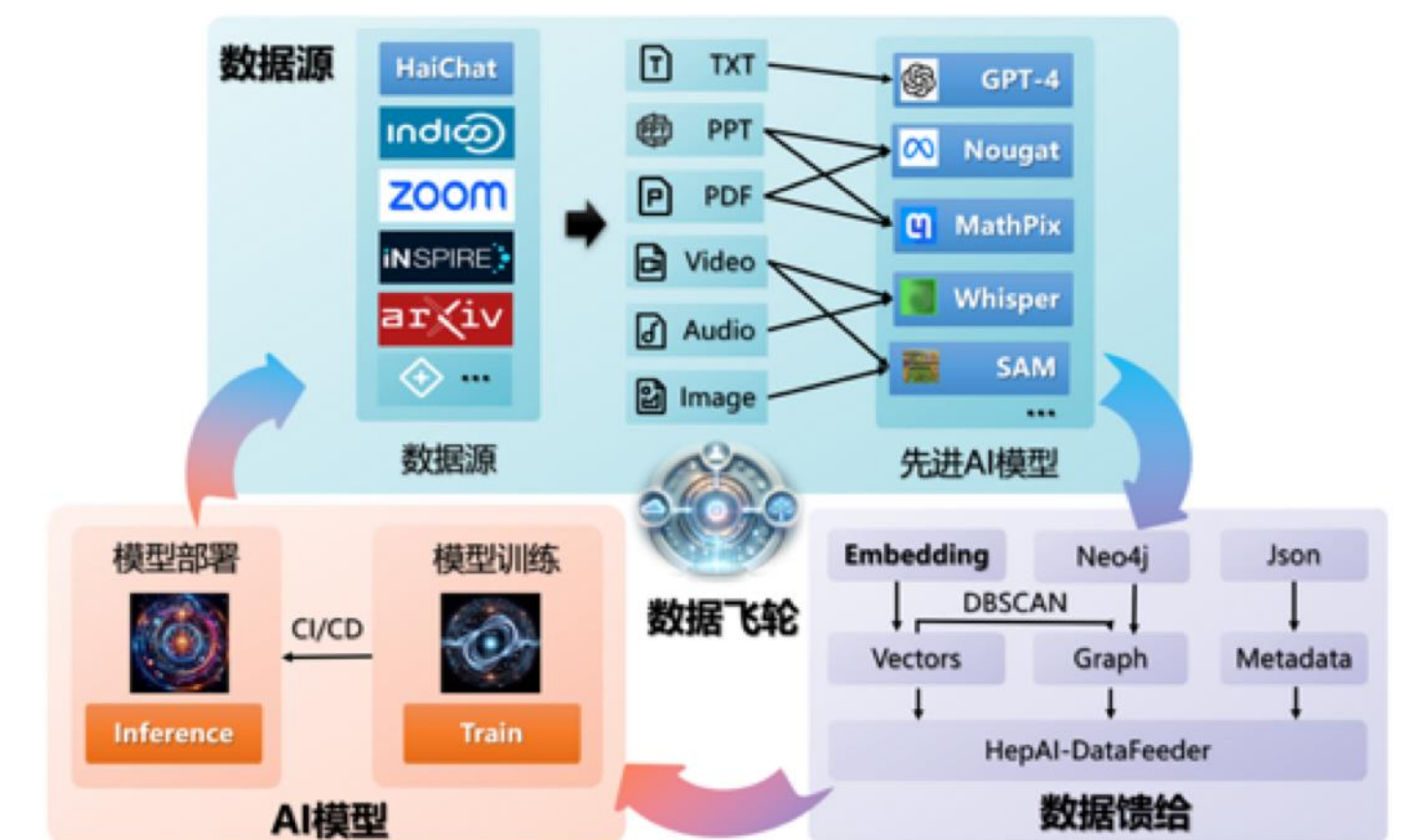
Summary

- Collected and cleaned BESIII internal data

Internal website information, example codes, BESIII (500+) memos and QAs (~2K). After incorporating the RAG model, Dr. Sai performs better in BESIII Q-A and Code generation.

- Near plans:

- Optimize the data cleaning process, regular maintenance and updates of BESIII website
- Collect more datasets, such as indico, BOSS codes, and Information in PDG, for training and RAG
- Develop graphRAG, based on knowledge graph
- Data fly-wheel



Thanks for your listening!