

# **BESIII Internal Data cleaning** and RAG for LLM

Junkun Jiao<sup>1</sup>, Zijie Shang<sup>2</sup> <sup>1</sup>Jilin University, <sup>2</sup>Lanzhou 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



W Dr. Sai



# Outline

## Introduction

Xiwu model for high energy physics (HEP)

 Retrieval-Augmented Generation (RAG) BESIII internal data cleaning

Scraping and document collection



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 Al Index technical performance benchmarks vs. human performance

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





### **High energy physics?**

## Xiwu: A LLM for HEP research

- Large language models, on the base of the tra 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 for more details).
  - XiWu 2.0, based on LLaMa3-8B-Instruct, still in training.



### Large language models, on the base of the transformer architecture, demonstrate impressive



## Dr. Sai

- Al agent: Al 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: Al assistant, helps scientist on data analysis.
  - Ultimate Goal: Al 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: <u>VectorStore</u>: 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:

 $- J/\psi \rightarrow \rho \pi$ 

- ...

 $-\psi(3686) \rightarrow \pi^+\pi^- I/\psi$ 

 $-e^+e^- \to Z_c(3900)\pi^{\pm}, Z_c(3900) \to J/\psi\pi^{\pm}$ 

**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. Fo	rum List by Category		
( )	Forums by Category Forums by Time Order Request a New Forum	Recent Postings Search in Forums Subscribe to Forums	Member Info Members List New Member	Overview Contact Admin

## **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/~lighthadron/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

{"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": "615044724d8252b7dc879db6139389cdbf89a4603dd5a869eb5c28426ebb032d", "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\u00a0;\nEc.m.= 4230 MeV: 1011 pb-1, taken from 2013-04-01 to 2013-05-25.the runno: 32239-- 33484 , read Elog\_Apirl/ 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"}



### For conditional search

### Link of the page

### content in the page





## **Document collection: Code example**

## **BESIII Code for signal MC**





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

### 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\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 \"\$TOFRECROOT/share/jobOptions\_TofRec.txt\"\n#include \"\$TOFENERGYRECROOT/share/TofEnergyRecOptions\_MC.txt\"\n#include \"\$EMCRECROOT/share/EmcRecOptions.txt\"\n#include "\$MUCRECALGROOT/share/jobOptions\_MucRec.txt\"\n\n#include \"\$EVENTASSEMBLYROOT/share/EventAssembly.txt\"\n#include \"\$PRIMARYVERTEXALGROOT/share/jobOptions\_kalman.txt\"\n#include \"\$VEEVERTEXALGROOT/share/jobOptions\_veeVertex.txt\"\n\n#include \"\$HLTMAKERALGROOT/share/jobOptions\_HItMakerAlg.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**

How to extract correct information from PDF files?

During the data cleaning, we found it hard to extract lacksquare

correct information from PDF files, especially for

equations and charts.

HEP-related papers contain **20%** more formulas per ullet

page and **27%** longer formulas than articles in other

scientific fields.



<sup>3/2</sup> were taken in Ref [15]

ratio of  $\psi'$  and  $\psi''$  partial widths to  $J/\psi\pi$ 

 $(\psi'' \rightarrow \gamma \chi_{c0}) = 145 \text{ keV} \cos^2 \phi (1.73 + \tan \phi)$ 

 $\Gamma(\psi'' \rightarrow \gamma \chi_{c2}) = 167 \text{ keV} \cos^2 \phi(0.17 + \tan \phi)$ 

 $(\psi' \rightarrow \gamma \chi_{c0}) = 67 \text{ keV} \cos^2 \phi (1 - 1.73 \tan \theta)$ 

 $\Gamma(\psi' \rightarrow \gamma \chi_{c1}) = 56 \text{ keV} \cos^2 \phi (1 + 0.87 \tan \phi)$ 

 $(\psi' \rightarrow \gamma \chi_{c2}) = 39 \text{ keV} \cos^2 \phi (1 - 0.17 \tan \phi)$ 

Lesults for  $\psi''$  radiative decays [5], for  $\sigma(e^+e^- \rightarrow \psi'')$ 

 $= 0.128 \pm 0.023$ 

 $\Gamma(\psi' \to e^+ e^-)$ 

Charmless final states and S- and D-wave mixing in the  $\psi$ 

Jonathan L. Rosner\* cs, University of Chicago, 5640 S. Ellis Avenue, Chicago

is expected to be mainly  $c\overline{c(1^{3}D_{1})}$ , but tensor forces and cparticle pairs can mix it with  $\psi'(2^3S_1)$  and other states. The implications of this mixing for decays of  $\psi''$  to noncharmed final states are discussed. (i) The ratio  $\Gamma(\psi'' \rightarrow \gamma + \chi_{c2})/\Gamma(\psi'' \rightarrow \gamma + \chi_{c0})$  is expected to be highly ressed if  $\psi''$  is a pure D-wave state, and is enchanced by mixing. (ii) The expected decay  $\psi' \rightarrow \rho \pi$  and her "missing" modes can appear as corresponding  $\psi''$  partial widths, enhanced by a factor depending on th ixing angle. General arguments then suggest a branching ratio of about 1%, give or take a factor of 2, f less hadronic decays of  $\psi''$ . (iii) Enhancements can appear in penguin amplitudes in B decays, i  $\rightarrow K \eta'$  branching ratios, and direct CP-violating asymmetries in  $B \rightarrow K \pi$  decays.

PACS number(s): 13.25.Gv. 12.39.Jh. 13.20.Gd. 14.40.Gz DOI: 10.1103/PhysRevD.64.094002

> turbative QCD also was adequate to explain the observe surbative QCD also was accquate to exprain the control leptonic width [12]. Probably both perturbative and non-perturbative (e.g., coupled-channel) effects are present.

The mixing of the  $\psi''$  with other states can affect both its

ys and those of the other states. In Sec. II we discuss a

ic and radiative partial decay rates of these states. The

d by mixing [5,7,13–15]. "missing decay modes" of the  $\psi'$  [16], such as  $\rho \pi$ 

and  $K^*\overline{K}$ +c.c., are a long-standing puzzle [17-21]. Re-

cently Suzuki [22] showed that if a  $\psi'$  decay amplitude du

wher in Sec. III using the  $\psi' - \psi''$  mixing model descri

earlier. We propose that as a result of coupled-channel effects

mixing the  $\psi'$  and the  $\psi''$ , and for significant effects or

non-charmed final states in decays of both particles, it is

ikely that virtual or real  $D^{(*)}\overline{D^{(*)}}$  pairs produced in lo

e cases are the decays of B mesons, which can invo

ation of the final cc pair can lead to an effectiv

TABLE I. Properties of the  $\psi'' = \psi(3770)$ 

 $(MeV/c^2)$   $\Gamma_{tot} (MeV)$   $\Gamma_{ee} (keV)$   $\mathcal{B}(D^0\overline{D^0})$   $\mathcal{B}(D^+D)$ 

expected decay width  $\Gamma(\psi' \rightarrow \rho \pi) \simeq 0.5$  keV and other

' modes could show up as corresponding partia

ling on the mixing angle. Since the latter sta

plified model for  $\psi' - \psi''$  mixing and its implications for

sed if  $\psi''$  is a pure

vidths in  $\psi''$  decays, possibly en

I. INTRODUCTION

The lowest resonance in electron-positron collisio bove the charmed particle pair production threshold is the  $\psi'' = \psi(3770)$ , discovered somewhat after the  $J/\psi(3097)$  and the  $\psi' = \psi(3686) [1]^1$  It provides a rich source of  $D^0 \overline{D^0}$  and  $D^+D^-$  pairs, as anticipated theoretically [2]. The largest data mple of  $\psi''$  decays studied so far, by the Mark III Collabo ration at the Stanford electron-positron collider SPEAR [3], has been  $9.56\pm0.48$  pb<sup>-1</sup>. Plans are under way to accumuate as much as 3 fb<sup>-1</sup> at the Cornell Electron Storage Ring CESR), which will permit much more incisive tests of a mber of open questions [4]. In the present paper we dis-ss several of these which involve observation of non-armed final states of the  $\psi''$ . These have been studied in previous papers [5,6] based on the Mark III data.

The  $\psi''$  is the only present candidate for a D-wave ( um level. (Strategies for finding the corre ponding bb levels have been noted in Refs. [7,8].) Although arily  $c\overline{c(1^3D_1)}$ , its leptonic width (quoted in Tabl bution from mixing with S-wave tates, such as the nearby  $\psi'(2^{3}S_{1})$  and to a lesser extent with  $J/\psi(1^3S_1)$  [10] and  $n \ge 3$  S-wave states above 4 GeV Early calculations of this mixing based on contri diate real and virtual states of cha irs [2] predicted a  $\psi''$  contribution to the  $e^+e^- \rightarrow D\overline{D}$  cross on which indicated the utility of this state as a "charm tory" and predicted its leptonic width quite well.<sup>3</sup> It was ter found that mixing due to a tensor force based on per

The numbers in parentheses indicate the masses of the particle

We shall use spectroscopic notation  $n^{2S+1}L_J$ = 1.2.3... is the radial quantum number: S=0 or 1 is the O pin;  $L = S, P, D, \dots$   $(l = 0, 1, 2, \dots)$  is the orbital ang m; and J=0,1,2,... is the total spin. For later discussions of mixing due to couple

HARMLESS FINAL STATES AND S AND D WAVE

©2001 The American Physical Society

Decay mode	$J/\psi$ decays [9] $\Gamma_{tot} = 87 \pm 5$ keV $\Gamma_{ee} = 5.26 \pm 0.37$ keV		$\psi'$ decays [33] $\Gamma_{tot}=277\pm31$ keV [9] $\Gamma_{ee}=2.12\pm0.18$ keV [9]		
	В	$\Gamma$ (keV)	B	$\Gamma$ (eV)	$\Gamma_{\rm pred}{}^{\rm a}~({\rm eV})$
ρπ	(1.27±0.09)%	$1.10 \pm 0.10$	<2.8×10 <sup>-5</sup>	< 8.6	443±63
K <sup>+</sup> K <sup>*-</sup> (892) <sup>b</sup>	(0.50±0.04)%	$0.44 \pm 0.04$	$<3.0 \times 10^{-5}$	< 9.2	$177 \pm 24$
Based on prescri Plus c.c. $\Gamma(\psi'' \rightarrow \gamma \chi_{c0}) =$	ption given in text. 510±190 keV,	(12)	higher-mass $\psi'$ into lo Then we expect for a	w-multipli	city final state
$\Gamma(\psi'' \to \gamma \chi_{c1}) =$	440±160 keV,	(13)	$\Gamma(\psi' \rightarrow f$	$= \Gamma(J/\psi)$	$\rightarrow f) \frac{\Gamma_{ee}(\psi')}{\Gamma_{ee}(\psi')}$

These partial widths scale as  $1/\sigma(\psi^{\eta})$ . So far it does not seem possible to reconcile the central values of these results with the values of  $\phi$  suggested earlier.<sup>4</sup> The model for mixing between  $\psi'$  and  $\psi''$  may be oversimplified, and relative arch for a 338 MeV monochromatic photon in the decays of the  $\psi''$  would represent a worthwhile first step in the deation of this interesting resonance's mixing param

### III. MISSING MODES OF THE $\psi'$

arized a wide class of hadroni ay modes of the  $\psi'$ , measured by the BES Collaboration at the Beijing Electron-Positron Collider (BEPC), which a emost is the  $\rho\pi$  final state, with  $K^+K^{*-}(892) + c.c.$ econd place. Let us review the expectations and the data for e two modes. [The decay  $\psi' \rightarrow K^0 \overline{K^*}^0$ (892)+ n observed with a branching ratio of (8.1  $\pm 1.6$   $\times 10^{-5}$  which indicates the contribution of a sig tual-photon contribution [18,19,22], and we sl

iscuss it further.] e summarize in Table III the total widths, 1 ratios, and derived partial widths for  $J/\psi$  and  $\psi'$  decays in  $\rho \pi$  and  $K^+ \overline{K^*}(892)^-$ , as well as the partial widths predicted or the  $\psi'$  decays to these final states. Both hadronic and leptonic decay rates are proportional to the square of the wave function at the origin  $|\Psi(0)|^2$ . Although one might expect an additional factor of  $1/M_V^2$ , where  $M_V$  is the mass of the decaying vector meson entering into the leptoni width, we shall ignore this effect, since it is probably offse by a (form) factor suppressing the hadronic decay of the

<sup>4</sup>The solution with  $\phi = 12^{\circ}$ , favored by coupled-channel calculation tions [29,30], predicts  $\Gamma(\psi'' \rightarrow \gamma \chi_{c(0,1,2)}) = (524,73,61)$  keV, implying that the  $\chi_{c1}$  signal of Ref. [5] should not be confirmed.

JONATHAN L. ROSNE

		-	- 3-	-	- 3-
Final	ω		Pure 'S <sub>1</sub>	Pure $D_1$	
state	(MeV)	С	$\Gamma({}^3P_J)/\Gamma({}^3P_0)$	С	$\Gamma(^{3}P_{J})/\Gamma(^{3}P_{0})$
${}^{3}P_{0}$	338	1/9	1	2/9	1
${}^{3}P_{1}$	250	1/3	1.22	1/6	0.30
'n	200	5/0	1.16	1/00	0.012

 $\rightarrow \overline{s}$  or  $\overline{b} \rightarrow \overline{d}$  penguin amplitude [19,23–25], which app to be needed in understanding large branching ratios for .  $\rightarrow K \eta'$  [26] and  $B \rightarrow K \pi$ . Moreover, Suzuki [22] has p posed that this reann ilation, at least in  $\psi'$  decays, is a posed that this realimination, at least in  $\psi$  decays, is as circle with a large final-state phase. We discuss implication of this suggestion for CP violation in B decays in Sec. while Sec. V concludes.

### II. RADIATIVE $\psi''$ DECAYS

relative branching ratios for radiative decays to  $1^{3}P_{1}$ ) states are very different for 2S and 1D states. The observation of radiative decays  $\psi'' \rightarrow \gamma + \chi_{c}$  can determine he degree to which the  $\psi''$  is mixed with an S-wave state  $c^{-2}$  22  $(2^{-2})^{-2}$ 



where  $e_Q$  is the quark charge (in units of |e|), = 1/137.036 is the fine-structure constant,  $\omega$  is the phot energy, and  $\langle r \rangle$  is the matrix element of r between initial and final radial wave functions. The coefficients C are summ rized in Table II, where we compare relative rates for E1 sumptions from  $\psi''$  to  $\chi_c$  states under the two extreme a sumptions of a pure S-wave or a pure D-wave. The distin tive pattern associated with the pure  ${}^{3}D_{1}$  configuration is ratio  $\mathcal{B}(\gamma + \chi_{c1})/\mathcal{B}(\gamma + \chi_{c0}) = 0.3$  and an almost complete suppression of the ratio  $\mathcal{B}(\gamma + \chi_{c2})/\mathcal{B}(\gamma + \chi_{c0})$ .

more detailed model can be constructed by assu that the  $\psi''$  is a mixture of a  $1^{3}D_{1}$  and a  $2^{3}S_{1}$  state [15]  $|\psi''\rangle = |1^3 D_1\rangle \cos \phi + |2^3 S_1\rangle \sin \phi$  $|\psi'\rangle = -|1^3D_1\rangle\sin\phi + |2^3S_1\rangle\cos\phi$ ptonic widths of  $\psi''$  and  $\psi'$  are then [27]  $\Gamma(\psi'' \rightarrow e^+ e^-)$  $= \frac{4 \alpha e_c}{M_{u''}^2} \sin \phi R_{2S}(0) + \frac{5}{2 \sqrt{2} m_c^2} \cos \phi R_{1D}''(0)$ 

avored value  $\phi \simeq 12^\circ$ . Either branching ratio

 $(\phi \pi) < 1.3 \times 10^{-3} \times [5 \text{ nb}/\sigma(\psi')] [5].$ 

ompatible with the current upper bound  $\mathcal{B}(\psi'')$ 

An alternative mechanism discussed by Suzuki [22] for

ntroducing an additional nonperturbative  $\psi'$  decay ampli-

ude is mixing with a vector glueball state (first discussed in entext of  $J/\psi$  decays [34]). In this case the  $\psi''$  is per-

nitted, but not required, to mix with the vector glueball, so

there is no particular reason for the missing partial widths for decays to show up as corresponding  $\psi''$  partial dec

ONATHAN L. ROSNER

ntly neglected to take account of relative signs o and D-wave contributions in the first three of the ], probably as a result of relativistic con The  $\psi'$  decays are expected to be particularly sense such corrections as a result of the node in the 2S function; it is possible that the  $\psi''$  predictions could be eliable, since neither the 1D nor 1P radial wave fund

### PHYSICAL REVIEW D 64 09400

erage I find to be 8.0±0.7 nb.5 This possible discrepand was a factor motivating the studies in Refs. [5,6]. Those and related searches need to be performed with greater sensitivi and with off-res grounds from such processes as  $e^+e^- \rightarrow \gamma^{*-}$ drons. In any event, the search for the "missing fina states" of the  $\psi'$  among the decay products of the  $\psi''$  is a able goal of foreseen studies [4]

### IV. IMPLICATIONS FOR B DECAYS

ment is a matter of some debate [42], since a rece

The enhancement of the inclusive branching rat

Gérard and Weyers [20] have proposed that the three gluon decay of the  $\psi'$  is absent or suppressed, and that the  $\psi'$  decays to hadrons instead mainly via a two-step process volving an intermediate  $c \overline{c(}^1 P_1)$  state. Feldmann and Kroll 1] have proposed that the  $J/\psi \rightarrow \rho \pi$  decay is enhanced ather than  $\psi' \rightarrow \rho \pi$  being suppressed) by mixing of the  $J/\psi$ with light-quark states, notably  $\omega$  and  $\phi$ . Both mechanism do not imply any special role for  $\psi''$  charmless decays. Ar-guments against them, based on data summarized in the last of Refs. [17] and in Ref. [33], include the appearance of essed light-quark decay modes of the  $\psi'$  and the lack of evidence for helicity suppression in  $J/\psi$  dec

nvolving a single virtual photon. As Suzuki has noted, the cases of suppressed hadronic final states of the  $\psi'$  cannot extend to all its decays; indeed the total hadronic width of  $\psi'$  exceeds estimates based on trapolating from the  $J/\psi$  using perturbative QCD by some -70% [22,35]. The non-perturbative effect of coupling to rtual charmed particle pairs, followed by the rennihilation of these pairs into non-charmed final states ust thus be re nsible for some tens of keV of the total width of the  $\psi'$  in Suzuki's scheme.

A corresponding effect in the decays of the  $\psi''$ , which is bout 85 times as wide as the  $\psi'$ , would contribute at most a ent to its total width. Present searches for non-charmed cays of the  $\psi''$  [5,6] are not sensitive enough to exclude as possibility since they did not compare on-resonance data with data taken off-resonance at a sufficiently close energy

 $\mathcal{B}(B \to \eta' X)$  [45] in comparison with theoretical expectations [46] can be explained A related method allows one to estimate the partial decay rate of  $\psi''$  to non-charmed final states. The branching ratio  $\mathcal{B}(J/\psi \rightarrow \rho \pi)$  is  $(1.27 \pm 0.09)$ %. Since about 1/3 of  $J/\psi$  decays can be ascribed to non-3g mechanisms, we expect  $\rho\pi$ to account for about 2% of all *hadronic J1*  $\psi$  decays, and thus no more than this percentage of  $\psi''$  hadronic charmless decays. (The availability of more final states undoubtedly re- $\rightarrow K^0 \pi^+$ , can be generated duces the  $\rho\pi$  fraction in comparison with  $J/\psi$  hadronic de-bly in  $B^+ \rightarrow \pi^0 K^+$  and  $B^0 \rightarrow K^+ \pi^-$ , a CP-violating asymptotic decays.) We thus estimate for hadronic charmless decays  $\mathcal{B}(\psi'') \ge 2 \times 10^{-4}/2\% \simeq 1\%$ , again give or take a factor of 2 depending on the sign of  $\phi$ . This is consistent with our pre-

vious estimate. experiment [47]. The enhancement of the penguin amplitude It is even possible that we have seriously underestimated by the intrinsically non-perturbative charm rescattering The verify possible has we have schooldy underestimated the role of non-charmed final states in hadronic  $\psi''$  decays. If so, there is a chance of reconciling the smaller cross section for  $e^+e^- \rightarrow \psi''$  measured by the Mark III Collaboration using a comparison of single-charm and double-charm production,  $\sigma(\psi'') = 5.0 \pm 0.5$  nb [3], with higher values obtained by <sup>5</sup>The same average was found in [5] without the data of [40].

Example of charts and formulas in a HEP domain paper

This relation has been used to predict the quantities  $\Gamma_{\text{pred}}$  in Table III. One sees that  $\psi' \rightarrow \rho \pi$  is suppressed by a factor of at least  $\sim 50$  with respect to naive expectations, while the ing factor for  $K^+K^{*0}(892)$  + 2] has proposed that the coupling of  $\psi'$  to vir tual pairs of charmed particles could provide an amplitud which interferes destructively with the perturbative QCD process  $\psi' \rightarrow 3g$  in the specific cases of  $\rho \pi$  and  $K\overline{K}^*(892)$ +c.c. hadronic decays. If this is the case, and if virtual ned particle pairs also play a role in mixing  $\psi'$  and  $\psi'$ 

would expect a similar amplitude to  $D^{(*)}D^{(*)} \rightarrow \rho \pi$  or  $KK^*(892) + c.c.$ In the absence of a detailed coupled-channel analysis, I ume that the main effect on  $\dot{\psi}'$  and  $\psi''$  of their mutu oupling to charmed particle pairs is precisely the mixing ussed in the previous section. Let us assume that this mixing and the couplings of  $\psi'$  and  $\psi''$  to  $\rho\pi$  and

 $K\overline{K^*}(892)$  + c.c. are such as to cancel the  $\psi'$  hadronic vidths to these final states [which are related to one another

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

ip instead as decay modes of  $\psi''$ , enhanced by the factor of

tios for two solutions of mixing angle $\phi$ .					
φ (°)	-27±2	12±2			
$1/\sin^2\phi$	4.8±0.6	22±6			
$\Gamma(\psi'' \rightarrow \rho \pi)$ (keV)	$2.1 \pm 0.4$	9.8±3.0			
$\mathcal{B}(\psi'' \rightarrow \rho \pi)(10^{-4})$	0.9±0.2	4.1±1.4			

ions undoubtedly play a role. Nevertheless, the

sive branching ratios  $\mathcal{B}(B \rightarrow K \eta')$  [45], in comparison wi the penguin contribution leading to  $B^0 \rightarrow K^+ \pi^-$  or B (5) In any  $B \to K\pi$  process in which the dom

uin amplitude leads to a net enhancement of  $\overline{b} \rightarrow \overline{s}$  and d transitions. The enhancement need not be large enoug itions, which are in the range of a few percent of all . (2) The number n, of charmed particles per average L decay can be reduced by the reannihilation of  $c\bar{c}$  to light quarks. The degree to which this improves agreement wit

n Refs. [19,23-25] in which the decay  $\overline{b} \rightarrow \overline{ccs}$  or  $\overline{b} \rightarrow \overline{ccs}$ butes to a penguin amplitude with a large strong phas veral implications of this possibility were reviewed i and others have been pointed out in [24]. These includ diminished with respect to the theoretical prediction if the

ely with the pertubative 3g contribution in the  $\rho\pi$  an  $K\overline{K^*}(892)$  + c.c. channels. If this new contribution is due tering into non-charmed final states through charme particle pairs, it is exactly the type of contribution prop

A key observation in Ref. [22] with regard to the add ys is that it is likely t e a large final-state phase, in order to interfere destru

 $\frac{1}{M_{\psi'}^2 \Gamma(\psi' \to e^+ e^-)} = \left| \frac{1}{0.734 \cos \phi - 0.095 \sin \phi} \right|$ ies agree with those of Kuang and Yan [28], whose  $\theta$  is the s our  $-\phi$ . As they note, the smaller  $-|\phi|$  solution  $\Gamma(\psi'' \rightarrow \gamma \chi_{c1}) = 176 \text{ keV} \cos^2 \phi (-0.87 + \tan \phi)^2$ 

 $= \frac{4 \alpha^2 e_c^2}{M_{\psi'}^2} \left| \cos \phi R_{25}(0) - \frac{5}{2 \sqrt{2} m_c^2} \sin \phi R_{1D}''(0) \right|^2$  $e_{1} = 2/3$   $R_{2}c(0) = (4\pi)^{1/2} \Psi_{2}c(0)$  is the radial 2 nction at r=0, and  $R''_{1D}(0)$  is the second derivation radial 2D wave function at the origin.  $M_{\mu''}^2 \Gamma(\psi'' \to e^+ e^-) = |0.734 \sin \phi + 0.095 \cos \phi|$ 

## **Document Collection: File in PDF format**

Special technology for document parse: HaiNougat

Develop a model for HEP (HaiNougat), the advanced iteration of the Meta  $\bullet$ Nougat model, facilitating the extraction of complex information from HEP

documents.

- A demo of HaiNougat accessible via: <u>https://ai.ihep.ac.cn/m/hai-nougat</u>  $\bullet$
- More details in <u>https://github.com/ai4hep/hai-nougat</u>  $\bullet$





## **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:	Choose a PDF	Chapdf	t Upload to	o HaiNougat	
File processed successfully					
<b>↓</b> Download					
\begin{table}					_
\begin{tabular}{I c c c c}	\hline \hline Mass (	MeV/\(c^{2}\))	& \(\Gamma_{\I	rm tot}\)	
(MeV) & \(\Gamma_{\rm	ee}\) (keV) & \({\cal	B}(D^{0}D	}^{0})\) & \({\ca	B}	
(D^{+}D^{-})\) \\ \hline					
3769.9\(\pm\)2.5 & 23	.6\(\pm\)2.7 & 0.2	26\(\pm\)0.04	& 58\% & 42\%	6 \\ \hline	
\hline \end{tabular}					
\end{table}					Table
Table 1: Properties of th	e \(\psi^{\prime}\)= \	(\psi(3770)\).			Table
\[\langle\rho\pi \psi^{\pr	ime}\rangle = \langle	e\rho\pi 2^{3}S_	_{1}\rangle\!\cos	\phi\!-	$\langle  ho$
\!\langle\rho\pi  1^{3}D_	{1}\rangle\!\sin\phi\!	=\!0,\] \			
[\langle\rho\pi \psi^{\pri	me}\rangle =				
\langle\rho\pi 2^{3}S_{1	}\rangle\!\sin\phi\!+\!	!\langle\rho\pi			
1^{3}D_{1}\rangle\!\cos\	phi \tag{16}\] \[= \lan	gle\rho\pi 2^{3	}S_{1}\rangle\!/\	sin\phi,\]	

### An example of using HaiNougat to parse HEP domain paper.



Mass $(MeV/c^2)$	$\Gamma_{tot}~({\rm MeV})$	$\Gamma_{ee}~({\rm keV})$	${\cal B}(D^0{ar D}^0)$	$\mathcal{B}(D^{+}D^{-})$
$3769.9 \pm 2.5$	$23.6\pm2.7$	$0.26\pm0.04$	58%	42%

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

$$egin{aligned} & & & \langle 
ho \, \pi | 2^3 S_1 
angle \cos \phi - \langle 
ho \, \pi | 1^3 D_1 
angle \sin \phi = 0, \ & & \langle 
ho \, \pi | \psi' 
angle &= \langle 
ho \, \pi | 2^3 S_1 
angle \sin \phi + \langle 
ho \, \pi | 1^3 D_1 
angle \cos \phi & & (16) \ & & = \langle 
ho \, \pi | 2^3 S_1 
angle / \sin \phi, \end{aligned}$$



## **Document Collection: File in PDF format**

### **BESIII** Memo lacksquare

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)** lacksquare

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



### HaiNougat

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



# Outline

## Introduction

- Xiwu model for high energy physics (HEP)
- Retrieval-Augmented Generation (RAG)
- BESIII internal data cleaning
  - Scraping and Document Collection







## 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)

### • 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 ( $\Delta R$ ) in the gamma conversion rate (R) can be expressed as:  $\Delta R = sqrt((\Delta R_stat)^2 + (\Delta R_sys)^2)$  where  $\Delta R_stat$ is the statistical uncertainty and AR\_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**

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 eta' to 4pi might be detailed in specialized research papers or reviews.

Using RAG

x 10^-4.

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



### • Q: What is the predicted decay rate of eta' to 4pi according to the broken-SU(6)xO(3)

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



## **Evaluation: Code generation**

Separate user requirement into a workflow consisting of subtasks.

Example: Zc(3900)

- Q "I want to analyze Zc(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 Zc(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):



I want to generate signal MC sample of Zc(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/\psi ) 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

- Near plans:
  - website
  - training and RAG
  - Develop graphRAG, based on knowledge graph
  - Data fly-wheel



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

### Optimize the data cleaning process, regular maintenance and updates of BESIII

### Collect more datasets, such as indico, BOSS codes, and Information in PDG, for



## Thanks for your listening!