

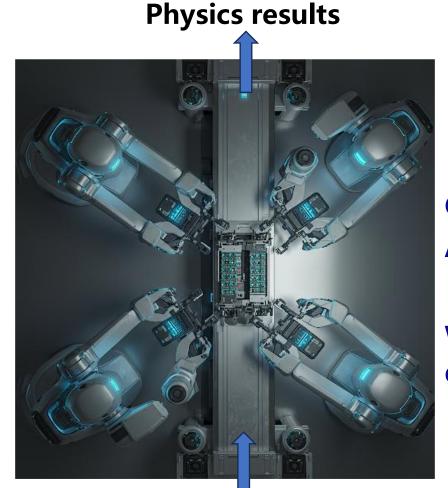
LLM-based AI assistant for HEP data analysis

Ke Li

on behalf of Dr. Sai working group

Outline

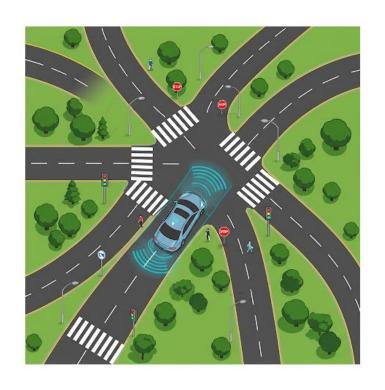
- Motivation
 - what is LLM
 - why we need LLM
- Dr. Sai project
 - architecture
 - training
- Results from beta version
- Summary and prospects

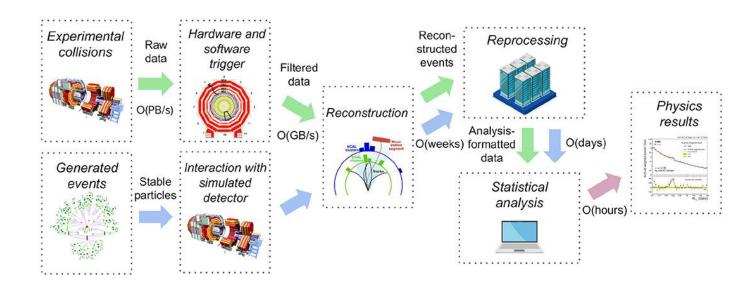


Collision data

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

Driving vs. HEP data analysis

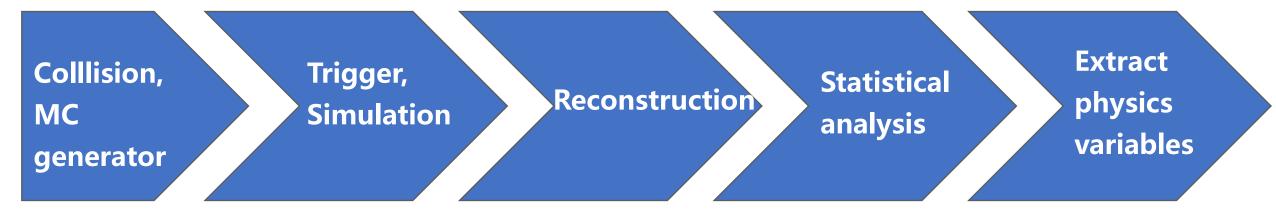




- Clear rules,
- Simple actions,
- Massive training data

- Complex rules, huge amount of degree-of-freedoms
- Lots of sub-steps, need lots of HEP knowledge
- Only ~10k experimental results delivered in 50 years

Data process workflow at HEP experiment



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

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

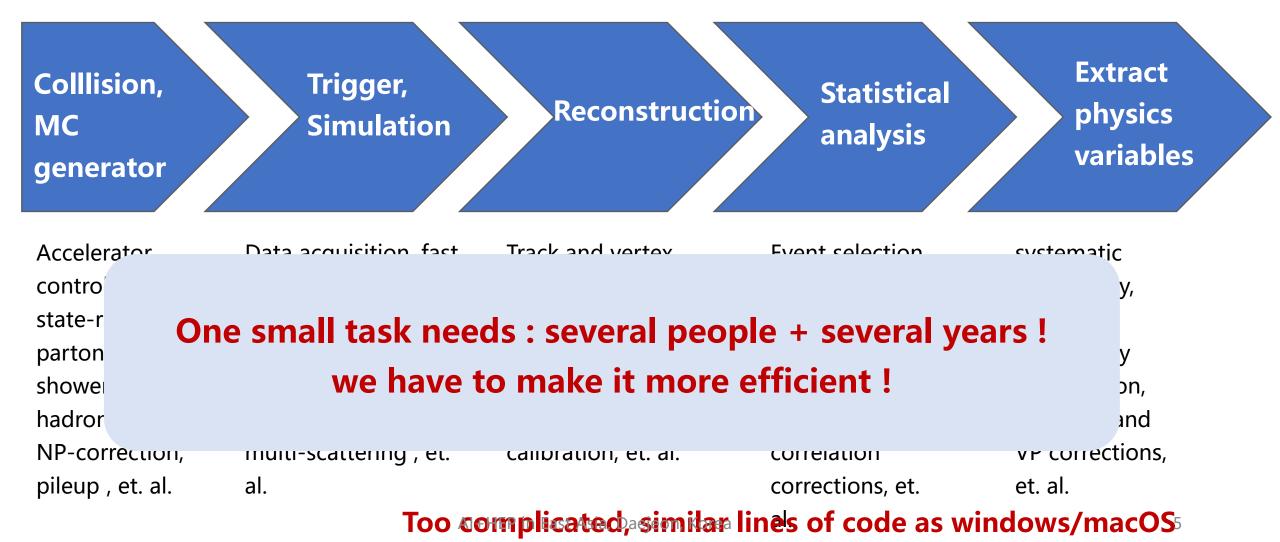
Track and vertex finding and fitting, clusterization and reconstruction of jet, jet tagging, kinematic fit, detector calibration, et. al.

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

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

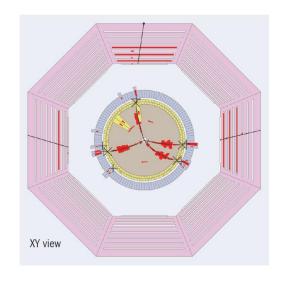
Too Acomplicated as imilar lines of code as windows/macOS4

Data process workflow at HEP experiment

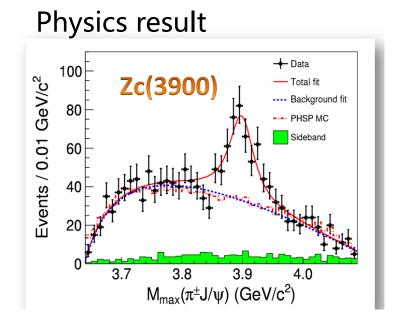


Focus on last two steps

Collision event







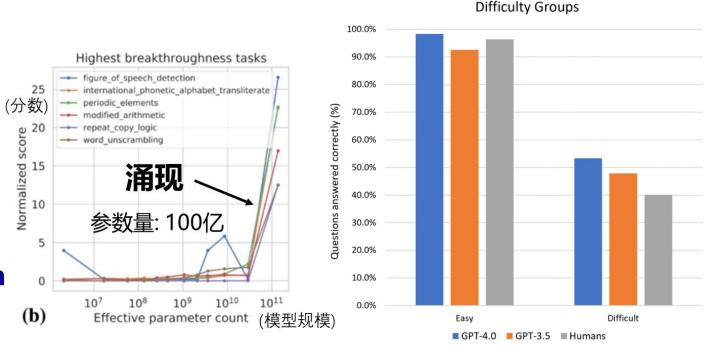
DAQ->Hits -> reconstruction -> selections and statistical analysis -> physics results extraction

Take BESIII as a starting point, given the analysis at LHC is much more complex

Can we use AI(LLM) to automate the workflow?

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
- 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**



Why we need LLM

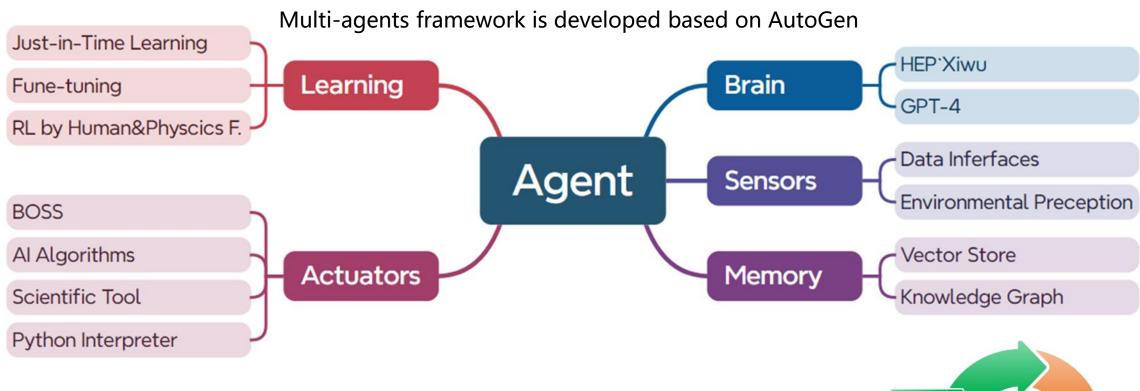
- Major effort in BESIII analysis is spent in writing/testing/updating code/text
 - LLM is good at code/text generation!
- Key problems for LLM at HEP
 - How LLM can understand and remember HEP knowledge better?
 - Logic, data structure, tools' guideline, validation and feedback ...
 - Current solutions:
 - More accurate and good quality data for training
 - More tests and validations
 - More proper architecture

Dr. Sai (赛博士) project for BESIII/HEP

Short for Dr. Science and Dr. Cyber in Chinese

- Al Agent: Al tools capable of autonomously performing complex tasks
 - LLM = brain -> Al agent = human
- Al agent based on LLM model (switchable: GPT/LLaMa/DeepSeek)
 - A demo: Xiwu model, based on Llama 3, trained with a part of BESIII internal data, e.g. memo/drafts, BOSS source code, Q-A during internal review
- One milestone: Al assistant (https://drsai.ihep.ac.cn)
 - Chatbot, MC generation, signal extraction, and a navigator inside BESIII
 - Capable of simple task
 - Ready for BESIII internal test!
- Final goal: Al scientist, it can analyze the collision data automatically
 - Developing new AI models, targeting at ~2027

Preliminary Dr. Sai



Key of this project:

make the results from AI more reliable

- Good quality data
- In-the-fly validation and test
- Multi-agents architecture

Main Agents:

• Planner: Planning and tasks decomposition

Scientist 👄

- Coder: Write BOSS code
- Tester: Using scientific tools for testing

Planner Agent Group Coder
Tester

Current training data

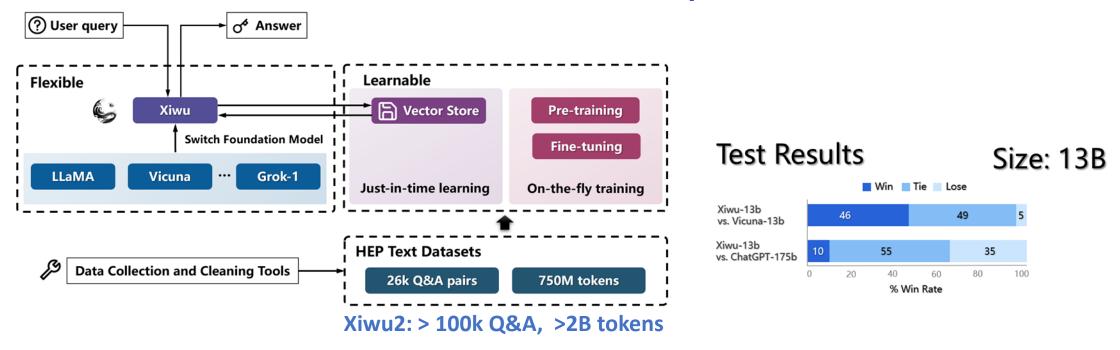
- Recent papers on arXiv
 - PDF files parser: HaiNougat, advanced iteration of the Nougat model
- Good quality chat history from IHEP-AI platform
 - The data is cleaned by human or AI (GPT4)
 - 180k Question-Answer pairs in 3 months

BESIII internal data

- Internal memo/notes, parsed by HaiNougat
- Question-Answer pairs from 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 RAG or used in training and fine-tuning

The brain of Dr. Sai - Xiwu (溪悟) 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 3 will base on DeepSeek-V3/R1



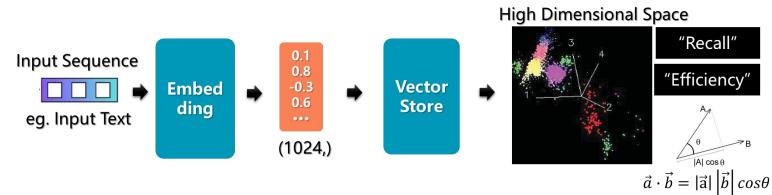
Xiwu 2.0 preliminary evaluation result

Model	MMLU_average	BoolQ	Humaneval	Hep_test
Baichuan2-7B-Chat	48.9	5.44	2.44	74.56
chatglm3-6b	51.82	72.6	61.59	71.52
vicuna-7b-v1.5-16k	33.03	45.81	17.07	62.67
Llama3-8b	58.34	83.61	21.34	76.94
Xiwu 1.0	41.66	0.8	12.2	48.06
Xiwu 2.0	58.23	83.55	19.51	78.06

- Xiwu 2.0: fine-tuned from LLaMa3-8B
 - Feasibility verified preliminarily
 - Improvements by introducing new domain data
 - Next: more smart way to represent and train the HEP data

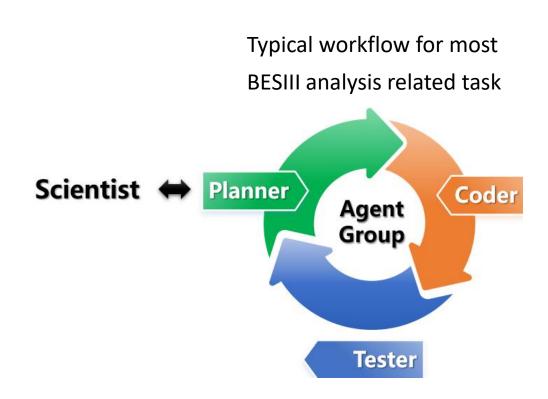
The Memory of Dr. Sai - RAG

- 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 multidimensional space
 - Usage: store BESIII internal data
 - user send BESIII related questions
 - RAG return question + BESIII internal data to LLM



Multi-agents managment system

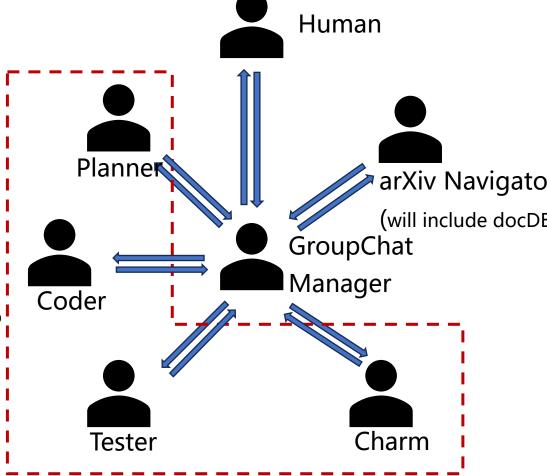
- Developed based on AutoGen framework
- Normally one agent is dedicated for one task, HEP data processing is very complicated, impossible for one agent
- Multi-Agents (foundation model is switchable):
 - GroupChatManager
 - Planner: Planning and tasks decomposition
 - Coder: Write C++/python/BOSS code
 - Tester: Using scientific tools for testing/execution
 - Charm: BESIII internal assistant
 - Common tools: arXiv navigator, 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

Groupchat for multi-agents

- 1. Human pass task to Dr. Sai by HumanProxy
- 2. GroupChatManager will select the proper speaker (agent)
 - 1. Human can correct the selection
 - The task will be decomposed into subtasks in the beginning and can be revised later
 - 3. All the messages including the chat history will be passed to the next speaker
 - 4. If one task failed, error message will send to GroupChatManager then restart the task
- If Tester need to execute a BESIII-related program, people have to set up a worker in BESIII computing environment first



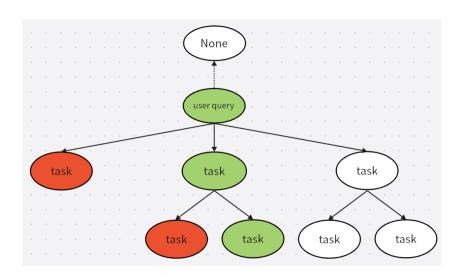
Only for BESIII

Status and prospects for Dr. Sai

- First version is Ready for BESIII members
 - webui: https://drsai.ihep.ac.cn (contact us if you can't access)
 - 3 main components
 - Preliminary BESIII Al assistant
 - automated data analysis: task decomposition, code/text generation, execution and test
 - Personal assistant
 - user can make their own vector store for personal data
 - Chatbot
 - switchable: GPT3.5/4, LLaMa3, Xiwu2 ...
 - can read PDF and image

Task decomposition

- One of the key part of Dr. Sai
 - Compose complex task into multiple small, simple, and well-defined sub-tasks
 - We will prepare more data to train the model again in future
- Sub-tasks will be executed in sequence
 - All sub-task succeed = task succeed



How to search for Zc(3900)

"Observation": "The user requests to decompose the task of searching for Zc(3900) from electron-positron collision data at a center-of-mass energy of 4.26 GeV.",

"Thoughts": "To decompose the task, we need to break down the process into manageable steps, including data collection, event selection, background suppression, signal extraction, and analysis.",

"Status": "CONTINUE",

"Plan": [

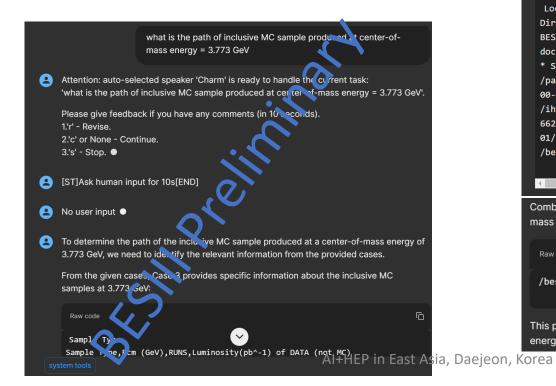
- "(1) Data Collection: Gather the electron-positron collision data at a center-of-mass energy of 4.26 GeV from the BESIII experiment.",
- "(2) Event Selection: Identify and select events that are likely to contain the Zc(3900) signal. This typically involves selecting events with the first state particles of interest, such as $\pi^{\pm}J/\psi$.",
- "(3) Background Suppression: Apply criteria to reduce oackground events that can mimic the signal. This may involve using kinematic cuts, particle identification, and other techniques.",
- "(4) Signal Extraction: Use invariant mass calculations and other methods to extract the Zc(3900) signal from the selected even s. This involves plotting the invariant mass spectrum of the $\pi^\pm J/\psi$ system and identifying peaks corresponding to Zc(3900).",
- "(5) Statistical Analysis: Perform statistical analysis to determine the significance of the observed signal. This includes fitting the invariant mass spectrum and estimating the signal yield and background.",
- "(6) Systematic Uncertainties. Evaluate and account for systematic uncertainties in the measurement. This may involve varying selection criteria, fitting methods, and other parameters.",
- "(7) Results and Interpretation: Interpret the results in the context of theoretical models and previous measurements. Compare the observed properties of Zc(3900) with theoretical predictions and other experimental results.",
- "(8) Documentation and Reporting: Document the analysis process, results, and conclusions. Prepare a report or publication to share the findings with the scientific community."

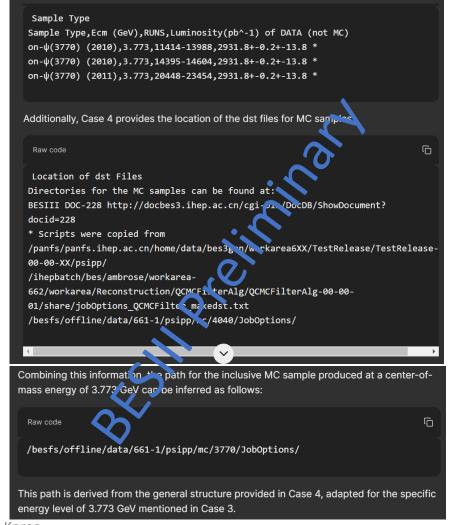
],

"Comment": "The task involves a detailed and systematic approach to search for the Zc(3900) particle in electron-positron collision data. Each step is crucial to ensure the

Example: BESIII Internal Navigator

Dr. Sai can assist people to find the correct information, the prompt is "what is the path of inclusive MC sample produced at center-of-mass energy = 3.773 GeV. ".





Towards Dr. Sai V2/V3

- V2 (end of 2025)
 - Better knowledge representation: vector store -> knowledge graph
 - Better agents interaction logic
 - Modeling of experimental analysis procedure:
 - Reorganize the analysis framework to reduce the amount the code/text to be generated
 - Preliminary DSL (domain specific languages) for HEP analysis
 - Comprehensive evaluation system
- V3 (end of 2026)
 - Extract chain-of-thoughts (CoT) from all HEP papers
 - Extract the analysis procedures from all experimental papers to DSL
 - Align the DSL with domain code and CoTs, and perform reinforcement learning based on DSL, similar to deepseek-r1, and get the feedbacks from human experts

Lots of on-going activities, stay tune

An extremely simplified example of CoT

Problem: We're looking for a hypothetical "dark photon" (A'), a particle that interacts with dark matter. One way to search for it is through its decay into an electron-positron pair (e+e-). We expect the dark photon to be relatively light.

Chain-of-Thought:

- 1. Theoretical Prediction: The dark photon (A') is predicted to decay into an e+e- pair.
- Data Collection: We analyze data collected by the ATLAS detector. We look for events with two energetic electrons (e+ and e-).
- Reconstruction: We measure the momentum and energy of the e+ and e-. We calculate the
 invariant mass of the e+e- pair. If the dark photon exists, the invariant mass of the e+e- pair
 should cluster around the mass of the A'.
- Observation: We see a small "bump" or excess of events in the invariant mass distribution of e+e- at a specific low mass value.
- 5. Inference: The "bump" might suggest that a particle with that mass is decaying into e+e-.
- 6. Simplest Decision: If the mass of the "bump" is consistent with the predicted mass range for the dark photon, we say we *might* have seen evidence for it.

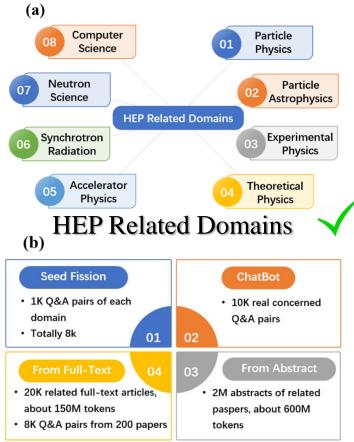
Summary

- LLM could be useful for HEP
 - Not just generate code/text preliminarily, but can be used to automate the analysis
- A demo: Al agents system for HEP Dr. Sai
 - It can help people on data analysis (good accuracy for small task)
 - Many on-going studies to improve accuracy especially for complex task
 - Easy to port to other lepton collider experiments
 - Bottleneck: lack of "data"
 - Modeling of experimental analysis procedure and knowledge
 - Personpowers are more than welcome
 - Next Dr. Sai V2: more data, state-of-art AI technologies
- IHEP formed a ML Collaboration to work on AI4HEP
 - Lots of on-going activities, more details in the <u>kick-off meeting</u>
 - Welcome to join/collaborate/discuss AI for HEP! Contact me (like@ihep.ac.cn)

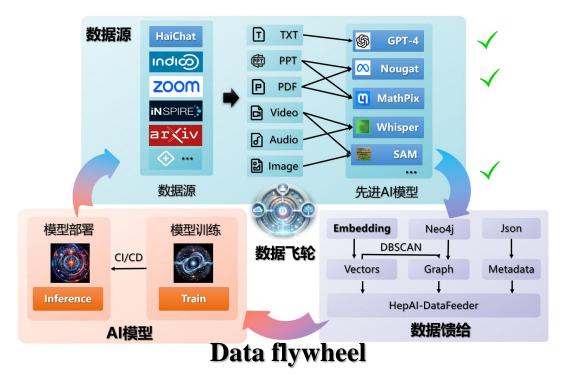
back-up

Sensors of Dr. Sai

Data flywheel enables continuous iterative evolution of models





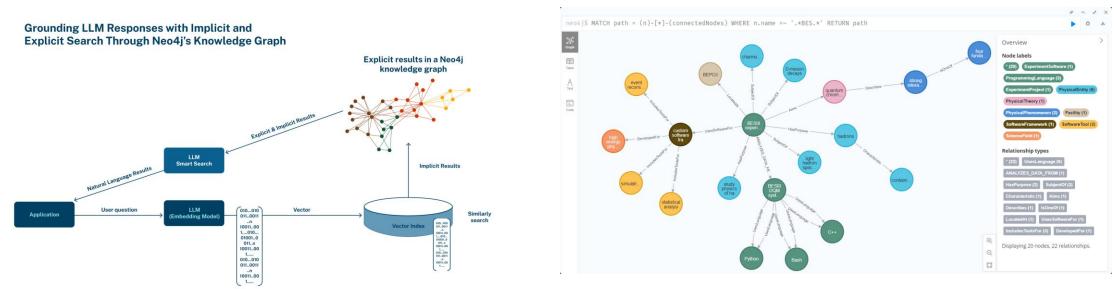


The "data-driven flywheel effect" improves the algorithm model by constructing a circular data path, attracting more users and generating more data during their use. The new data is then used to improve the algorithm model, forming a positive feedback loop.

Memory of Dr. Sai: knowledge graph

Association relationship retrieval based on knowledge graph

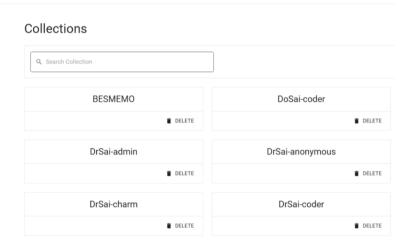
- Building BESIII Knowledge Graph:
 - Entity Relationship Construction → Eliminating Ambiguity → Adding Attributes and Labels → Storage
- Search:
 - User's question → Question Embedding → Search for Top K nodes → Obtain node information
- Progress: Built a BESIII knowledge graph demo based on 500 Q&A pairs

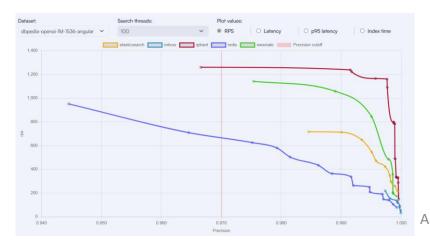


Next step: Test knowledge retrieval and inference performance, develop KG's API to serve Dr Sai

Memory of Dr. Sai

• Hepai worker+llama index+qdrant



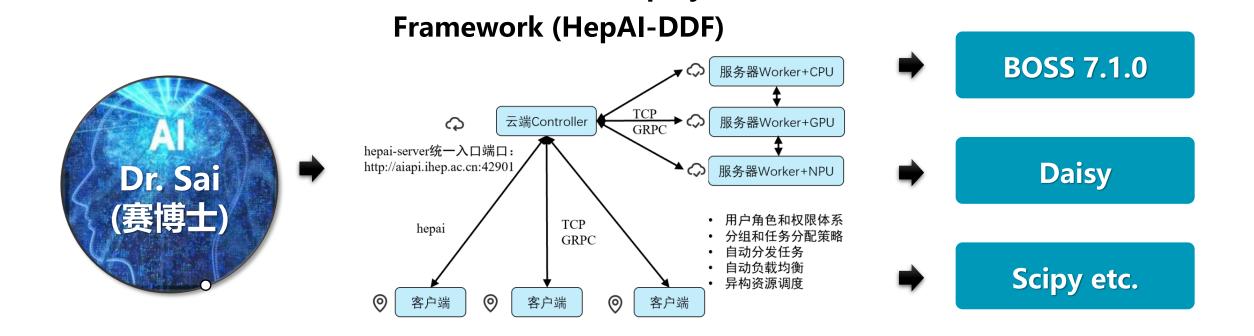




Actuators of Dr. Sai

Based on HepAI DDF, it is easy to expand Dr Sai's actuator component.

The Distributed Deployment



Example: fitting

Prompt:

"I want to do two things:

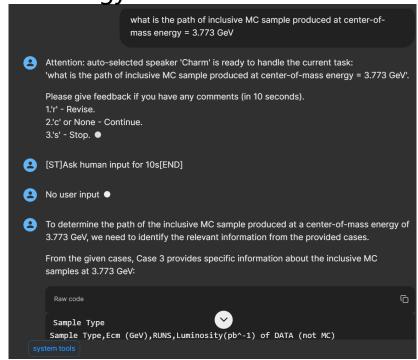
1. generate a fitting script to fit a Gaussian distribution from the output file 'fake_data.root'. There is a TH1F name 'h_gaus' in the file.

2. use Tester to execute this code."

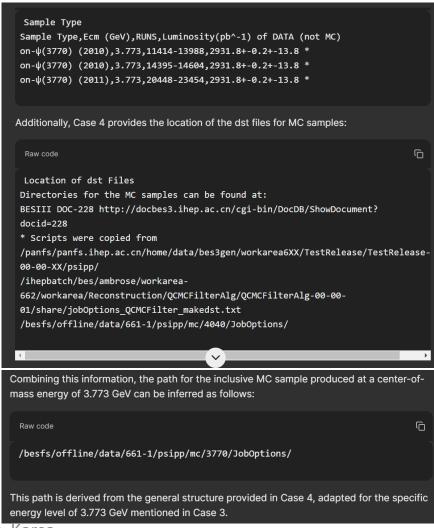
Example: BESIII assistant

Prompt:

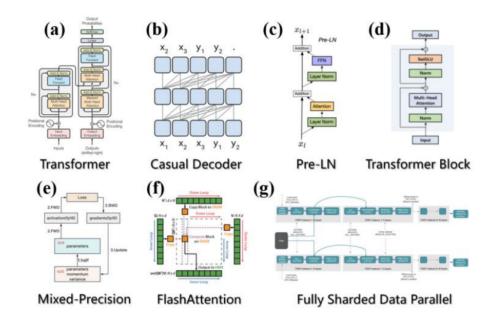
"what is the path of inclusive MC sample produced at center-of-mass energy = 3.773 GeV."



Please refer to Pan's talk for more details



Xiwu 2.0: Training techniques



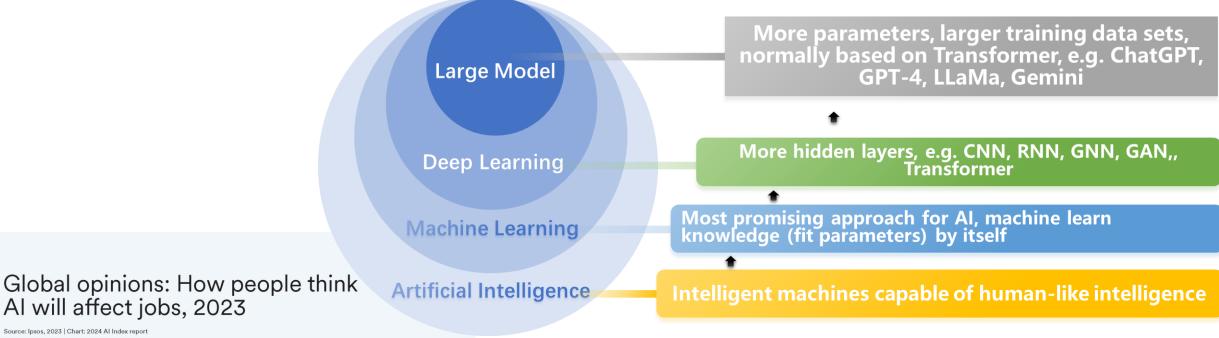
Reduce computational cost and training time, while improving model's performance

Frequently used techniques in LLM training

- Mixed-precision Training reduce memory usage and improve computational throughput without sacrificing model accuracy
- FlashAttention
 enhance floating-point operations per second (Flops)
- Fully Sharded Data Parallelism (FSDP)
 parallel training across multiple devices
- offload certain computations from GPU to the CPU
- Torchrun Parallelism
 Training on single node with Multi-GPU
- Bitsandbytes
 vector quantization and mixed-precision decomposition
- Deepspeed utilize mixed precision, gradient checkpointing, and offloading memory to the host CPU
- ZeRO Series (ZeRO-2, ZeRO-3, ZeRO-offload)
 memory optimization strategies, such as activation
 checkpointing and optimizer state partitioning
- Low-Rank Adaptive (LoRA)

Approximate the weight matrices with low-rank factors while preserving the model's representational capacity 7

Machine learning and Al



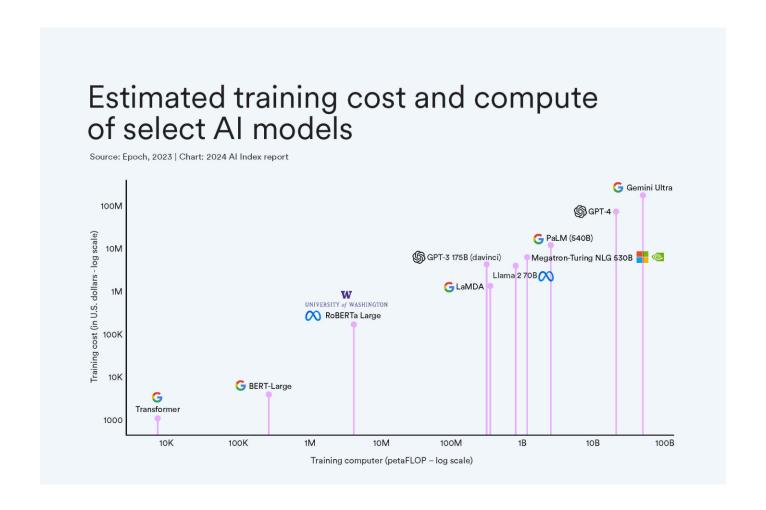
Al will replace your current job in the next 5 years

Al will replace your current job in the next 5 years

Al will replace your current job in the next 5 years

Al will replace your current job in the next 5 years

the next industrial revolution



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 well as outputting results.

This is a C++ code for a data analysis algorithm called Gam4pikp. The algorithm is designed to analyze data from high-energy particle collisions and identify specific patterns 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 calculations 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 study the properties of these particles.

Simple test: Dr. Sai vs GPT4

As expected, better performance in HEP

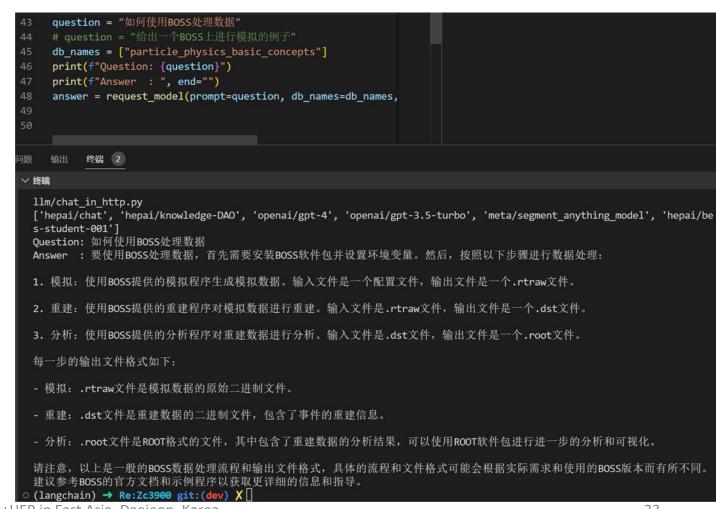
Q: 如何使用BOSS处理数据?

Answer from Dr.

Sai

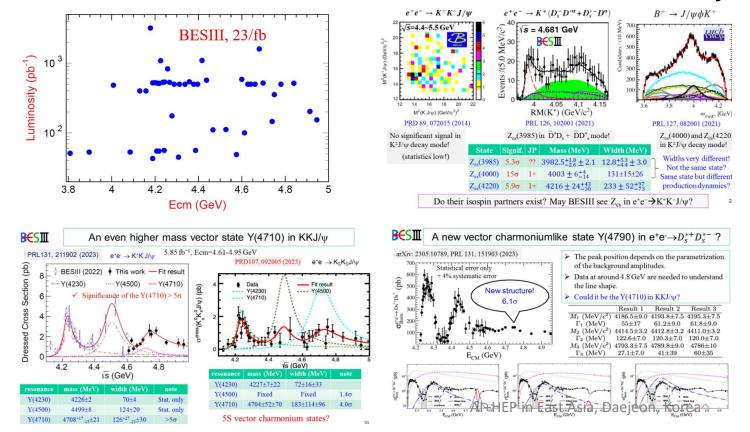
Answer from GPT-4





Status and prospects for Dr. Sai

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



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

Status and prospects for Dr. Sai

XYZ studies at BESIII:

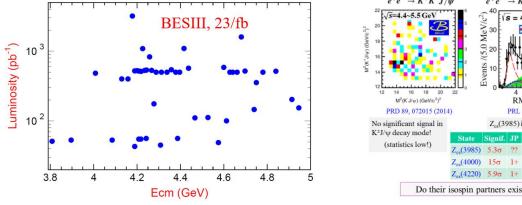
An even higher mass vector state Y(4710) in KKJ/ψ

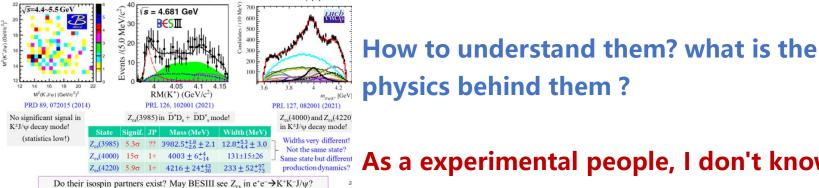
PRL131, 211902 (2023) e⁺e⁻ → K⁺K⁻J/ψ

Significance of the Y(4710) > 50

124±20

- Many (intermediate) resonants, many decay channels -> interferences
- Difficult to retrieve useful information correctly





As a experimental people, I don't know.

physics behind them?

A new vector charmoniumlike state Y(4790) in $e^+e^- \rightarrow D_s^{*+}D_s^{*-}$? > 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? 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±3.2 122.6±7.0 120.3±7.0 120.0±7.0 M_2 (MeV/ c^2) 4793.3+7.5 4789.8+9.0 4786+10 E_{CM} (GeV) Γ₃ (MeV) 27.1±7.0 41±39

But the cross section measurements of

ALL channels should be one necessary condition.

how many papers published from BESIII

The number of papers published by the BESIII collaboration can vary over time as new research is conducted and published. As of my knowledge cutoff date in October 2023, I don't have the exact count of papers published by BESIII.

To get the most up-to-date and accurate information, you can:

- 1. **Visit the BESIII Collaboration Website**: The official website of the BESIII collaboration may have a list of their publications or a link to their recent papers.
- Check Scientific Databases: Use databases like the INSPIRE HEP (High Energy Physics) database, the arXiv preprint service, or journal publishers' websites to search for papers published by the BESIII collaboration.
- 3. **Contact the Collaboration Directly**: You can reach out to the BESIII collaboration or one of its representatives for detailed information on their publications.

These methods should provide the latest count and specifics of the academic papers published by the BESIII collaboration.

tion has published over 1,200 papers in peer-reviewed dings. This number is constantly growing as new data is

iment at the Beijing Electron Positron Collider (BEPC) in dying the properties of exotic hadrons and other aspects ation is one of the largest in the world, with over 1,000 itutions in 20 countries.

cover a wide range of topics, including the properties of tates, the search for new physics beyond the Standard c reactions and processes.

Summary:

- · Production Channel:
 - e+e- collisions with ISR: (e^+ e^- \rightarrow \gamma \pi^\pm X(3872))
- Decay Channels:
 - Predominantly hadronic decays like (X^\pm(3872) \rightarrow \pi^\pm J/\psi)
 - Radiative decays like (X^\pm(3872) \rightarrow \gamma \pi^\pm J/\psi)
 - Multi-body decays where the final state includes charmonium.

By focusing on these channels, you can efficiently search for the charged partner of X(3872) at BESIII.

Data preparation: HaiNougat

- During the data clean for Xiwu LLM system, we found it is hard to obtain correct information from PDF, especially for Math equations and Tables
- For HEP-related papers, 20% more formulas per pages and 27% longer formulas than typical academic documents (computer science, chemics ...)
- Develop a model for HEP (HaiNougat) from the Visual Transformer model Nougat. Also an important part for dataset preparation from Xiwu LLM system!
- A demo accessible via https://ai.ihep.ac.cn/m/hai-nougat
- More details in https://github.com/ai4hep/hai-nougat

Data preparation: HaiNougat

