



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

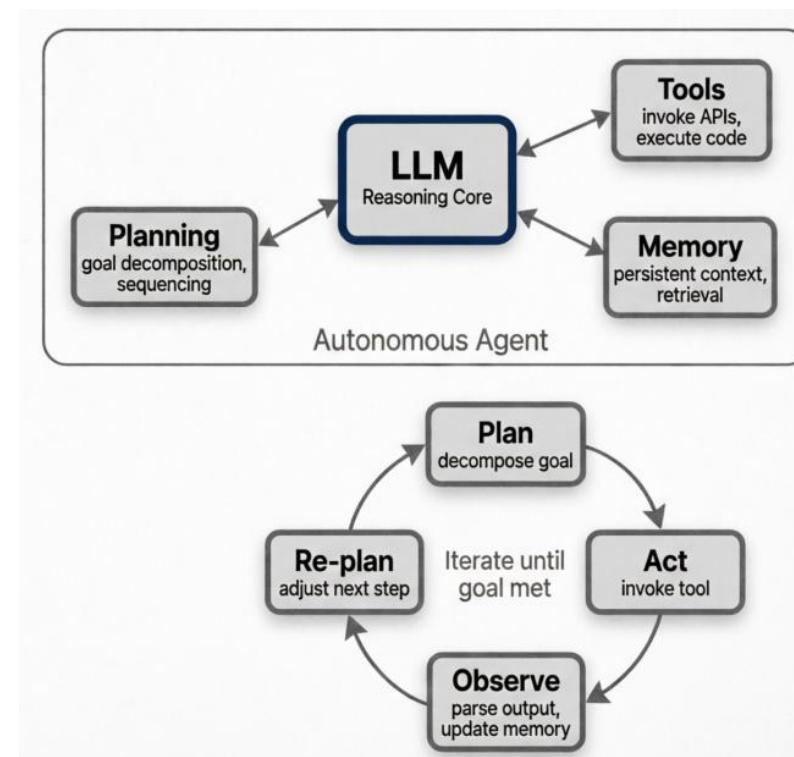
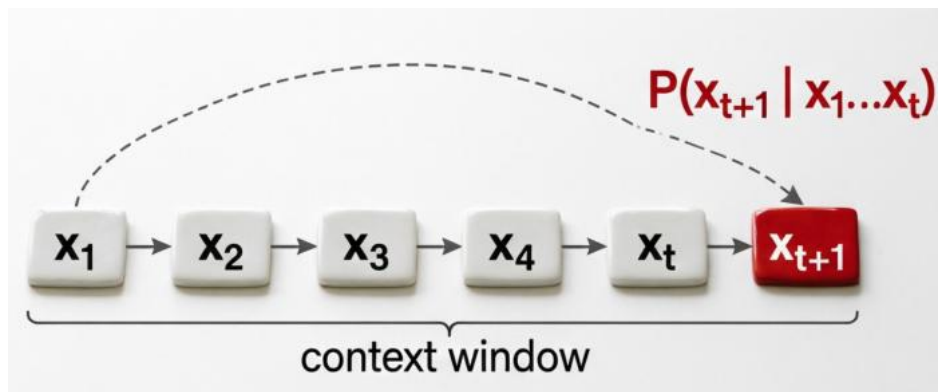
AI partner at BESIII: DrSai

Part I: Introduction – ML and LLMs&Agents

Part II: Construction of DrSai, application at BESIII

Ke Li (like@ihep.ac.cn)

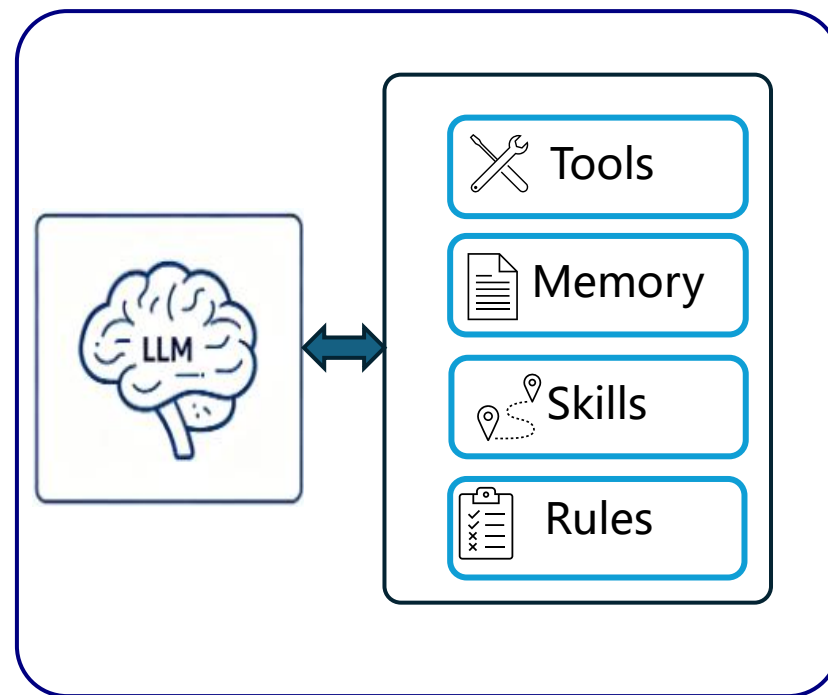
Recap: LLM&Agent



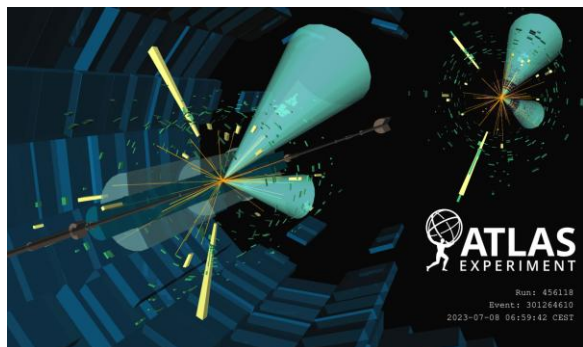
Outline

- Goal-driven AI Scientists
- The DrSai project
- IHEP AI infrastructure
- Status and plan
- Summary and outlook

Agent



新粒子发现的重要指标



× (>数百亿事例)



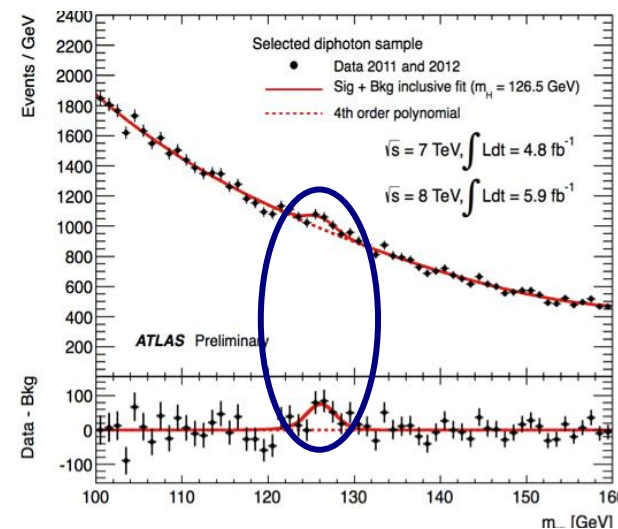
数据分析

$$Z \approx \frac{S}{\sqrt{B}}$$

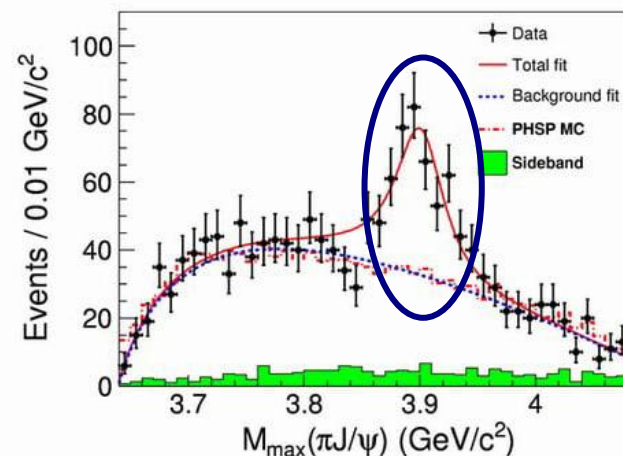
新粒子显著性

AI赋能方式:

- 提升S
- 降低B
- 加速数据分析

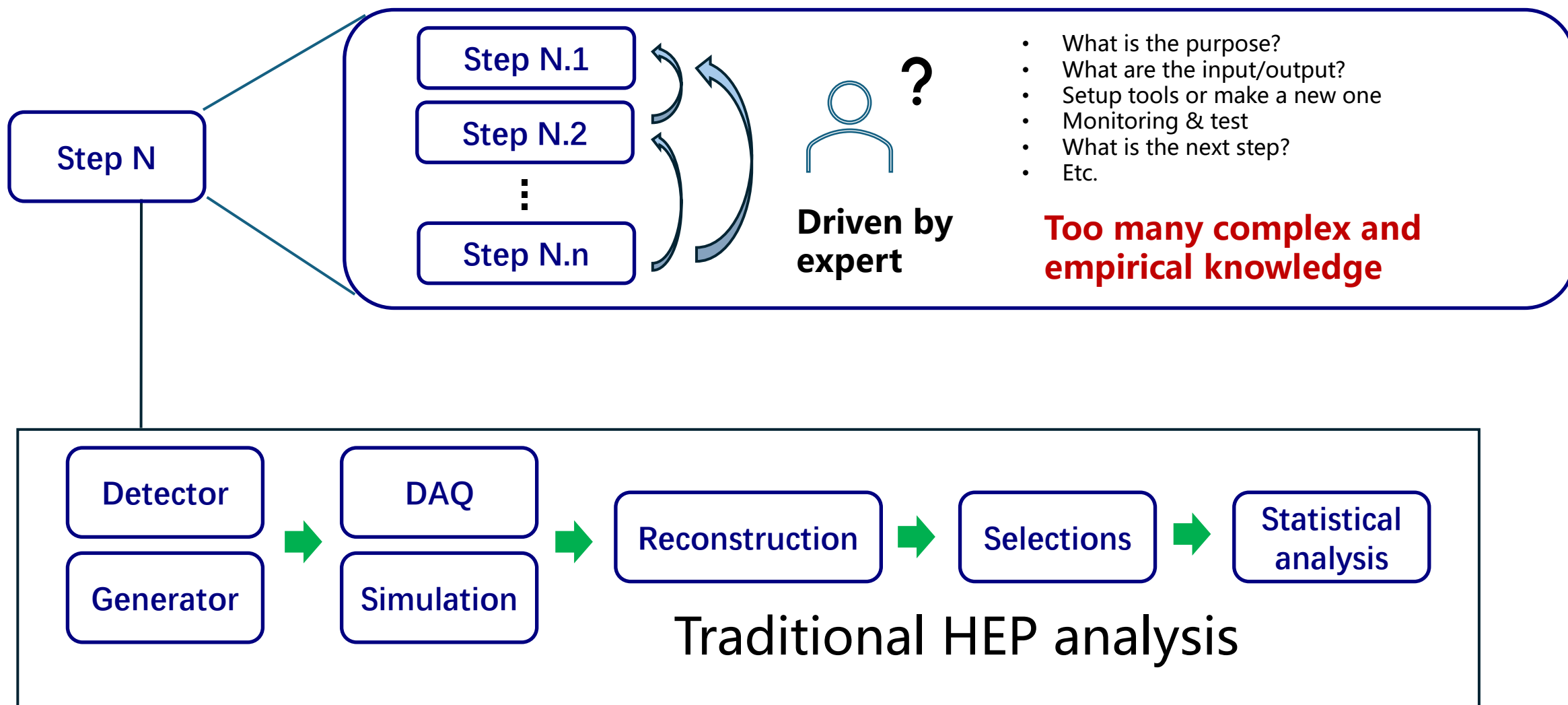


新粒子: Higgs

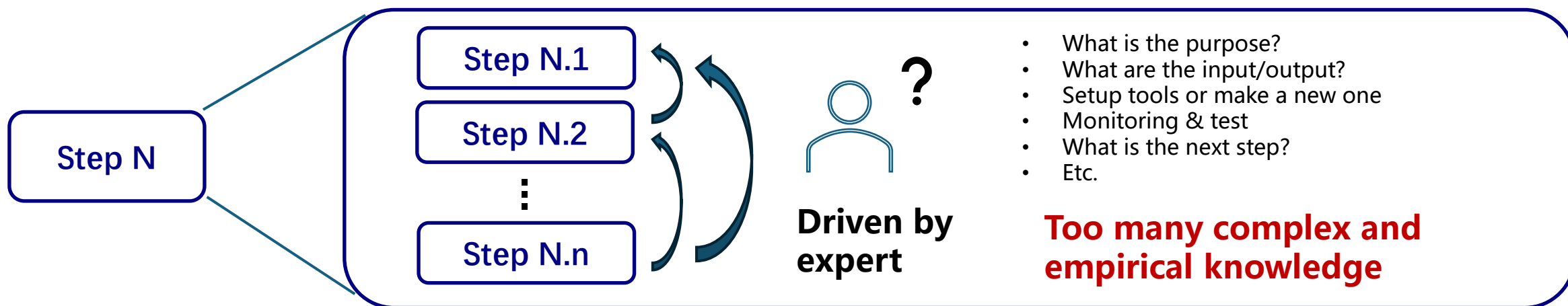


新粒子: Zc(3900)

Towards Agentic AI for HEP



Towards Agentic AI for HEP



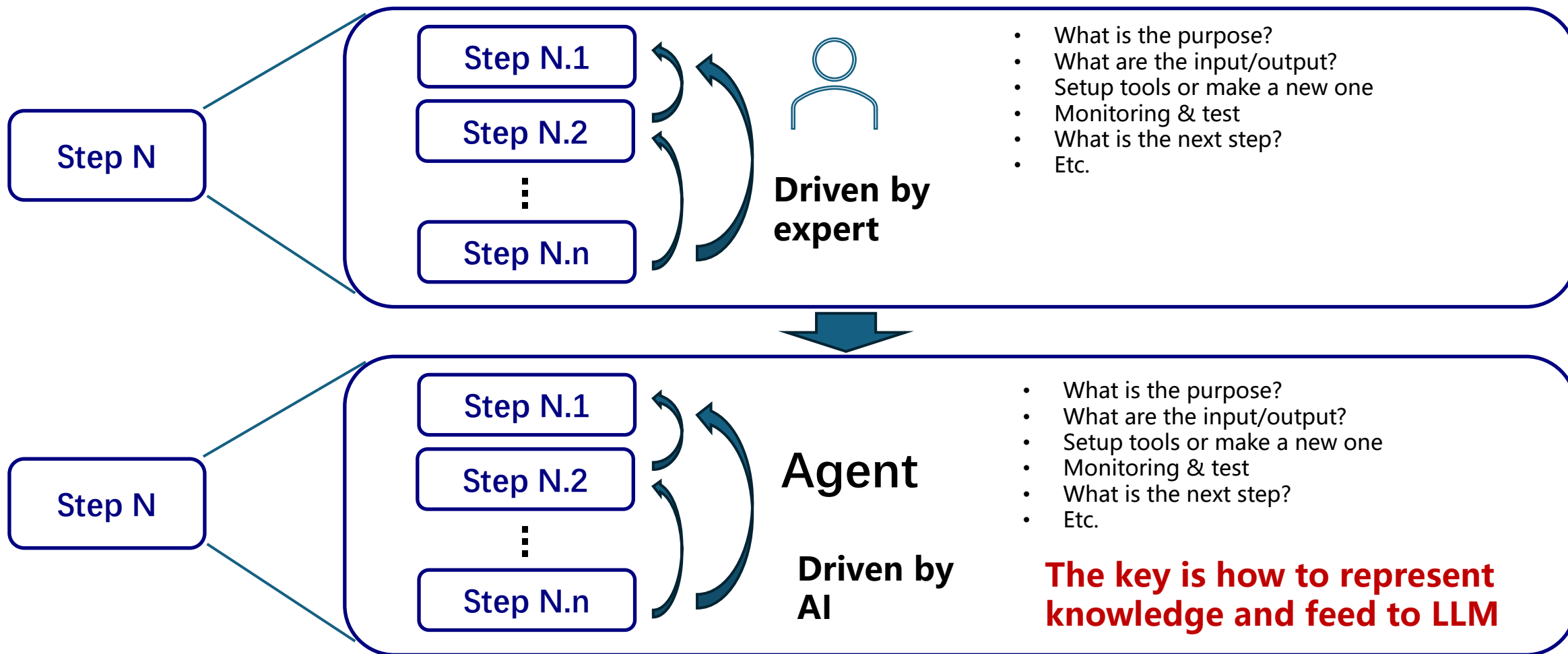
- Expertise is often unstructured, though some can be represented by natural language.
- Today, we can leverage LLMs to build an AI partner for scientists.

The AI partner designs and conducts the entire workflow based on human commands and feedback.

Already completed

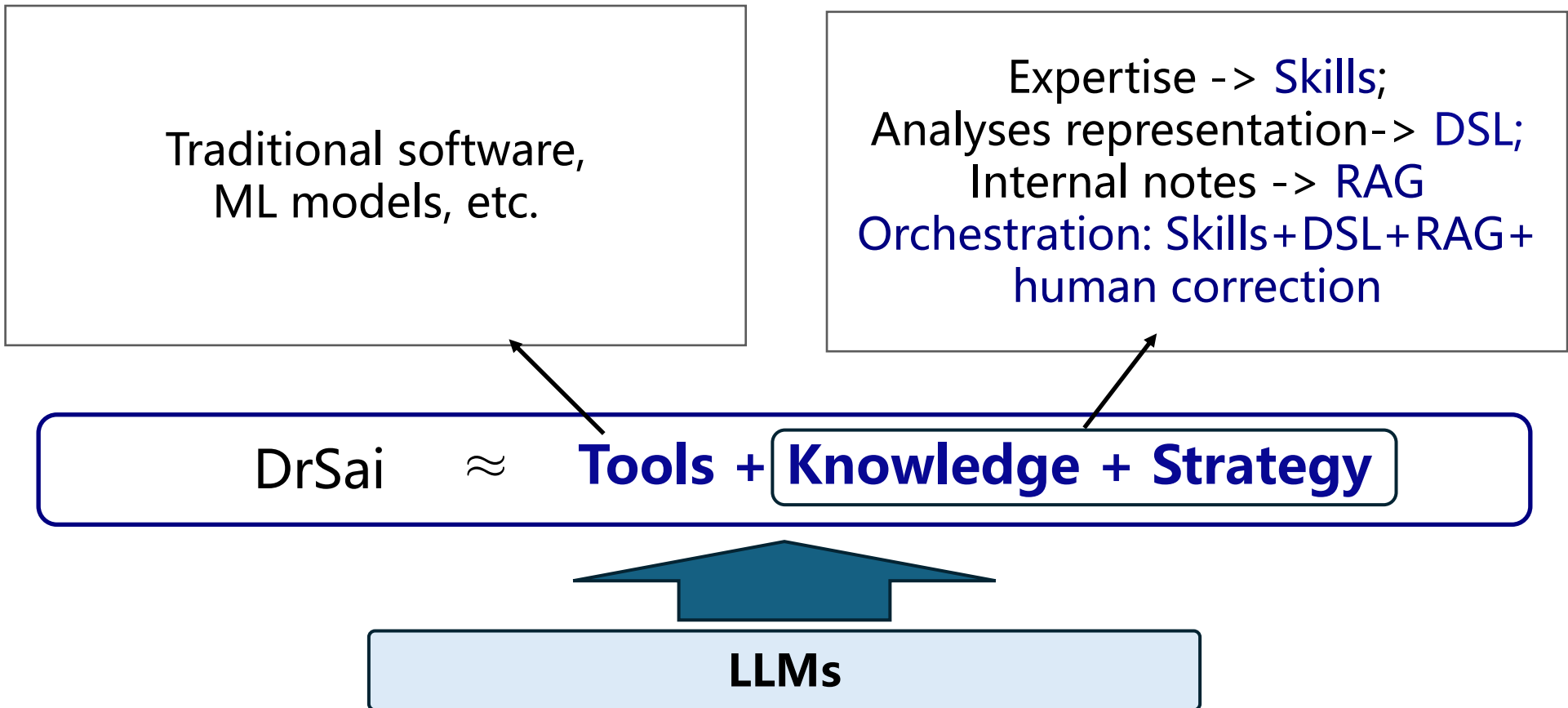
- In the future, with a well-defined representation of knowledge, we can further develop an **AI scientist**.
 - **Humans would only need to propose a goal — and leave everything else to the AI.**

Towards Agentic AI for HEP

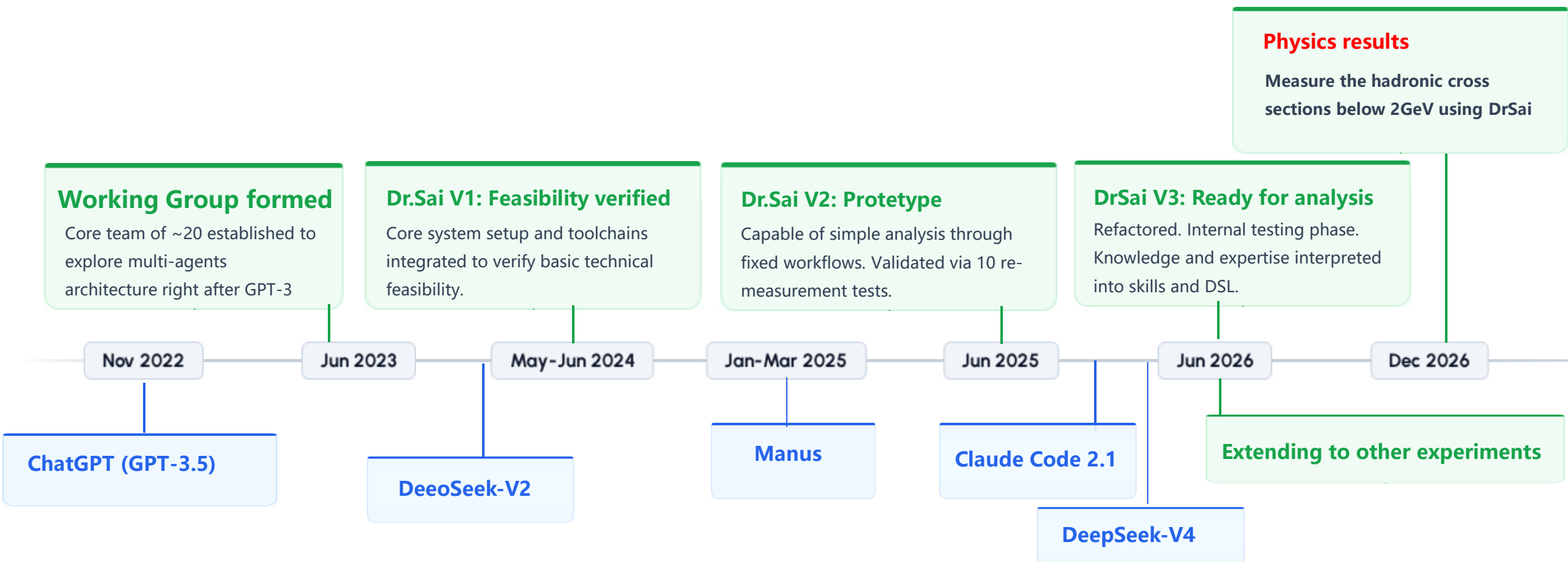


The DrSai project

Purpose: Pioneering autonomous physics discovery at **BESIII (and beyond)** through Agentic AI

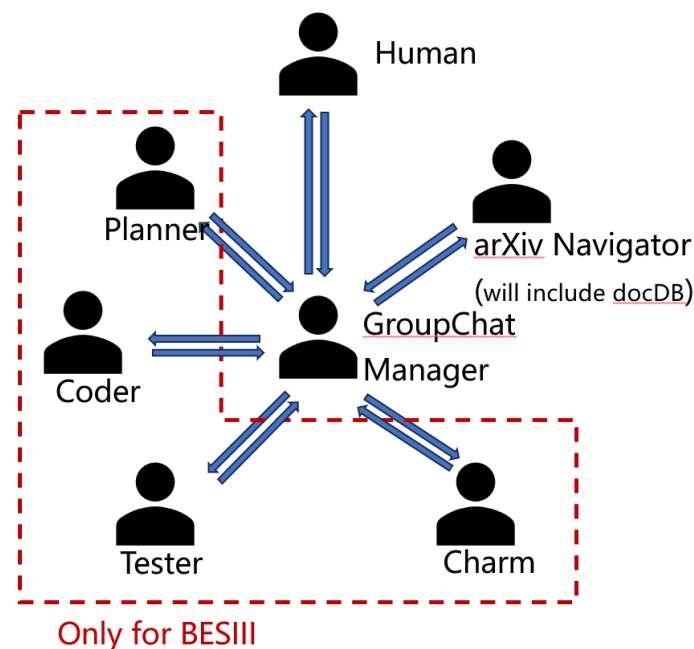
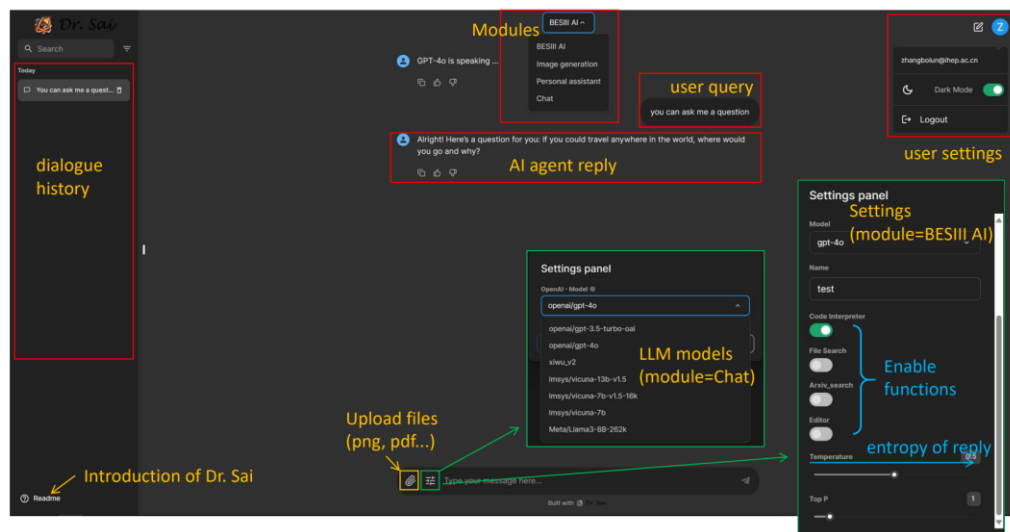


DrSai Evolution Milestones



DrSai V1

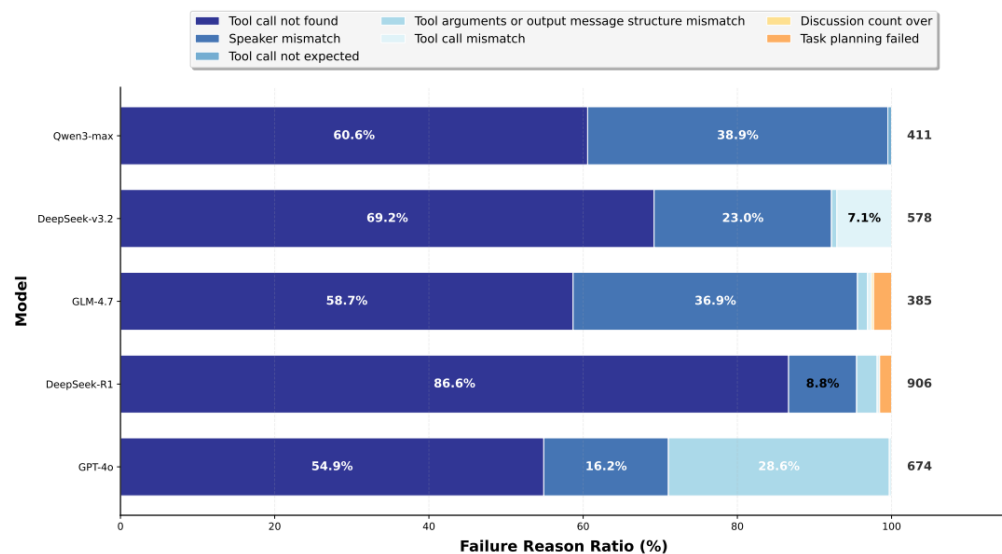
- Developed based on open-source multi-agents framework: AutoGen.
- Deploy workers to wrap-up the tools
- Tested LLMs on intent understanding, orchestration, tool invocation, etc.
- **Verified feasibility**



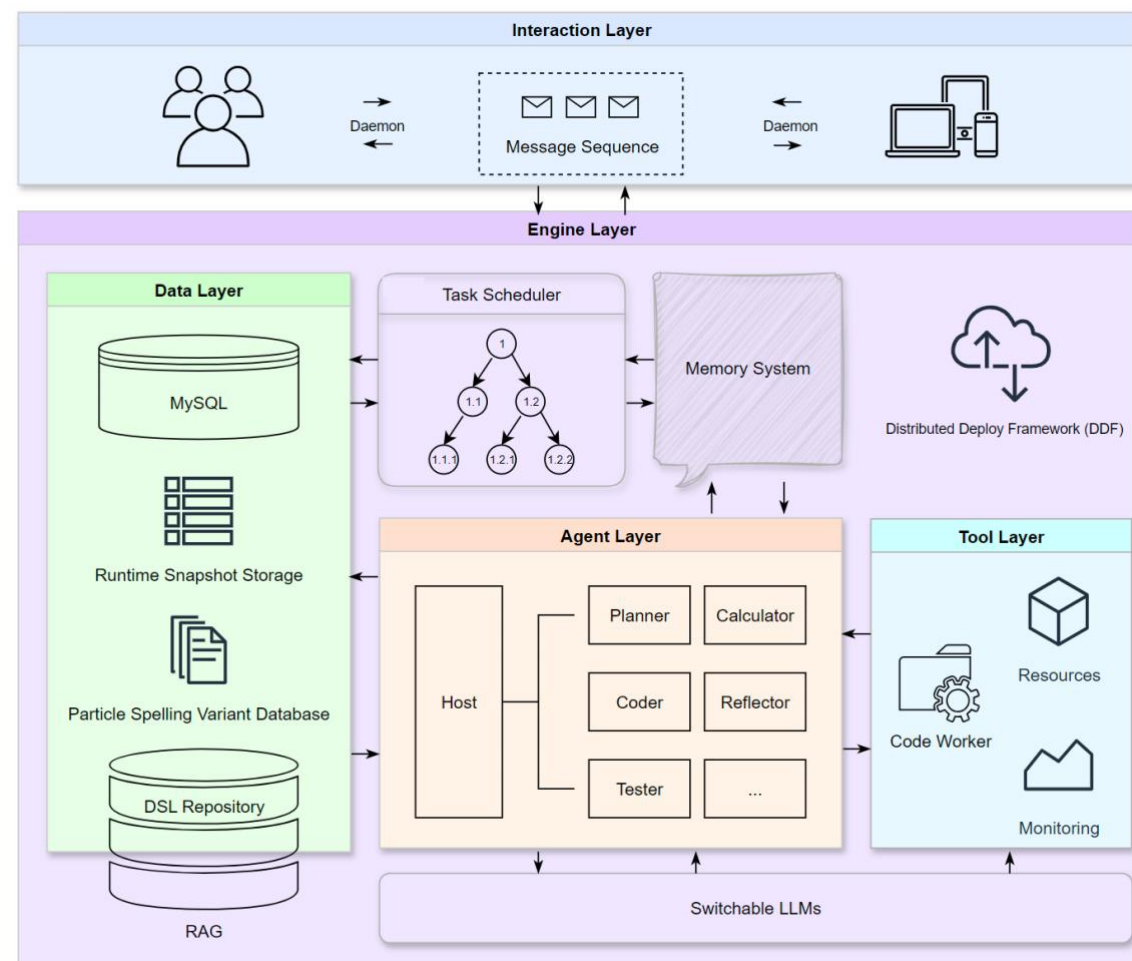
More details in [the talk at ICHEP2024](#)

DrSai V2

- Make BESIII analysis agent-based
- Analysis represented via DSL (preliminary)
- Built a benchmark for BESIII analysis

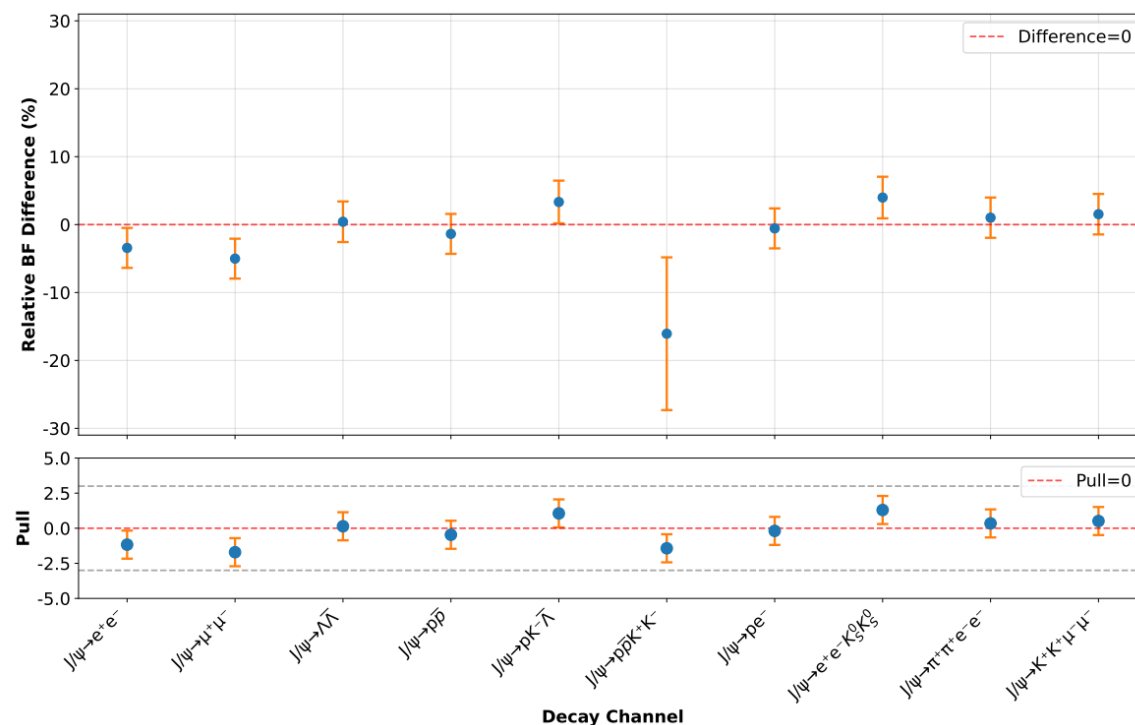


Tested several LLMs



DrSai V2

- The whole system was **validated by re-measuring** known branching fractions (BFs) of the J/ψ decays !
- Built a sample using MC with all known decays
- Use $\psi(3686) \rightarrow \pi^+ \pi^- J/\psi$ process to measure the BFs of J/ψ
- DrSai V2 performed the whole analysis
 - From simulation to preliminary systematics study
- All the measured BFs showed **good consistency** with injected values
- First full chain validation of DrSai on a realistic physics analysis pipeline



Identified Core Challenges

Knowledge & Self-Evolution

Often unstructured (experience, intuition). How to make it computable, reasonable, and continuously growing?

Solution

Domain-Specific-Language (DSL) represent analysis with an on-the-fly extraction and iterative refinement from expert feedback

Reasoning & Task planning

High level goals need complex and hallucinations-free decomposition + environment-aware reflections

Solution

Expert guides the system to orchestrate workflows for multiple analyses → distills into Skills → next workflow iteratively refined

Tools/codes & Execution

Complex software system, and must align with physics goals, as well as auto-checking

Solution

Error-Aware GraphRAG – stores long connections. DSL/Skill align with code blocks or tools, enabling multiple checkpoints (e.g., injection tests).

Infrastructures & Safety

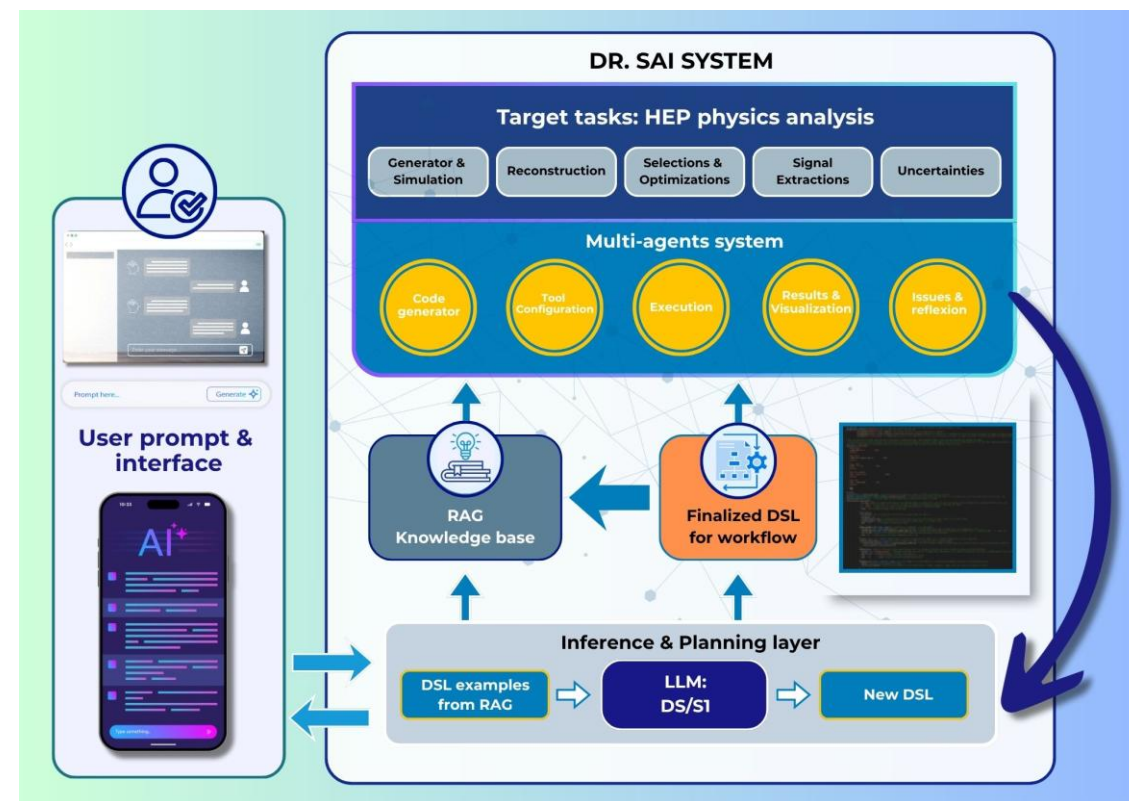
Scaling bottlenecks (GPU costs) vs. data privacy and autonomous safety risks

Solution

Local deployment of open-source LLM (DeepSeek) at IHEP + Cloud support.
Only allow predefined operations.

DrSai V3

- With DrSai V2, we have identified four core challenges and developed corresponding solutions
- Major updates
 - AutoGen-based framework -> Claude Code agent runtime
 - Also DeepSeek-TUI, Cline, etc.
 - A set of analysis Skills – built and refined
 - Hundreds of DSLs extracted from published papers
 - Error-aware GraphRAG for code generation
 - Agentic RAG for internal notes of previous analyses



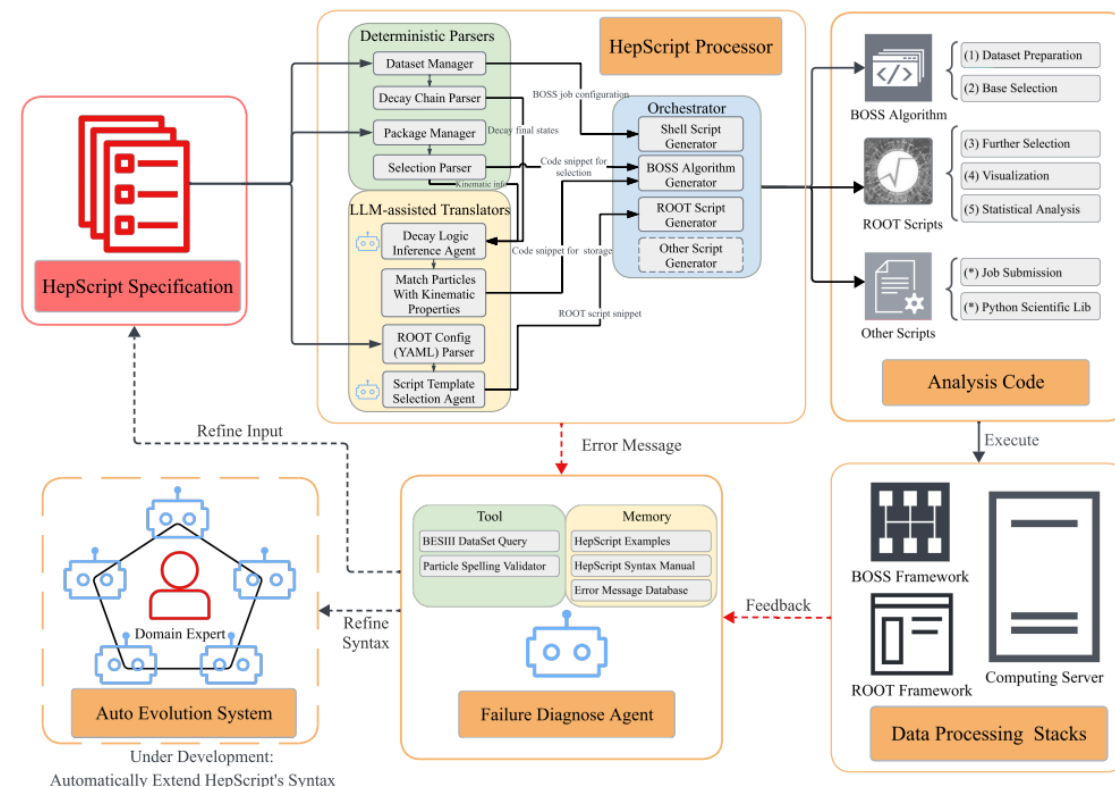
[Front. Phys., 2026, 21\(9\): 096401](#)

Analysis representation: HepScript (a DSL)

- Represent BESIII analysis
 - Pseudocode in Ruby/YAML format
- Bridges high-level intent with code/tools
- Defines a constrained action space
- The majority are mapped into code blocks
 - Only a small part need conversion by LLM
- **Scalable:** automatically extracted from published paper or generated on-the-fly for new analysis

HepScript	LLM	N_S	N_F	SR (%)	One Retry SR (%)	Three Retries SR (%)
LLM-generated (72 Packages)	Deepseek-R1	34	38	47.3 ± 5.8	87.8 ± 3.8	94.6 ± 2.6
	GLM-4.7	31	41	43.2 ± 5.7	90.5 ± 3.4	95.9 ± 2.3

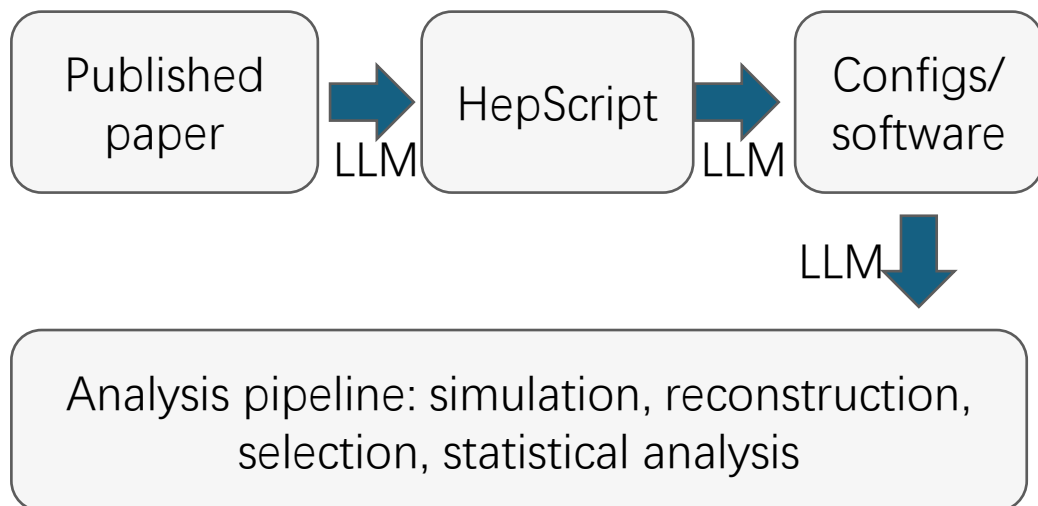
Good accuracy for generation of analysis algorithms



arXiv:2605.01423

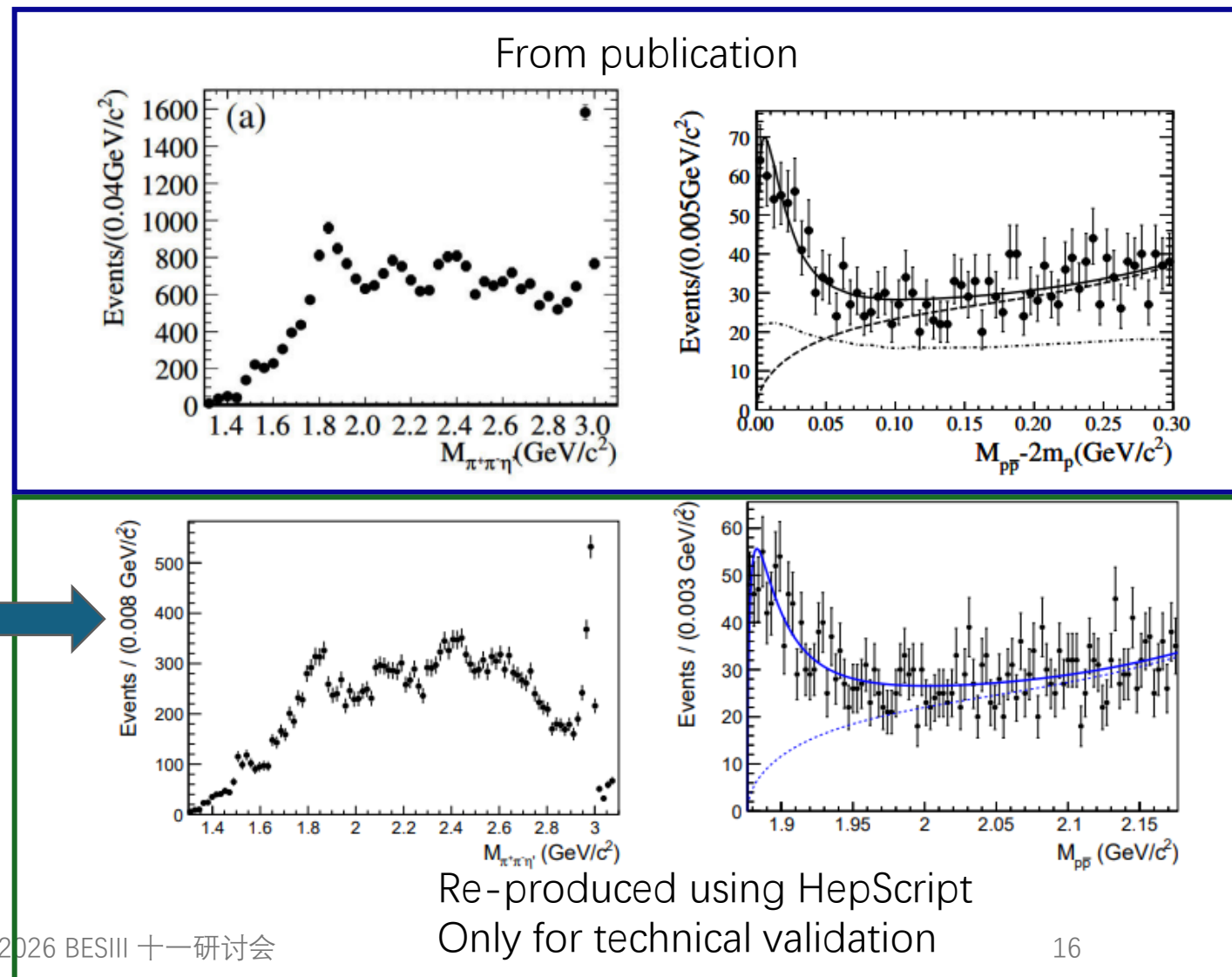
HepScript-driven analysis reproduction

- Necessary for real-world complex analysis
- A showcase: re-produce the previous analysis



- Used in the DrSai V3 for intent understanding, information completion, and code generation

2026/10/06



16

Now: reasoning for systematic uncertainty

- With RAG integrated, the system can accurately deduce most sources of systematic error and recommend suitable evaluation approaches.
- Now doing more test

Skill vs. Paper (Table II, arXiv:1210.2831)

Source	Σ (paper)	Ξ (paper)	Skill status	Note
Tracking+PID	2.8%	4.8%	✓ correct	Correctly decomposed into 4 sub-sources
Photon PR	5.0%	1.0%	✓ correct	Paper counts all 5γ (Σ); skill correctly uses 1 non- $\pi^0 \gamma$
π^0 selection	1.0%	—	✓ identified	Cat-1 formula gives ~2%; paper uses control sample
Mass windows (Σ, Λ, Ξ)	0.6%	0.3+0.3%	✓ identified	Cat-2, needs ROOT files
Vertex fits	—	0.5%	✓ correct	2 separate rows: Λ VF + Ξ VF (Ξ mode only)
Kinematic fits	4.3%	3.8%	✓ identified	Cat-2, needs ROOT files
Signal fitting	4.7%	6.4%	✓ identified	4 sub-sources: signal shape, bkg, range, η_c params
Event generators	0.4%	2.8%	✓ identified	mc_angular + mc_jpsi_decay_model (Cat-4)
Peaking background	0.3%	1.3%	✓ identified	RAG-inferred; new trigger rule added
N_J/ψ	1.2%	1.2%	✓ exact match	Paper value used; RAG default is 0.56% (updated)
Intermediate BFs	23.5%	23.6%	✓ standalone	Correctly flagged as 3rd PDG uncertainty

Five-Step Pipeline

Step 0 Parse input

HepScript .rb (Mode A) or NL description (Mode B) → Component Summary

Step 1 Formula-first reasoning

Write formula → trace variables → enumerate operations (Pass 1) → RAG audit (Pass 2) → syst_reasoning.md

Steps 2–3 Manifest + checkpoint

YAML records every uncertainty & its estimation method · Analyst reviews & approves before sub-agents proceed

Step 4 Directly cited lookup

Sub-agents look up values of directly cited uncertainties from the RAG database

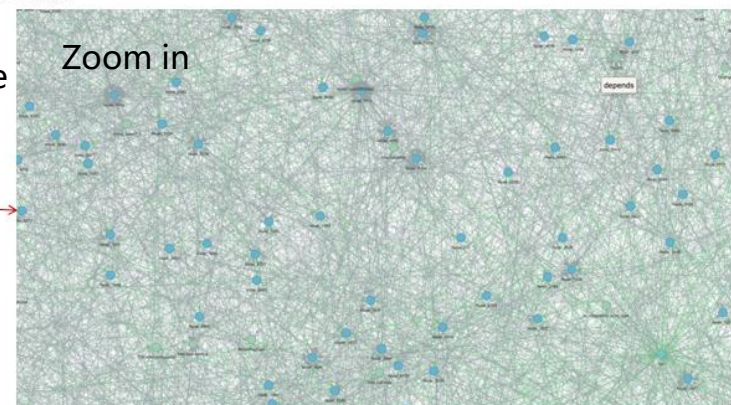
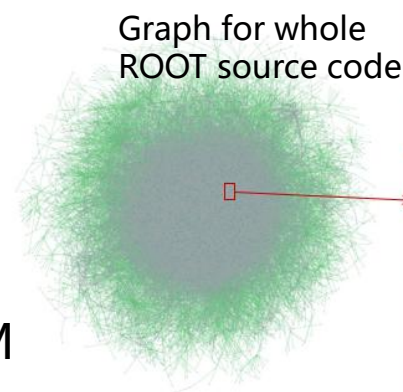
Step 5 Sub-agent evaluation

Generate ROOT macros for estimating the remaining (Cat-2–4) sources → uncertainty table

Domain code/script generation

- Complex code dependencies at HEP
 - LLMs struggle to reason over large-scale, intricate dependency correctly and efficiently
- We start with open-source software - ROOT
 - **Massive interconnections -> a large graph**
 - Nodes : class, package, etc.
 - Edges : contains, extends, depends
 - Obvious improvement on accuracy
 - Further improvement from **Error-Aware RAG**
 - First find the related subGraph, then feed LLM
- This method can be easily applied to HEP software development

- **Total nodes: 47851**
- **Total edges: 163795**
- REPO nodes: 1 | - PACKAGE nodes: 2609 | - FILE nodes: 19347
- TEXTFILE nodes: 5255 | - CLASS nodes: 20639

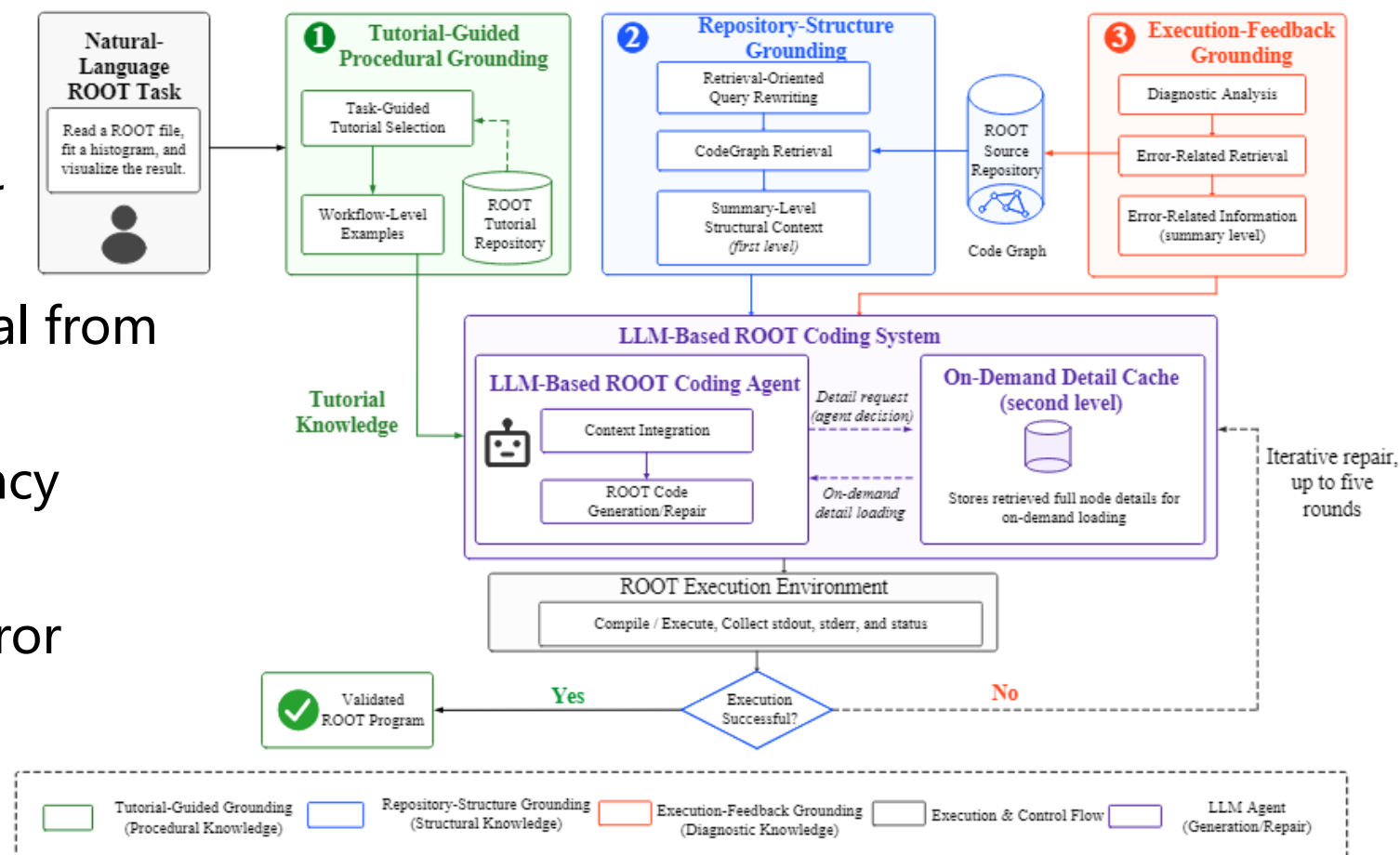


Github Repo:
<https://github.com/yues36067-cmd/ROOTER>

Agentic generation

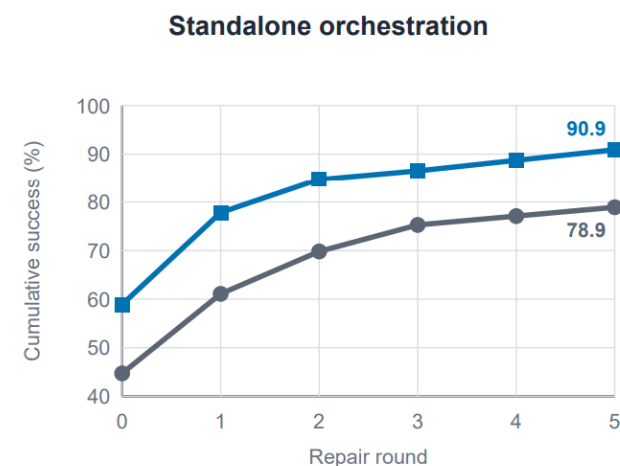
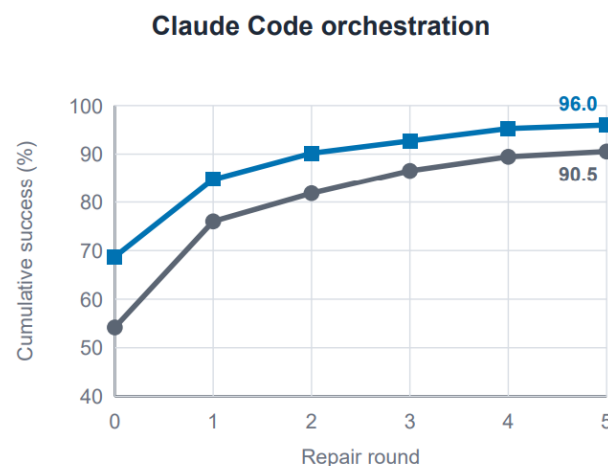
- Overall Workflow

- Intent understanding & semantic conversion
- Similar-example retrieval from tutorials
- Intent-driven dependency extraction
- Code execution with error handling



Domain code/script generation

- Two benchmarks and evaluation methods have been developed
 - Good improvement on accuracy
 - LLM: deepseek V4 pro
 - Testing claude code with opus4.8/fa
- Integrated to DrSai, works well
- This work can be expand to software development at HEP



More details in arXiv:[2609.01095](https://arxiv.org/abs/2609.01095)

Knowledge Query System

Knowledge for Physics Analysis



- After a few years investigation,
- We found the most useful case for RAG is to help decision making

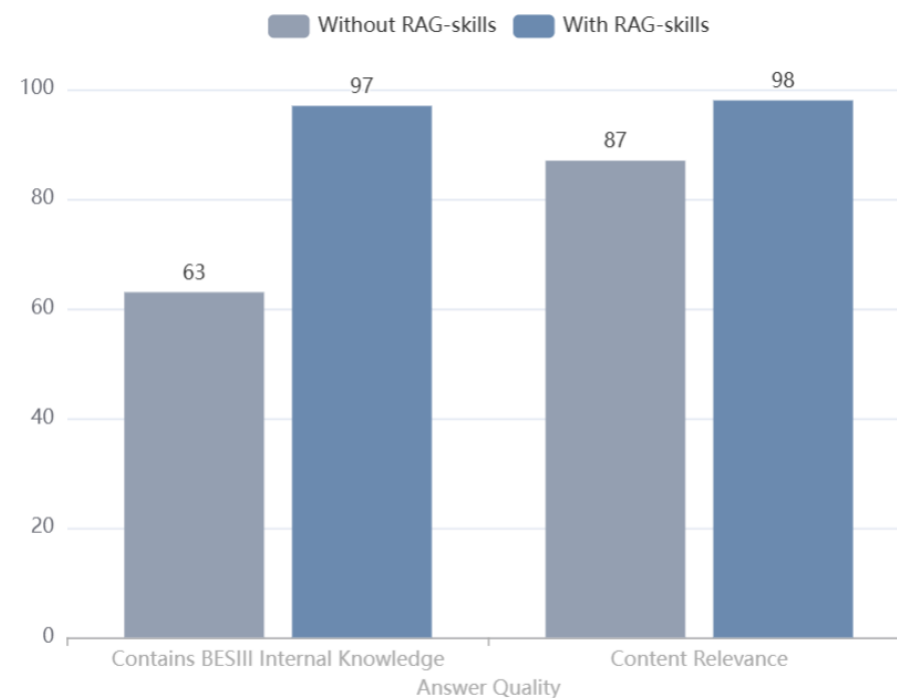
- BESIII experiment has a lot of documentation, spread over myriad forms (Hypernews, BESIII official/internal website, Indico, Gitlab...)
- Analyses are documented in thousands of Memos: **Questions are heterogeneous** (a cut value, a specific background, a systematic uncertainty) all need different retrieval.

<i>Knowledge Type</i>	<i>What it matches</i>
<i>Lexical</i>	<i>Literal values — energy points, numbers, symbols</i>
<i>Procedural</i>	<i>Methodological templates of an analysis type</i>
<i>Semantic</i>	<i>Conceptual reasoning within one analysis context</i>
<i>Synthetic</i>	<i>Aggregative relations across multiple analyses</i>

- **Agent + Knowledge-base**: LLM agents with RAG can, in principle, answer "how was this done before?" — if they retrieve the right chunk for the right question.

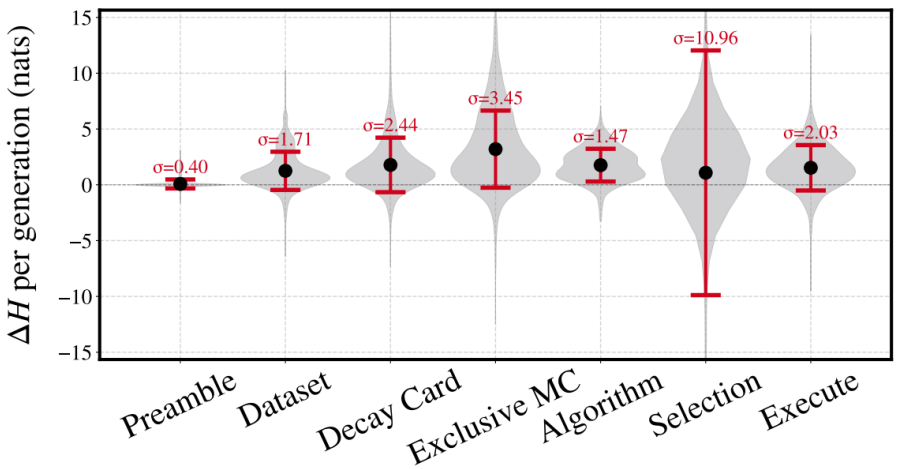
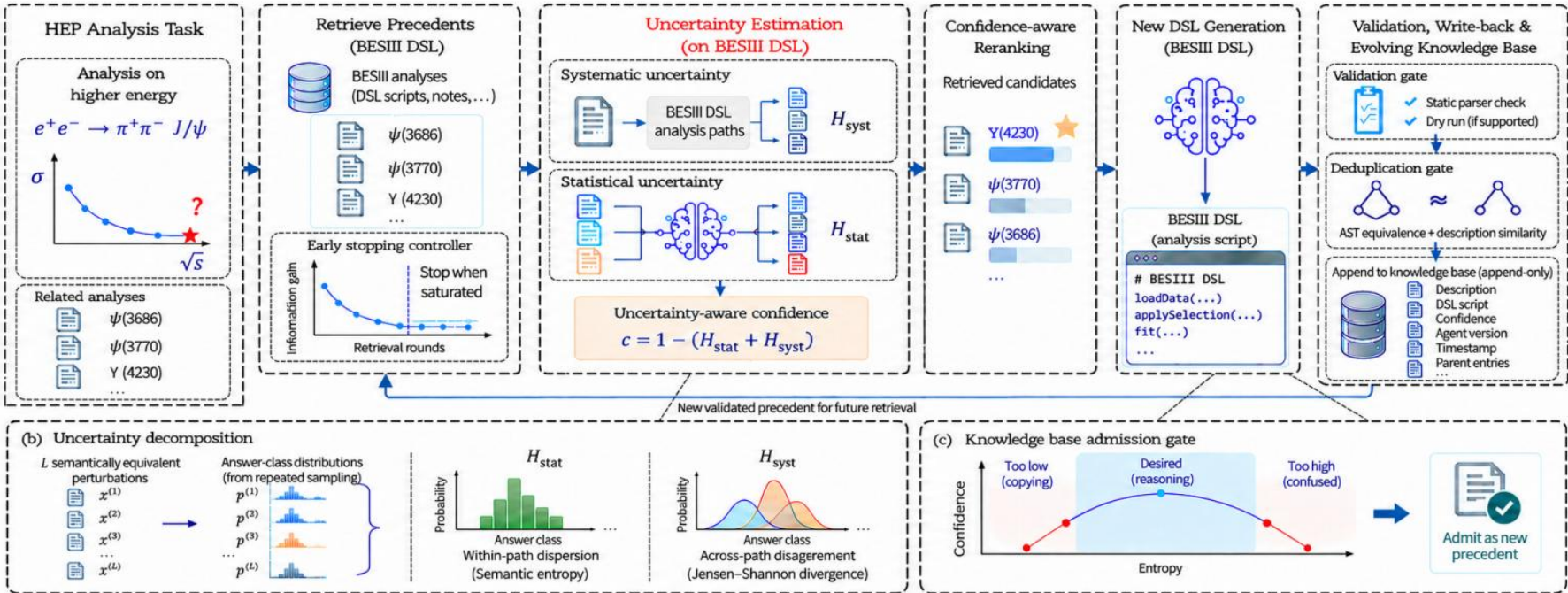
Query system with Agentic RAG/Wiki

- Multiple sources: [internal notes](#), [websites](#), [talks](#), etc.
 - Multi-modal parsing via Miner-U; Embedding model using Bge-M3
- Previous analyses at BESIII are chunked by analysis steps and sub-steps
 - Object selection, optimization, signal extraction, systematic uncertainty, etc.
- [Metadata](#) are defined for each analysis
 - Goal, dataset, observables, etc.
- Agentic retrieval with Skills
 - Hybrid: [Grep + RAG \(dense and sparse channels\)](#)
 - Routed by LLM
- User can ask the previous analysis details
 - How to deal with xxx systematics ?
 - How to suppress xxx background ?
 - Etc.
- [The system will be used to provide suggestions for new analysis](#)



Recent study on RAG

- Inspired by uncertainty from HEP
- Define an uncertainty (or confidence) for each message in RAG

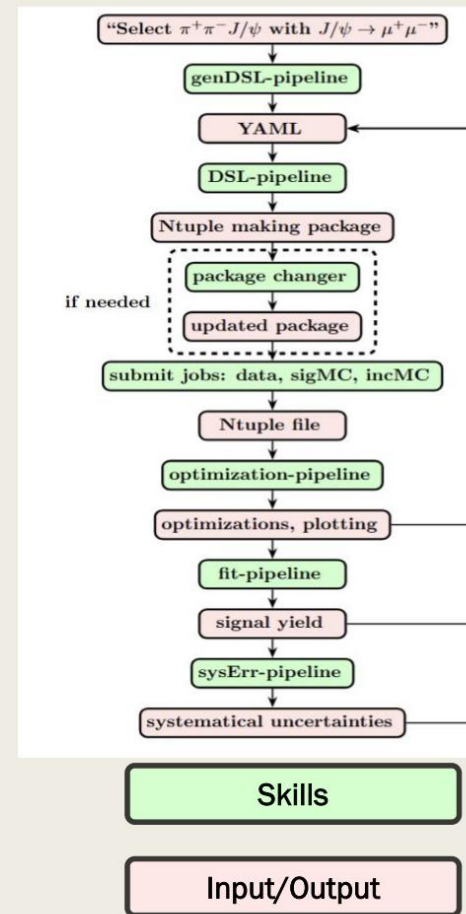
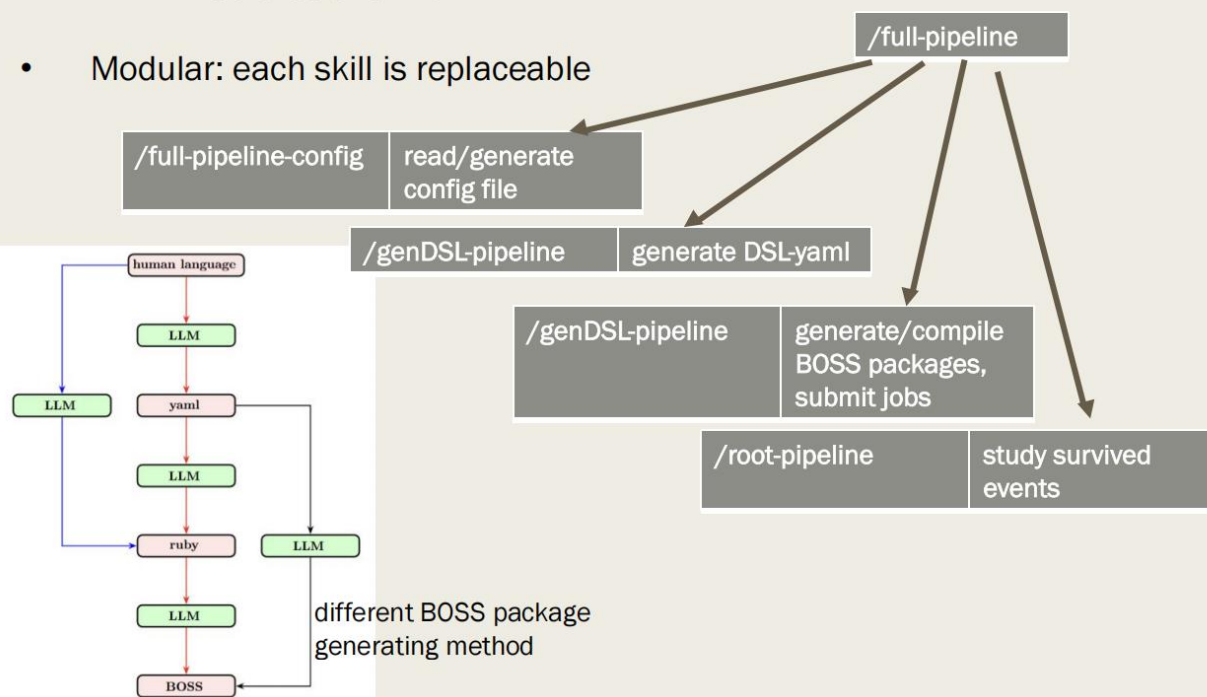


Perturbation	Strict (%)			Struct (%)			Phys-Acc (%)		
	NR	Rel	Conf	NR	Rel	Conf	NR	Rel	Conf
$k = 3$	71.8	<u>76.2</u>	76.7	89.3	<u>91.0</u>	91.0	57.6	<u>62.4</u>	63.4
$k = 10$	71.8	<u>75.7</u>	76.3	89.3	<u>90.6</u>	90.8	57.6	<u>62.6</u>	63.3
paraphrase	71.8	<u>75.2</u>	76.0	89.3	<u>90.3</u>	90.7	57.6	<u>61.8</u>	63.9
ORACLE	100			100			100		

Skills for BESIII analysis

Workflow Based on Skills

- General-purpose Agent + DSL (as a skill) + Skills + RAG (as a skill)
- Three tires of skills
- Modular: each skill is replaceable



Another part of DSL

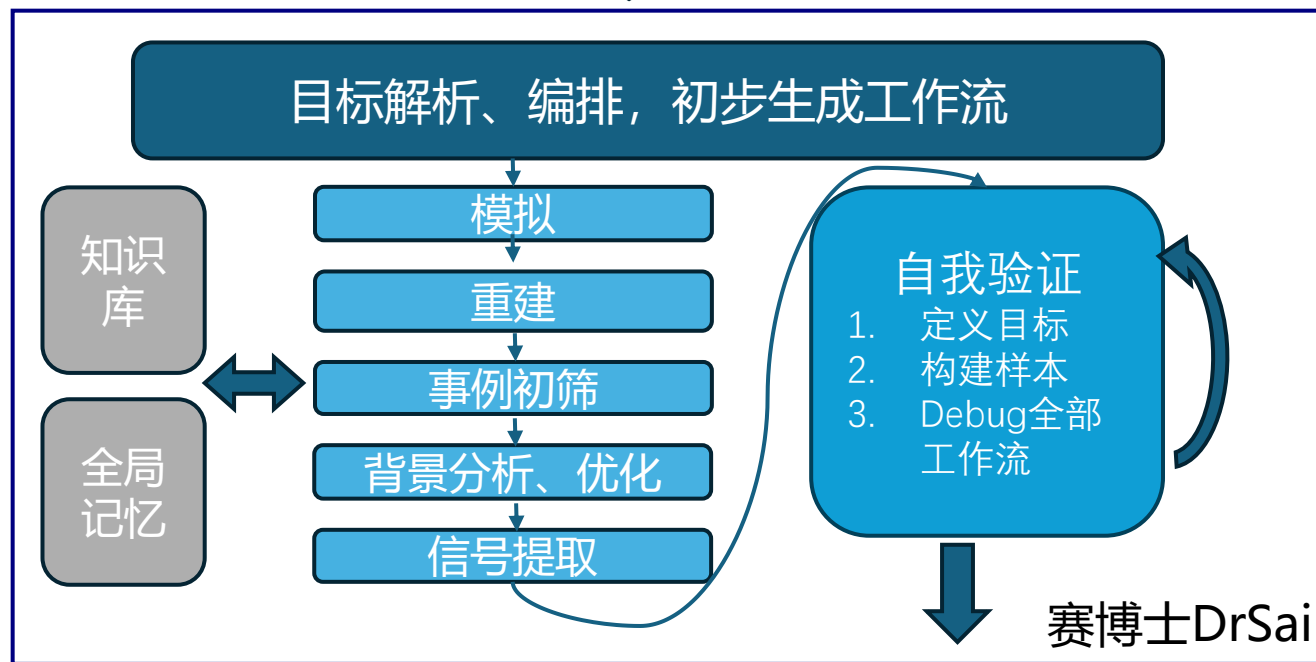
- DSL-yaml as work "note": represent the analysis pipeline
 - Use existing work to extract knowledge for preliminary DSL-yaml generation
 - YAML format, parsed by LLM, high readability
 - Interactive work, revisable at any time (manually or via LLM).
 - Finally obtain "note" that contain all the information of this analysis
 - Write MEMO automatically

```
metadata:
  name: pip psi
  BAM: "000"
  dataType: 4260
  comment: "Absolute branching ratio of J/psi to lepton pair via radiative return at 4.26 GeV. The signal yield of e+e- -> pip
  pim J/psi, J/psi -> ep em (mup mum) is determined by fitting the invariant mass distribution of the lepton pair. The signal y
  fields N_signal are determined with the fit to data. The efficiencies are given by signal MC samples."
  selections:
    - analysis_1:
      - "e+e- -> pip pim J/psi"
      - "J/psi -> ep em"
    - analysis_2:
      - "e+e- -> pip pim J/psi"
      - "J/psi -> mup mum"

analysis_1:
  dir: .
  dir_data: None
  dir_sigMC: None
  dir_incMC: None
  dir_draw: None
  sample:
    data: None
    incMC: None
    sigMC:
      - num events: 50000
      process: e+e- -> pip pim J/psi, J/psi -> ep em
  base_selection:
    comment:
      tagged_data: "nCharge>3, nCharge<5"
    tag:
      common: pip pim ep em
      other:
    basic_pid:
      - pi: "momentum less than 0.7 GeV/c"
      - e: "momentum larger than 1 GeV/c, EMC deposition less than 0.8 GeV"
      - mu: "momentum larger than 1 GeV/c, EMC deposition less than 0.4 GeV"
    kmfit:
      - var: 4C
        setting: setChisqCut(200,0.005)
      - var: 5C (4C, M_jpsi)
        setting: setChisqCut(9999,0.005)
    further_selection:
      chisq:
        - {var: chisq4C, optimize: 1D, strategy: FoM, default: (0,200), Nbin_region: "100, (0,200)"}
      costheta:
        - {var: polarAngle_pip, optimize: False, strategy: FoM, default: (-0.8,0.8)}
        - {var: polarAngle_pim, optimize: False, strategy: FoM, default: (-0.8,0.8)}
      theta:
        - {var: openAngle_pip_pim, optimize: False, strategy: FoM, default: (0.2, 3.2)}
      Mass:
        - {var: "Mjpsi: p4_ep + p4_em", optimize: False, strategy: FoM, default: (3.09,3.11), Nbin_region: "100, (3.,3.2)"}
      RM_pi:
        - {var: "maxRMpi: max(RM(pi+), RM(pi-)), ecms=4.25795", optimize: False, strategy: FoM, default: (3.6,4.2), comment: "di
          agnostic only, NRM=100(3.6,4.2), sideband defined as (3.05,3.08)U(3.12,3.15)"}
    fit:
  analysis_2:
  systematic_uncertainties:
  result:
```

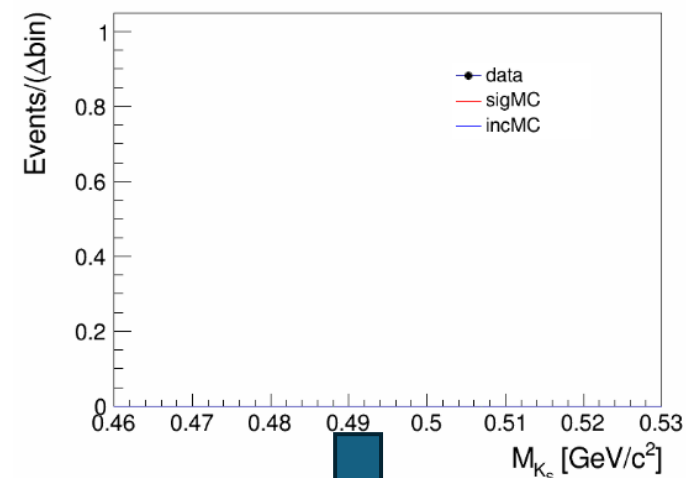
简单案例

提示词：请用质心系能量为4.26GeV的数据分析
 $e+e \rightarrow \pi+\pi^- J/\psi$ 过程，并寻找 $\pi+J/\psi$ 上新的粒子

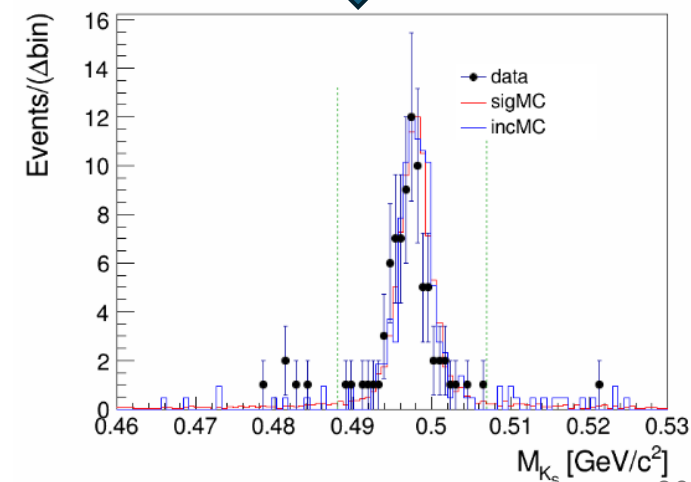


显著提升稳定性，目前可并行处理>30个物理过程

自我验证前，很多错误结果

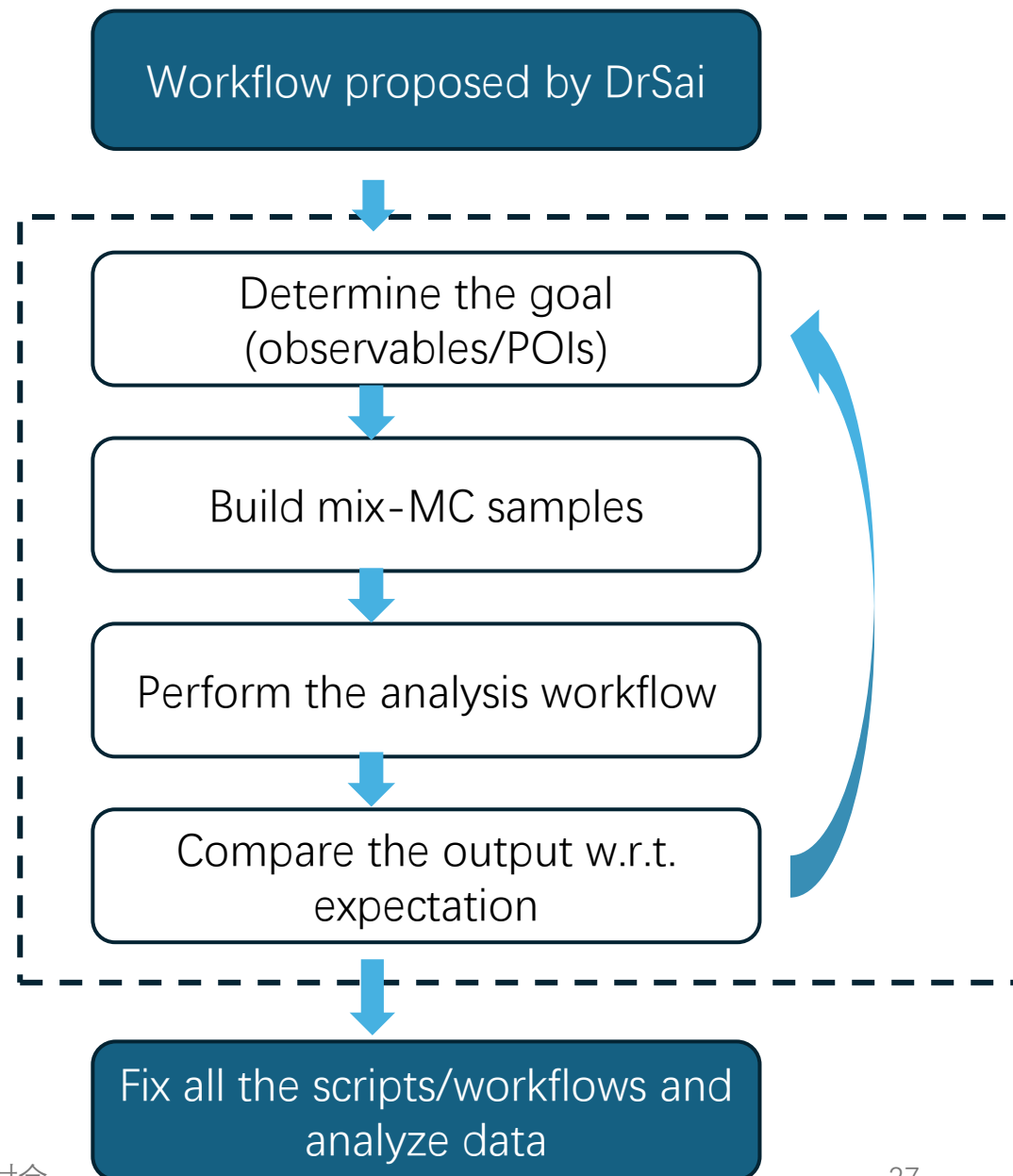


自我验证后



Self-validation

- A new module inspired by **Recurrent Self-Improvement (RSI)** from AI and Input/Output test at BESIII
 - A preliminary version is integrated
 - Automatic I/O check
 - **Fix most of the bugs**
- Improve the stability obviously
 - Especially helpful for analysis with multiple sub-agents
- Will add more physics reasonability check



Introduction of DeepSeek-Harness

- Deepseek has long been known as an effective LLM, whereas its corresponding harness is released recently.
- Deepseek harness (DSH) is an agent framework built around the DeepSeek LLM family.
- Its domain knowledge lives in a skill / plugin layer — plain specification documents executed step by step by a single agent, plus a plugin runtime such as an SSH plugin for remote cluster access.
- DSH is **open-source** and **the community is now very active and well maintained**— a knowledge package like DrSai can be transplanted by simply copying the skill definitions.

Introduction of DeepSeek-Harness

- Both agent frameworks share a "knowledge package" — same name, different nature:

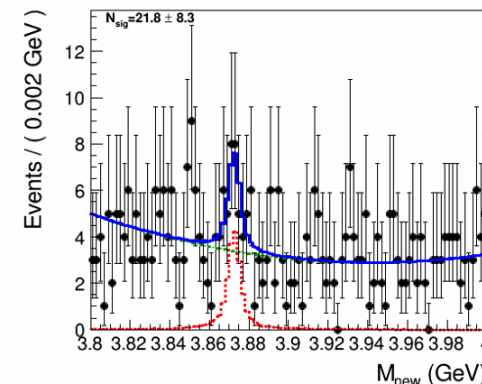
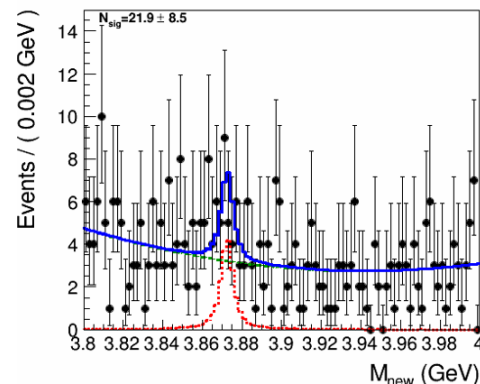
Item	Claude side	DSH side
Knowledge package	"DrSai HEP analysis package"	Same name, fully ported (formerly command, now renamed skill)
Object type	skill / command layer (sub-agents + slash-command system)	skill / plugin layer (skill directory + plugin runtime)
Essential difference	Centered on "command / sub-agent orchestration", multi-agent division of labor	Centered on "skill spec documents", a single agent executes skills one by one
Pipeline organization	Relies on manual multi-round direction of sub-agents	Built-in full-pipeline / dsl-pipeline / root-pipeline orchestration skills

- Claude: **"command + sub-agent" orchestration**; DSH: **"skill directory + plugin" specs** — a paradigm difference, not mere configuration.

Claude code -> DeepSeek Harness

- Previously we developed DrSai on Claude code agent runtime
 - LLM is ported to DeepSeek-V4
- Now open-source framework
 - DeepSeek Harness (DSH)
 - Everything is plugin
 - More suitable to us
- We have tested DrSai on DSH
 - DSH is deployed locally
 - Validated by performing re-measurements
 - Consist results w.r.t. DrSai on Claude Code
 - LLMs are the same
 - Skills/DSL/RAG are the same

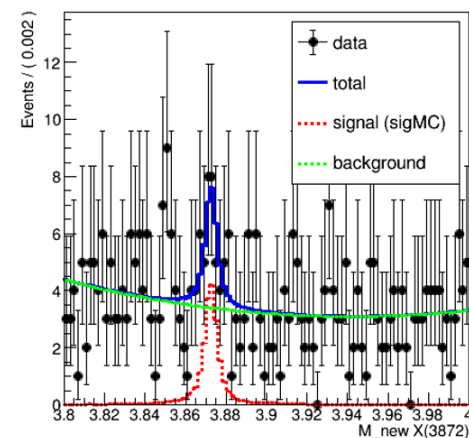
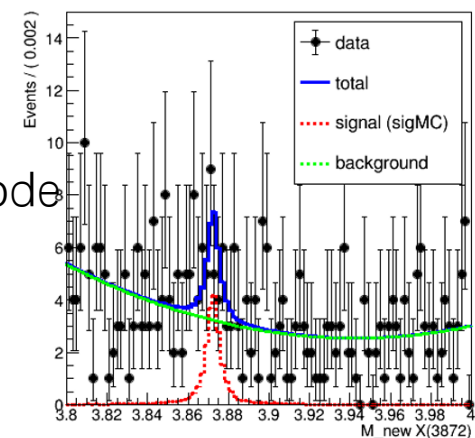
DrSai on
DSH



X(3872) ee

X(3872) mu mu

DrSai on
Claude Code



Please refer to [Mingze's talk](#) for more details

Claude code -> DeepSeek Harness

Metric	Claude (deepseek-v4)	DSH (deepseek-v4)
Input token	251,334	1,271,856
Output token	316,616	603,053
Cache read	47,180,416	10,028,160
Shadow (DSH compaction count)	—	791,586
Total raw tokens (incl. shadow)	≈ 47.7M	≈ 12.7M
Cache hit rate	99.47%	88.74%

- We have compared the computing performance preliminarily
- Slightly less tokens for DSH
- We plan to use DSH as default harness

Cost conversion

Unified price model input : output : cache-read : shadow = 1 : 5 : 0.1 : 1.25

Claude ≈ 6.55M input = ¥0.97

DSH ≈ 6.28M input = ¥0.92

input-equivalents

DSH is in fact ~4% cheaper (Claude costs ≈1.04×)

A physics analysis (incl. debug) usually consumes 100M + tokens

Comparisons

Dimension	Claude	DSH	Notes
Engineering efficiency	≈	≈	Claude: single session, zero compaction, but cache-token overhead grows explosively; DSH sustains context via 32 compactions (+40 prunes) in one session, with information-loss risk
Physics results	≈	≈	Yields agree within errors (ee 21.9 ± 8.5 vs 20.6 ± 8.3 ; $\mu\mu$ 21.8 ± 8.3 vs 20.7 ± 8.1); both joint fits needed identical manual fixes (skill-layer knowledge gap, not framework)
Portability	×	✓	DSH skills are spec documents, portable with the prompt alone; Claude commands are "commands + sub-agents", migration needs multi-agent orchestration
Extensibility	×	✓	DSH has built-in end-to-end full-pipeline orchestration skills, naturally extensible; Claude relies on manual round-by-round direction
Cost	×	✓	DSH slightly better (6.55M vs 6.28M input-equivalents, ~4% cheaper)
Human-agent interaction	×	✓	DSH supports plugin-style DSL, with different levels of physics info managed as different file forms; Claude relies only on the context session

Instructions

- Say “analysis with DrSai, see config.yaml”
 - A yaml-like style configuration
 - Or generate config.yaml via human-AI interaction
- config.yaml contains:
 - Directories
 - Confirmation
 - subAgents-parallel
 - sysErr estimation
 - BOSS package generation
 - Physical processes

```
# ===== Global Fields =====

outdir: ./logR          # directory for analysis output.
runBOSS_outdir: ./logR/proj # directory for BOSS job projects. omitted → interaction
workarea: /workfs2/bes/liutong2016/boss713/workarea/Analysis/R # BOSS package workarea. omitted → interaction

confirm_mode: minimal    # detailed (ask at every step) | minimal (interact only for required items)
parallel: yes             # yes | no. Whether to enable parallel agent execution.

# If both yes, LLM will combine the outputs
sysErr: yes               # yes | no. Whether to generate systematic error code via RAG
sysErr-ruby: yes          # yes | no. Use the systematic error analysis benchmark developed for HepScript

genBOSSCode: direct      # ruby | direct. BOSS code generation method.
                        # ruby - generate BOSS code via Ruby/HepScript
                        # direct - generate directly

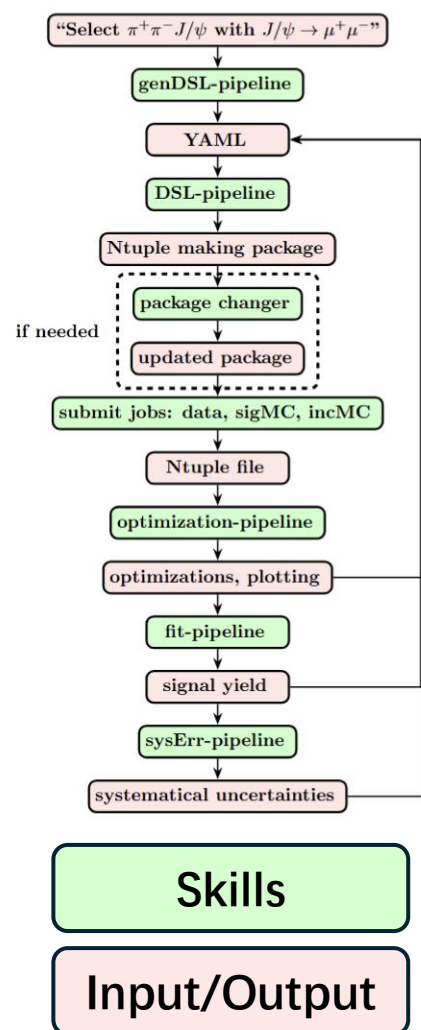
# ===== entries Mode and Entries =====

mode: single              # single | multi
                        # single - one analysis
                        # multi - multiple independent analyses

ww: Rvalue                # Optional. Valid only when mode=single and entries≥2
                        # Ignored in other modes (multi / single+entries=1)

entries:
  # Format per line: process threshold_cut=<GeV> description
  # - process          - channel name & algorithm base name, used for subdirectory and analysis package naming.
  # - threshold_cut    - energy threshold for KKMC+BesEvtGen (GeV). If missing, full-pipeline will interactively p
  # - description      - brief description of the analysis procedure: final-state particles, etc.
  # - prompt at load time (default 3.0).
  - "bhabha threshold_cut=0.001022 select e+e- at 1.9 GeV"
  - "dimu threshold_cut=0.2113 select mu+mu- at 1.9 GeV"
  - "pipi threshold_cut=0.280 select pi+pi- at 1.9 GeV"
  - "pipipi0 threshold_cut=0.415 select pi+pi-pi0, pi0->gamma gamma at 1.9 GeV"
```

A recording for Zc(3900) re-discovery



```

(drSai) [liutong2016@beslogin005 DSL]$
(drSai) [liutong2016@beslogin005 DSL]$
(drSai) [liutong2016@beslogin005 DSL]$
(drSai) [liutong2016@beslogin005 DSL]$echo $PATH
/hpcfs/bes/gpupwa/liutong2016/conda/miniconda3/envs/drSai/bin:/hpcfs/bes/gpupwa/liutong2016/conda/miniconda3/condabi
n:/cvmfs/bes3.ihep.ac.cn/bes3sw/Boss/7.1.3/InstallArea/share/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/Gaudi/Gau
di-v27r1-g4107-python2.7/InstallArea/x86_64-centos7-gcc49-opt/scripts:/cvmfs/bes3.ihep.ac.cn/bes3sw/Boss/7.1.3/Insta
llArea/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/Gaudi/Gaudi-v27r1-g4107-python2.7/Inst
allArea/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/lcg/LCG_84/pytools/1.9_python2.7/x86_
64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/lcg/LCG_84/sqlite/3.11.0/x86_64-centos7-gcc49-op
t/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/lcg/LCG_84/mysql/5.5.38/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihe
p.ac.cn/bes3sw/Externallib/lcg/LCG_84/XercesC/3.1.1/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Ext
ernallib/lcg/LCG_84/castor/2.1.13-6/x86_64-centos7-gcc49-opt/usr/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/lcg/L
CG_84/ROOT/6.28.02/x86_64-centos7-gcc49-opt-python2.7/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/lcg/LCG_84/xroot
d/4.2.3/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/lcg/LCG_84/Python/2.7.10/x86_64-centos7-gcc49-opt/Ext
ernallib/lcg/LCG_84/CMT/v1r25/Linux-x86_64:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/contrib/gdb/7.6.1/bin:/hpcfs/be
s/gpupwa/liutong2016/conda/miniconda3/envs/drSai/share/rubygems/bin:/hpcfs/bes/gpupwa/liutong2016/conda/miniconda3/b
in:/afs/ihep.ac.cn/users/l/liutong2016/topoana-02-05-03/bin:/cvmfs/container.ihep.ac.cn/bin:/afs/ihep.ac.cn/soft/com
mon/sysgroup/hep_job/bin:/usr/local/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/cvmfs/common.ihep.ac.cn/software/cctools
:/opt/puppetlabs/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Externallib/lcg/LCG_84/cernlib/2006a/x86_64-centos7-gcc49-opt/bin
(drSai) [liutong2016@beslogin005 DSL]$C1
Claude Claude2
(drSai) [liutong2016@beslogin005 DSL]$Claude2
Claude Claude v2.1.143
glm-5.1[1m] · API Usage Billing
/besfs5/groups/tauqed/liut/R/DSL

> 做pipi7psi分析. 配置在config_pipi7psi.yaml

• Explore(Find config_pipi7psi.yaml)
  Done (3 tool uses · 13.9k tokens · 22s)
  (ctrl-c to expand)

Read 1 file (ctrl-c to expand)

• 配置已读取. 启动 full-pipeline.

• Skill(/full-pipeline)
  Successfully loaded skill

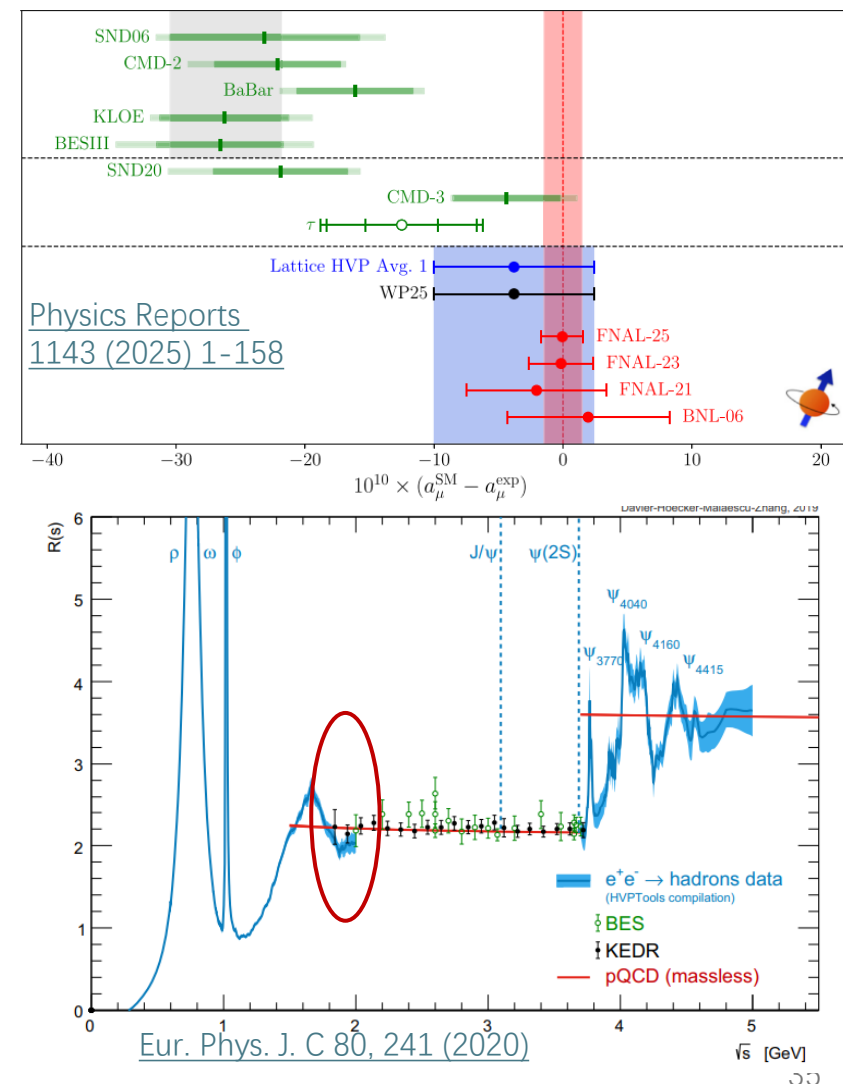
+ Creating... (47s · 1.654 tokens · almost done thinking)
  Tip: Use /btw to ask a quick side question without interrupting Claude's current work

>
  >> auto mode on (shift+tab to cycle) · esc to interrupt
  
```

begin

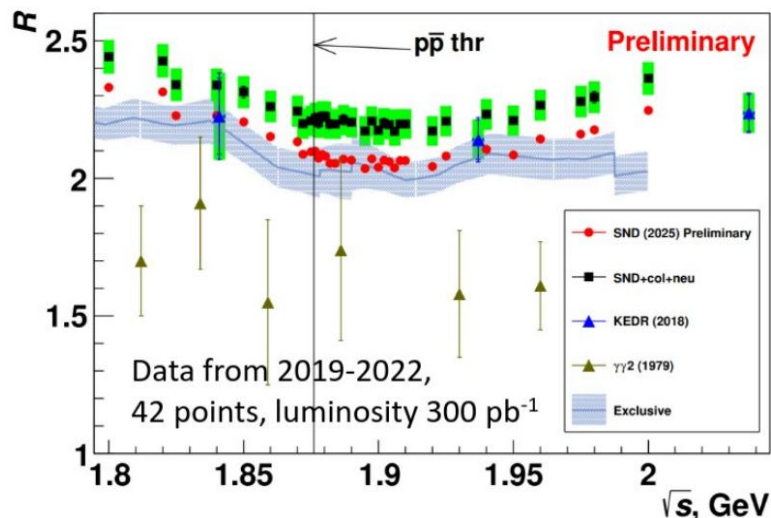
Analysis using DrSai at BESIII

- DrSai V3: **ready for BESIII physics analysis !**
 - **Vibe research – no coding needed now**
- **Now analyzing real data at BESIII:**
 - Measure all the hadronic cross section exclusively below 2 GeV, **a painful analysis for human thus long delayed**
 - A key input for HVP in muon g-2 calculation
 - Dispersive evaluation differ from LQCD
 - Experimental measurements of $e^+e^- \rightarrow \text{hadrons}$ are particularly important.
 - **Starting from 1.9 GeV:**
 - Experts steer the setup and input essential parameters, e.g. generators
 - DrSai design, execute and verify the entire analysis under expert supervision
 - Experts will **cross-check** all the measurements



Analysis using DrSai at BESIII

- Exclusive R value measurements at 1.9 GeV
 - Difference between inclusive and exclusive results



Recent results from the SND experiment

Signature	Final states (preliminary, published)
2 charged	$\pi^+\pi^-, K^+K^-, K_S K_L, p\bar{p}$
2 charged + γ 's	$\pi^+\pi^-\gamma, \pi^+\pi^-\pi^0, \pi^+\pi^-\eta, K^+K^-\pi^0, K^+K^-\eta, K_S K_L \pi^0, K_S K_L \eta, \pi^+\pi^-\pi^0\eta, \pi^+\pi^-\pi^0\pi^0, \pi^+\pi^-\pi^0\pi^0\pi^0, \pi^+\pi^-\pi^0\pi^0\pi^0\pi^0$
4 charged	$\pi^+\pi^-\pi^+\pi^-, K^+K^-\pi^+\pi^-, K_S K^\pm \pi^\mp, K_S K_L \pi^+\pi^-, K_S K_S \pi^+\pi^-$
4 charged + γ 's	$\pi^+\pi^-\pi^+\pi^-\pi^0, \pi^+\pi^-\eta, \pi^+\pi^-\omega, \pi^+\pi^-\pi^+\pi^-\pi^0\pi^0, K^+K^-\eta, K^+K^-\omega$
6 charged	$\pi^+\pi^-\pi^+\pi^-\pi^+\pi^-, K_S K^\pm \pi^\mp \pi^+\pi^-, K_S K_S \pi^+\pi^-$
6 charged + γ 's	$3(\pi^+\pi^-\pi^0)$
Neutral	$\pi^0\gamma, \eta\gamma, \pi^0\pi^0\gamma, \pi^0\eta\gamma, \pi^0\pi^0\pi^0\gamma, \pi^0\pi^0\eta\gamma$
Other	$n\bar{n}, \pi^0 e^+ e^-, \eta e^+ e^-$
Rare decays	$\eta', D^*(2007)^0, f_1(1285)$

CMD-3 analysis

Phys. Rev. D **97**, 114025 (2018)

Channel	This work (KNT18)	HLMNT11 [9]	Difference
Chiral perturbation theory (ChPT) threshold contributions			
$\pi^0\gamma$	0.12 ± 0.01	0.12 ± 0.01	0.00
$\pi^+\pi^-$	0.87 ± 0.02	0.87 ± 0.02	0.00
$\pi^+\pi^-\pi^0$	0.01 ± 0.00	0.01 ± 0.00	0.00
$\eta\gamma$	0.00 ± 0.00	0.00 ± 0.00	0.00
Data based channels ($\sqrt{s} \leq 2$ GeV)			
$\pi^0\gamma$	4.46 ± 0.10	4.54 ± 0.14	-0.08
$\pi^+\pi^-$	502.99 ± 1.97	505.77 ± 3.09	-2.78
$\pi^+\pi^-\pi^0$	47.82 ± 0.89	47.51 ± 0.99	0.31
$\pi^+\pi^-\pi^+\pi^-$	15.17 ± 0.21	14.65 ± 0.47	0.52
$\pi^+\pi^-\pi^0\pi^0$	19.80 ± 0.79	20.37 ± 1.26	-0.57
$(2\pi^+2\pi^-\pi^0)_{\text{no } \eta}$	1.08 ± 0.10	1.20 ± 0.10	-0.12
$3\pi^+3\pi^-$	0.28 ± 0.02	0.28 ± 0.02	0.00
$(2\pi^+2\pi^-\pi^0)_{\text{no } \eta\omega}$	1.60 ± 0.20	1.80 ± 0.24	-0.20
K^+K^-	23.05 ± 0.22	22.15 ± 0.46	0.90
$K_S^0 K_L^0$	13.05 ± 0.19	13.33 ± 0.16	-0.28
$K\bar{K}\pi$	2.80 ± 0.12	2.77 ± 0.15	0.03
$K\bar{K}2\pi$	2.42 ± 0.09	3.31 ± 0.58	-0.89
$\eta\gamma$	0.70 ± 0.02	0.69 ± 0.02	0.01
$\eta\pi^+\pi^-$	1.32 ± 0.06	0.98 ± 0.24	0.34
$(\eta\pi^+\pi^-\pi^0)_{\text{no } \omega}$	0.63 ± 0.15	-	0.63
$\eta 2\pi^+2\pi^-$	0.11 ± 0.02	0.11 ± 0.02	0.00
$\eta\omega$	0.31 ± 0.03	0.42 ± 0.07	-0.11
$\omega(\rightarrow \pi^0\gamma)\pi^0$	0.88 ± 0.02	0.77 ± 0.03	0.11
$\eta\phi$	0.45 ± 0.04	0.46 ± 0.03	-0.01
$\phi \rightarrow \text{unaccounted}$	0.04 ± 0.04	0.04 ± 0.04	0.00
$\eta\omega\pi^0$	0.42 ± 0.10	-	0.42
$\eta(\rightarrow \text{npp})K\bar{K}_{\text{no } \phi \rightarrow K\bar{K}}$	0.01 ± 0.01	-	0.01
$p\bar{p}$	0.07 ± 0.00	0.06 ± 0.00	0.01
$n\bar{n}$	0.06 ± 0.01	0.07 ± 0.02	-0.01
Estimated contributions ($\sqrt{s} \leq 2$ GeV)			
$(\pi^+\pi^-\pi^0)_{\text{no } \eta}$	0.53 ± 0.05	0.60 ± 0.05	-0.07
$(\pi^+\pi^-\pi^0)_{\text{no } \eta}$	0.25 ± 0.25	0.28 ± 0.28	-0.03
$K\bar{K}3\pi$	0.08 ± 0.03	0.08 ± 0.04	0.00
$\omega(\rightarrow \text{npp})2\pi$	0.10 ± 0.02	0.11 ± 0.02	-0.01
$\omega(\rightarrow \text{npp})3\pi$	0.20 ± 0.04	0.22 ± 0.04	-0.02
$\omega(\rightarrow \text{npp})K\bar{K}$	0.01 ± 0.00	0.01 ± 0.00	0.00
$\eta\pi^+\pi^-\pi^0$	0.11 ± 0.05	0.11 ± 0.06	0.00

Analysis using DrSai at BESIII

- Exclusive R value measurements at 1.9 GeV
 - Difference between inclusive and exclusive results
- Neglect small processes, still >30 channels, an excellent scenario for DrSai
 - Codes are generated with subAgents simultaneously
 - Strategies are suggested by RAG
 - Only human-AI interaction, no manual coding!

```

>> auto mode on · 25 local agents · i to manage

• main
general-purpose BOSS code gen: bhabha (j2=1) 6m 52s · 1 69.8k tokens
general-purpose BOSS code gen: digam (j2=2) 6m 3s · 1 73.4k tokens
general-purpose BOSS code gen: dimu (j2=3) 6m 3s · 1 75.8k tokens
general-purpose BOSS code gen: eeee (j2=4) 6m 3s · 1 61.9k tokens
general-purpose BOSS code gen: eeeta (j2=5) 6m 3s · 1 53.4k tokens
general-purpose BOSS code gen: eeetap (j2=6) 6m 3s · 1 58.9k tokens
general-purpose BOSS code gen: eekk (j2=7) 5m 17s · 1 31.8k tokens
general-purpose BOSS code gen: eepipi (j2=8) 5m 17s · 1 51.8k tokens
general-purpose BOSS code gen: eemumu (j2=9) 5m 17s · 1 54.7k tokens
general-purpose BOSS code gen: pi0 gamma (j2=10) 5m 17s · 1 56.5k tokens
general-purpose BOSS code gen: eta gamma (j2=11) 5m 17s · 1 55.8k tokens
general-purpose BOSS code gen: pi+pi- (j2=12) 5m 17s · 1 47.3k tokens
general-purpose BOSS code gen: pi+pi-pi0 (j2=13) 4m 22s · 1 48.4k tokens
general-purpose BOSS code gen: pi+pi-pi+pi- (j2=14) 4m 22s · 1 49.9k tokens
general-purpose BOSS code gen: pi+pi-pi0pi0 (j2=15) 4m 22s · 1 38.5k tokens
general-purpose BOSS code gen: pi+pi-pi+pi-pi0 (j2=16) 4m 22s · 1 39.7k tokens
general-purpose BOSS code gen: pi+pi-3pi0 (j2=17) 4m 22s · 1 35.7k tokens
general-purpose BOSS code gen: 6pi (j2=18) 4m 22s · 1 50.6k tokens
general-purpose BOSS code gen: 4pi2pi0 (j2=19) 3m 18s · 1 11.5k tokens
general-purpose BOSS code gen: pipi4pi0 (j2=20) 3m 18s · 1 30.3k tokens
general-purpose BOSS code gen: omega eta pi0 (j2=21) 3m 18s · 1 38.3k tokens
general-purpose BOSS code gen: K+K- (j2=22) 3m 18s · 1 29.2k tokens
general-purpose BOSS code gen: K_S0 K_L0 (j2=23) 3m 18s · 1 46.8k tokens
general-purpose BOSS code gen: KKpipi (j2=24) 3m 18s · 1 31.5k tokens
general-purpose BOSS code gen: KKpi (j2=25) 3m 18s · 1 49.4k tokens

```

2026/10/06

nCharge=2	
$\pi^+ \pi^-$	
$\pi^+ \pi^- \pi^0$	BAM-546
$\pi^+ \pi^- 2\pi^0$	
$\pi^+ \pi^- 3\pi^0$	
$\pi^+ \pi^- 4\pi^0$	
$\pi^+ \pi^- \eta$	
$\omega \eta \pi^0$	BAM-00672
$K^+ K^-$	
$K^+ K^- \pi^0$	BAM-376, BAM-655
$K^+ K^- \pi^0 \pi^0$	BAM-229, BAM-346
$K^+ K^- \pi^0 \pi^0 \pi^0$	$KK3\pi \sim 0.08$
$K^+ K^- \eta$	$\phi \eta$ dom.
$K_S^0 K_L^0$	
$K_S^0 K_L^0 \pi^0$	BAM-376, BAM-655
$K_S^0 K_L^0 \pi^0 \pi^0$	
$K_S^0 K_L^0 \pi^0 \pi^0 \pi^0$	$KK3\pi \sim 0.08$
$K_S^0 K_L^0 \eta$	

2026 BESIII 十一研讨会

nCharge=4	
$2(\pi^+ \pi^-)$	
$2(\pi^+ \pi^-) \pi^0$	BAM-416, BAM-596
$2(\pi^+ \pi^-) 2\pi^0$	
$2(\pi^+ \pi^-) 3\pi^0$	
$2(\pi^+ \pi^-) \eta$	~ 0.11
$K^+ K^- \pi^+ \pi^-$	BAM-229, BAM-346
$K^+ K^- \pi^+ \pi^- \pi^0$	$KK3\pi \sim 0.08$
$K^+ K^- \pi^+ \pi^- \pi^0 \pi^0$	$< 3(\pi^+ \pi^-) \sim 0.28$
$K^+ K^- \pi^+ \pi^- \pi^0 \pi^0 \pi^0$	$< 3(\pi^+ \pi^-) \sim 0.28$
$K^+ K^- \pi^+ \pi^- \eta$	$< 0.11 ?$
$K_S^0 K_L^0 \pi^+ \pi^-$	BAM-376, BAM-655
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0$	$KK3\pi \sim 0.08$
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \pi^0$	$< 3(\pi^+ \pi^-) \sim 0.28$
$K_S^0 K_L^0 \pi^+ \pi^- \eta$	
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0$	$KK3\pi \sim 0.08$
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \pi^0$	< 0.08
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \pi^0 \pi^0$	< 0.08
$K_S^0 K_L^0 \pi^+ \pi^- \eta$	$< 0.11 ?$

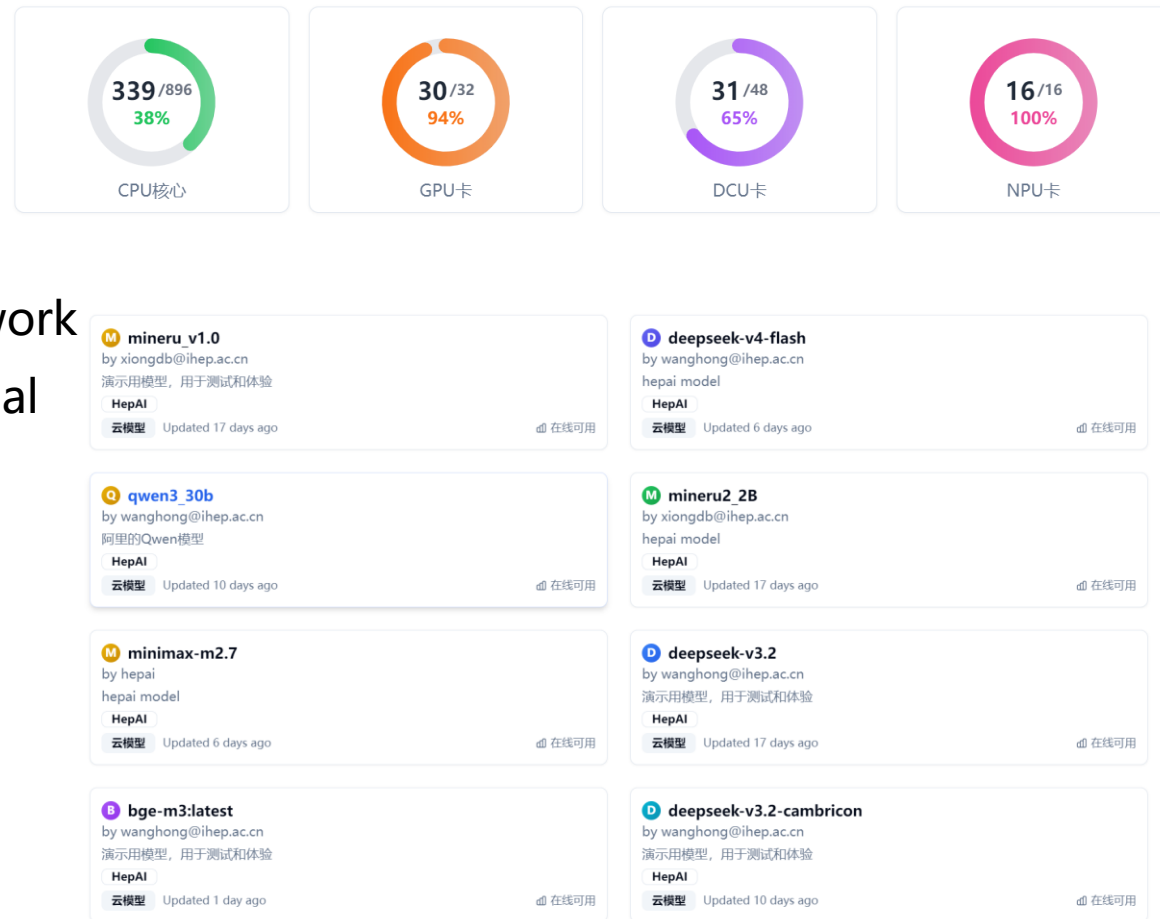
nCharge=0	
$\gamma \pi^0$	
$\gamma \eta$	

nCharge=6	
$3(\pi^+ \pi^-)$	~ 0.28
$3(\pi^+ \pi^-) \pi^0$	
$3(\pi^+ \pi^-) \eta$	$< 2(\pi^+ \pi^-) \eta \sim 0.11$
$K^+ K^- 2(\pi^+ \pi^-)$	$< 3(\pi^+ \pi^-) \sim 0.28$
$K^+ K^- 2(\pi^+ \pi^-) \pi^0$	$< 3(\pi^+ \pi^-) \sim 0.28$
$K^+ K^- 2(\pi^+ \pi^-) \eta$	$< 2(\pi^+ \pi^-) \eta \sim 0.11$
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0$	$KK3\pi \sim 0.08$
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \pi^0$	< 0.08
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \eta$	< 0.08
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \pi^0 \pi^0$	$< 3(\pi^+ \pi^-) \sim 0.28$
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \pi^0 \pi^0 \pi^0$	$< 3(\pi^+ \pi^-) \sim 0.28$
$K_S^0 K_L^0 \pi^+ \pi^- \pi^0 \pi^0 \pi^0 \eta$	$< 2(\pi^+ \pi^-) \eta \sim 0.11$

37

AI Infrastructure at IHEP

- New dedicated AI platform:
 - 2240 CPU cores, 40 GPU, 48 DCU, 16 NPU
 - More (2X) are coming
- LLMs access and deployment
 - Unified API via Distributed Deployment Framework
 - **Hybrid solution:** local deployment + commercial cloud
 - Open-source LLMs (e.g. DeepSeek)
 - High concurrency
 - Ensure data security
 - Closed-source LLMs accessible **via API**
 - Cache-hit techniques reduce the cost to 2-10%



DrSai Collaboration

- DrSai is already yielding physics results
 - Physicists interact with DrSai using natural language, only a few times with **zero coding!**
- Now extending to other experiments:
 - Jiangmen Underground Neutrino observatory (JUNO)
 - Large High Altitude Air Shower Observatory (LHAASO)
 - Neutron/Light source, etc.
- We are forming a DrSai collaboration
 - Solving Agentic AI challenges together

Welcome to join !

Summary and outlook

- DrSai: an agent-based system advancing HEP experiment
 - Project launched at June 2023
 - Major challenges identified → multiple solutions proposed & developed
 - Current approach:
 - Industrial coding agent + LLM
 - Skills + DSL + RAG : encode expertise, analysis representation, and technical memory
- **AI partner is being realized today**
 - First physics results expected by the **end of 2026**
- Many challenges remain towards AI scientist – but we are actively solving them
 - How to add physics knowledge/rule/check in a reasonable approach?
 - As AI evolves rapidly, we are adapting our strategy accordingly

Thanks for your attention !

Extra info

- Agentic AI 培训: <https://indico.ihep.ac.cn/event/29693/>
- AI workshop at next week: <https://indico.ihep.ac.cn/event/29575/>
 - There will be another tutorial for Agentic AI
- DrSai V2 paper: <https://arxiv.org/abs/2604.22541>
- DSL paper: <https://arxiv.org/abs/2605.01423>
- Another 2 papers will be released soon

Many challenges remain

- Agentic AI for HEP is a critical and challenging endeavor
 - Novel ideas are essential to drive progress
- Ongoing studies are plentiful – collaboration is warmly welcomed
 - Agentic AI for new ML models
 - Detector design (layout and chip)
 - Continuous evaluation with expert feedback
 - Reasonableness testing
 - Knowledge representation via reinforcement learning
 - Migration to other HEP experiments



Back-up

DrSai – development roadmap

- **2023.06: Working group launched (~20 people)**
- **2024.06: V1**, Feasibility tested and verified
 - Setup the system, all tools integrated
- **2025.06: V2**, The prototype was conducted
 - Capable of very simple analysis with fixed workflow
 - Validated by performing re-measurements
- **2026.06: V3**, under internal testing
 - Knowledge extracted into DSLs
 - Expertise distilled into Skills
 - Analyzing real data now, expect to have the **first “AI-driven” experimental results by end of 2026 !**

Error-Aware Retrieval and Skills

- Another retrieval by identifying the sources of errors
 - Improve the accuracy obviously
- Further improvement by introducing a set of Skills
- Two benchmarks and evaluation methods are provided

Average Score



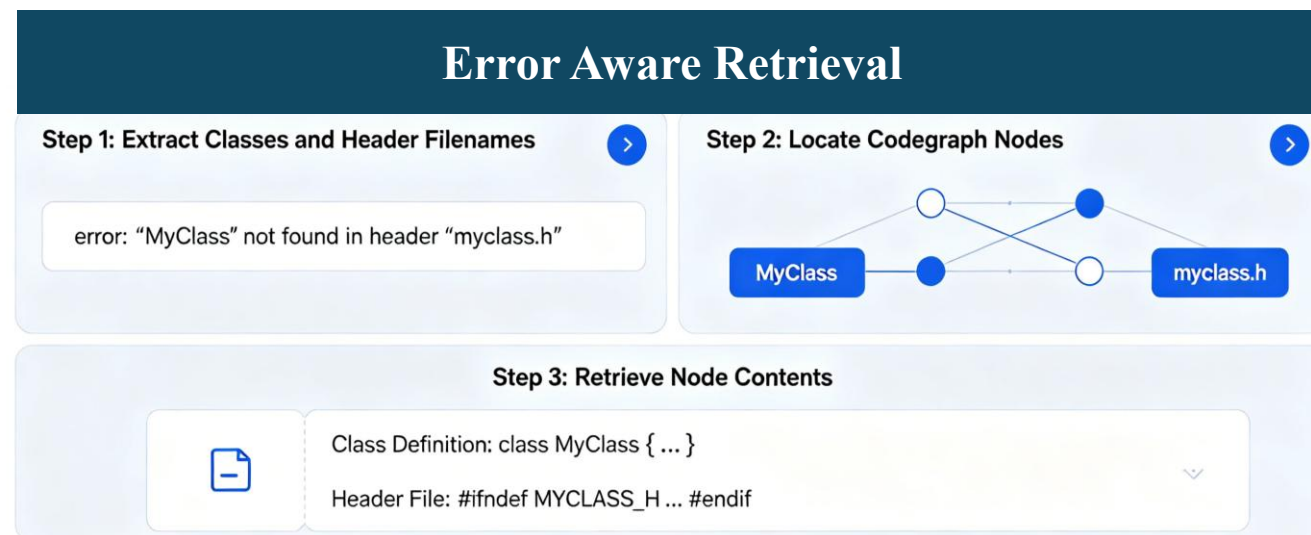
Files evaluated: 50 base, 47 skill-version outputs.

Language uplift

C +6.57

Python +3.20

+4.99 avg score



Significant improvement

Codegraph_RAG

27.3% to 35%

After five rounds of error feedback

58.33% to 69.33%

Error_RAG

69.33% to 73.67%

Design tutorials as skills (To be tested)

AI scientists: Goal-driven autonomy

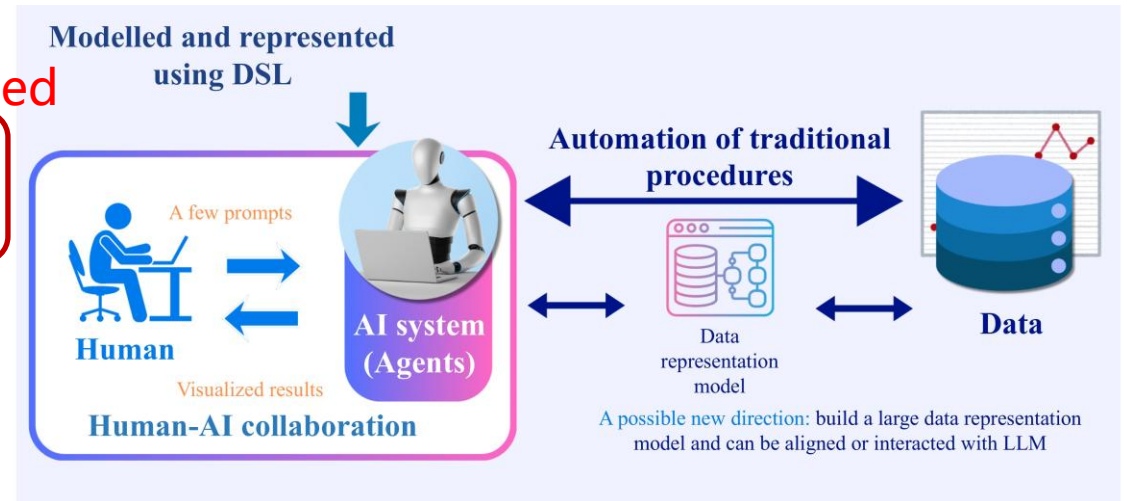
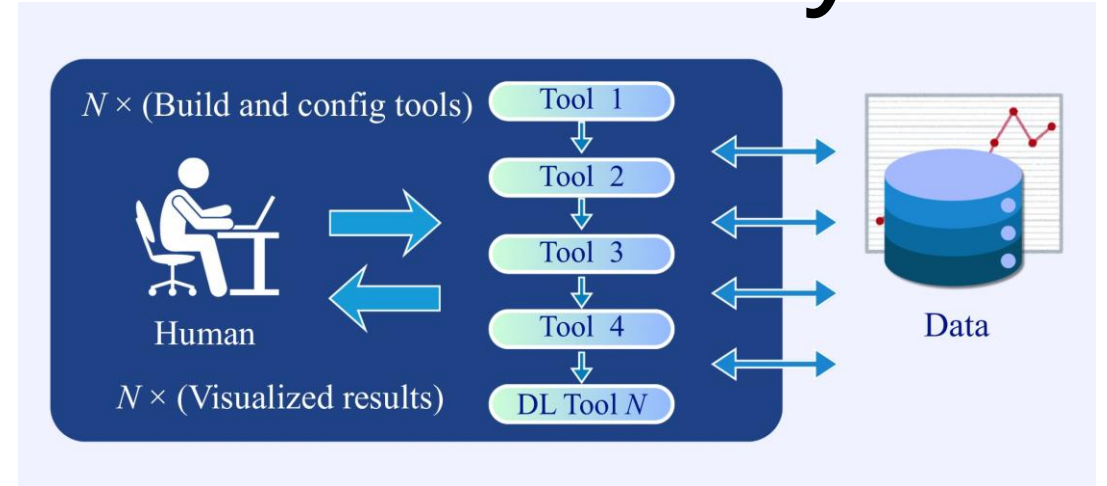
- An AI system with structured knowledge representation and autonomous research capabilities
- Can perform the full research loop and continue revolution:

- Goal → Design → Execute → Analyze
→ Generate new knowledge

- Key distinction from AI Partner:

- AI Partner: relies on human commands and operate through “vibe research”
 - AI Scientist: human only need to sets high-level goals

Already
completed



Milestone developments

Setup the system, standardize and integrate all the tools, including BESIII Offline Software System(BOSS) and HTCondor

**Done at
2024.06**

Represent the knowledges on how to used tools and analysis procedures into a DSL, automatically extracted from published papers

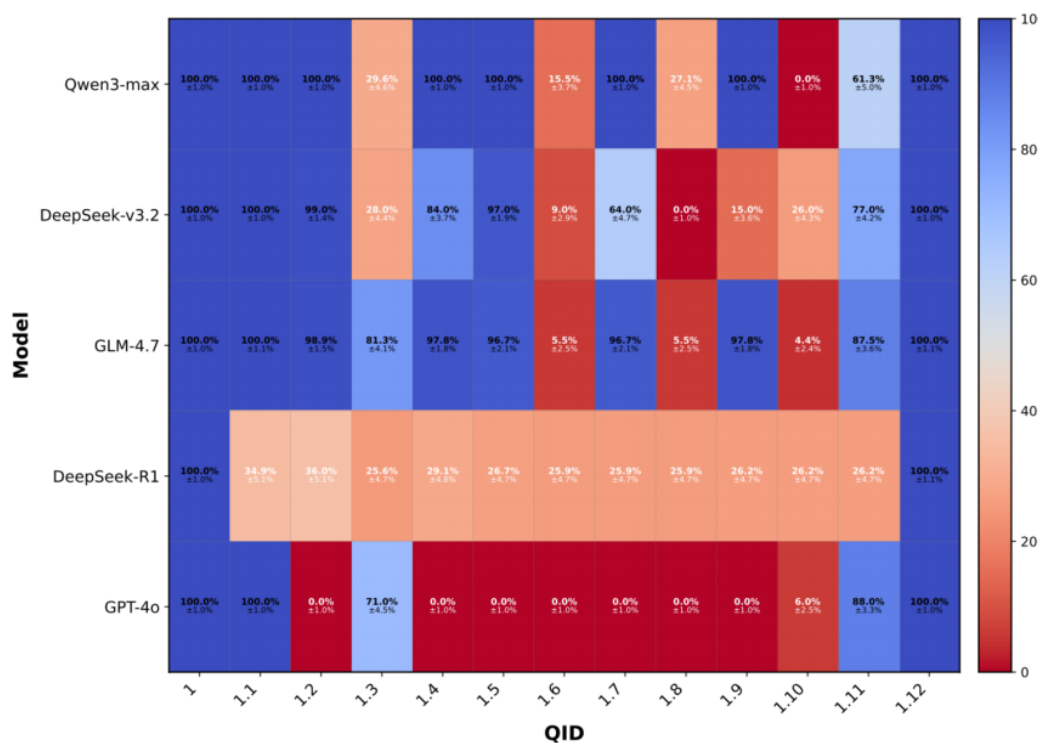
**Done at
2025.06**

Distill human expertise into Skills on how to perform analysis, including workflow definition and tests (i.e. input/output test)

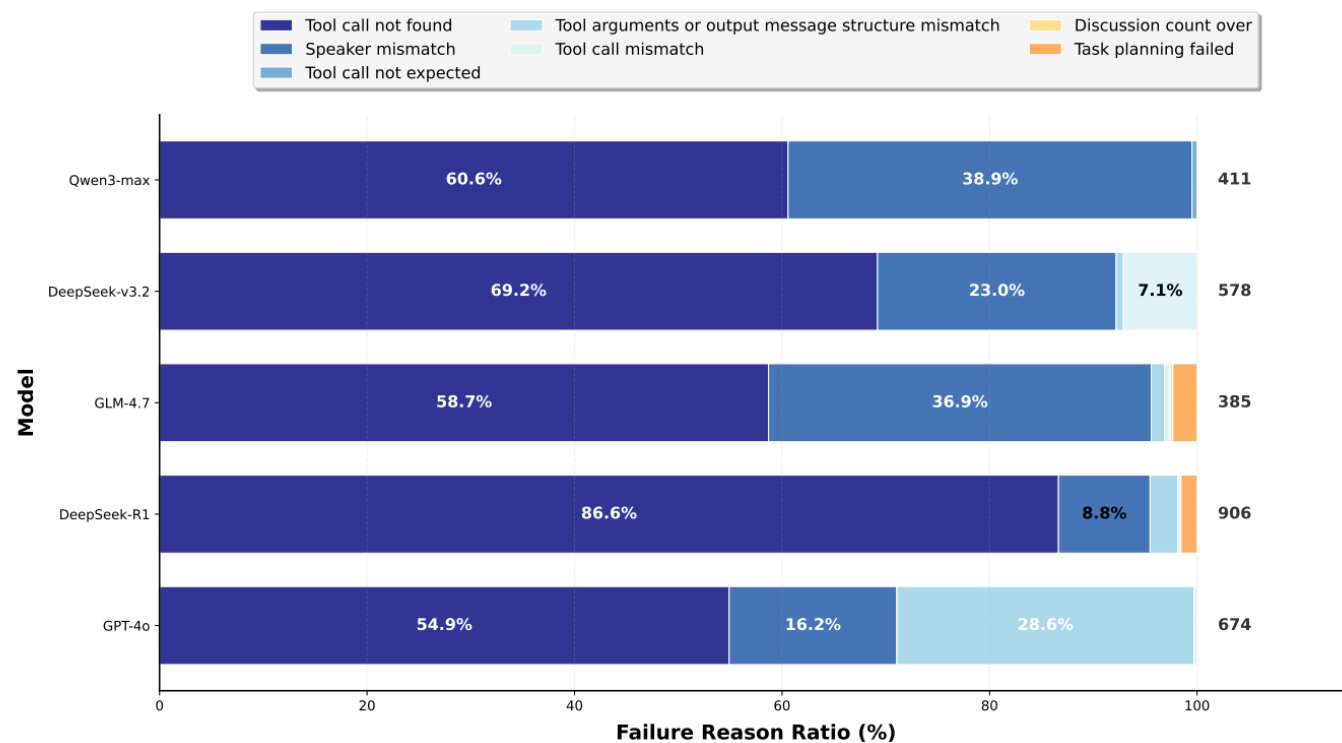
**Done at
2026.05**

Demonstration of DrSai V2

- Performance study and error analysis were conducted using LLMs at 2025
 - Tool call was the bottlenecks
- We are finalizing the DrSai V3 now, and will make a more comprehensive study



2026/10/06



2026 BESIII 十一研讨会

48

AI infrastructures at IHEP