

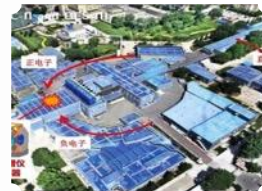
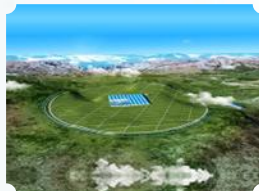


AI Platform & Agents

for Large Scientific Facilities

17th FCPN/L Workshop, Lyon, France

Zhengde Zhang
Institute of High Energy Physics, CAS
National High Energy Physics Data Center
2026.07.02





Useful AI is not just a model (ML/DL/LLM). It needs a platform, and it needs agents that can connect models with real scientific workflows.



Large collaborations already share detectors, software and computing. The next reusable layer can be scientific agents.

Data-intensive

PB-scale data streams require search, selection, validation and reuse across experiments.

Software-intensive

Physics workflows depend on ROOT, BOSS, GSAS-II, simulation, fitting and experiment-specific code.

Collaboration-intensive

A shared agent layer can preserve knowledge and make analysis workflows more reproducible.

Key message: not an AI chatbot, but an AI-enabled collaboration layer for large facilities.



HepAI Platform (HAI)

- Model services
- AI-ready data
- Elastic computing
- Secure execution and audit
- Agent runtime environment



OpenDRSAI / Dr.Sai

- Agent harness
- Scientific tools
- Memory and RAG
- Domain skills
- Strategy

Infrastructure makes agents useful; agents make infrastructure accessible to scientists.



A one-stop platform **designed** for real facility workloads, not only interactive chat.

Next-gen AI Infrastructure

Models

Unified access to local and cloud LLMs/VLMs through a common API layer.

Data

Metadata, provenance, data discovery and AI-ready data pipelines.

Computing

Elastic GPU/DCU/NPU resources and isolated environments for long tasks.

Agents

OpenDrSAI, remote workers, tools, skills and workflow orchestration.

HaiChat Assistant →

Dr.Sai Physics Agent

HaiAcademic

AI Computing Cluster

The platform connects models, data, computing resources and experiment software into one operational environment.

Real-time Data Processing

Intelligent Model Training

High Performance Computing

Model Infrastructure @ HAI platform



- Model infra provides unified LLMs & VLMs API services and billing in IHEP.
- Achieved through a self-developed **Distributed Deploy Framework (DDF)** software.

HepAI Model Plaza

Welcome to Model Plaza

Discover, compare and use cutting-edge AI models to accelerate high energy physics research

65+ Available Models 10+ Providers

Search models...

2026-05-20 模型广场全新改版上线, 新增概览页、分类浏览、热门榜单等功能

Featured

deepseek-v4-pro Free

deepseek-v4-flash Free

claude-opus-4-7 Anthropic

gpt-5.5 OpenAI

kimi-k2.6 Moonshot

glm-5.1 Zhipu

Deepseek-v4-flash/pro, glm-5.1, kimi-k2.6, minimax-m2.7, panshi-s1, intern-s1, etc

base_url: <https://aiapi.ihep.ac.cn/apiv2>

- Support self-deployed models to ensure data security.

- Access to external models to enable cutting-edge research

- Similar to OpenRouter, but includes other functionalities (distributed agents connections based on ours proposed Infinite Function protocol)

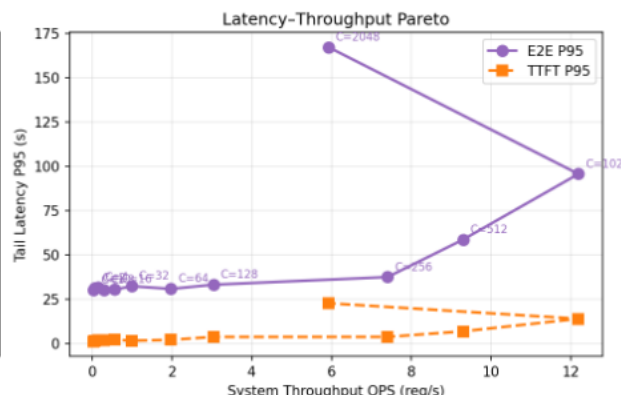
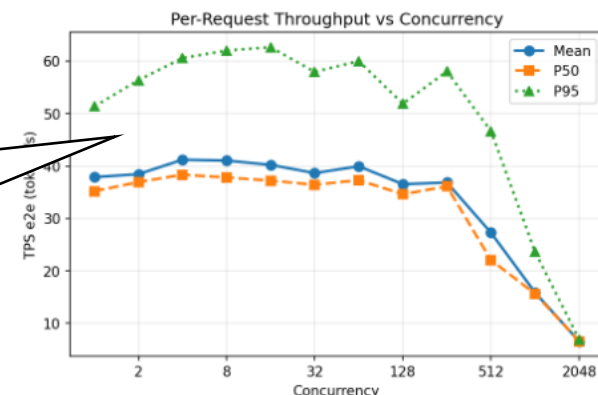
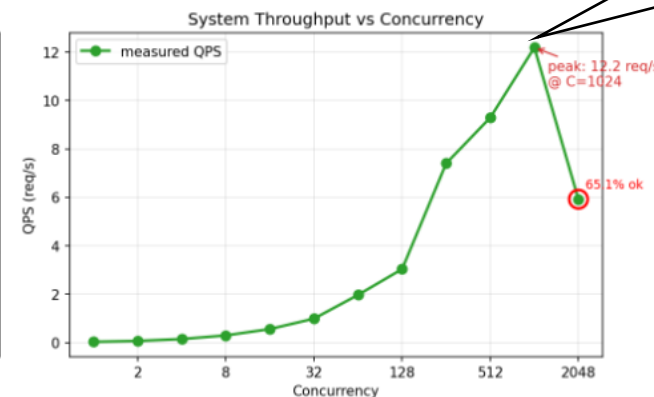
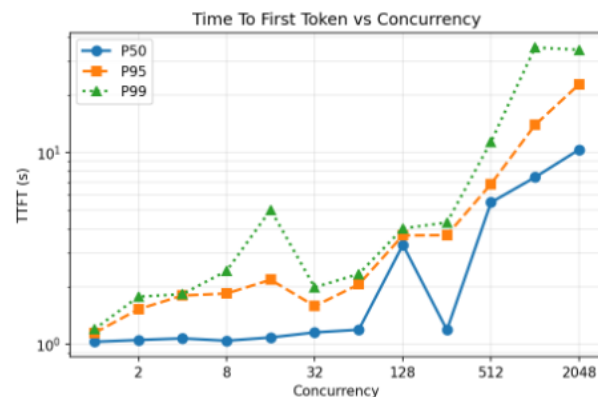
Capacity of Model Infrastructure



■ Performance and Capacity (deepseek-v4-pro for example)

Up to 1024
concurrent

Concurrency Sweep — ddf_sweep_20260520_000931



~50 TPS (Token
per seconds)

Test Result

- Once agents become widely used, model infrastructure becomes a production system.
- It must care about concurrency, latency, throughput and reliability, not only model quality

Usage Report of Model Infrastructure





IHEP LLM Startup
¥1000 LLM credit has been issued to your platform balance

May 22, 2026

65+
Available Models

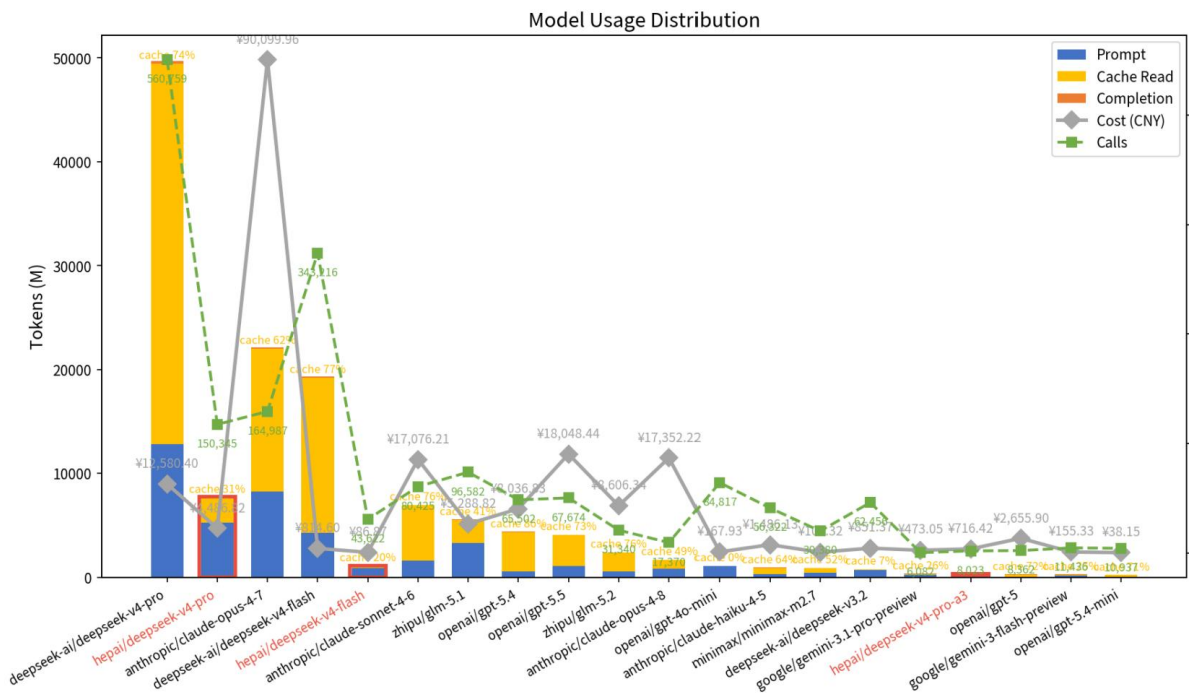
10+
Providers

Browse All

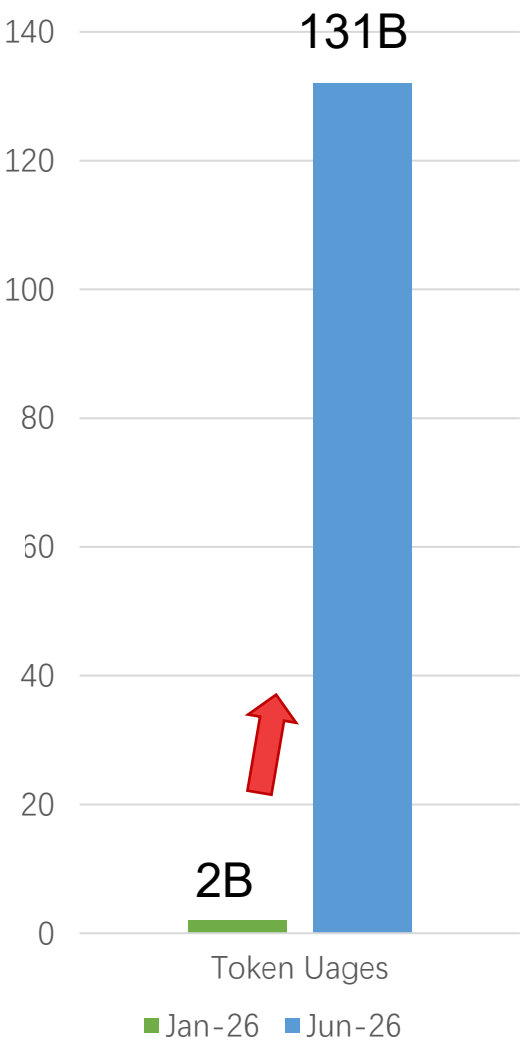
Docs

- For every employee, post-doc and students.
- Equivalent to ~55M tokens of deepseek-v4-pro

Static data (2026-06-01 00:00 UTC -- 2026-06-30 23:59 UTC)

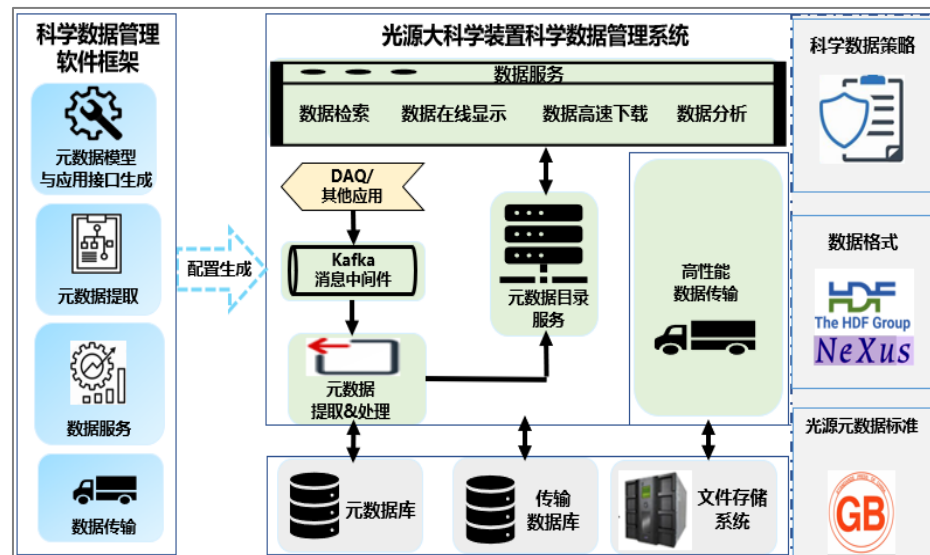


- Users: 1335
- Requests: 2.3M
- Tokens: 131B





A **large-scale facility data system** refers to the comprehensive framework and mechanisms for the collection, storage, processing, analysis, and utilization of data within large scientific facilities. It encompasses the entire lifecycle of data, from its generation to its final application, and ensures data quality, integrity, security, and availability.



Domas Architecture



Fusion Data for AI Scientist

- Agents need to find data, understand metadata, trace provenance, access right files, and produce reusable results.



Providing AI elastic computing services on-demand to eliminate resource bottlenecks

Hardware Power:



Deployment Flexibility:

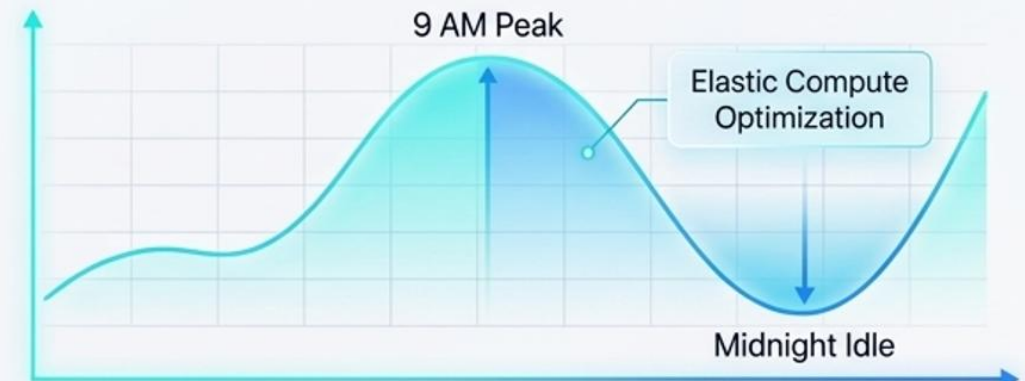
GPU: Single-machine, single-card training and deployment.

DCU: Large-scale model training.

NPU: Efficient model deployment.

Supports diverse workflows: EShell, Jupyter, Online VSCode, and VNC.

Compute Utilization Over Time



The Agent Matrix



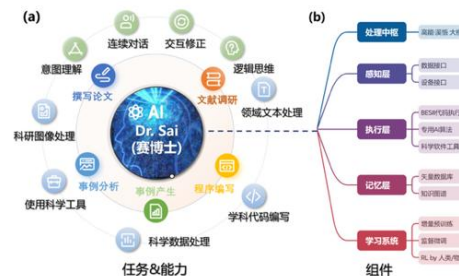
Dr.Sai Dora



AI agent for AI-Ready data.



Dr.Sai BESIII



For particle physics analysis.



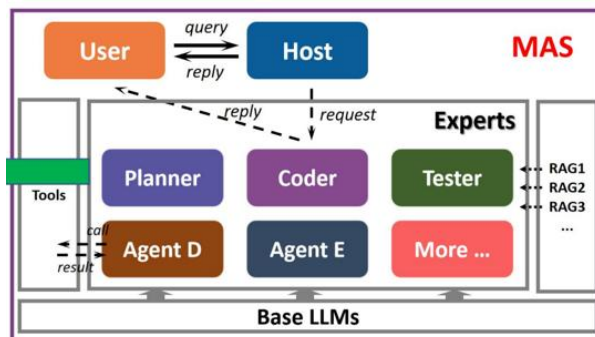
Dr.Sai Rongzai



For neutron data processing.



OpenDrSai



AI agent **framework** for scientific research on large facilities.

Claude Code (external)
OpenClaw (external)
Hermes (external)
and more..

One framework, multiple facilities, different scientific workflows.



■ DORA(Data ORchestration Agent) for AI-Ready data



Dr.Sai Dora
Data Orchestration

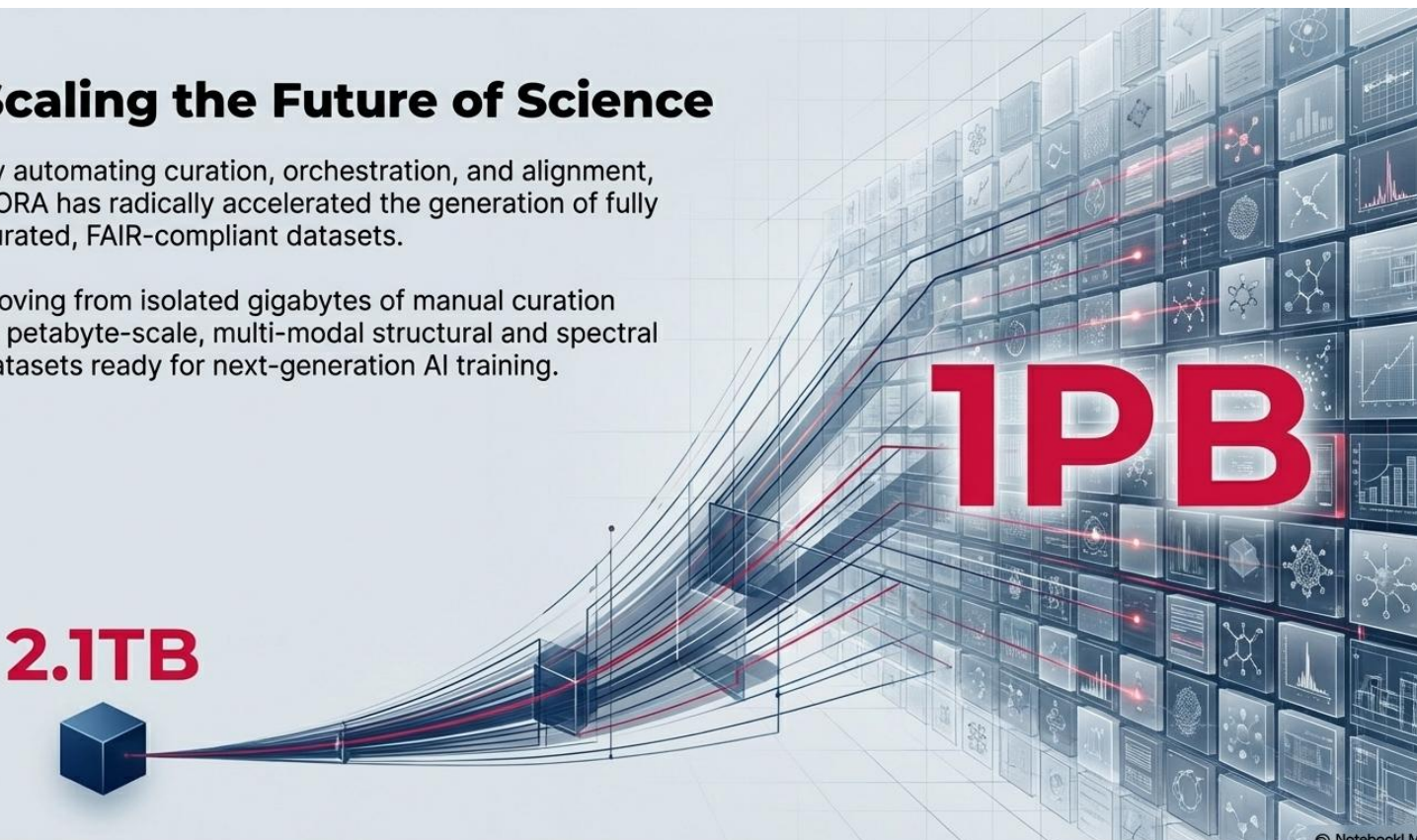
Focus: AI-Ready Data Pipeline.

Role: Automates the orchestration, standardization, and fusion of scientific datasets across multiple experimental stations.

Scaling the Future of Science

By automating curation, orchestration, and alignment, DORA has radically accelerated the generation of fully curated, FAIR-compliant datasets.

Moving from isolated gigabytes of manual curation to petabyte-scale, multi-modal structural and spectral datasets ready for next-generation AI training.



2.1TB

1PB

NotebookLM



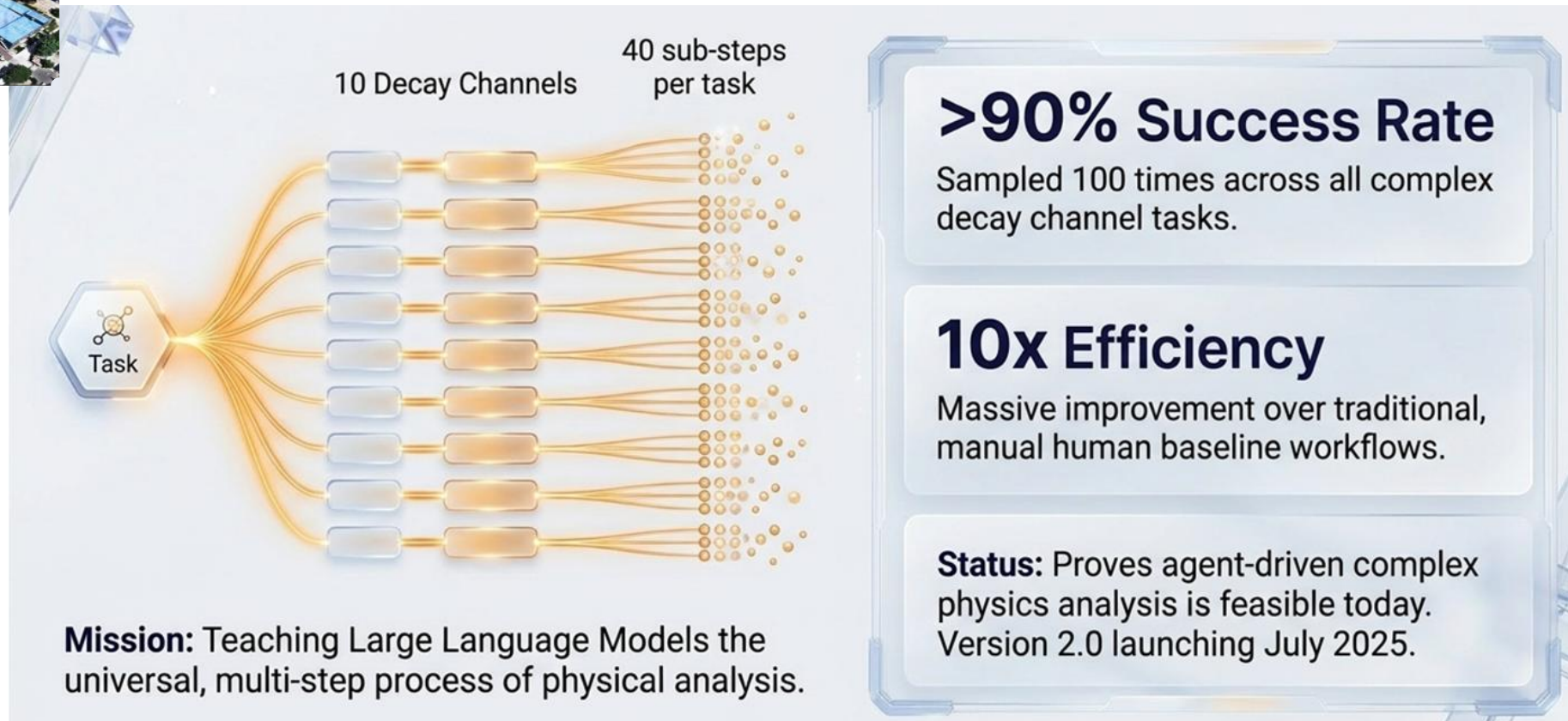
arXiv:2604.22541, submitted 24 Apr 2026

<https://drsai.ihep.ac.cn>

Real-world physics analysis from natural-language goals



For BEPC (Beijing Electron Positron Collider) & BESIII Experiment



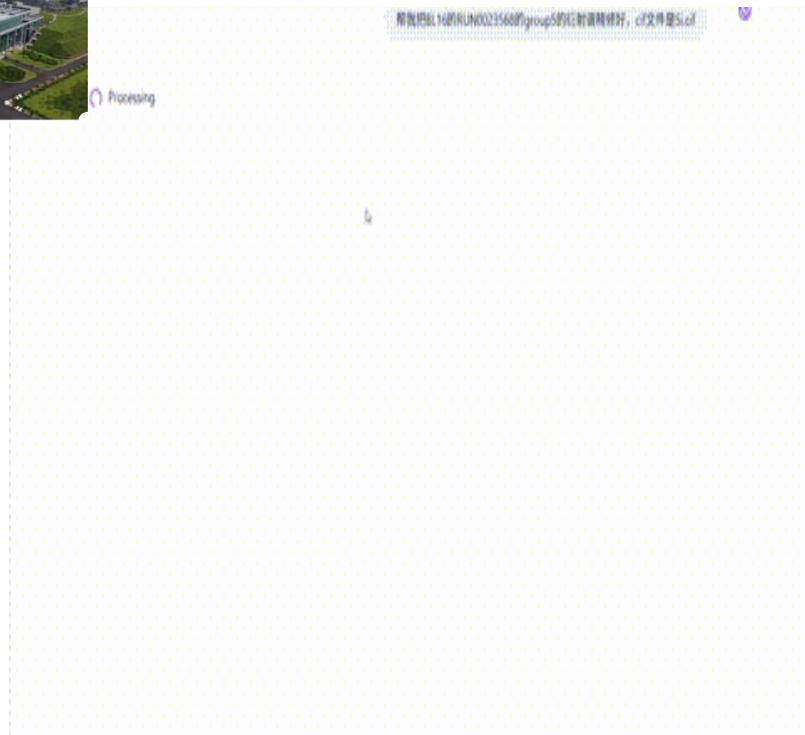


Autonomous Rietveld refinement of neutron diffraction data

arXiv:2605.13911, submitted 13 May 2026

<https://rongzai.ihep.ac.cn>

For CSNS (China Spallation Neutron Source)



Agent design

LLM + specialist refinement knowledge base
+ GSAS-II execution interface.

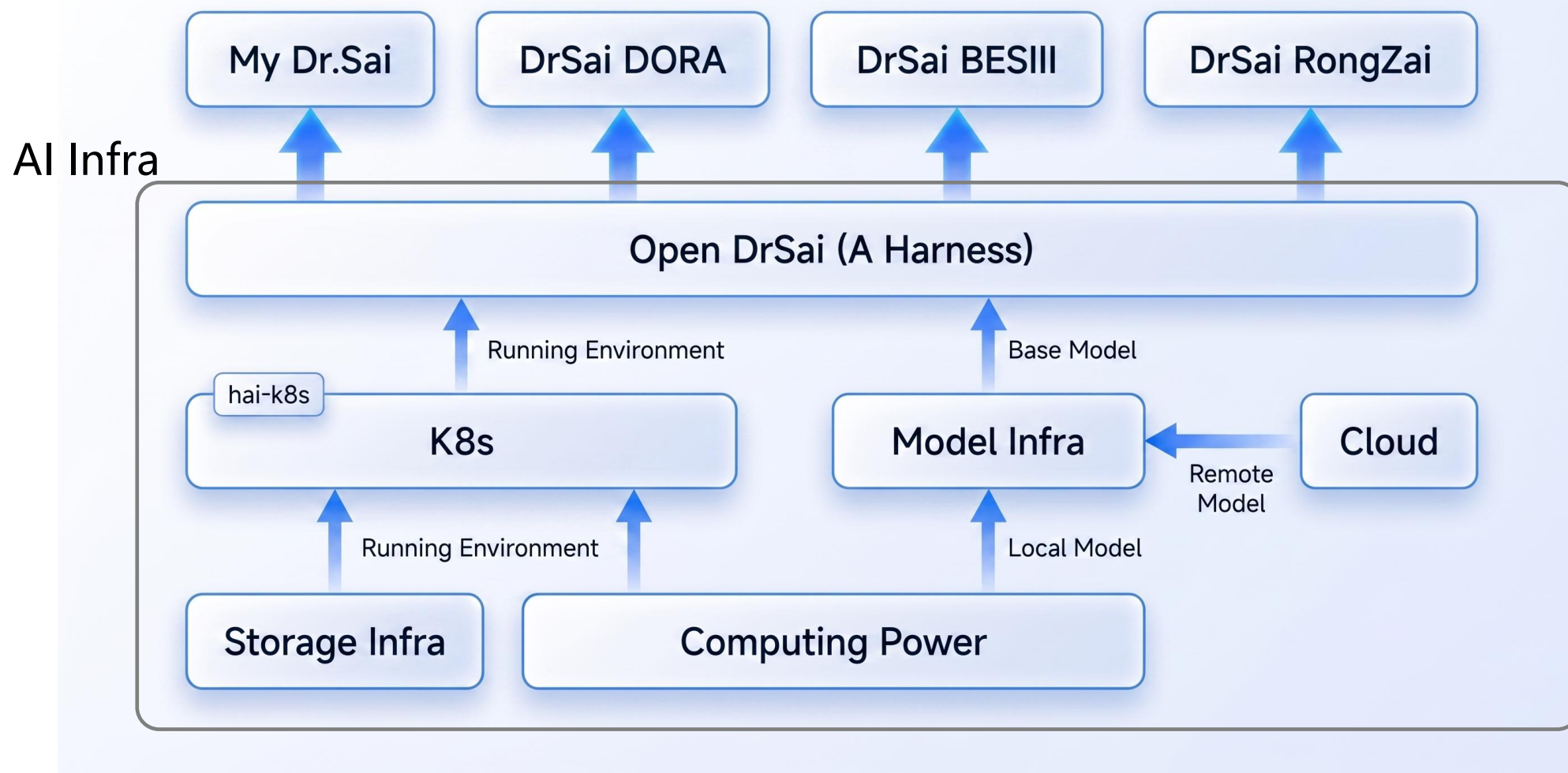
Deployment

Deployed at CSNS and open for external user
registration.

Closed loop

Plan, refine, inspect intermediate results,
adjust strategy, back up and roll back.

Human level results ($R_{wp} \sim 5\%$), $\sim 5x$ faster



- Agent capability is a system capability.
- It depends on the model, but also on data, tools, knowledge, runtime environments and platform integration.



A practical next step: build one French-Chinese scientific workflow agent prototype.

1. Joint benchmarks

Define reproducible HEP workflow tasks for agent evaluation.

2. AI-ready datasets

Share metadata standards, provenance and validation practices.

3. Tool integration

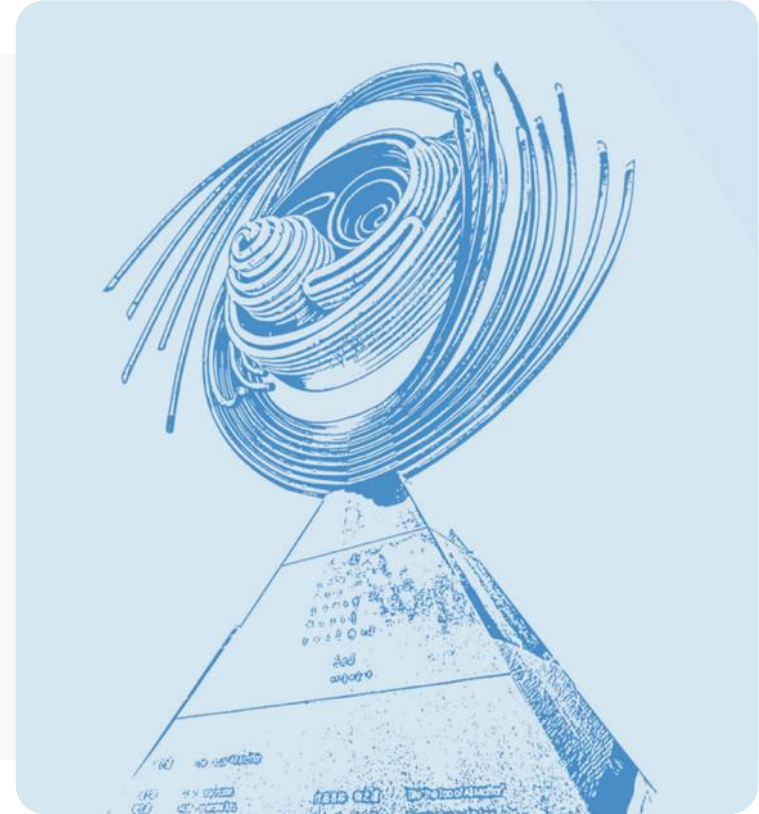
Connect agents with experiment software and computing centers.

4. Pilot workflow

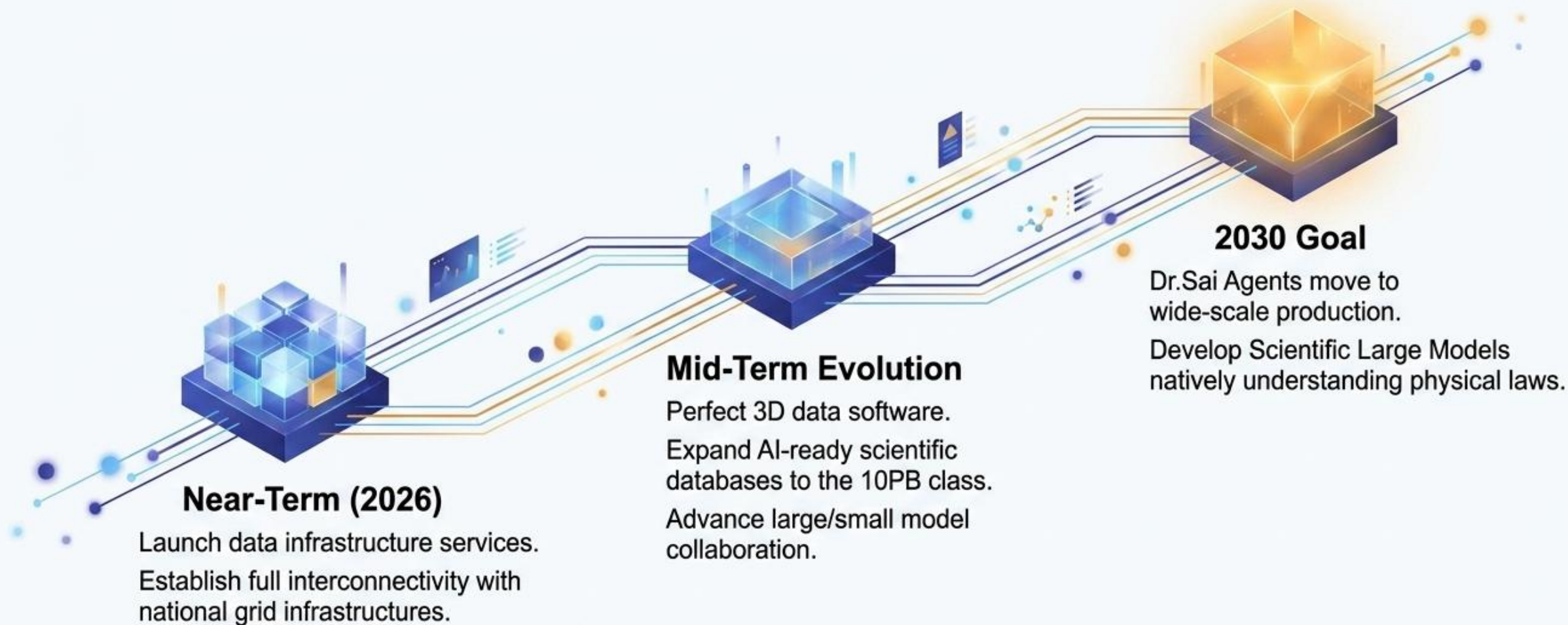
Select one analysis workflow and turn it into a reusable agent skill.

Conclusion: the future is not a chatbot for physicists, but a trusted platform where agents use models, data, computing and experiment software to perform real scientific workflows.

Thank You!



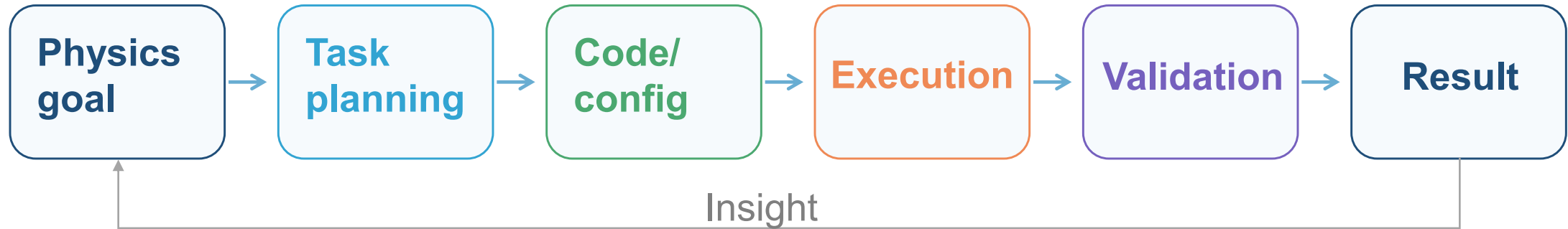
Roadmap to 2030: The High Energy AI Ecosystem



Connect & Build With Us:

- HepAI Platform: ai.ihep.ac.cn
- Dr.Sai Agent: opendrsai.ihep.ac.cn
- Open Source: github.com/hepai-lab/drsai
- Contact: zdzhang@ihep.ac.cn

The bottleneck is the workflow



Scientific work is not a single question-answer turn. It is a long chain of domain decisions, software calls, data checks and expert validation.

LLMs alone are not enough. They must be connected to data, tools, computing and domain knowledge.