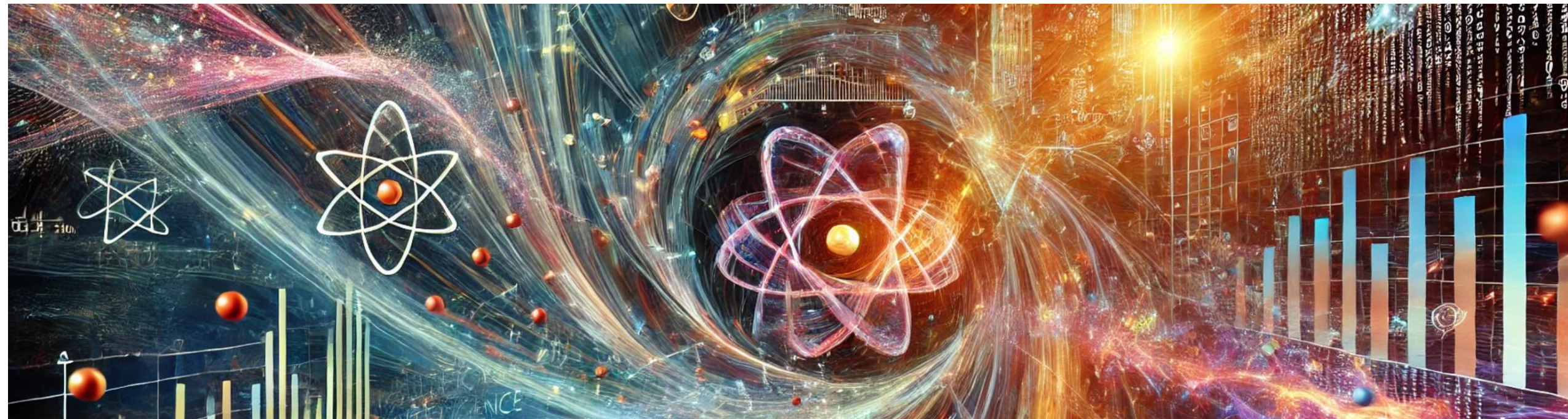




Computing Center, IHEP, CAS  
National HEP Data Center



# Dr. Sai物理分析智能体系统

Bolun Zhang, Zhengde Zhang

On behalf of Dr.Sai working group

2025.8.25 in Changchun



# Part 1

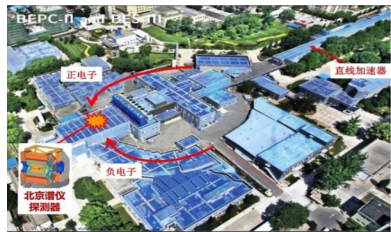
## Introduction

# Challenges in HEP



- Data Magnitude:  $\sim 10+$  PB
- Growth Rate: Accelerated with new techniques
  - BEPCII  $\rightarrow$  BEPCII-U (May 17, 2025): Luminosity Increase  $\sim 300\%$  at 2.35 GeV
  - HEPS (2025):  $\sim 800$  TB / day

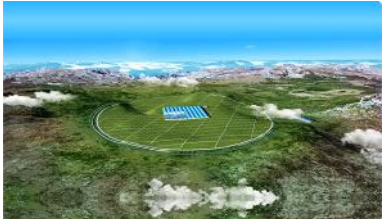
## Particle physics



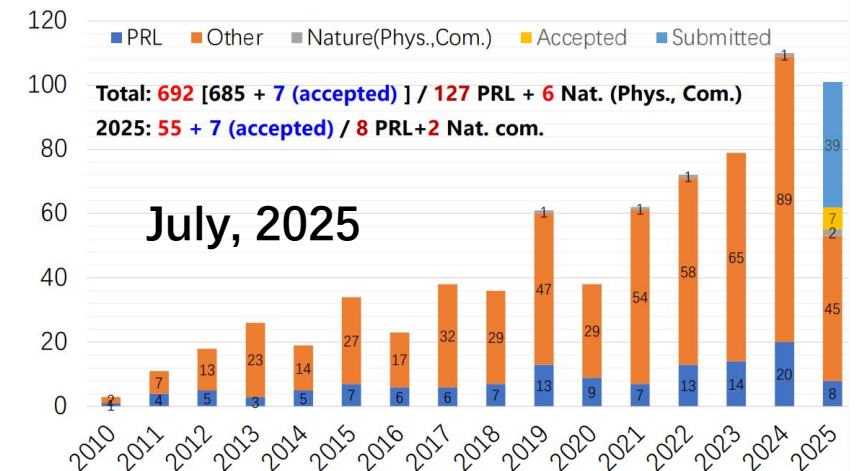
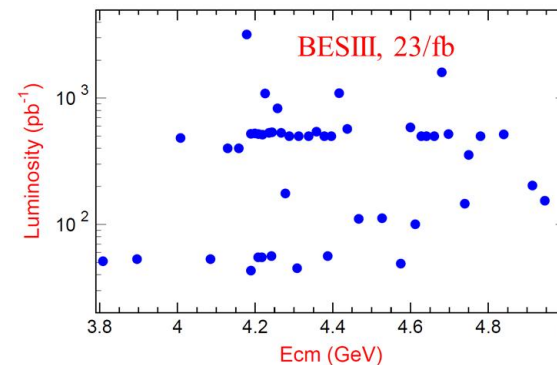
## Neutron science



## Astrophysics



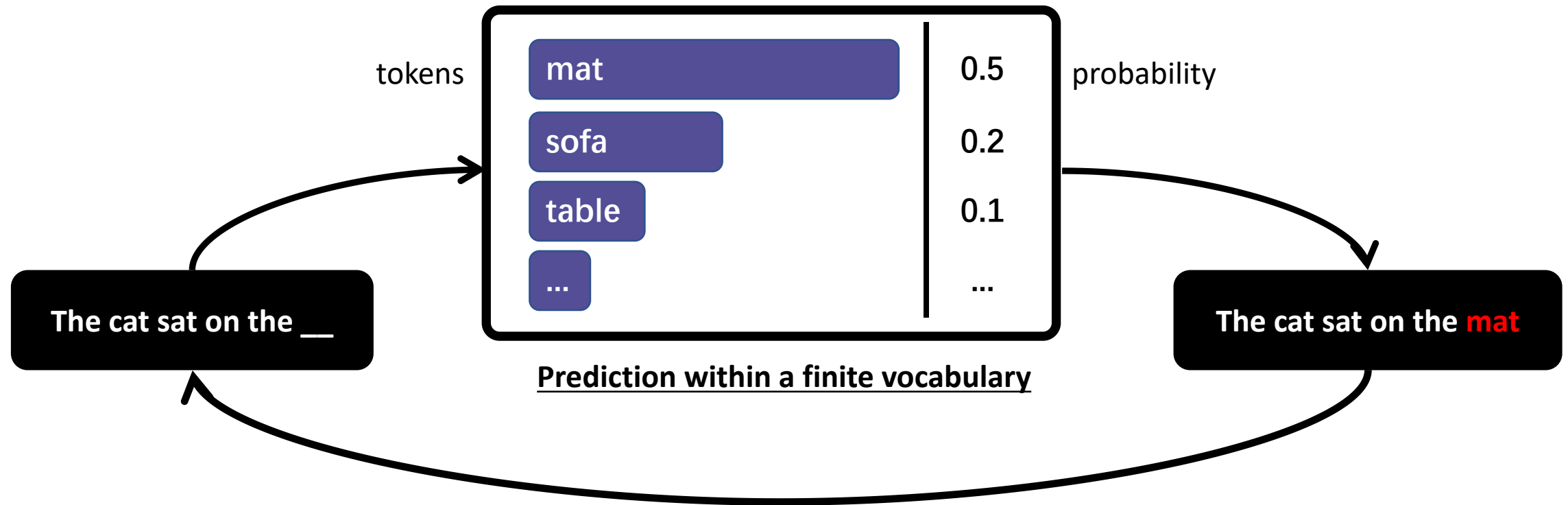
## Photon science



- $\sim 700$  physics results with  $\sim 700$  people in 16 years
  - One result normally took  $\sim 3$  years
- HEP needs a more efficient tool to investigate data



- Large Language Model, which refers to a deep learning model trained on extensive textual data. The core mechanism of most LLMs is to **predict the next word**.



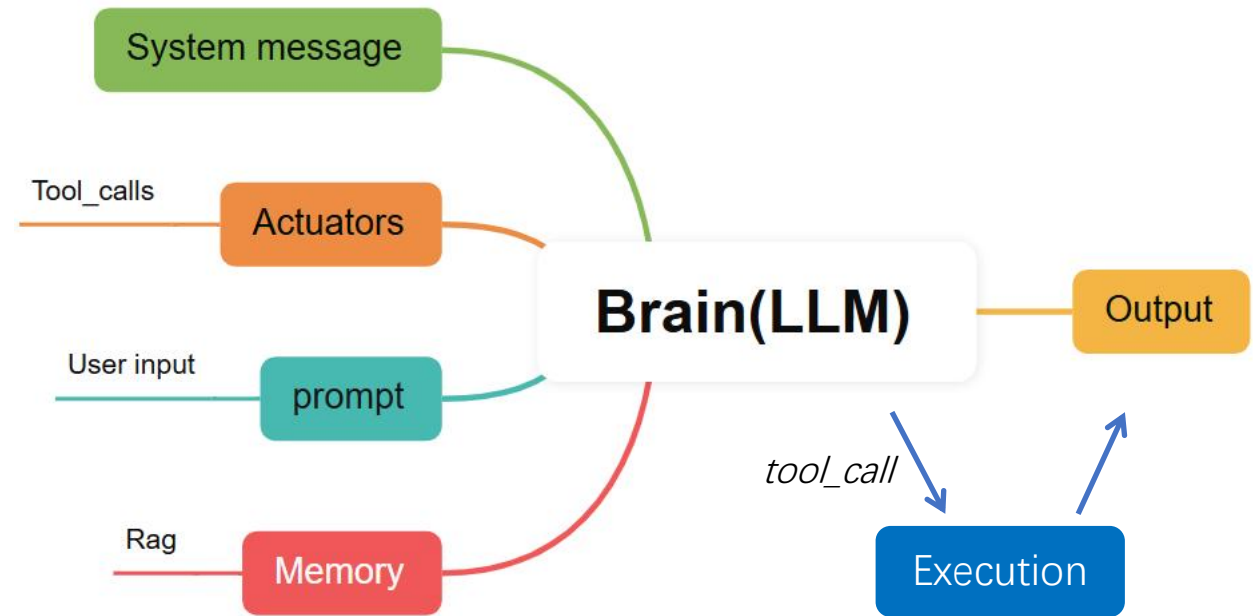
# From LLM to MAS



An entity that can perceive its environment, make decisions, and take actions in order to achieve certain goals or sets of goals.

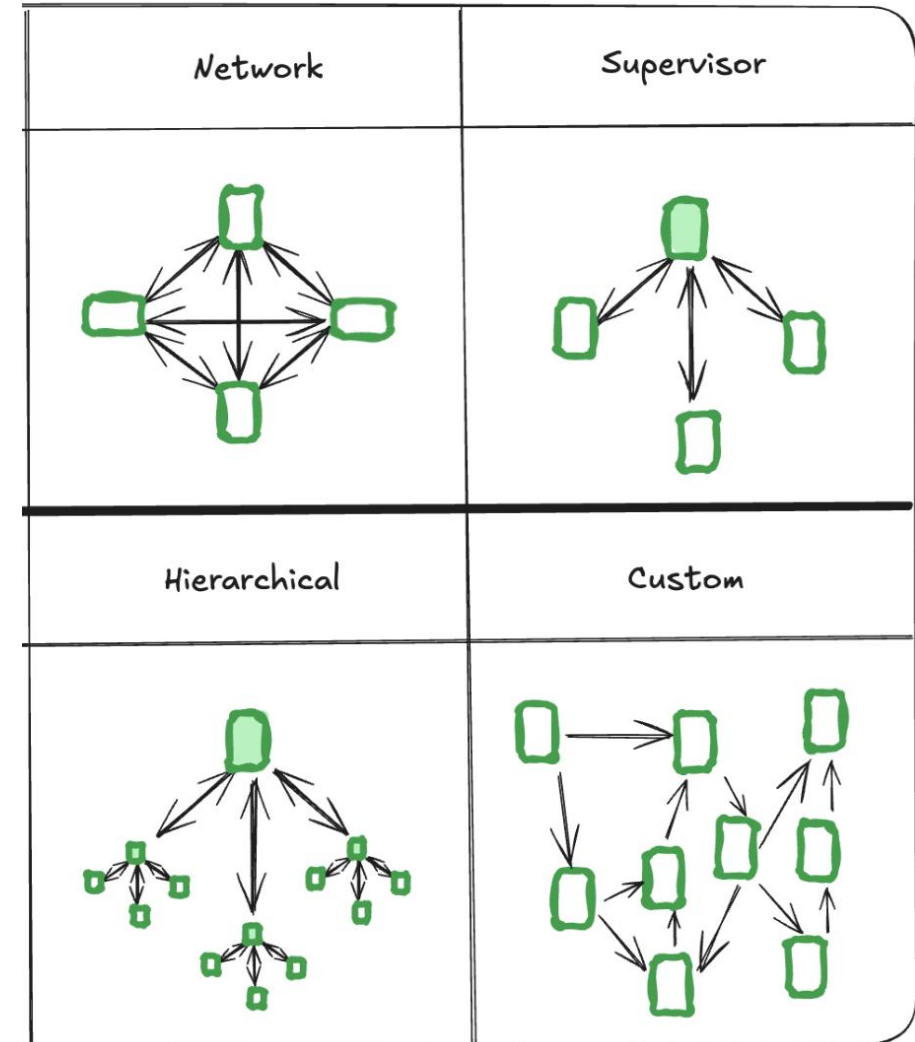
- Perception: multimodal -> text
- Decision: process text
- Action: reply or tool\_call

- Agent = LLM + tools
- DIY reply\_function to design the LLM output procedure



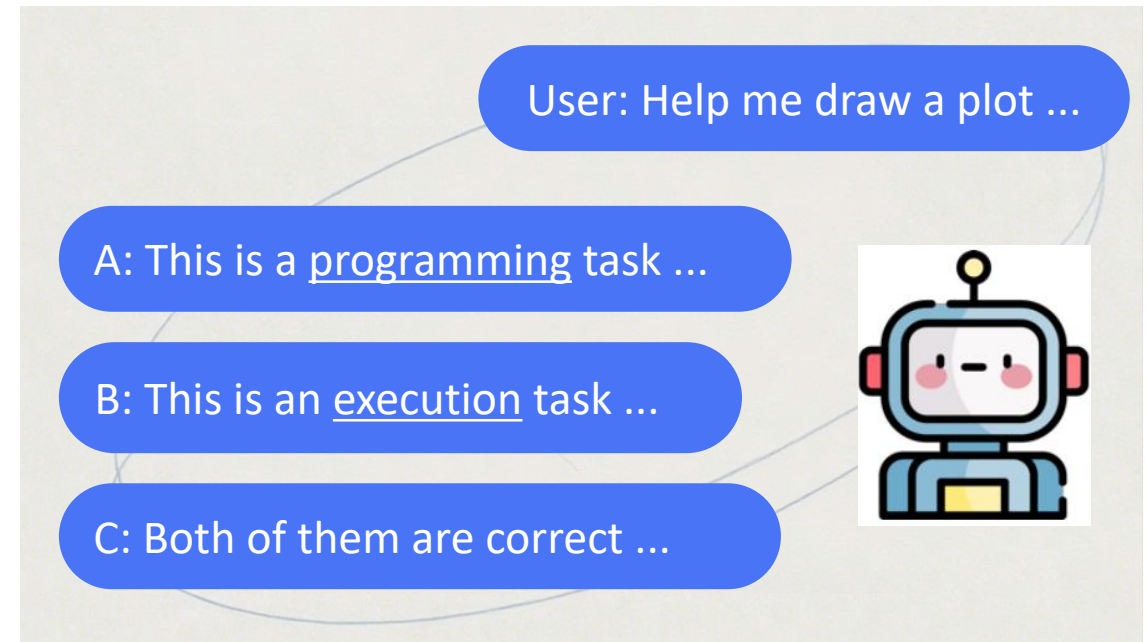
## Multiple theoretical frameworks:

- **Round-table:** Members speak in a fixed sequence.
- **Supervisor:** Each agent communicates with only one supervisor agent, which selects the next agent to be invoked.
- **Hierarchical:** Supervisor can act as an intermediate layer.
- **Custom:** Each agent interacts only with predefined specific agents.
- **Network:** Each agent can freely choose to interact with any agent.



## \* Why Do Multi-Agent LLM Systems Fail?

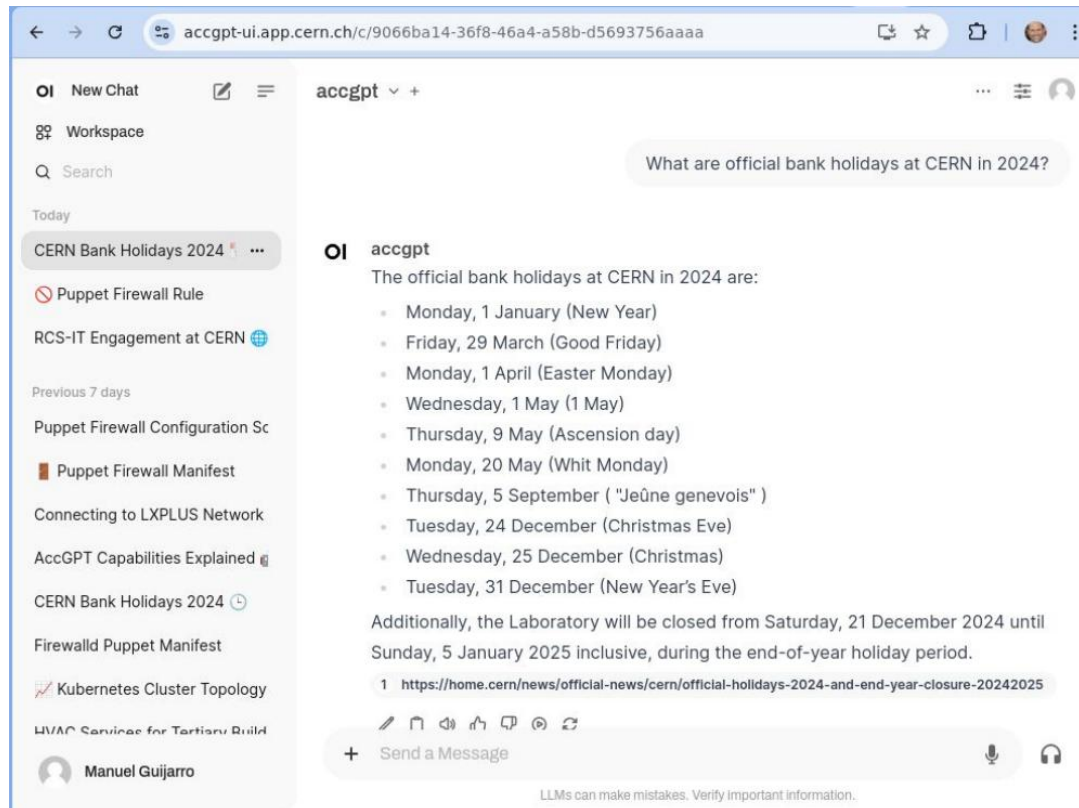
- Specification Issues
  - Unclear role definitions
  - Flawed interaction protocols
- Inter-Agent Misalignment
  - Persisting false assumptions
  - Ignoring critical inputs
- Absence of Task Verification
  - Surface-level validation only
  - No error-correction protocols



# LLM applications



- AccGPT: A CERN Knowledge Retrieval Chatbot
- ChatLas: An AI Assistant for the ATLAS Collaboration

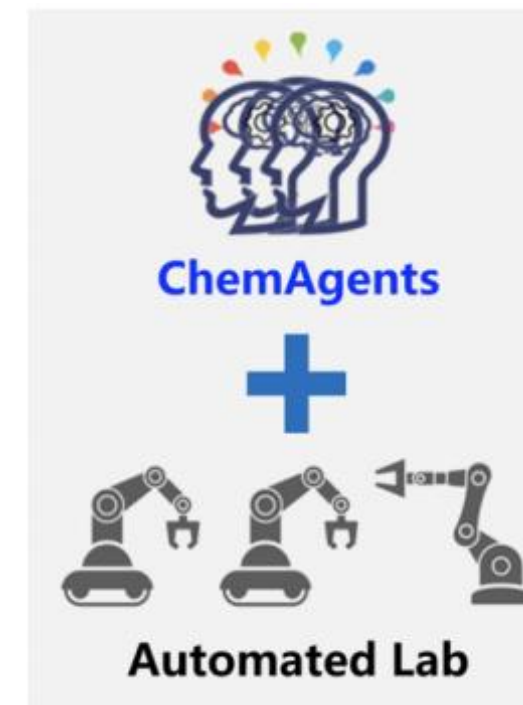
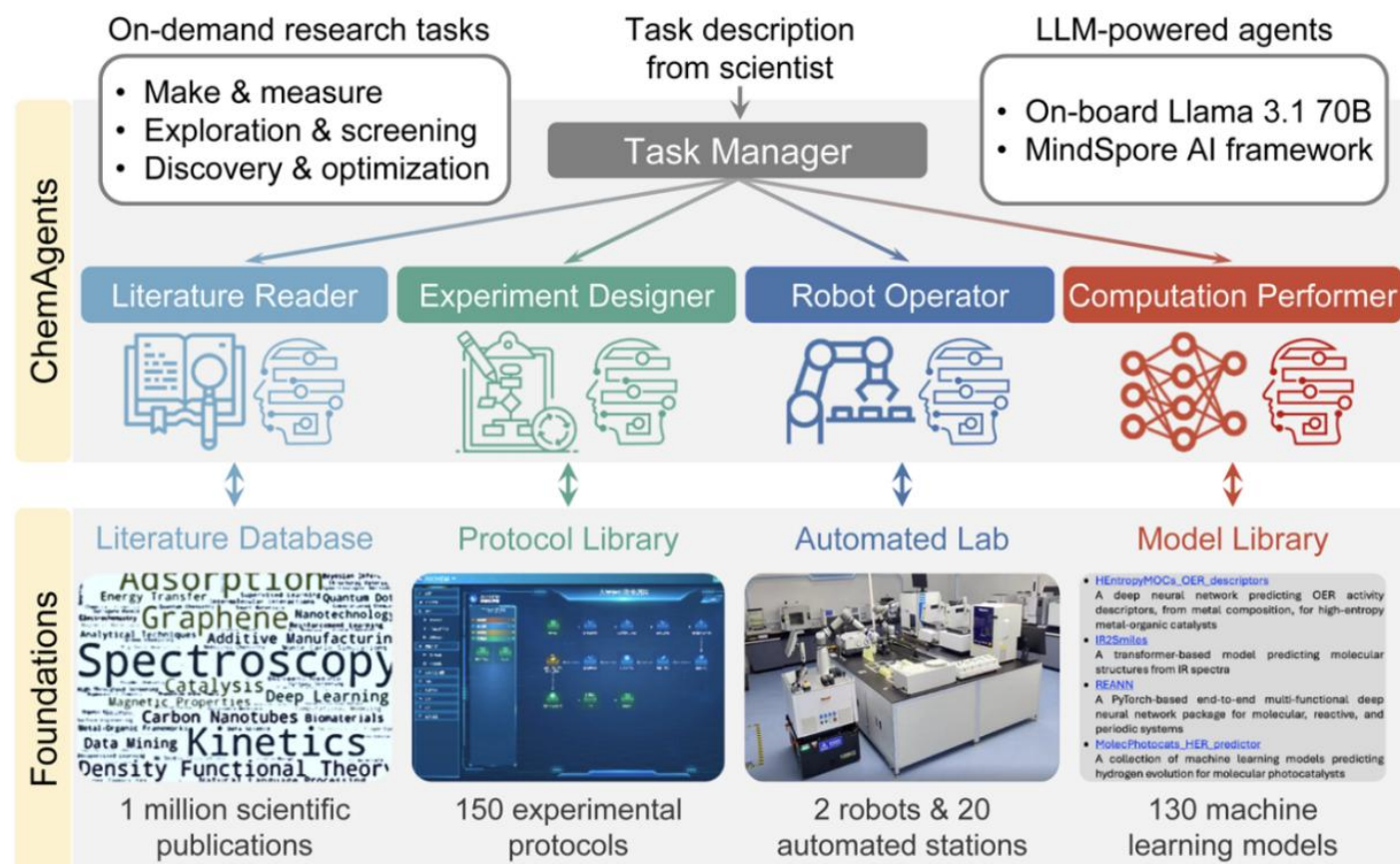


- RAG + LLM
- Internal knowledge retrieval

# LLM applications



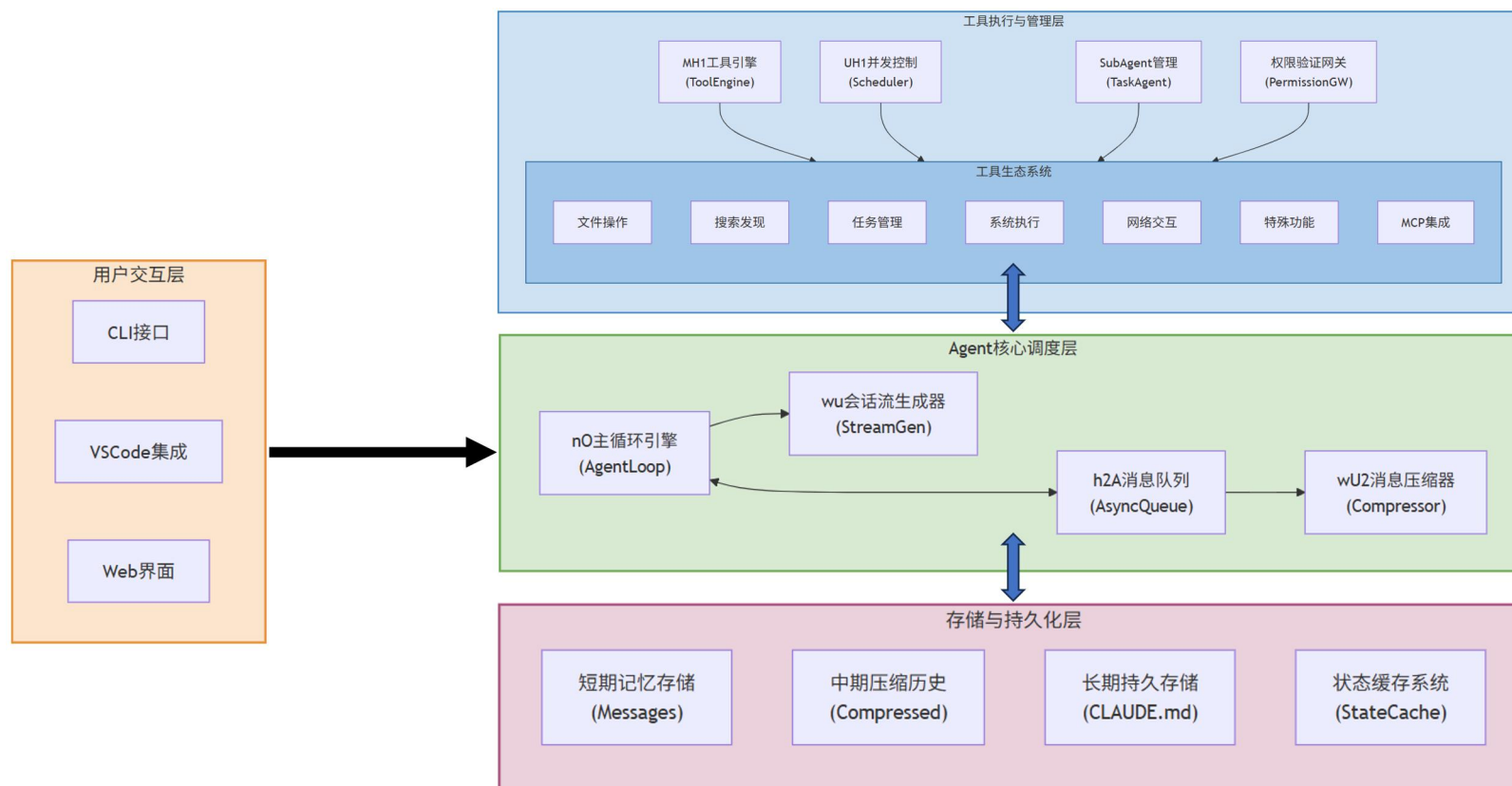
- A Multi-Agent AI (USTC) + automated Lab for Chemistry
- Component: Reader + Designer + Robot Operator + Executor, + DSL



# LLM applications



- Claude-Code: A revolutionary intelligent programming tool
- Flexible task planning, agent configuration, and tool calling



Claude

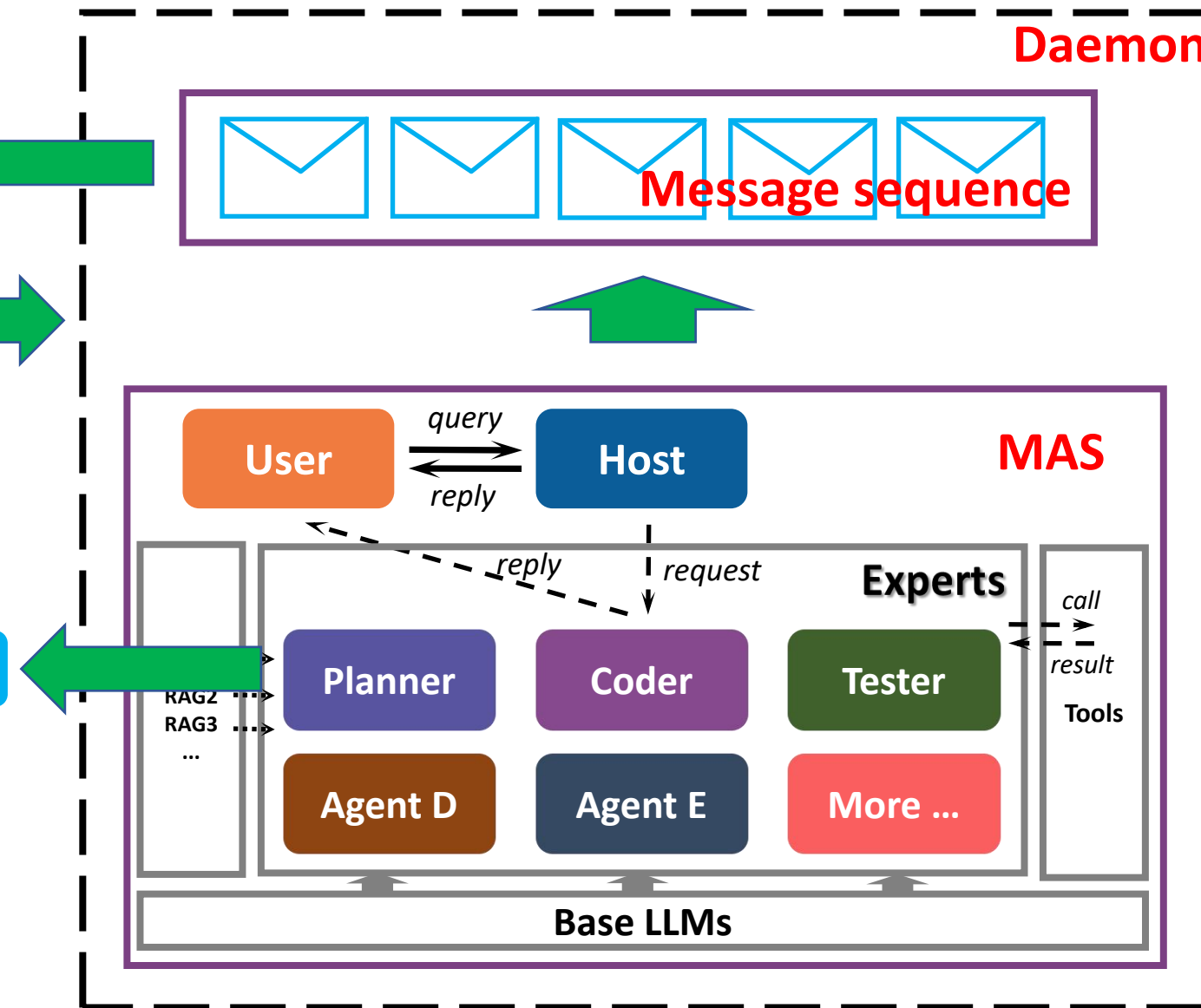
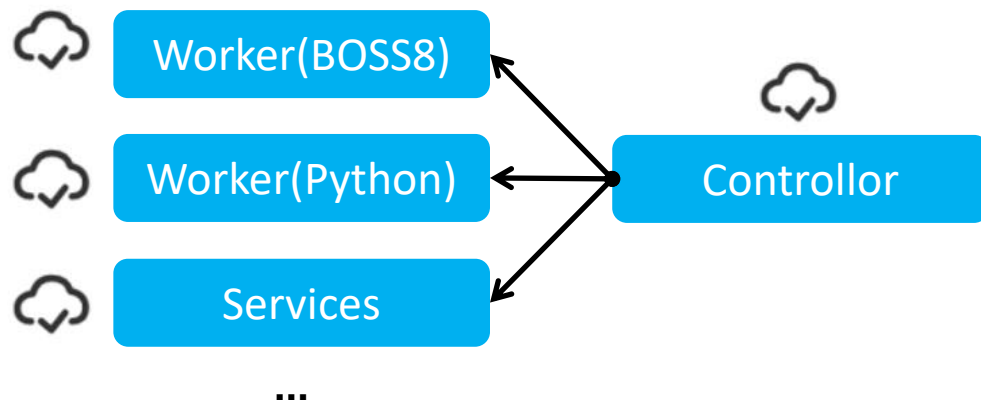
CLAUDE  
CODE



# Part 2

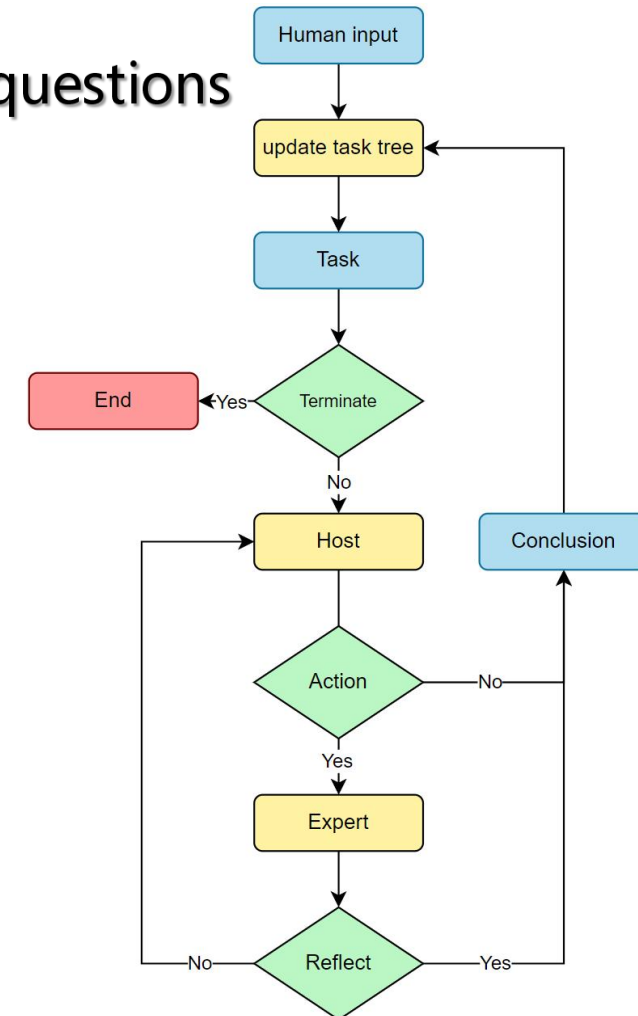
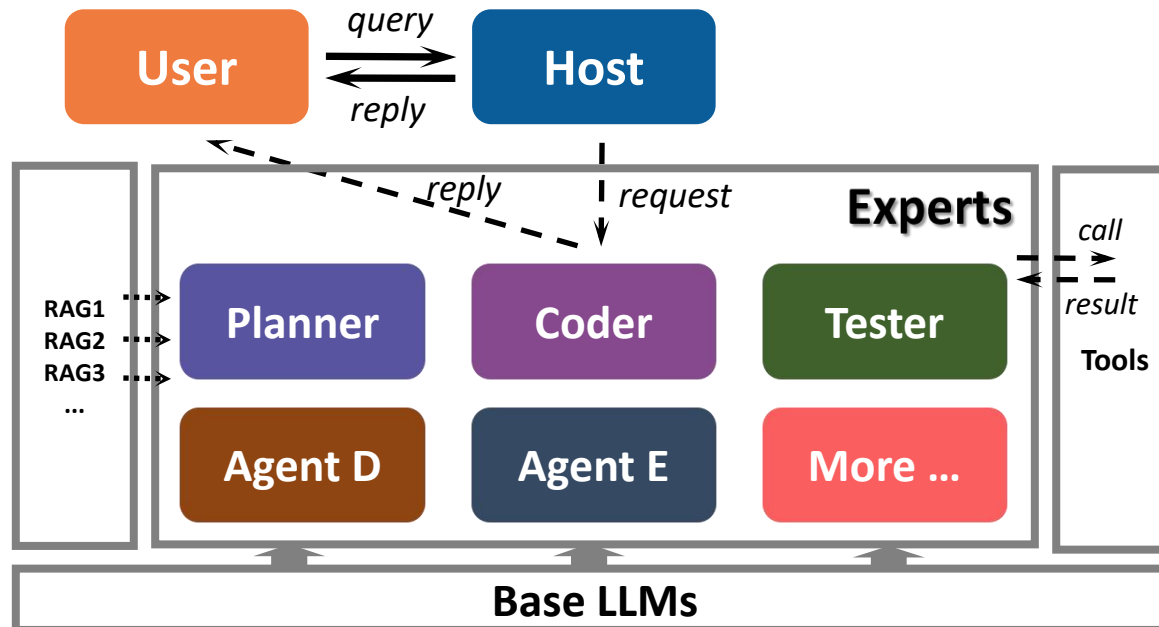
## Dr. Sai - assistant for BESIII

# Overview of Dr.Sai Agent



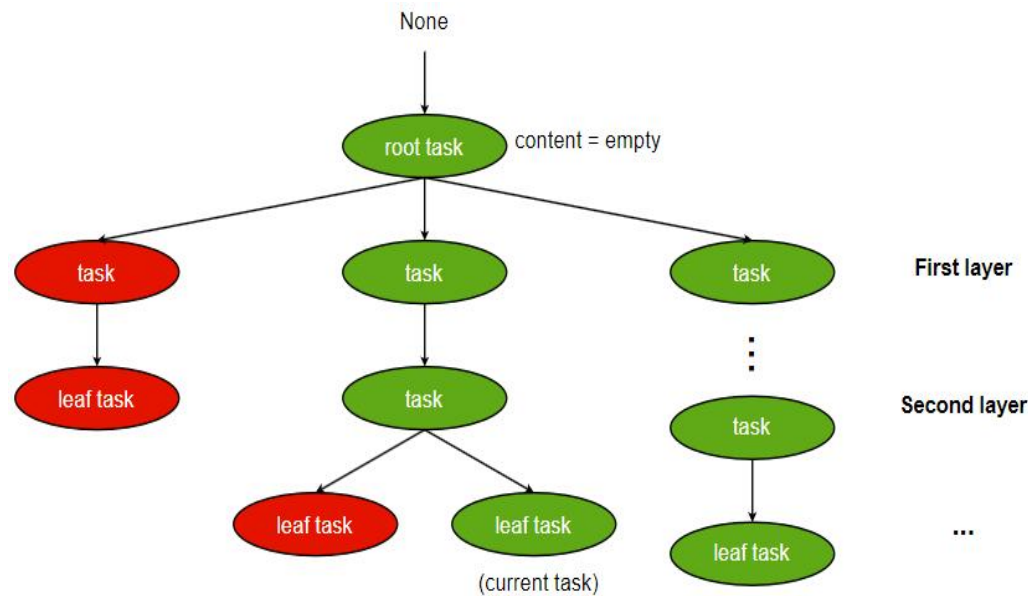


- [Talk-reasoner architecture](#) based on [AutoGen](#):
  - `Host` answers simple questions, `Expert` handles professional questions
  - Advantage: quicker response + more precise task assignment
- Task System:
  - Simulate questioning behavior

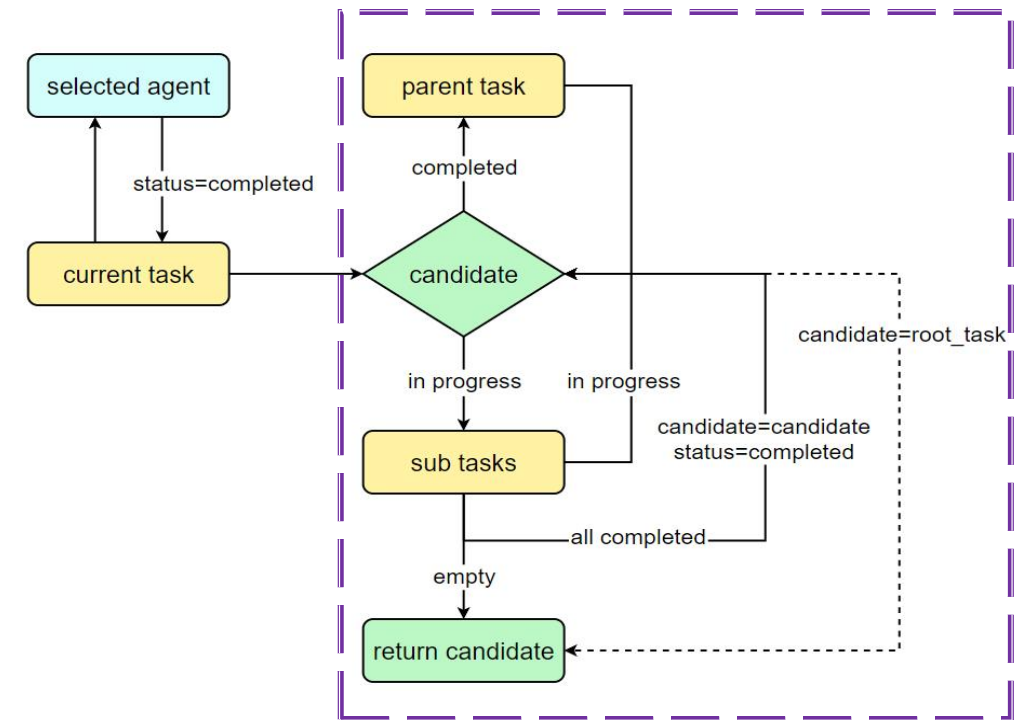




- Hierarchical task system
  - Suitable for correlated tasks in high-energy physics
  - Planner&Human can launch/modify task properties



**Task tree structure**

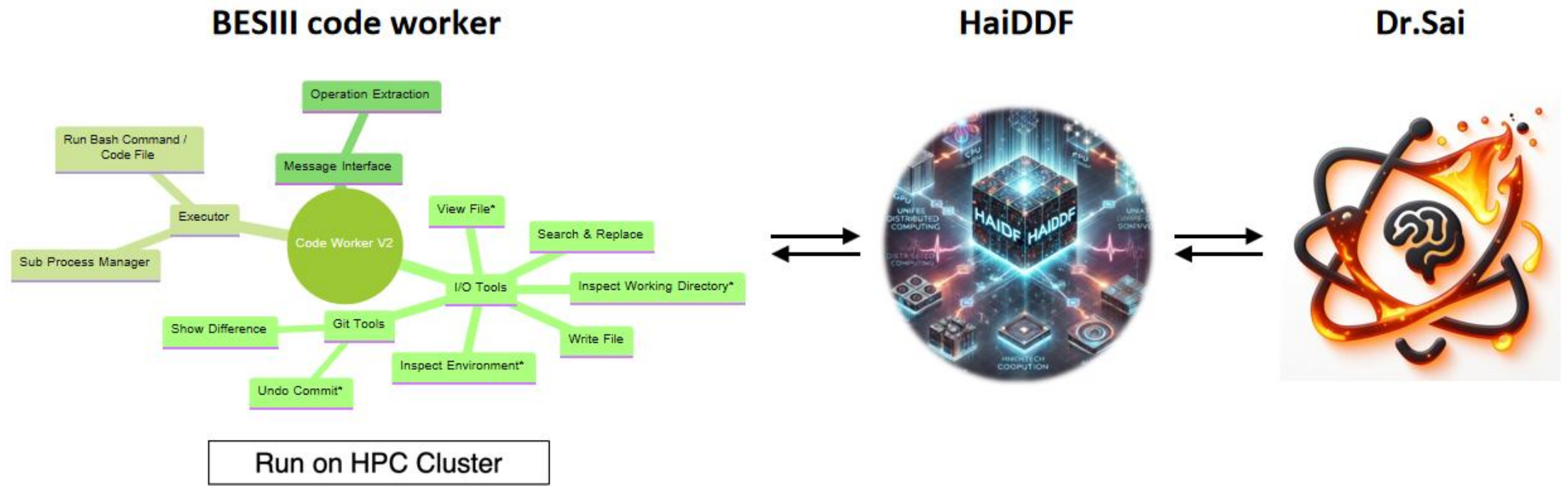


**Task proceed flowchart**

# Dr.Sai Agent - Worker



- I/O Tools (files and data), execute code blocks and manage jobs
- Implemented through HaiDDF, supporting distributed one click deployment
- Connecting to BOSS8 and python environment





- Base model:
  - Deepseek-v3 - good at filling a template without generating any unexpected content.
  - Deepseek-r1 - more complex tasks
  - Xiwu - a high-energy physics mode under active development, not yet available
- RAG: retrieve extra info to teach agent how to proceed domain tasks.
- Tool\_call: call functions to ensure a valid result
  - Mapping – arguments to full scripts (analysis algorithm, joboption scripts, drawing scripts)

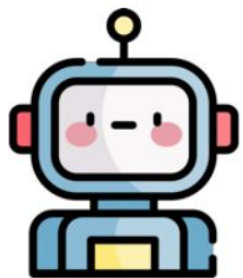
name	Base model	System prompt	RAG	Tool call
Host	Deepseek-r1	✓	-	Select_expert
Planner	Deepseek-v3	✓	-	-
Coder	Deepseek-r1	✓	DSL (Domain specific language)	Mapping_algorithm Mapping_joboption Mapping_drawPlot
Tester	Deepseek-r1	✓	-	Get_branch_name

# Constraints - Mapping



Problems in HEP algorithm generation:

- **C++ dependency on compilation environment:** Both header files and scripts need to be generated simultaneously. The code must be compiled before execution and cannot be run directly.
- **Code complexity:** Average length exceeding 1,000 lines, the code incorporates numerous domain-specific definitions. General models lack the relevant physical knowledge.



Tool call



```
json
1 {
2   "MetaData": {
3     "AlgorithmName": "VJpsiToMuMu",
4     "DecayChain": "J/psi -> mu+ mu-",
5     "Type": "algorithm"
6   },
7   "Sample": [
8     {
9       "Ecms": 3.097
10    }
11  ],
12  "Selection": {
13    "KinematicFit": [
14      {
15        "Constrained_Resonance": [],
16        "MissTrack": "",
17        "Method": "km_fit"
18      }
19    ],
20    "SaveType": "Event"
21  }
22 }
```

Necessary arguments



Analysis algorithm

Mapping



Joboption scripts

Drawing scripts

# Constraints - Anchor word



- Use keywords to enhance the model ability of task assignment. (Agent description <-> prompt)

<Role definitions>

<Instructions>

<Memories from RAG>

...

<Tools>

You can consult these experts with their descriptions:

- **Planner:** an agent to plan ...
- **Coder:** an agent to create code block ...
- **Tester:** an agent to execute code block ...

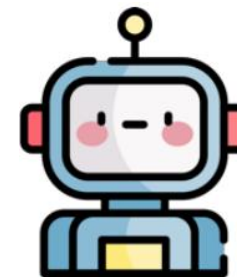
<Context>

## Anchor word

Plan  
Create  
Execute

## User Prompt

Generate a code  
Execute this code block  
Give me a plan



Host

I know who is the best one to handle this task ...

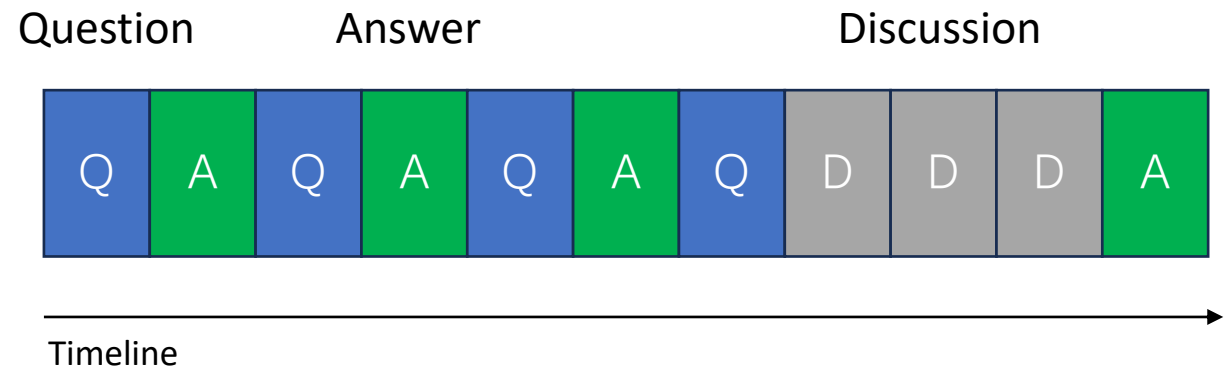
A text example to be processed by LLM

# Constraints - Messages



To address the issue of model attention being distracted by long text, we have designed message offloading mechanism

- Host: sees **all** the QA pairs
- Experts: sees the **latest** QA pair and global info
- Global info: key info extracted from chat history



**A typical conversation**

# Evaluation



## Key comparison

- speaker
- tool name
- tool call arguments / output\_msg -- Dict data structure

### 1. Task assignment:

- select wrong expert
- not use tool

### 2. Algorithm generation:

- generate wrong json card

### 3. Job submission:

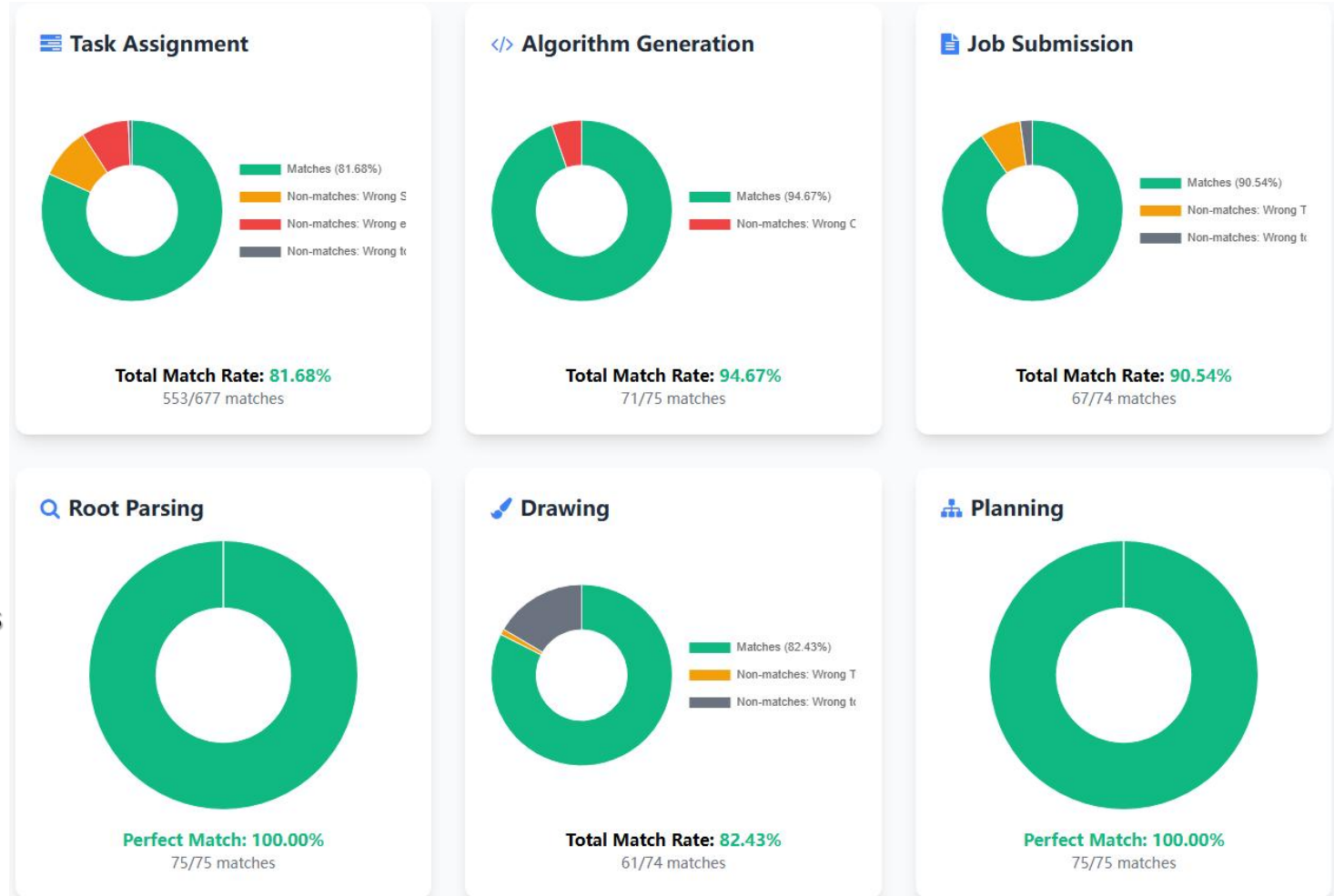
- generate wrong json card / get wrong arguments

### 4. Root parsing: -

### 5. Drawing:

- wrong order to select expert

### 6. Planning: -





- We have initially developed an AI assistant for BESIII analysis, which focuses on modeling the physical analysis process.
- Currently, it can automate some preliminary data preprocessing steps.
- The strength of this application lies in its excellent scalability, which allows it to accommodate various usage scenarios as well as a wider range of tools and tasks.
- Welcome to [drsai.ihep.ac.cn](https://drsai.ihep.ac.cn) to start your conversation!

Next to do:

- Support more decay modes
- Develop Xiwu, more component agents & tools
- Use [Magentic-UI](#) for more human interaction

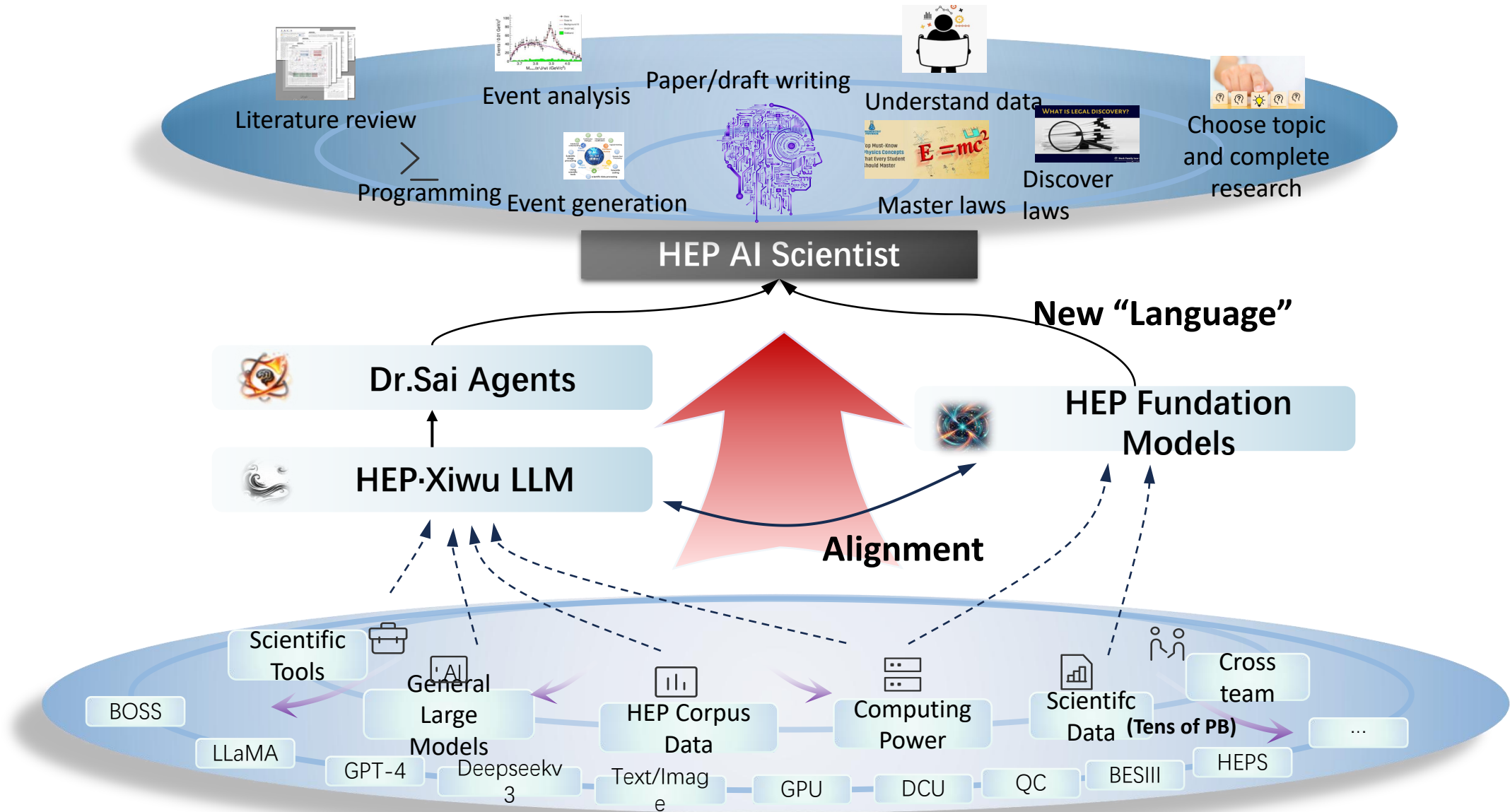
# Thanks



# Part 3

## Backup

# Roadmap





## Mapping details

- Using BOSS8 framework
- Take e/mu/pi/K/p + pi0/eta/K\_S0/Lambda0 as stable particle
- Do not support D-tag package
- Available variables: charge/E\_emc/MUC\_depth/momentum/energy/cosTheta/phi
- Only support second-order decay for now
  - $\psi(4260) \rightarrow \pi \pi [J\psi \rightarrow \mu \mu]$  ✓
  - $\psi(4260) \rightarrow \pi \pi [J\psi \rightarrow [\eta' \rightarrow \pi^+ \pi^- \eta] K^+ K^-]$  ✗
- Total 19 steps involved LLM in the whole workflow, expect  $0.95^{19} \sim 38\%$  probability to succeed.



- Human can access and interrupt at any time based on Magentic-UI

Dr.Sai

Current Session

Saved Plans

Sessions 1

+ New Session

Today

帮我测量psi(4260) -> pi+...

Step 2 of 7: 利用生成的psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]过程在4.26 GeV能量点上的截面，并且绘制J/psi (mumu) 的不变质量。先规划后执行。

帮我测量psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]过程在4.26 GeV能量点上的截面，并且绘制J/psi (mumu) 的不变质量。

Plan Accepted

Step 1: 创建一份特定的固定格式BESIII实验专用JSON变量卡代码，用于编写针对psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程的分析算法程序。

The Coder has completed the creation of the JSON variable card code tailored for the psi(4260) -> pi+ pi- [J/psi -> mu+ mu-] physical process analysis algorithm program. The JSON structure includes necessary fields such as event type, particle properties, and analysis parameters to facilitate the accurate measurement of the cross-section and the invariant mass of J/psi (mu+ mu-).

Step 2: 利用生成的psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程分析算法程序所对应的JSON变量卡来执行相关内置脚本，从而生成psi(4260) -> pi+ pi- [J/psi -> mu+ mu-]物理过程的分析算法程序。

Error code: 400 - {'detail': 'Exception: No worker with model `hepai/code-worker-v2-mapping` was found.\nException: No worker with model `hepai/code-worker-v2-mapping` was found.'}

Regenerate the mapping algorithm for psi(4260) -> pi+ pi- [J/psi -> e+ e-]

We need to create a new plan. The current step cannot be completed due to the unavailability of the required worker model. A new plan is needed to address this issue.

Here's a plan. You can edit it directly or through the chat.

Step 1

创建一份特定的固定格式BESIII实验专用JSON变量卡代码，用于编写针对psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程的分析算法程序。

Step 2

利用生成的psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程分析算法程序所对应的JSON变量卡来执行相关内置脚本，从而生成psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程的分析算法程序。

Step 3

生成固定格式BESIII实验专用JSON变量卡代码，用于创建psi(4260) -> pi+ pi- [J/psi -> e+ e-]物理过程的模拟、重建及分析的JobOption脚本，需生成100个事例，并同时提交实验数据、inclusive蒙特卡洛模拟数据和exclusive蒙特卡洛模拟数据。

Type your response here and let Dr. Sai know of any changes in the browser.

Dr. Sai can make mistakes. Please monitor its work and intervene if necessary. (Powered by Magentic UI)

Tasks progress

Single task's status and output/error

Human interruption

# Overview of Dr.Sai Agent



- Timeline

