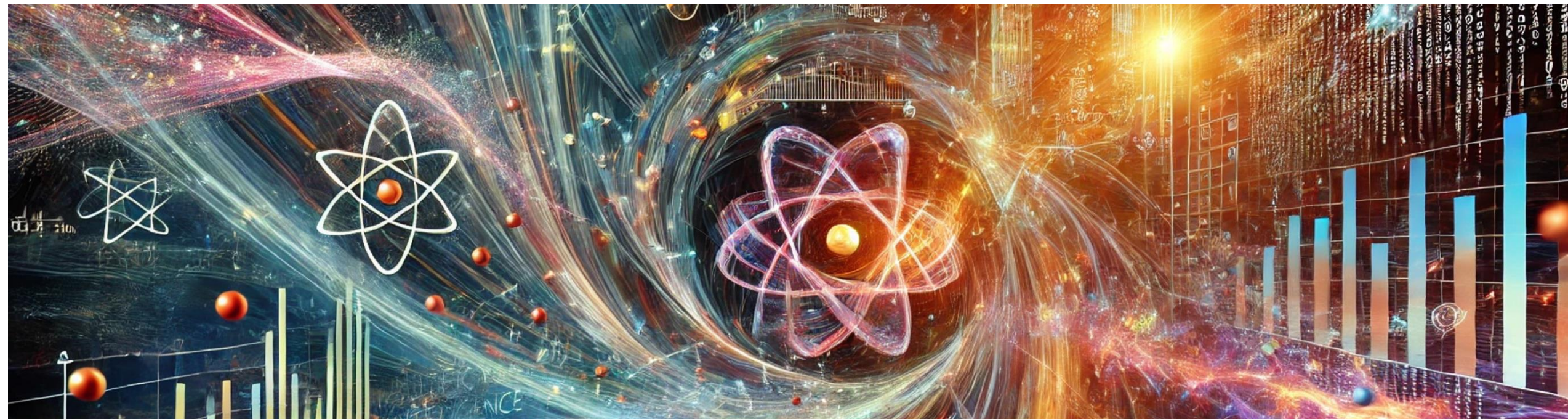




Computing Center, IHEP, CAS
National HEP Data Center



基于大模型的物理分析智能体和思考

张正德 etc. 高能所 计算中心&实验物理中心

2025年1月11日

合肥

Outline

01 引言

02 Dr.Sai物理分析智能体

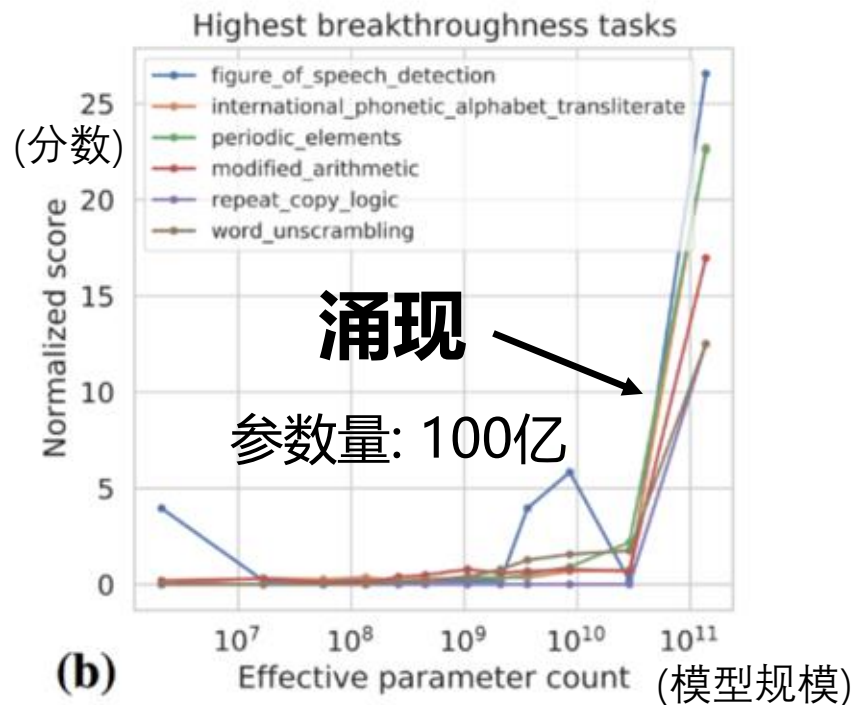
03 几点思考

04 总结与展望



GPT-4在所有自然语言处理下游任务上超过了专用AI模型，是通向通用人工智能的可能路径。¹

证明了**预训练、指令微调 and 强化学习大模型**的模式是实现更高层次智能的有效途径。



■ 大模型的四大能力

- 通用的意图理解能力
- 强大的连续对话能力
- 智能的交互修正能力
- 较强的逻辑推理能力

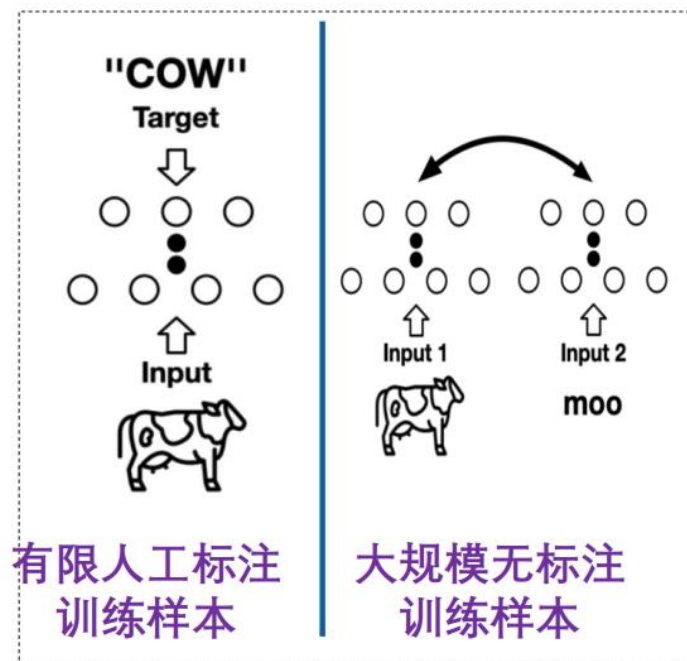
¹ 10.48550/arXiv.2303.12712 大模型的涌现现象

大模型技术有望突破瓶颈



- 开启了基于自监督学习的“**大数据+大模型**”新范式，从大规模的**无标注数据**中挖掘隐含的监督信息进行**通用知识**学习，成为迈向通用人工智能的重要途径。
- ChatGPT进一步联合**人类反馈的强化学习**，实现可与人类意图对齐的基础模型。

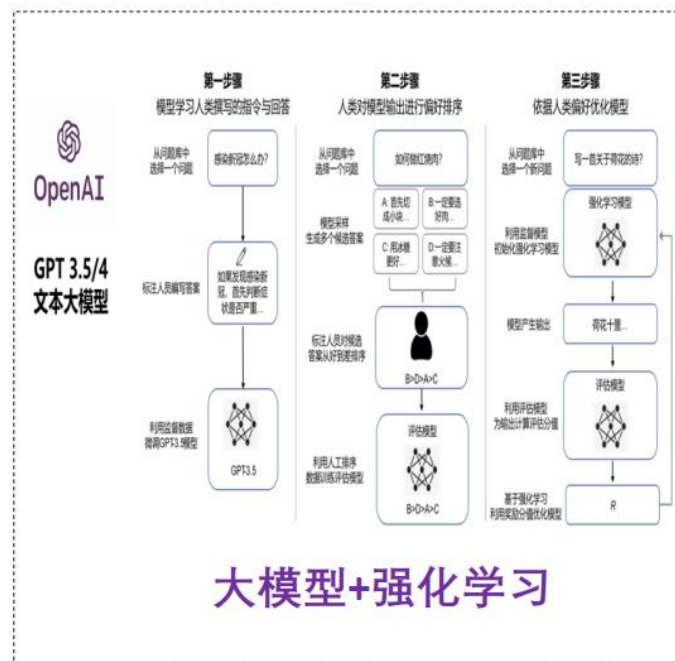
1 从有监督到自监督



2 从专用小模型到通用大模型

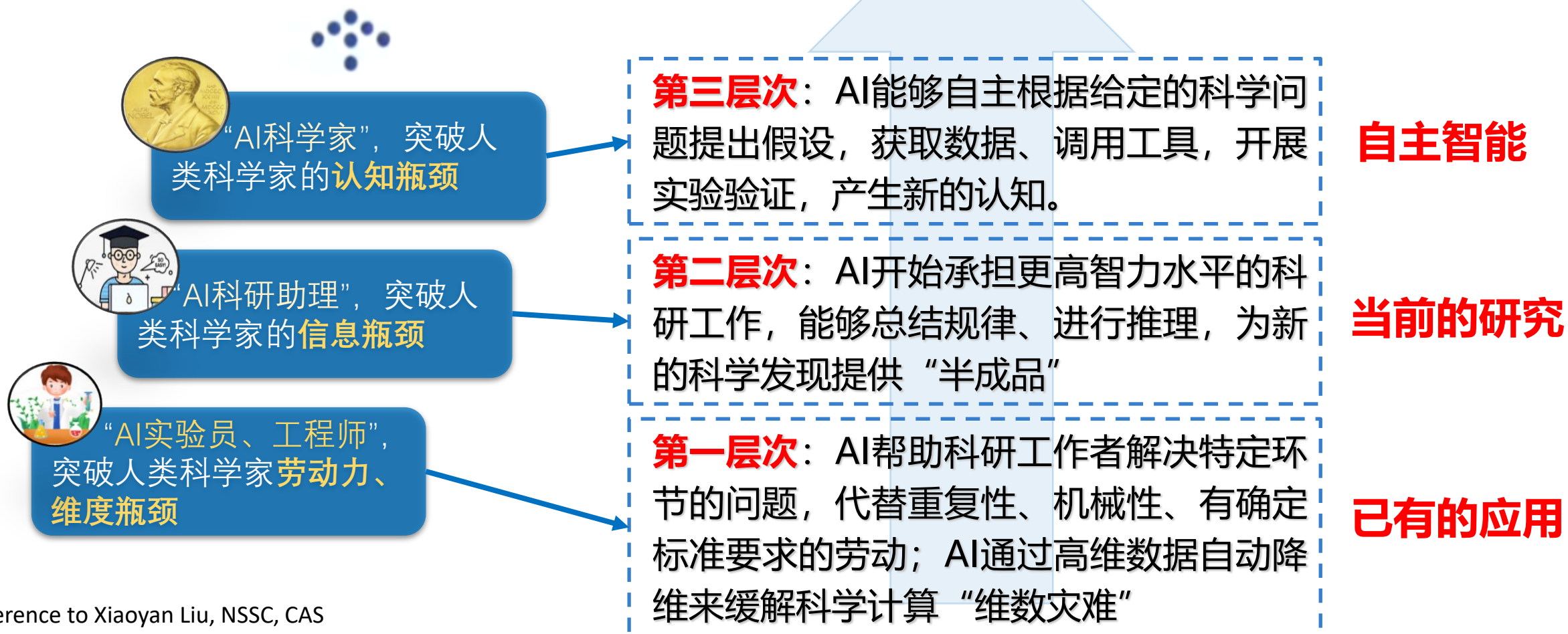


3 从数据驱动到人机交互学习

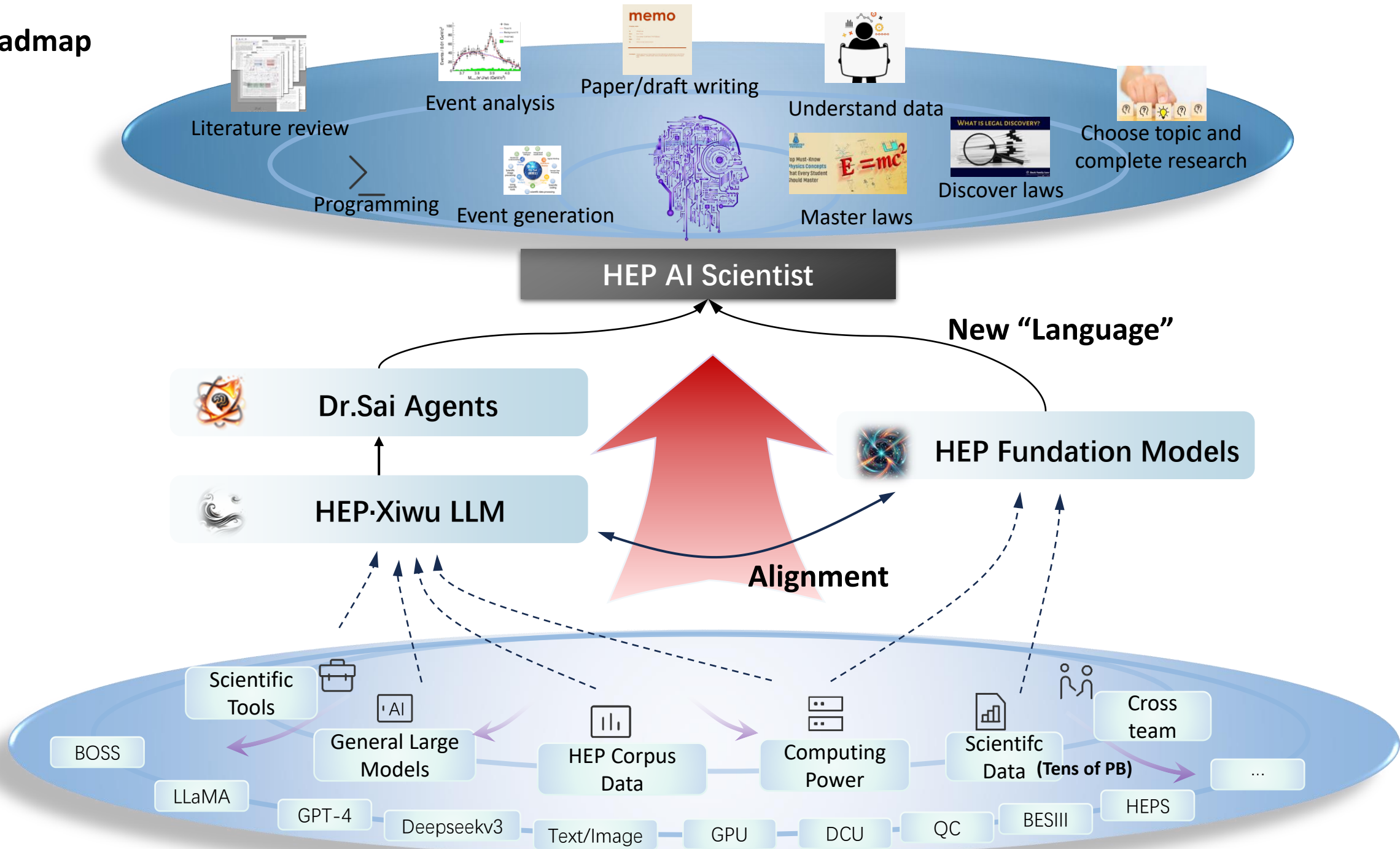


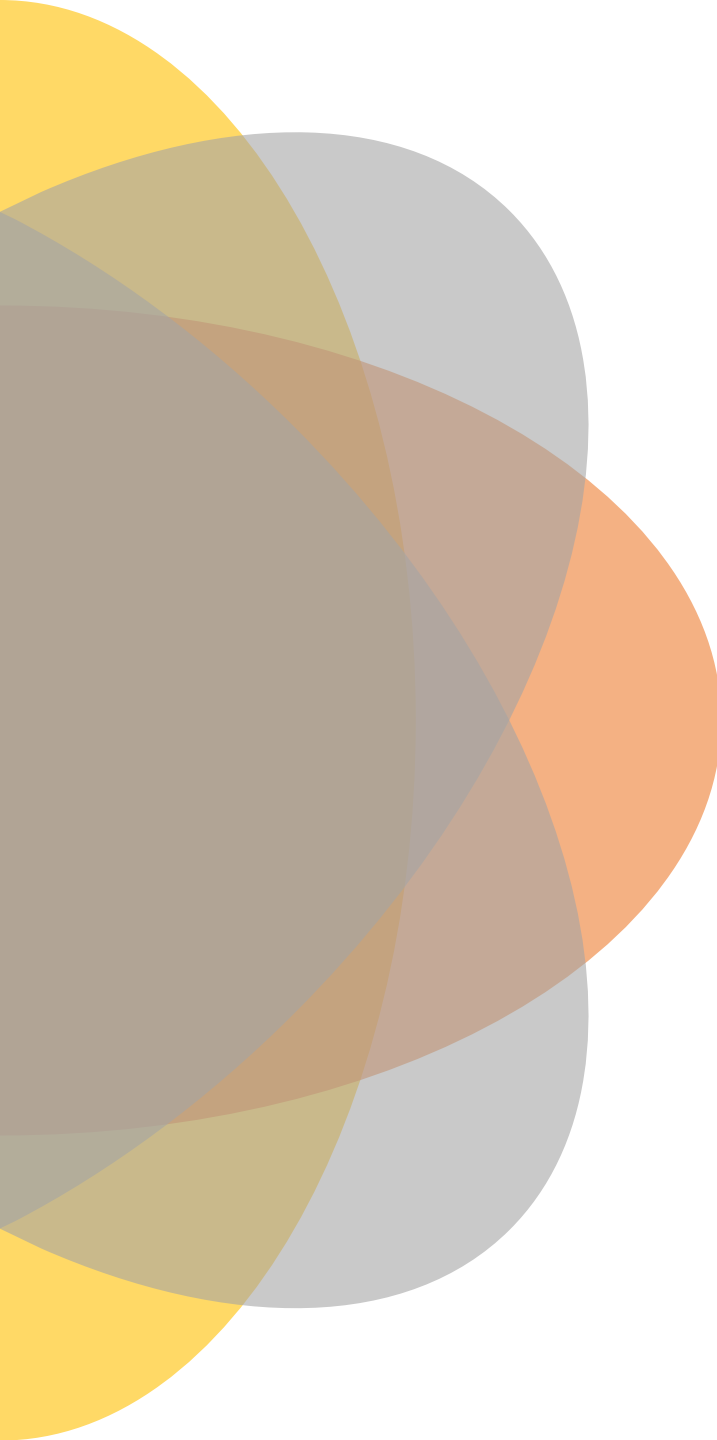


GPT-4的问世带来了“AI For Science”如何从原来帮助科学家突破**劳动力瓶颈**、**维度瓶颈**的阶段到过渡更高层次的突破**信息瓶颈**甚至**认知瓶颈**的巨大挑战和新的机遇。



Roadmap

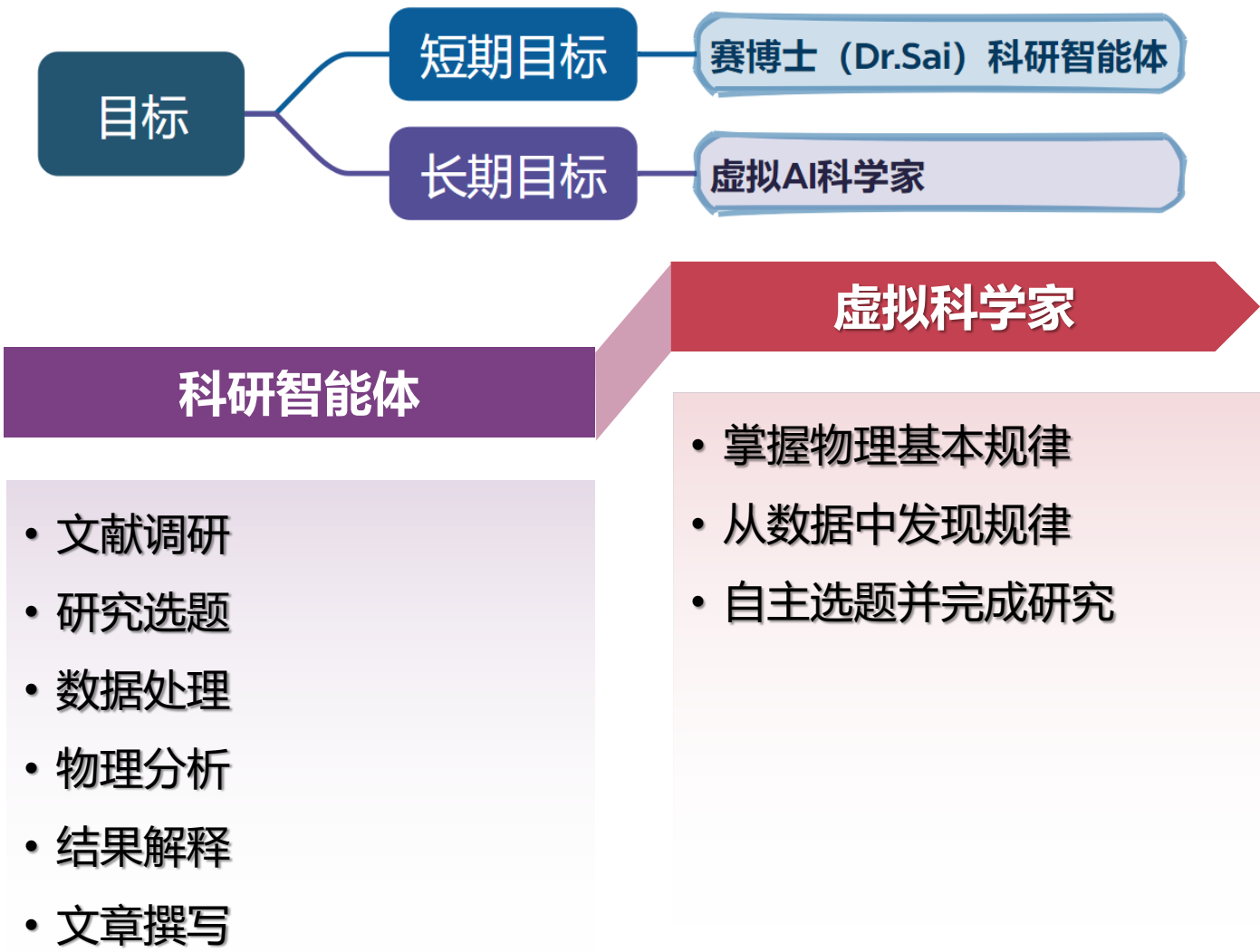




1

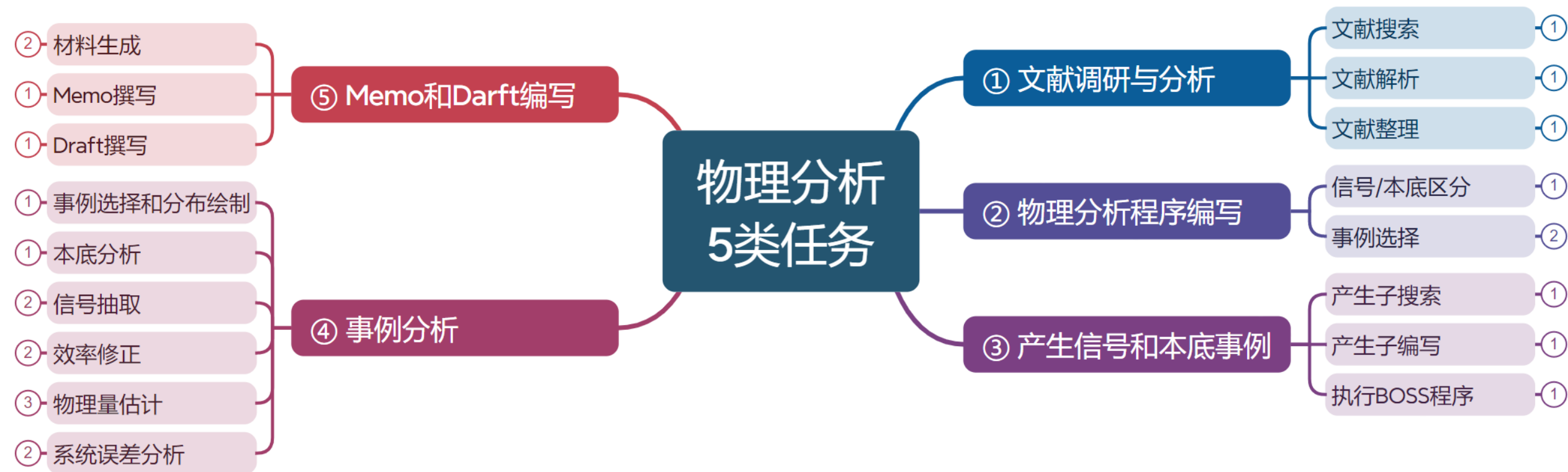
物理分析智能体

AI大模型驱动高能物理科学发现的研究





□ 物理分析是科学发现的关键环节之一

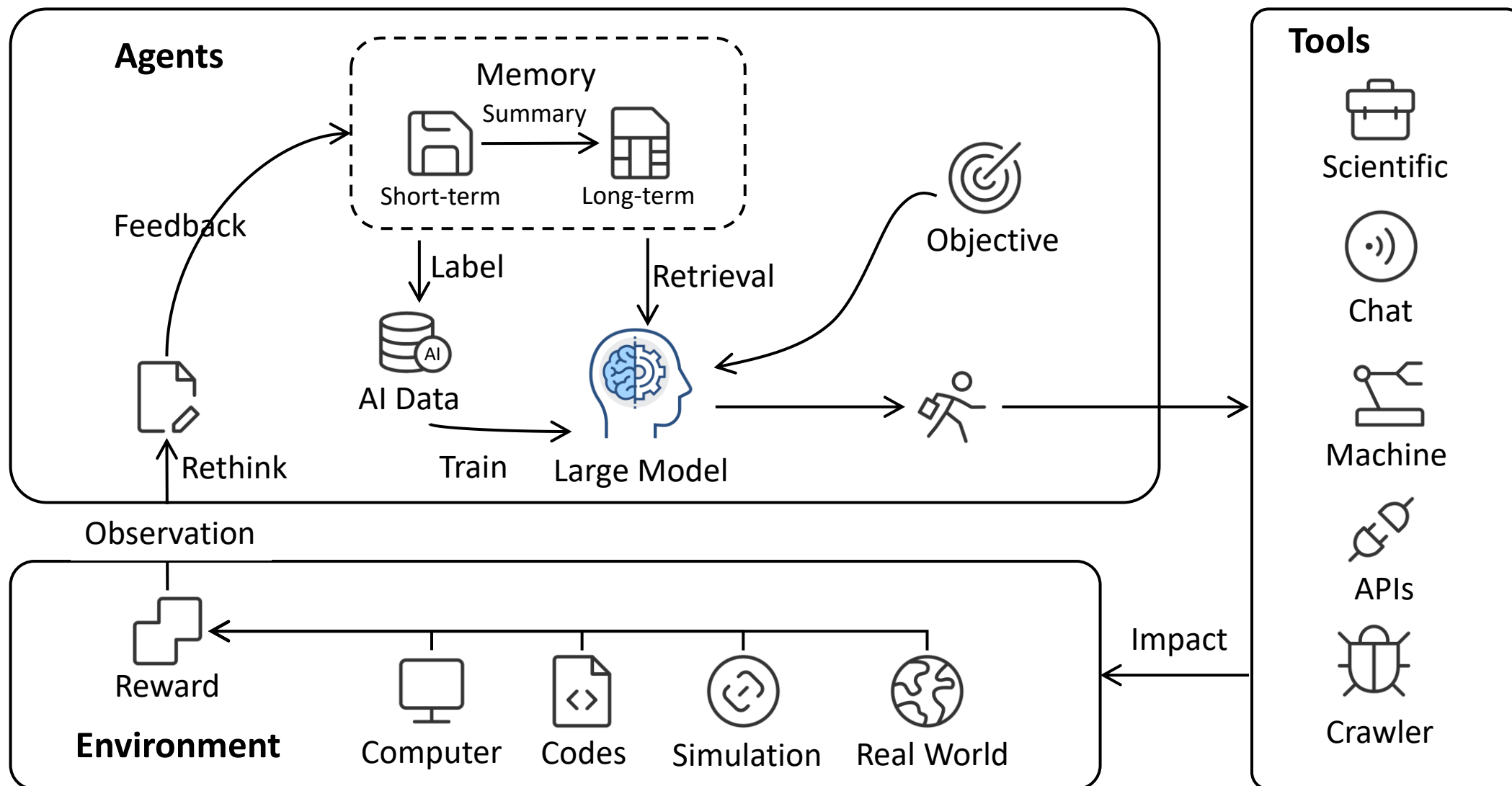


➡ □ 让**智能体**来**学会**复杂的物理分析方法

智能体的基本原理



智能体是指能够根据其知识、环境和输入，自主作出决策或执行行动的系统。

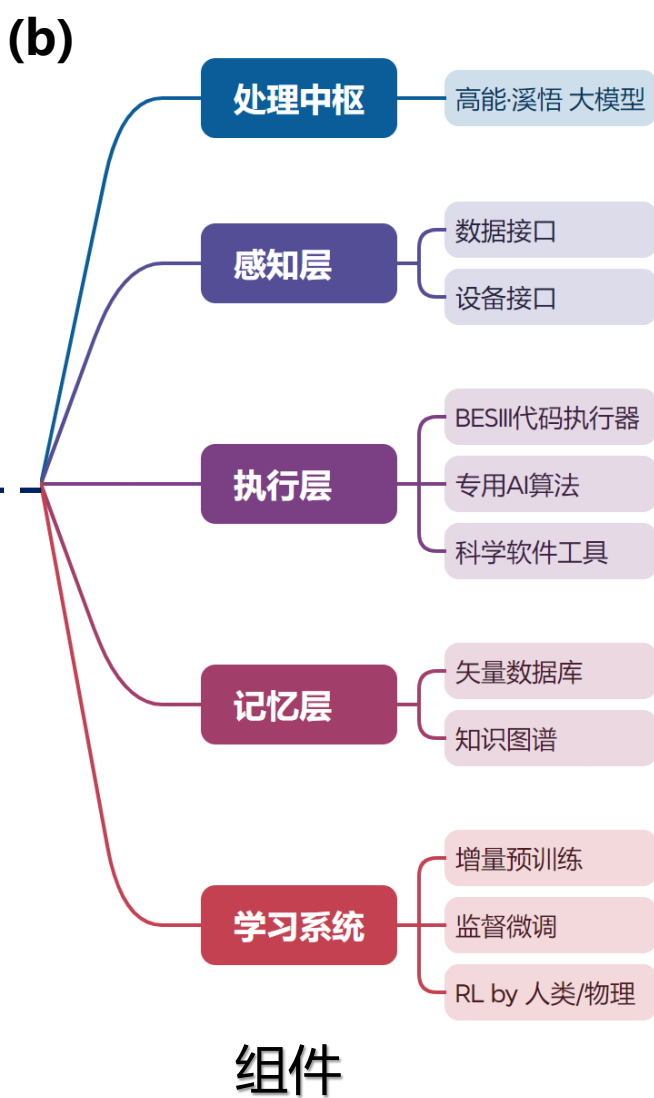
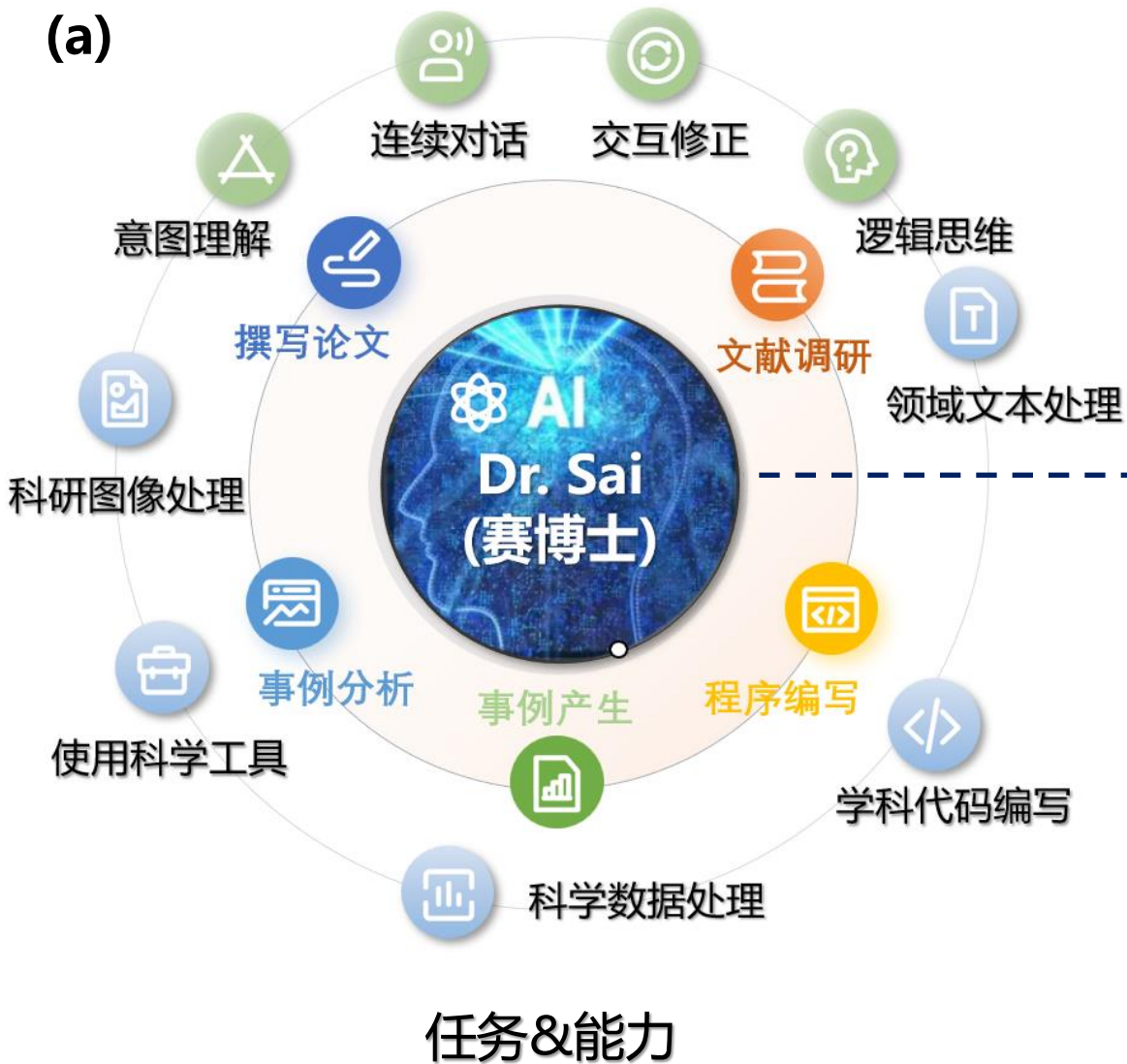


Dr.Sai智能体（专家级智能助理）

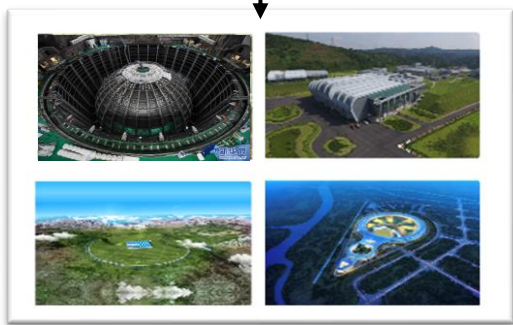
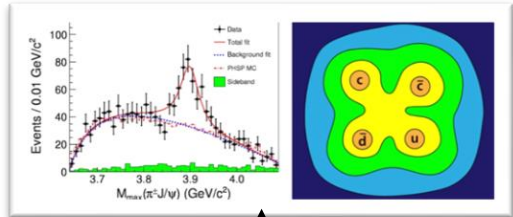


思路：让大模型学会通用的物理分析

本质：对研究过程的建模



目标：
重新发现 $Z_c(3900)$ 粒子

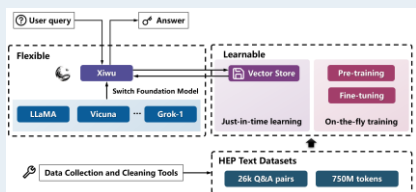


应用

Dr.Sai单智能体的各个组件



处理中枢

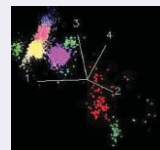


高能·溪悟 LLM

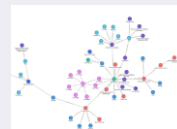
arXiv:2404.08001

- 目前基于LLaMA3-8B、70B
- 在 A800 和 国产 DCU K100上训练
- 领域问答增强
- 专业代码增强

记忆层



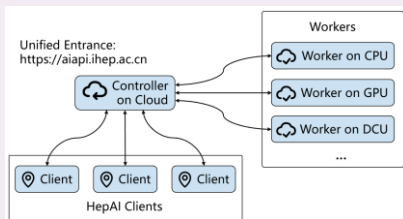
矢量数据库



知识图谱

- 矢量数据库
 - BAAI BGE-M3 嵌入模型
 - LLaMA Index
- 知识图谱
 - GraphRAG

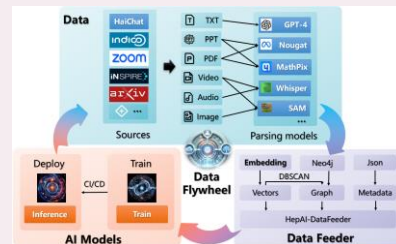
执行层



HaiDDF

- BESIII代码执行器
 - BOSS Worker
- Daisy科学软件框架
 - HEPSC T
 - HEPSPtycho
 - Etc
- 跨语言、跨系统、异构部署和负载均衡

感知层



数据飞轮

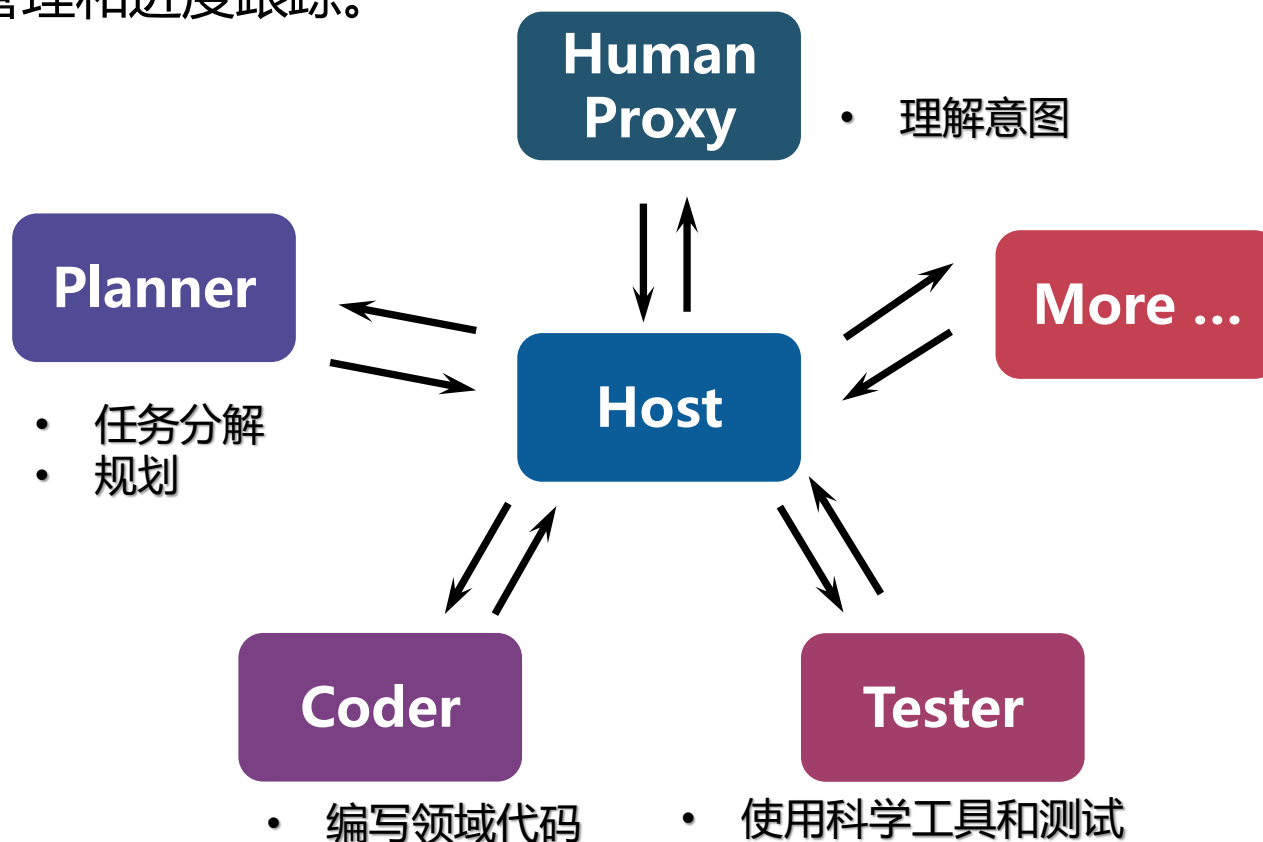
- 数据飞轮系统
 - HaiNougat
 - HaiChat
 - Whisper, SAM
 - ArXiv, Indico
- Domas数据管理框架
 - 科学数据元数据

Dr.Sai单智能体→多智能体协同系统



□ Dr. Sai 的多智能体协同系统 - 处理复杂任务

- 基于 AutoGen 框架。
- 每个智能体都配备特定的知识、工具和大语言模型 (LLM) 。
- 引入一个主持人 (Host) 智能体来管理群聊，便于扩展更多智能体。
- 引入一个人类代理智能体 (Human Proxy)，实现**人在环路**，允许人类随时打断、介入。
- 分层任务模型来进行任务管理和进度跟踪。



workflow:

Msg → host → LLM

Planner → **RAG** → LLM → host

Host → LLM

Autogen: arXiv.2308.08155)

Dr.Sai用户界面

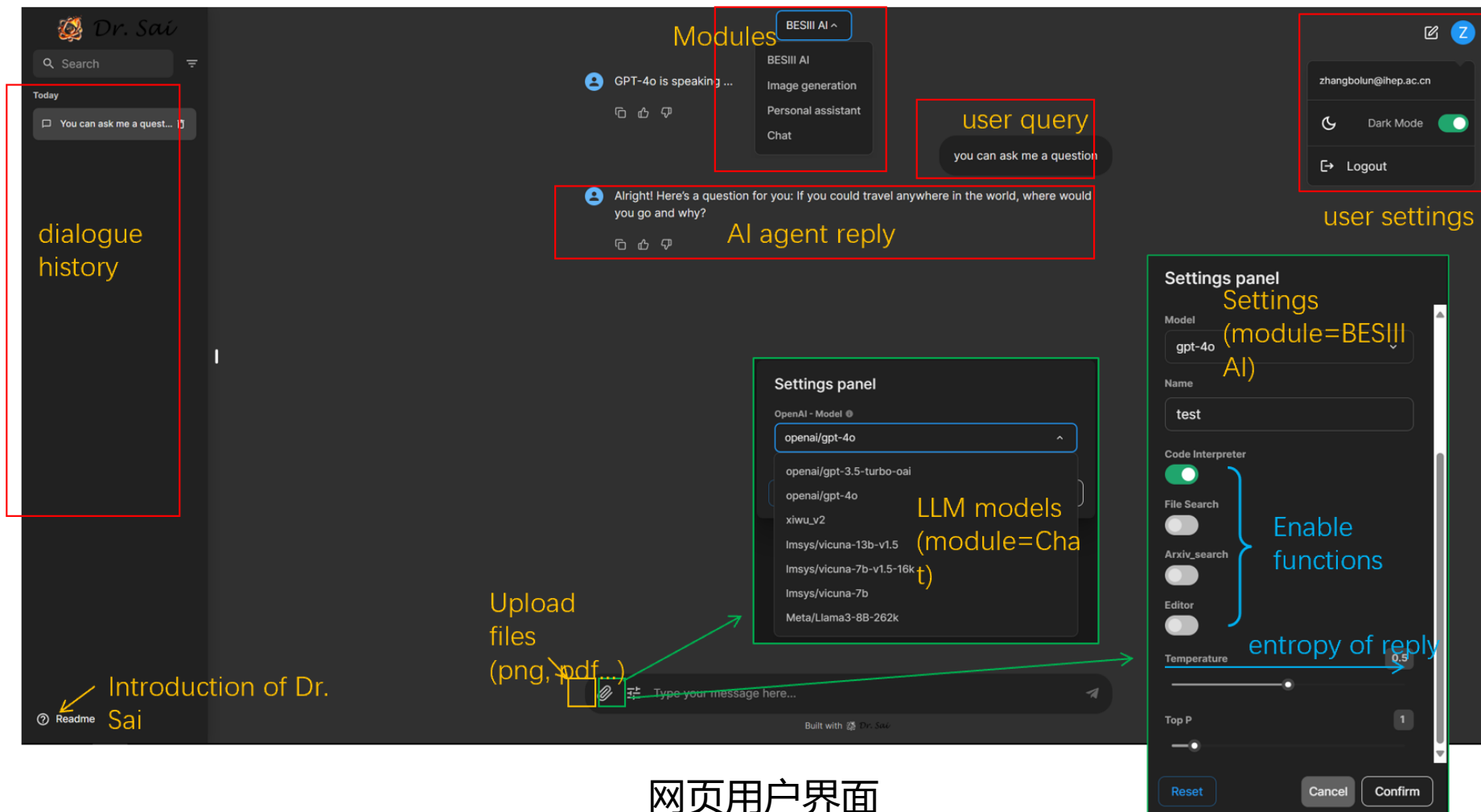


Dr.Sai v1.0.0 2024.07.01上线

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

(Based on Chainlit)

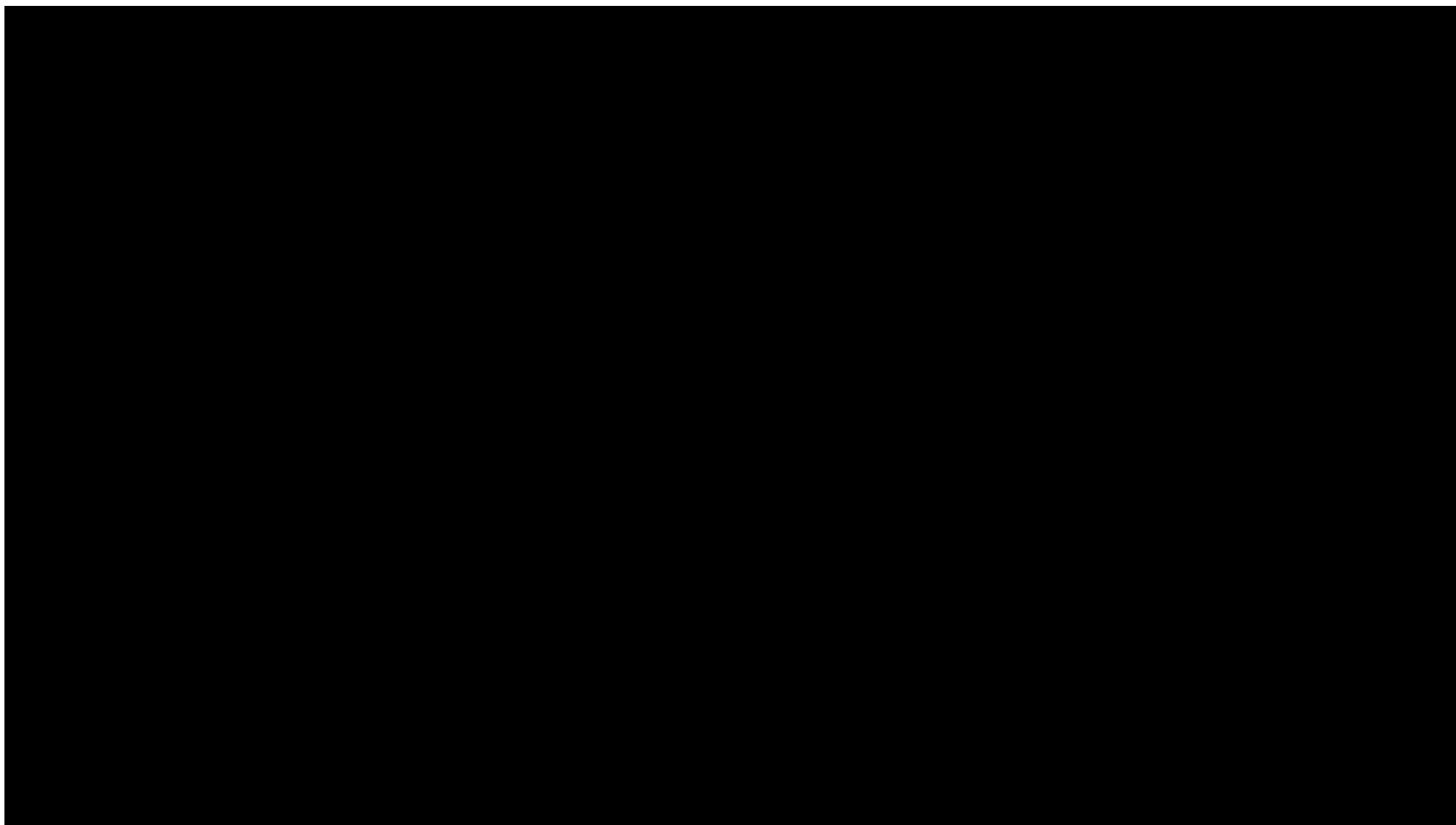
- 四个部分:
- BESIII AI Assistant
 - 搭载BESIII内部知识
 - 通过专业任务分解、代码生成、调用BOSS程序实现智能化物理分析
- Personal Assistant
 - 允许个人用户上传知识
- Image Generation
 - 基于AI的绘图
- Chatbot
 - 纯大语言模型: Xiwu, LLaMA3, GPT etc
 - 支持PDF和图像输入



网页用户界面



- **Host智能体**: 选择合适的发言者
- **Coder**: 领域代码生成
- **Tester**: 调用BOSS(BESIII离线分析软件)执行分析
- 绘制信号直方图



进展 (2025.01)



- 0 模糊的任务描述
- 1 明确的任务描述 (Planner)
- 2 生成模拟和分析程序 (Coder)
 - ◆ 检索增强得 (变量说明信息) + 特定提示词 → 变量卡
 - ◆ 4个Template + 变量卡 → (mapping) →
 - ◆ Decay card: 衰变卡
 - ◆ Sim+Rec: 数据生成脚本
 - ◆ Ana: 事例挑选脚本
- 4 执行任务 (Actuator)
 - ◆ 任务管理系统 → Tester执行任务, 得到 (sim → 模拟的电子学信号.rtraw, rec → 径迹信息.dst, ana → 物理粒子级别信息 (动量、能量、质量) .root)
- 5 分析结果
 - ◆ mc.root, data.root, 利用mc的信号形状指导在data中发现真正的信号。
 - 如: 利用pi+pi-J/psi过程模拟的信号形状, 拟合实验数据, 判断是否存在Zc(3900)

多智能体协同Coder+BOSS Actuator

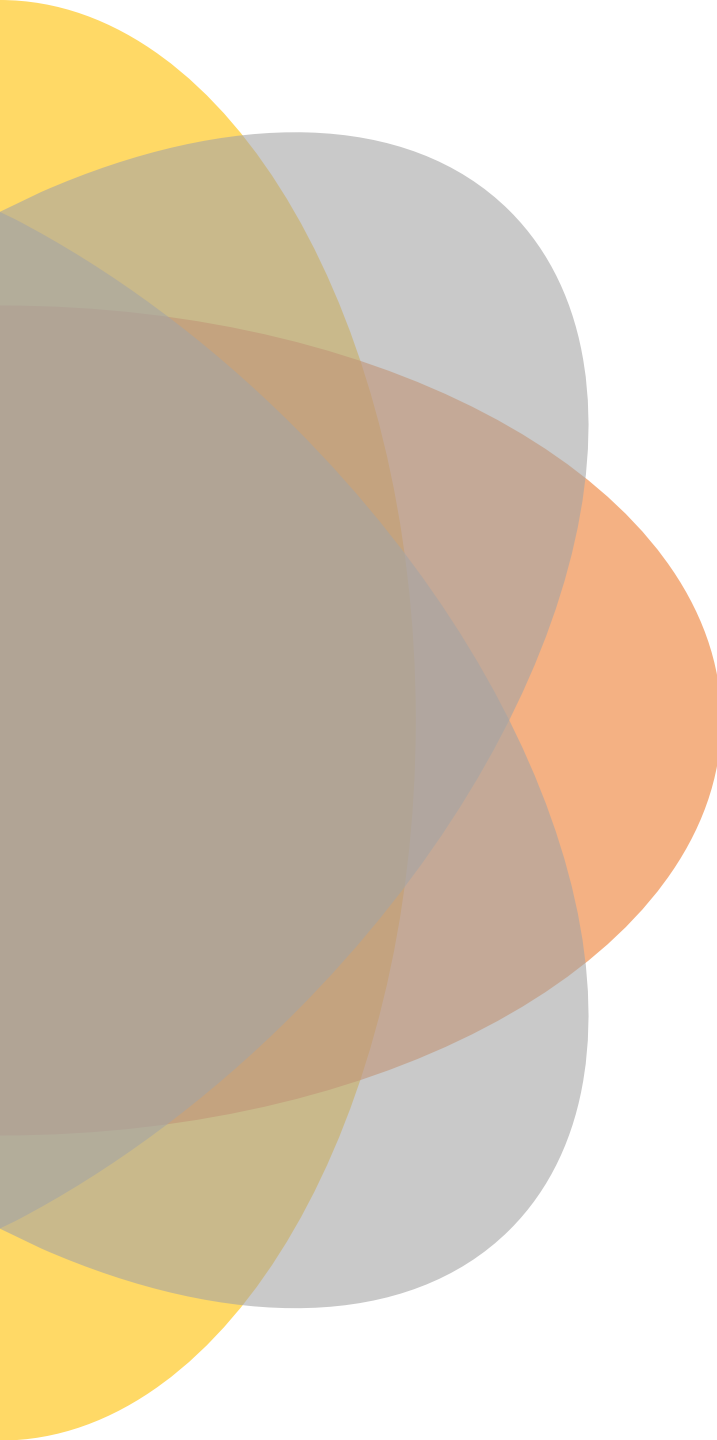
The screenshot displays a web interface for the BESIII Agent. The top section shows a list of new files generated by the agent, including setup and cleanup scripts. Below this, a 'Summary table' is visible, showing the status of various jobs. The table has columns for Job Key, Status, Create Time, Start Time, Finish Time, and Artifacts. One job is listed as 'finished'.

JOB KEY	STATUS	CREATE TIME	START TIME	FINISH TIME	ARTIFACTS
job_2025-01-10_193906_709794	finished	2025-01-10 19:39:06	2025-01-10 19:39:07	2025-01-10 19:39:22	Link

Below the table, there is a section for 'title_text' and a list of artifacts. At the bottom, there is a chat interface with a user 'admin@localhost' and a message about generating JSON files for the analysis of the pi+pi-J/psi process.



- 单个智能体能可完成单步物理分析过程
- 多智能体协同可将多步骤连接起来
 - 难点1：复杂物理分析任务的准确分解，需要专家输入（**专业数据**）
 - 难点2：涉及物理逻辑推理过程，思维链+强化学习技术能否实现尚待验证（**算法研究**）
- 促进粒子物理学发现的前景
 - 海量多维数据处理的复杂，AI大模型驱动的多智能体系统有望将极大提升**物理分析效率**
 - 科研范式变革：传统**的人类规划任务、写代码分析总结**→**智能体辅助高级劳动**



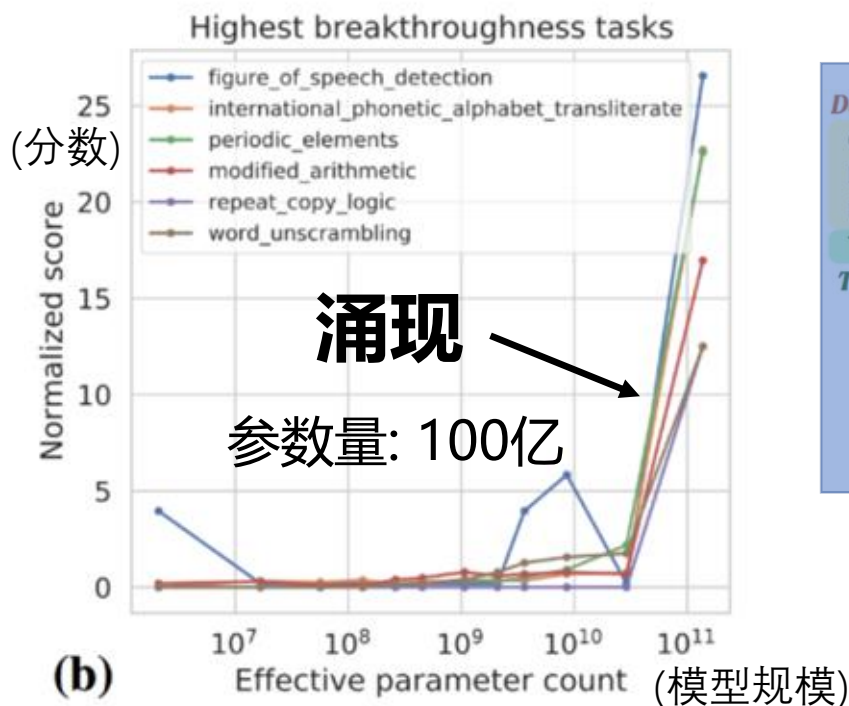
2

几点思考

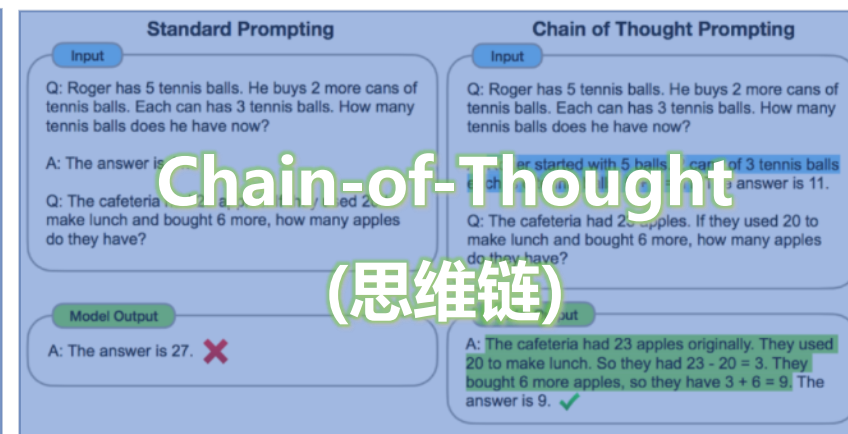
从语言模型迈向科学模型



预训练、指令微调 and 强化学习大模型的模式被现象学地证明是实现更高层次智能的有效途径。



大模型的涌现现象



大模型涌现的特殊能力

可行路径:

海量数据
无监督预训练

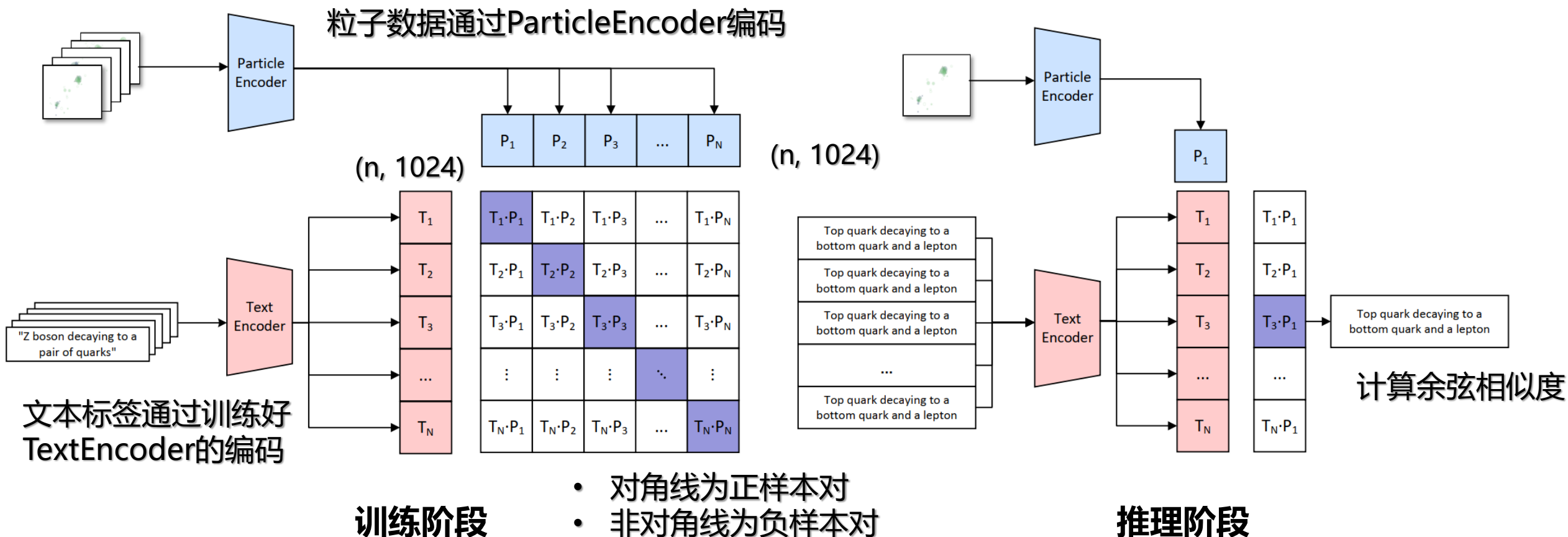
下游任务
指令微调

更高层次
的智能

从语言模型迈向科学模型



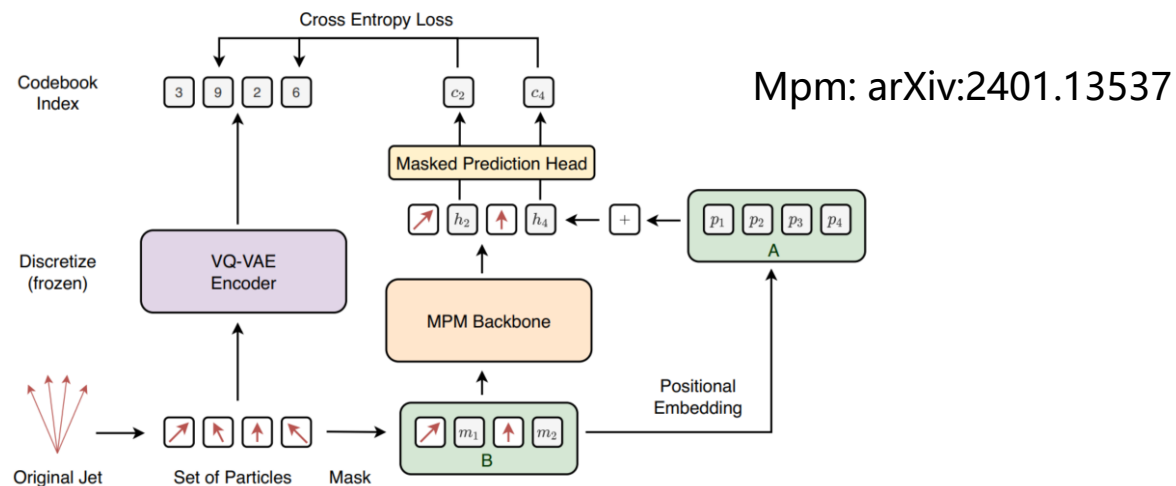
基于对比学习的**跨粒子物理**和**文本模态**自监督预训练方法。



从语言模型迈向科学模型



基于码本的生成预训练方法探索



输入数据:

$\Delta\eta$
 $\Delta\phi$
 $\log p_T$
 $\log E$
 $\log \frac{p_T}{p_T(\text{jet})}$
 $\log \frac{E}{E(\text{jet})}$
 ΔR



编码器



码本



解码器



数据表示
512维度

分类结果



分类头



生成重建:

$\Delta\eta$
 $\Delta\phi$
 $\log p_T$
 $\log E$
 $\log \frac{p_T}{p_T(\text{jet})}$
 $\log \frac{E}{E(\text{jet})}$
 ΔR



“The limits of my language mean the limits of my world.”

我的语言之局限，即我的世界之局限

——Wittgenstein

维特根斯坦

奥地利哲学家





■ 什么是AI-Ready数据集？

- 有明确的任务（输入、输出）
- 有科学价值
- 有明确的标准（真值、评价方法）

■ 数据的重要性：

Gabage In, Garbage Out!

■ 什么是好的AI数据集？

- 数据量大、维度高
- 多样化
- 高质量
- 机器可用的数据管道

■ AI数据基本要求

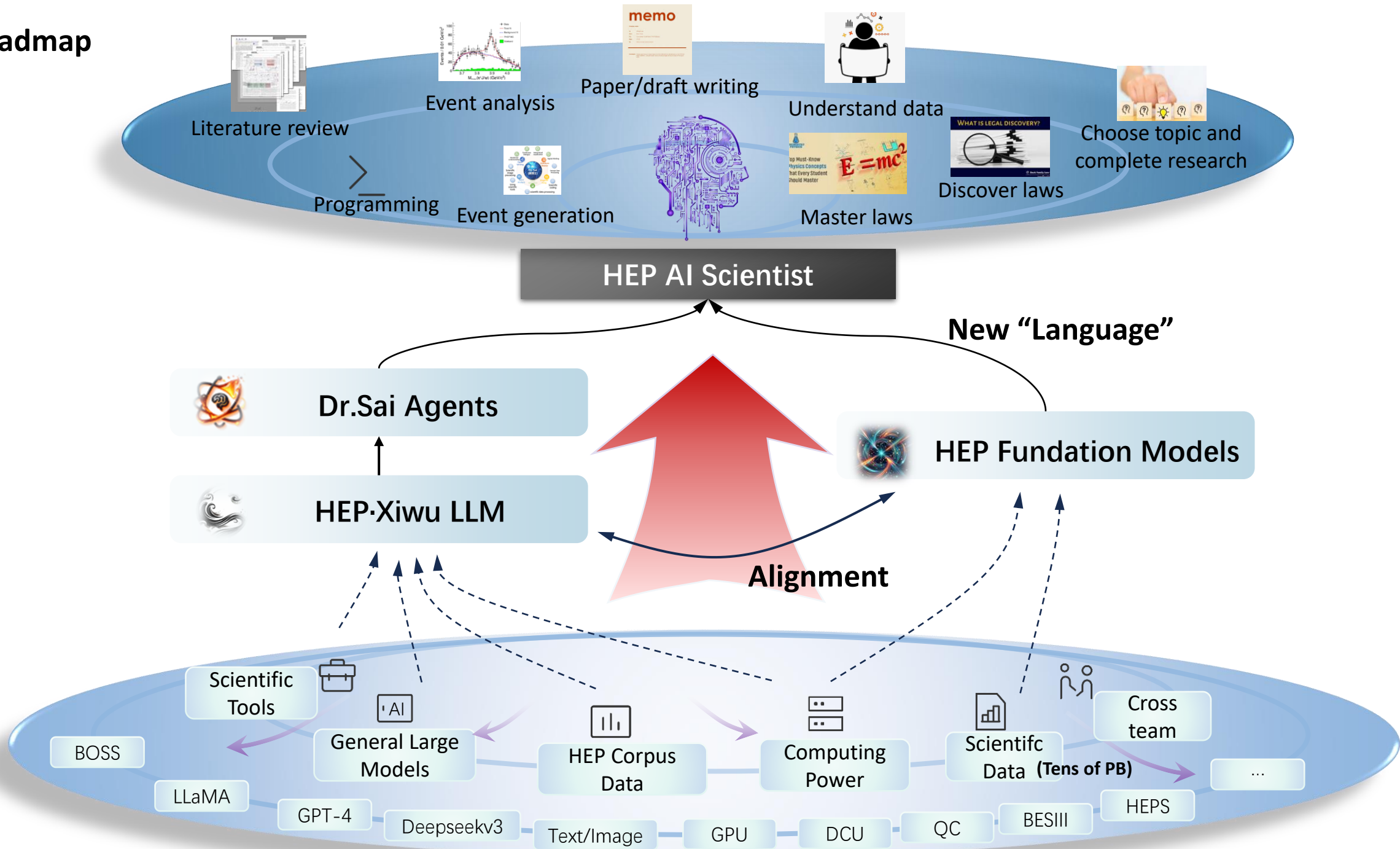
- 满足FAIR原则
- 统一的格式、标准



- 科学领域专家
- AI领域专家
- 交叉研究人员

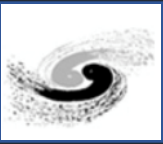
- AI大模型加速科学研究范式的向**智能范式**变革
- 充分利用大模型技术的两条思路
 - AI大模型 → 智能体 → **多智能体系统**：学会复杂的物理分析过程，显著提升高级劳动的效率，促进新粒子发现
 - 探索从头训练科学大模型：从数据驱动的角度，理解数据、发现规律
 - 语言模型和科学模型融合
- 物理分析逻辑推理数据集和思维链+蒙特卡洛树增强推理能力
- 简化BOSS代码和代码增强的模型
- 粒子物理基础
- 建立粒子物理的AI数据体系（语料数据、实验数据、模拟数据）

Roadmap





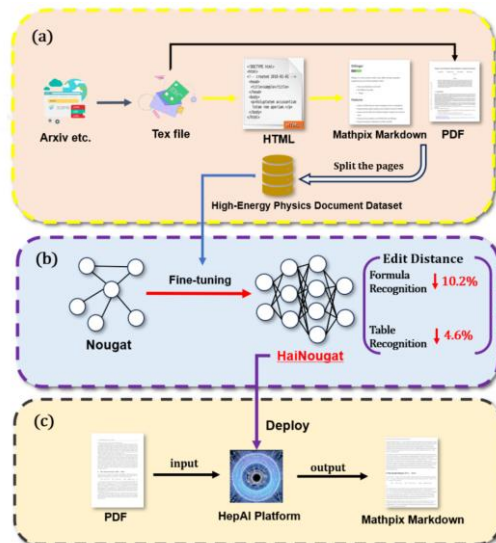
~~Thank you for listening~~



Backup

HaiNougat: A More Accurate PDF parser for High Energy Physics

罗建文



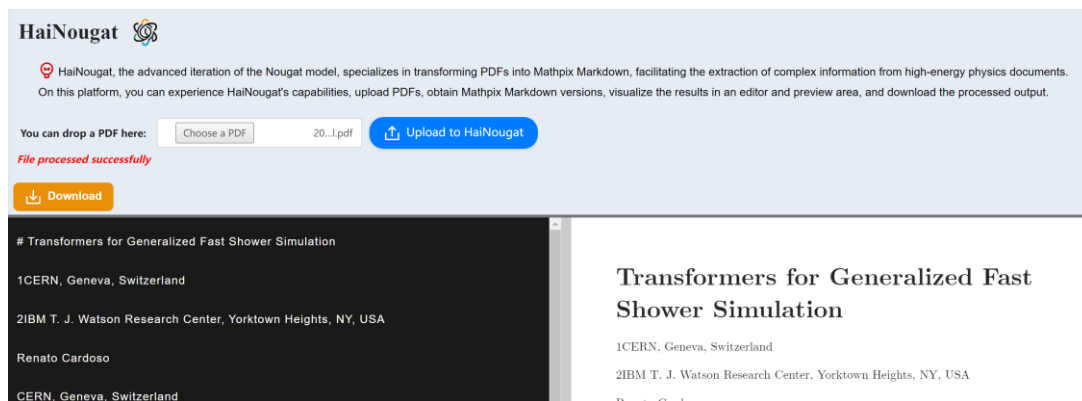
方法

- 收集清洗了10万论文图片和markdown格式的AI-Ready数据
- 基于Nougat微调训练了HaiNougat模型
- 基于HepAI分布式部署框架部署了HaiNougat服务

结果

- 更准确地将高能物理PDF论文解析为MathpixMarkdown格式
- 保持公式、表格和图片
- 是智能体的感知层的组件之一
- 部署成独立网页demo
- 解析速度：8页一个batch，20秒(RTX 4090)
- 论文投稿中...

HaiNougat UI <https://chat.ihep.ac.cn:5557/>



任务举例：用ROOT绘图



提示词:

理解意图生成代码:

1-shot

tot_prompt = ""

Generate a part of C/C++ code to solve the problem: {topic}.

You will use cern ROOT.

This is an example of the code: {self.example_code}

Please generate as clearly as possible.

topic: {topic}

draw the variable of 'Mpipm', the range of histogram is from 0.2 to 1.2, and should change the label of X-axis, and change the name of output figure.

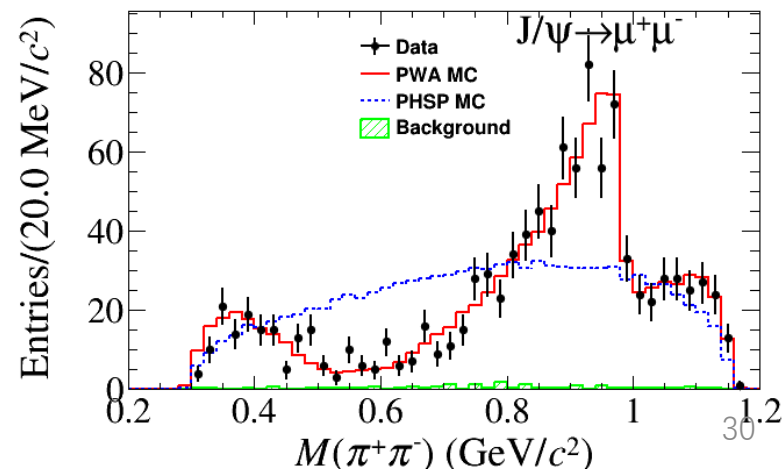
""

学会了修改指定代码，成功的利用1-shot画出不变质量谱

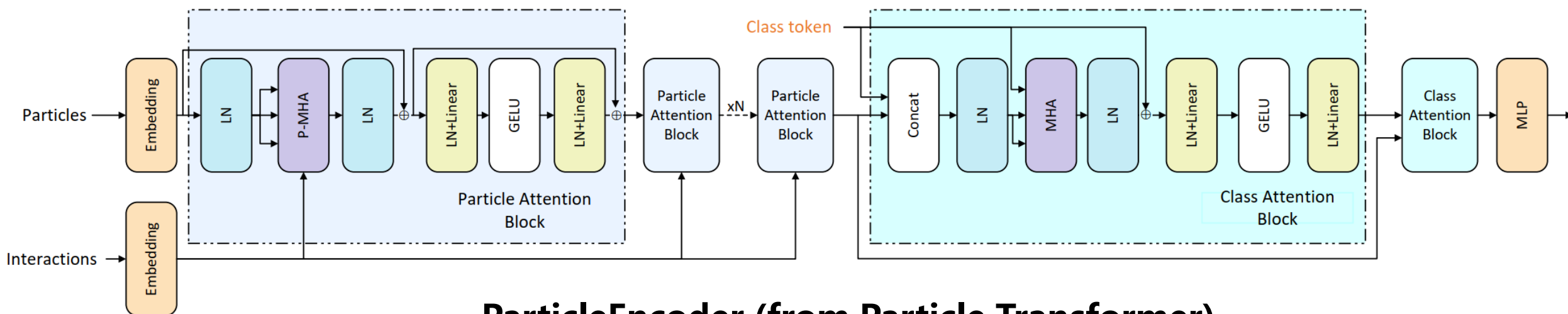
2025/1/11

```
// Define the lower and upper limits of the x-axis
Double_t xlow=0.2, xup=1.2;
// Define the variable to plot on the x-axis
TString plotx="Mpipm";
// Set the titles and styles for the histograms
h1->SetXTitle(Form("#it{M}({#it{#pi}}^{+}{#it{#pi}}^{-})
(GeV/{#it{c}}^{2}))");
// Save the plot as a PNG file
TString modified_pngfile = pngfile;
modified_pngfile.ReplaceAll(".png", Form("_%d.png", flag));
c->Print(modified_pngfile);
```

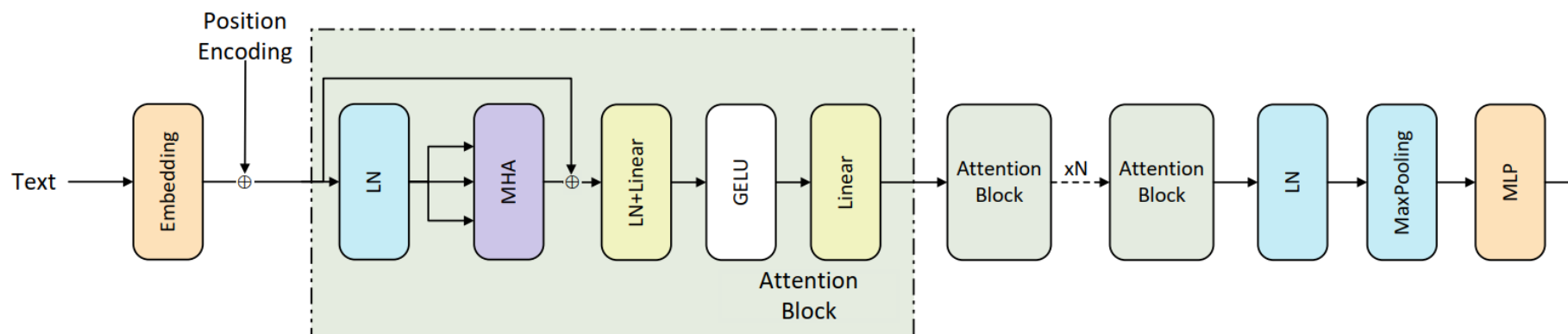
执行代码绘图结果:



CLPP: 神经网络结构

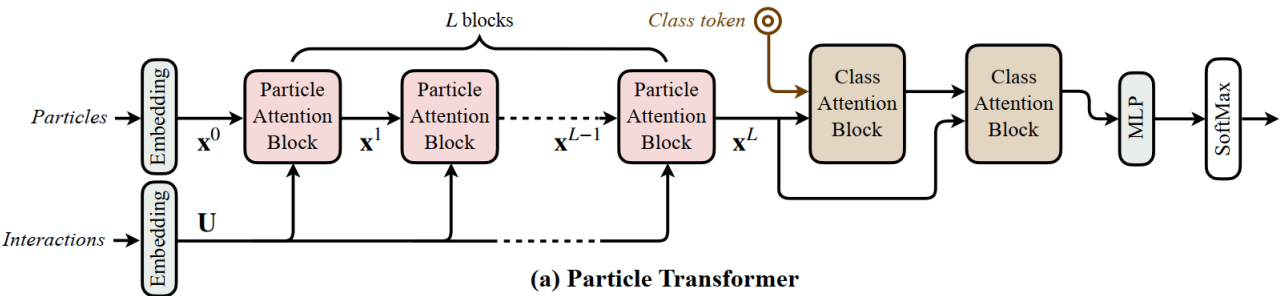


ParticleEncoder (from Particle Transformer)



TextEncoder (from BAAI/large)

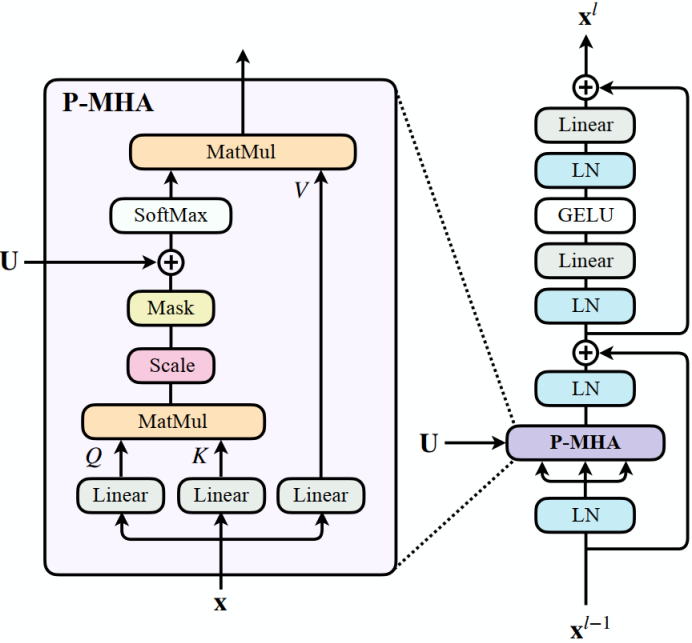
基准模型：Particle Transformer (ParT)



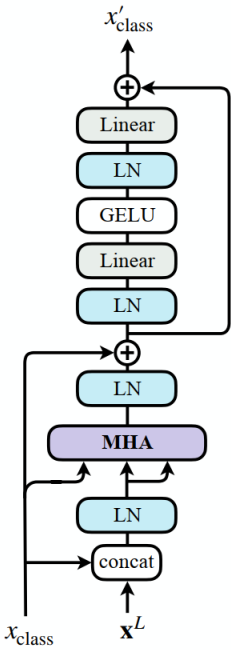
(a) Particle Transformer

JCTopQG

JetclassTopLandScapelQuarkGloun



(b) Particle Attention Block



(c) Class Attention Block

- 优点：
 - 添加了相互作用U，当前SOTA
 - 分别支持三种数据集
- 缺点：
 - 无法同时处理三个数据集
 - 不具备任务的扩展性

Figure 3. The architecture of (a) Particle Transformer (b) Particle Attention Block (c) Class Attention Block.

什么是AI智能体?



An AI agent refers to a **system** or **software** that can make autonomous decisions or perform actions on behalf of its users based on its knowledge, programming, environment, and inputs.



A LLM $\neq$ A person

An agent $\approx$ A person

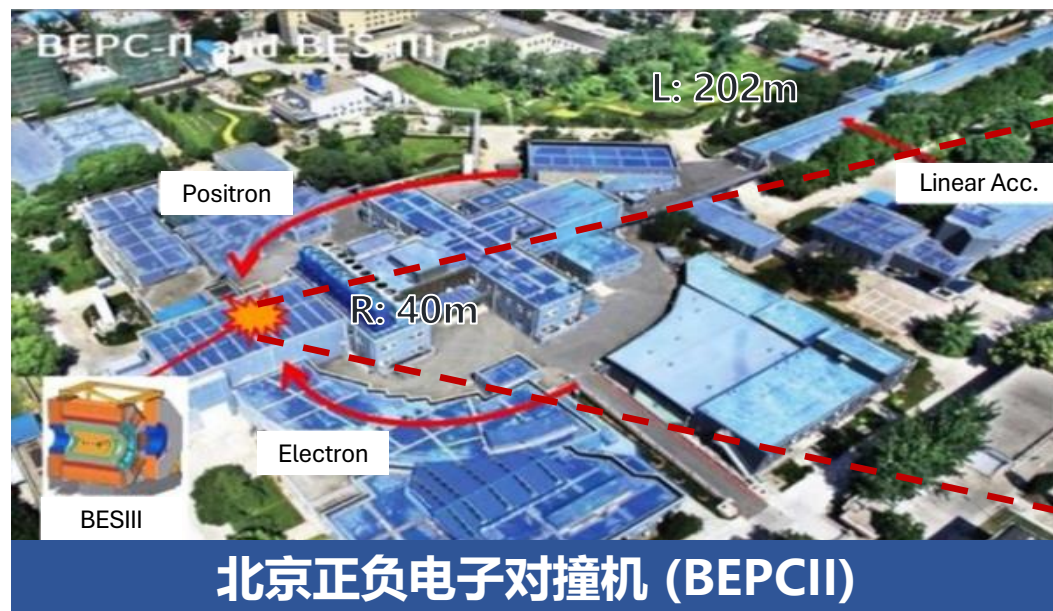
The LLM is the **processing core** (the Brain) of agent.

- 能学会领域专业知识
- 能学会科学分析工具

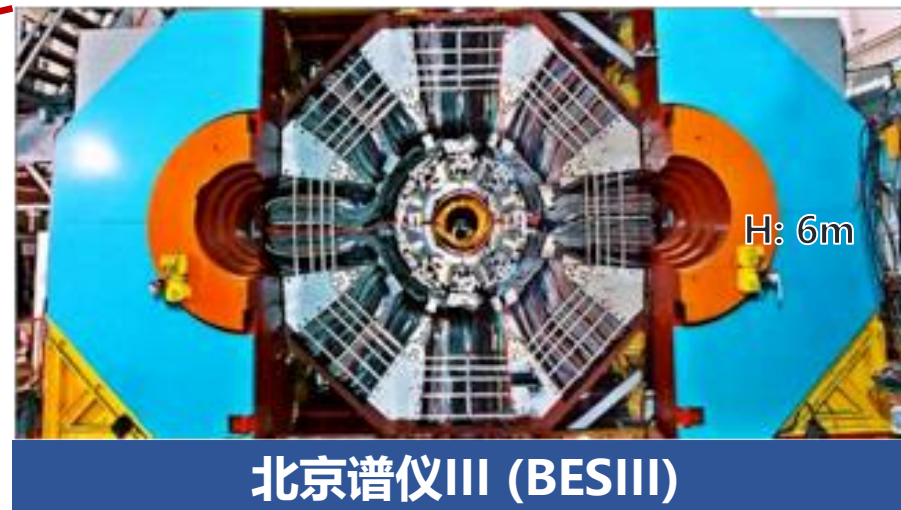
- 能感知数据和元数据
- 能编写代码和执行程序

- 能学会任务规划、角色分配和协同

BESIII实验



安装在
BEPC上的
探测器



BESIII

- 丰富的物理
- 超过600篇论文
- 16个国家，超过600科学家参与

数据

- 世界上最大陶-粲能区的数据样本
- 2-5.6GeV, 共振峰和扫描

充分利用

- 物理潜力尚未充分发掘
- AI能帮助挖掘海量高维数据中的数据潜力。

物理分析任务细分



物理分析任务

⑥
文献调研与分析

⑤
物理分析程序编写

⑥
产生信号和本底事例

⑪
事例分析

⑦
Memo和Darft编写

①文献调研与分析

文献搜索

数据库、爬虫、加载知识库

文献解析

PDF解析、物理语义、图片、表格、公式解析

文献整理

整理单/多篇文献的背景、方法、进展、结果，形成新的研究方法

②物理分析程序编写

信号/本底区分

根据指定信号过程，判断几种可能的本底过程

事例选择

产生合适的事例选择条件，以流程图或者表格形式展示

将选择条件转换为C++代码（语义转换），提取合适的变量存入ROOT文件

③产生信号和本底事例

产生子搜索

外部文档加载、技能库、粒子性质判断、搜索

产生子编写

Decay card、KKMC、BesEvtGen、RunID、RootIO

执行BOSS程序

模拟、重建、分析脚本

物理分析任务细分



物理分析任务

⑥ 文献调研与分析

⑤ 物理分析程序编写

⑥ 产生信号和本底事例

⑦ 事例分析

⑦ Memo和Darft编写

④ 事例分析

- 事例选择和分布绘制 — 物理量直方图边界、物理量分布、信噪比分析、进一步选择事例
- 本底分析 — 结合信号和本底MC分布，提高数据信噪比
- 信号抽取
 - 利用信号参数化和本底参数化得到信号数和统计误差
 - 分析拟合参数的关联矩阵和协方差
- 效率修正
 - 中间态过程的分波分析
 - 根据重建效率进行MC估计信号效率修正
- 物理量估计
 - 根据公式计算信号数、信号效率、中间引用值、估计截面、分支比
 - 误差估计、上限估计
 - 感兴趣物理量计算
- 系统误差分析
 - 分析误差项及其关联
 - 误差估计：引用误差、人为偏置系统误差、DATA和MC系统误差等

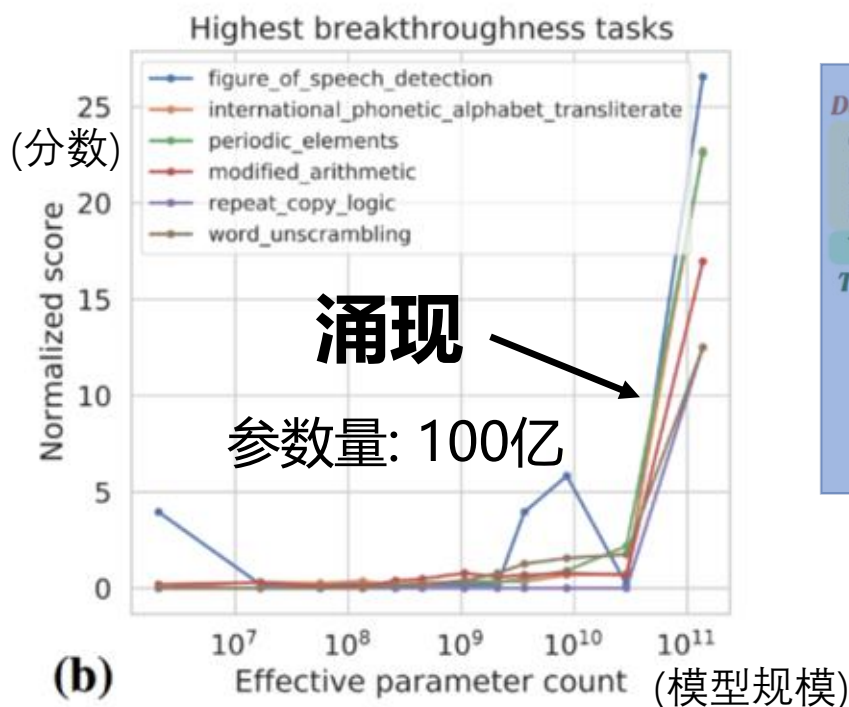
⑤ Memo和Darft编写

- 材料生成
 - 封面图生成，数据绘表格、绘制图片
 - 公式生成
- Memo撰写 — 撰写BESIII分析文档
- Draft撰写 — 给定期刊格式，从Memo撰写Draft

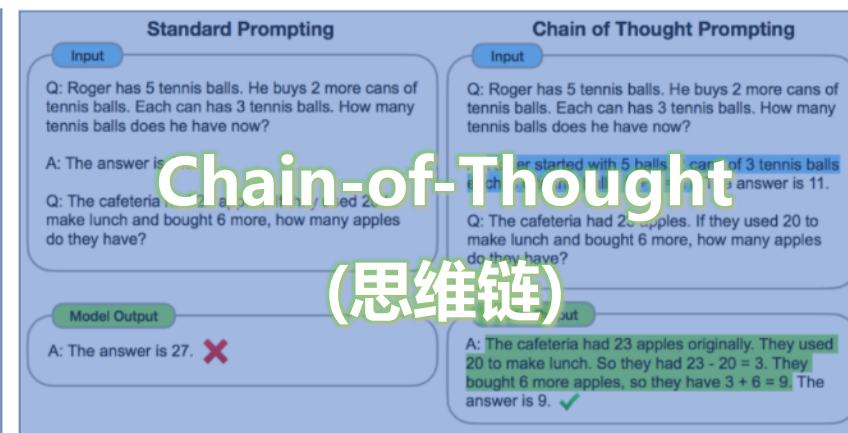
GPT-4 & 大模型



GPT-4在所有自然语言处理下游任务上超过了专用AI模型，是通向通用人工智能的可能路径。¹
证明了**预训练、指令微调 and 强化学习大模型**的模式是实现更高层次智能的有效途径。



大模型的涌现现象



大模型涌现的特殊能力

可行路径:

海量数据
无监督预训练



下游任务
指令微调



更高层次
的智能

¹ 10.48550/arXiv.2303.12712

- ❑ ChatGPT基于大规模语言模型GPT3.5，通过**人类反馈学习微调**而来的对话生成大模型。不再是传统意义的人机对话系统，而是以自然语言为交互的通用语言处理平台
- ❑ 超出预期的交互体验

- 通用的意图理解能力
- 强大的连续对话能力
- 智能的交互修正能力
- 较强的逻辑推理能力



推出**2个月**即达到**1亿**活跃用户
历史上**增长最快**的消费者应用程序



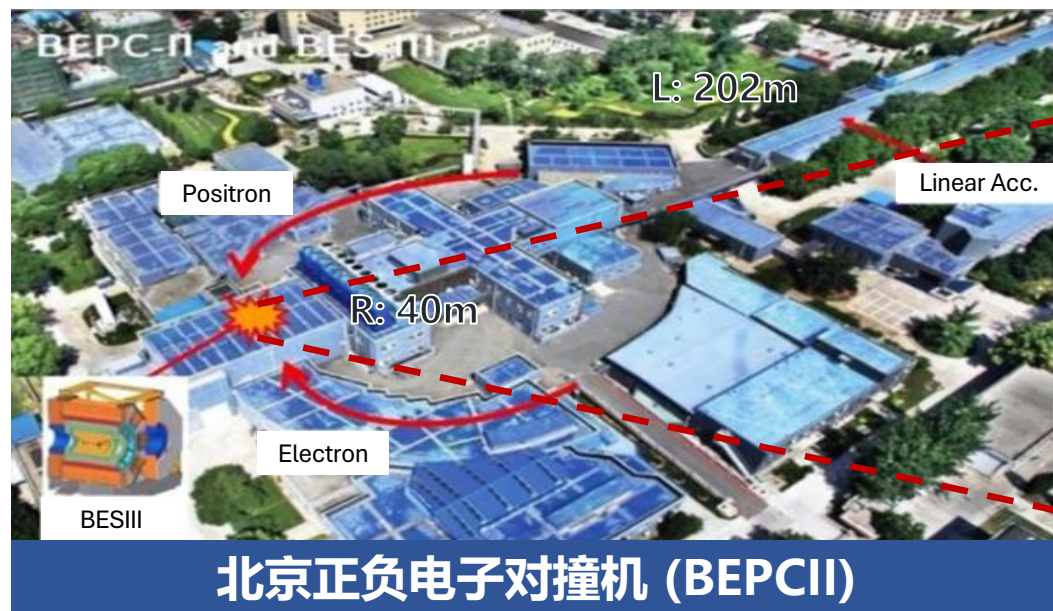
将对文字编辑、程序编译、智能问答等行业带来巨大冲击

GPT-4o

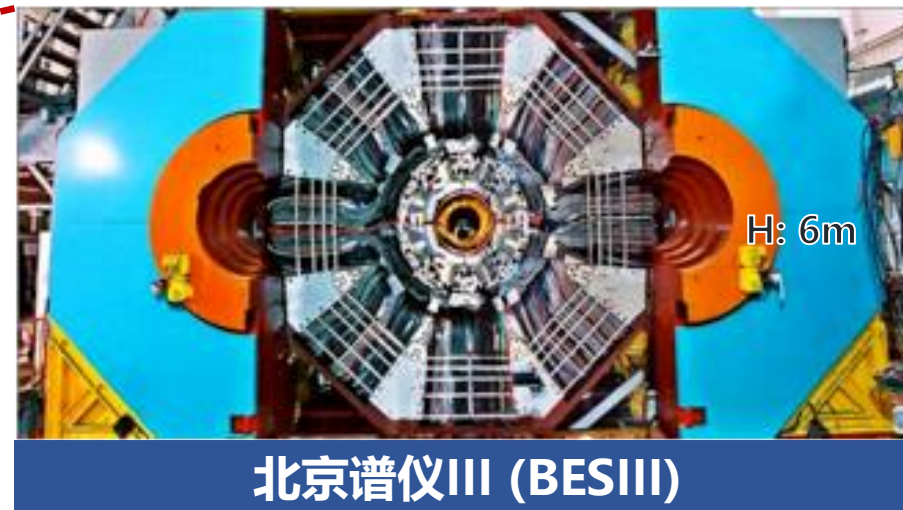
GPT-o1

Deepseek v3

大模型科学研究发展



安装在
BEPC上的
探测器



BESIII

- 丰富的物理
- 超过600篇论文
- 16个国家，超过600科学家参与

数据

- 世界上最大陶-粲能区的数据样本
- 2-5.6GeV, 共振峰和扫描

充分利用

- 物理潜力尚未充分发掘
- AI能帮助挖掘海量高维数据中的数据潜力。

整体路线



高能AI科学家

Research Agent



Dr.Sai(赛博士)

Large Language/Image Model



HEP·Xiwu(溪悟)

Large Science Model



HEP·LSM(科学大模型)

- 文献调研
- 研究选题
- 数据处理
- 物理分析
- 结果解释
- 文章撰写

- 理解物理数据
- 掌握物理规律
- 发现规律
- 自主选题
- 完成研究

General Large Model

LLaMA

GPT-4

Deepseekv3

...

HEP Text/Img Data

21K Q&A pairs

750M tokens

Computing Power

DCU or NPU

GPU

HEP Scientific Big Data
(Tens of PB)



...