

请勿外传
LM4HEP



大模型在高能物理领域的 应用思考和进展

张正德

中国科学院高能物理研究所 计算中心

2023年6月15日

目录

01 大模型概述

02 大模型在高能物理应用的进展

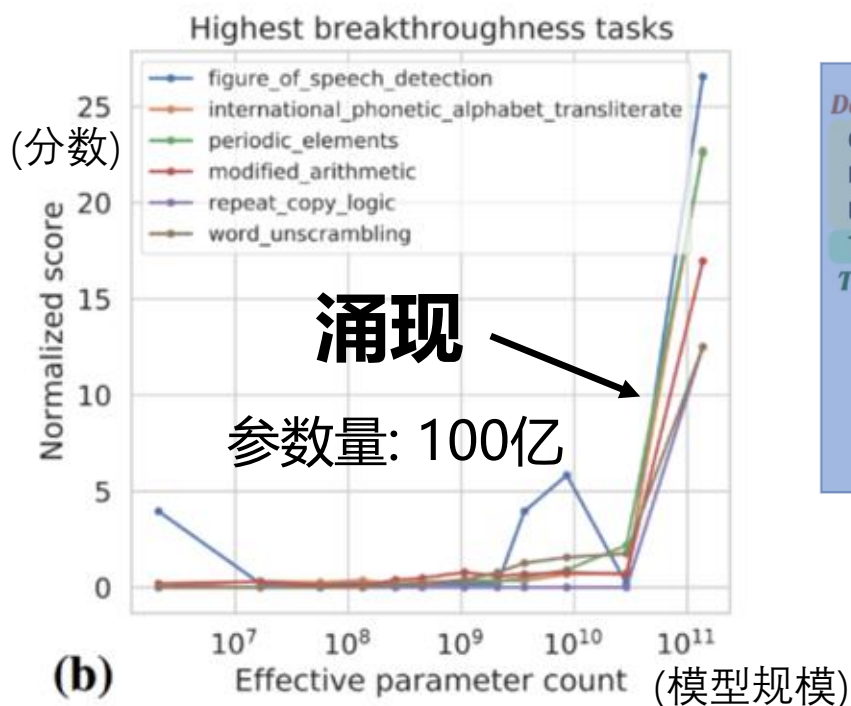
03 HepAI人工智能平台

04 总结

GPT-4 & 大模型



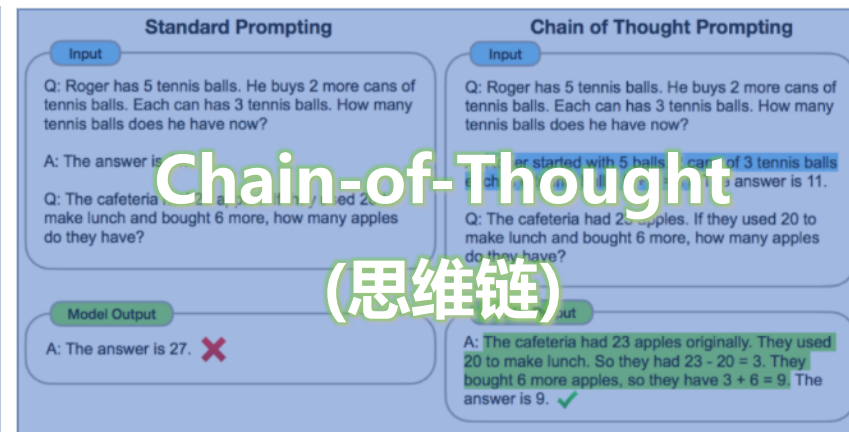
GPT-4在所有自然语言处理下游任务上超过了专用AI模型，是通向通用人工智能的可能路径。¹
证明了**预训练和指令微调大模型**的模式是实现更高层次智能的有效途径。



大模型的涌现现象



In-Context Learning
(文本内学习)



Chain-of-Thought
(思维链)

大模型涌现的特殊能力

可行路径:



¹ 10.48550/arXiv.2303.12712

目前的“高层次”智能

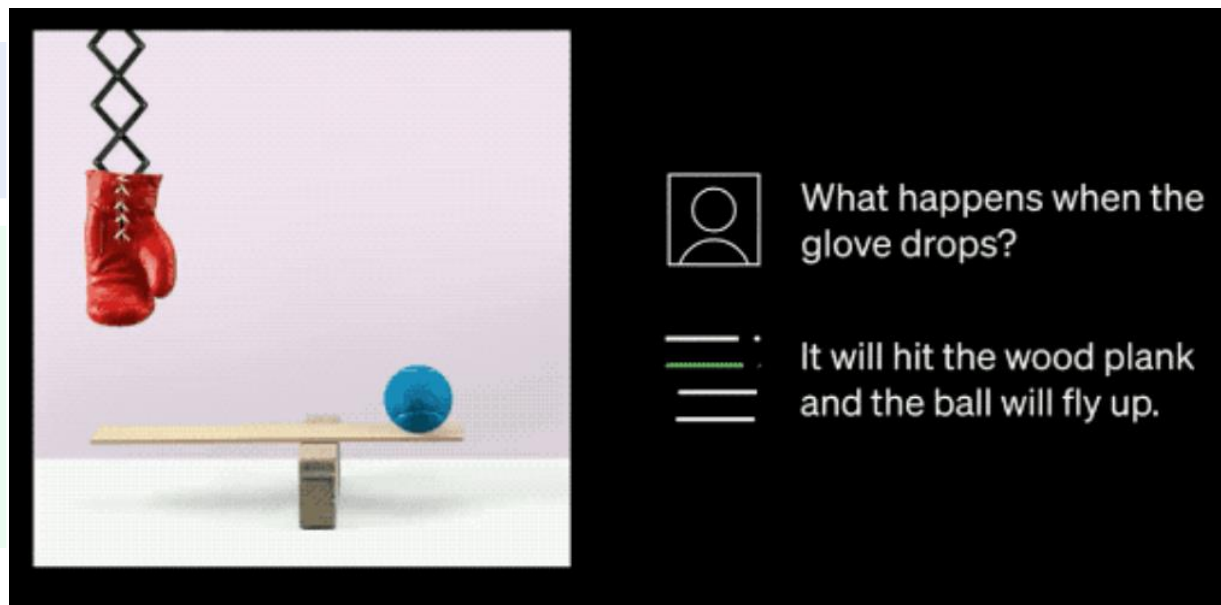


👤 有两栋大厦，A大厦和B大厦，两个大厦在3层连通，我在A大厦7层办公，想去B大厦9层，需要坐几次电梯，一共坐多少层？

🧠 由于A大厦和B大厦在3层连通，因此可以从A大厦的7层乘电梯到3层，然后从3层走到B大厦，最后再从B大厦的3层乘电梯到9层。因此，需要坐2次电梯，一共坐了：
 $(7 - 3) + (9 - 3) = 10$ 层
总共经过10层。

通用意图理解能力
强大连续对话能力
智能交互修正能力
较强逻辑推理能力

ChatGPT体验地址：<https://ai.ihep.ac.cn>

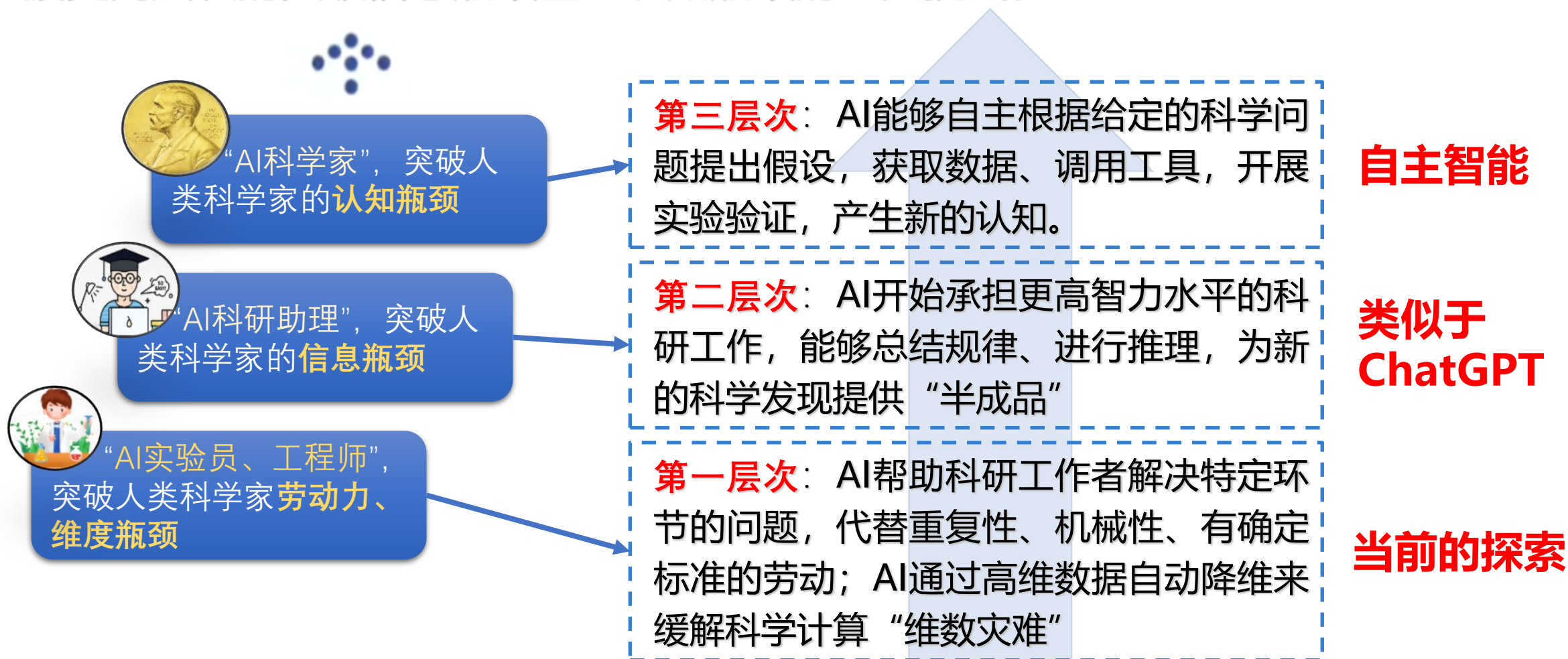


- 律师模拟考试排名前10%
- SAT(美国高考)排名前10%(ChatGPT在倒数10%)
- 物理所考研《量子力学》77/150→100/150

GPT-4带来的挑战



GPT-4带来了人工智能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。



“小模型” 和 “大模型” 对比



高层次智能如何而来? **大数据+大模型+大算力**

十万(ParticleNet), 千万(YOLOv8)

十亿(SAM), 千亿(ChatGPT)

参数量

11万图像(YOLOv8)

1100万图像(SAM), 1.4万亿Tokens(LLaMA)

训练数据量

不需要人工标注!

10G显存 10TFlops(YOLOv8)

10000G显存 2400TFlops(ChatGLM-1300亿)

所需算力

小模型

大模型

优缺点

- 速度快
- 性能依赖于大量数据
- 无可解释性

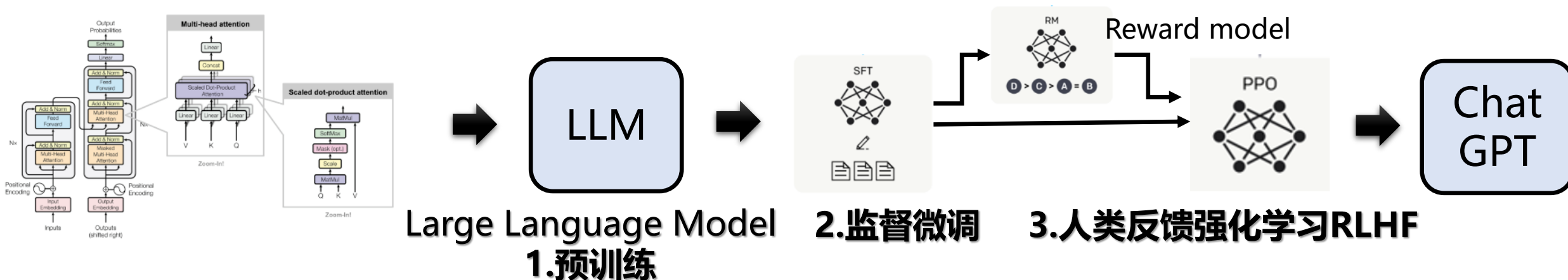
优缺点

- 速度慢
- 强大泛化能力, 具体任务极低的数据需求
- 带来“可解释”的新思路

BESIII 6PB原始数据大约3千亿Tokens。

高能物理领域的**海量大数据**提供了**发展大模型的绝佳基础**。

ChatGPT是如何炼成的



基于Transformer(2017)

本质上, LLM是能“**预测下一个词**”的“词语接龙”模型

博览群书——训练数据多
过目不忘——模型能力强

USER: Please introduce the Institute of High Energy Physics. 10 tokens

LLM: The Institute of High Energy Physics (IHEP) was established in 1973, It is one of the research institutes under the Chinese Academy of Sciences ... 32 tokens

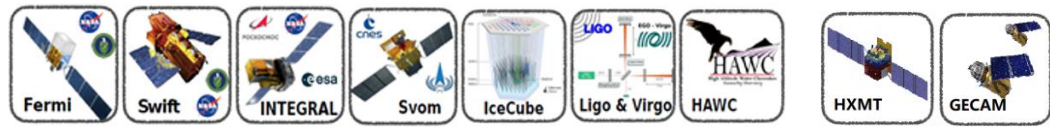
Token(标记): 单词或字母在词表中的索引(词表大小50257); Token是模型能辨别的最小单元。
Tokenizer: 把句子转换为标记的工具。

大模型在高能物理的应用进展

大模型带来的机遇



“一站式”天文警报信息汇集平台



```

TITLE: GRB230311.82
DATE-OBS: 2023-03-11T19:46:31.000
RA: 16h 33m 41.21s
DEC: -47d 38m 32.3s
...
    
```

平台多、信息孤立、零散

多源信息整合关键技术

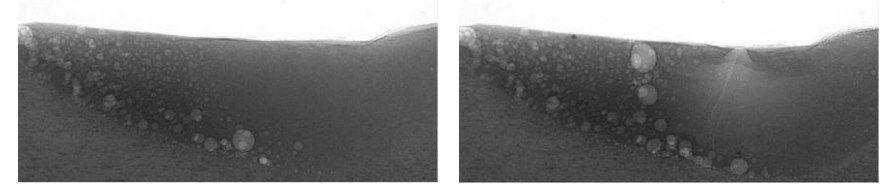
自然语言处理大模型: ChatGPT

```

source_name : Swift GRB230311.82
source_type : Gamma Ray Burst
RA : 16h 33m 41.21s
Dec : -47d 38m 32.3s
Error_R : 0.05
Observation_time : 2023-03-11 19:46:31 UT
Instrument_name : MASTER-SAAO robotic telescope
    
```

X射线增材制造缺陷智能分析

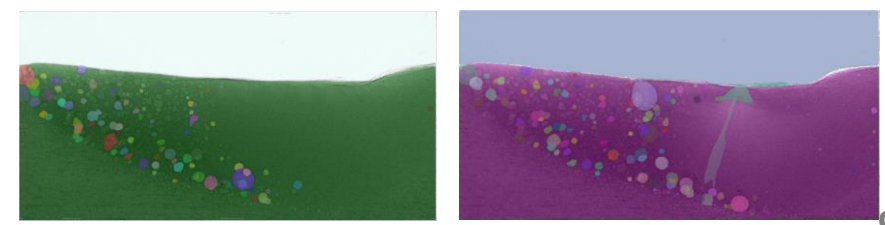
对航空航天国防等战略性高端装备研发与快速制造有重要意义。微观结构对材料性能有显著影响。



复杂、低分辨、动态

缺陷检测和跟踪关键技术

计算机视觉大模型: Segment Anything Model



大模型的泛化能力加速始数据到AI-Ready数据集的过程



洋枪大炮 VS 大刀长矛

发展高能物理领域大模型的路线

应用层面

文本/图像任务涌现

领域知识问答、学科代码、学科图像处理

科学任务涌现

Jet分类、Shower模拟等复杂多样的科学任务

系统涌现

提出问题、验证问题、发现知识

算法层面


HEP文本和图像大模型


HEP科学数据大模型

AI科学家

数据层面

 领域文本数据

 领域图像数据

 领域科学数据

JetClass数据集

JetImages数据集等

$$\begin{aligned} \mathcal{L}_{\text{StandModel}} &= \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{Higgs}} \\ &+ \mathcal{L}_{\text{fermion}} + \mathcal{L}_{\text{Yukawa}} \end{aligned}$$

算力层面

8*A100 (80G) GPU

GPU → NPU

NPU、DCU

2023

2028

2035



大模型在高能物理领域的初步探索



HaiChatGPT, 基于OpenAI API的免费体验版, 无需梯子, 性能与官网一致。



Web界面



命令行界面

- 开源地址: <https://github.com/zhangzhengde0225/HaiChatGPT>
- 网页: <https://ai.ihep.ac.cn>(内网、公网)

2023年3月8日上线以来, 共计403位ihep用户提供了2.3万个有效问答(截止6月8日)。

大模型在高能物理领域的初步探索



私有化、定制化高能物理文本大模型ChatHEP

模型：基于开源Vicuna-130亿参数语言大模型

算力：8*A100 (40G)

数据：

- HaiChatGPT收集并筛选的问答数据1200条；
- **种子裂变技术**引导GPT-3.5形成的高能物理、粒子物理、天体物理、同步辐射、中子科学问答数据2500条；
- GPT-3.5+人工从各中心文献的摘要和正文中提取的信息生成的问答数据1126条；
- IHEP官网公开信息提取的数据；
- arXiv上高能物理文献：引用大于10的2万篇。

技术：

- 采用Float16和量化技术减少显存占用
- 采用FlashAttention提升显卡FLOPs的利用率
- 采用FSDP全分片数据并行实现全模型训练
- 采用LoRA低秩自适应技术显著降低微调需求
- 采用Zero系列技术降低显存需求

发展方向：

- 能处理文本数据 ✓
- 能处理图像模态
- 能处理科学数据

大模型在高能物理领域的初步探索



提出**种子裂变(Seed Fission)**技术

Seed:

“高能物理”

Newbee

种子裂变

Checker

Answers

Expert

智能体:

- Newbee: 根据输入提出2个问题
- Checker: 挑选问题
- Expert: 回答问题

仅需一个“种子”就能引导语言模型输出**有深度、多样化**的数据集，极大地降低标注成本。

Questions:

- 高能物理研究的是什么?
- 高能物理学家使用什么来探索宇宙的基本结构和演化?
- 费米子具有什么样的自旋?
- 引力波是如何产生的?
- 红移可以用来估算什么?
- 什么是强相互作用?
- 如何检测暗物质?
- 夸克有几种“味道”? 分别是什么?
- 弱相互作用在核反应中扮演了什么样的角色?
- 为什么中子星密度非常高?
-

大模型在高能物理领域的初步探索

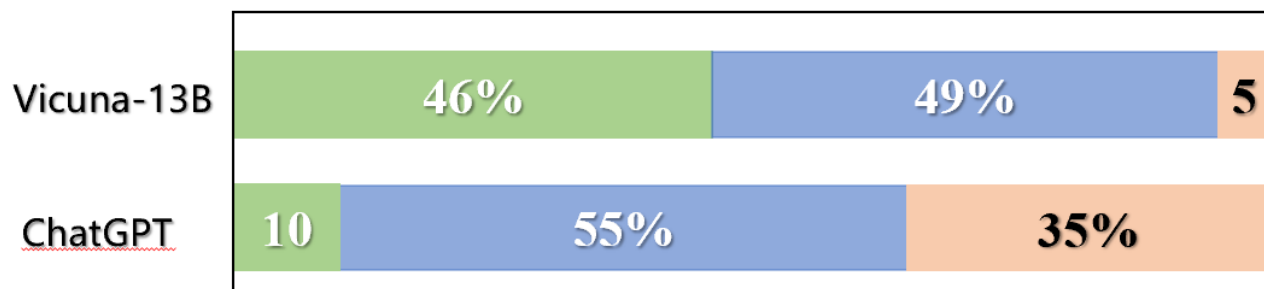


ChatHEP-130亿文本大模型测试结果

测试集问题抽样:

- BESIII的Topology 分析是什么?
- 为什么我们无法直接观测到暗物质?
- HEPAI是什么
- Why is particle therapy more expensive than X-ray treatment?
- What is the potential of the Higgs portal cosmology research field?

ChatHEP wins Tie ChatHEP loses



Human preference evaluation, Compare ChatHEP-13B with Vicuna-13B and ChatGPT-175B

测试数据:

- 粒子物理、天体物理、同步辐射和中子科学领域随机采样数据, 领域内部知识。

结果:

- 与基准模型Vicuna-13B相比, ChatHEP-13B 战胜或平局率达到95%, 在领域问答上**性能远优于基准模型**。
- ChatHEP-13B的性能达到了ChatGPT-175B的65%, 而Vicuna仅不到10%。

结论:

- 通过预训练、微调和Embedding等方法能有效的提升模型的领域概念准确问答的能力

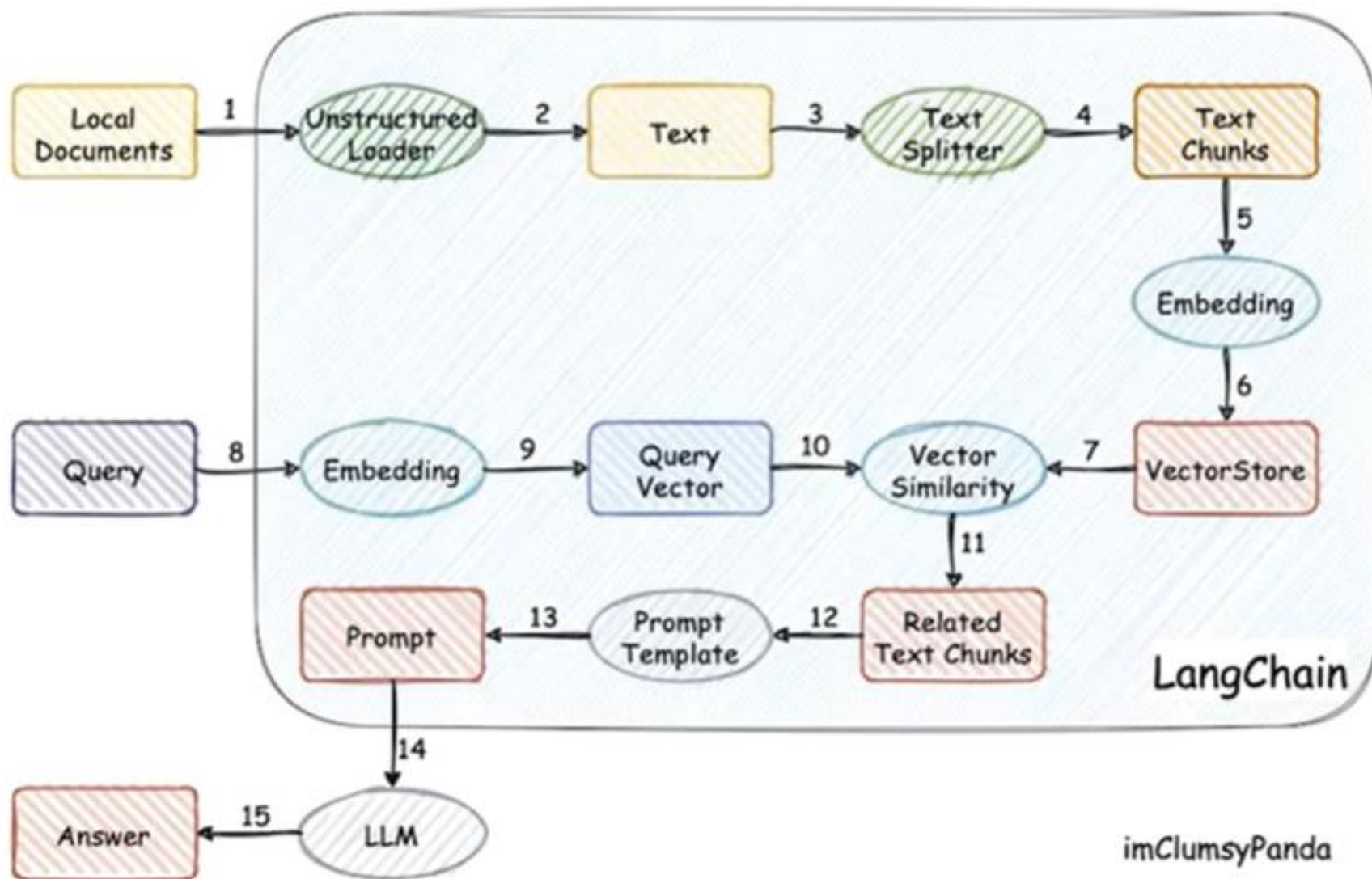
未来通过迭代在领域上能超越ChatGPT

大模型的能力扩展



Chain链!

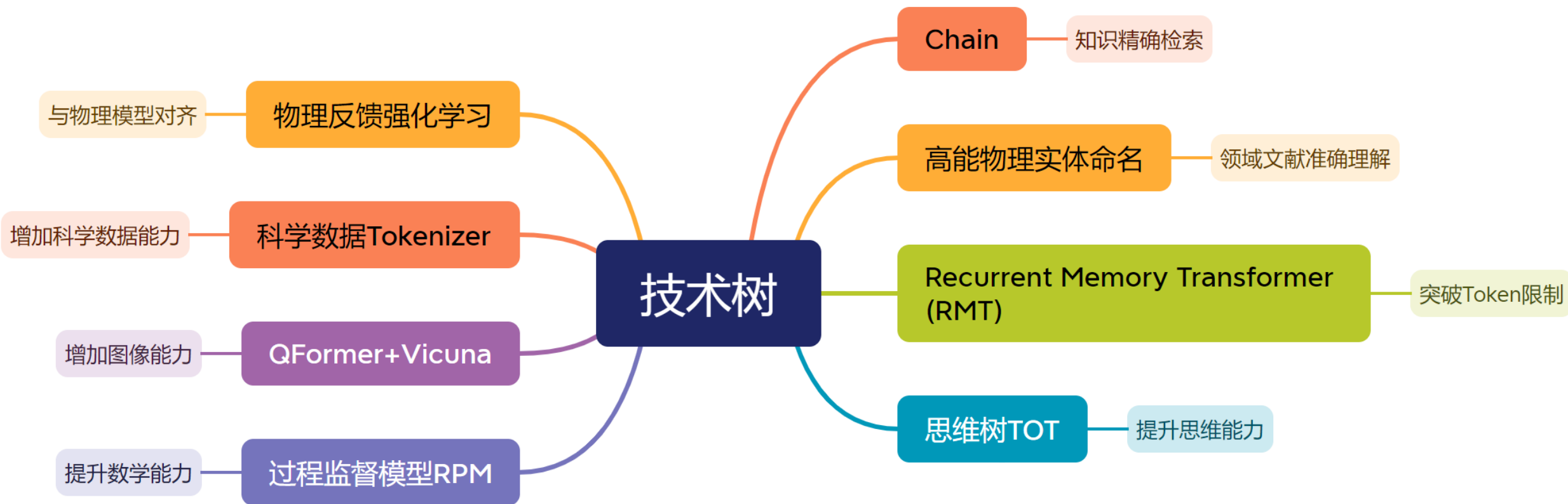
- 外部文档飞行时加载
- 知识精确检索



大模型的能力扩展



从技术层面ChatHEP/HepLM的更多能力扩展



高能物理人工智能平台HepAI

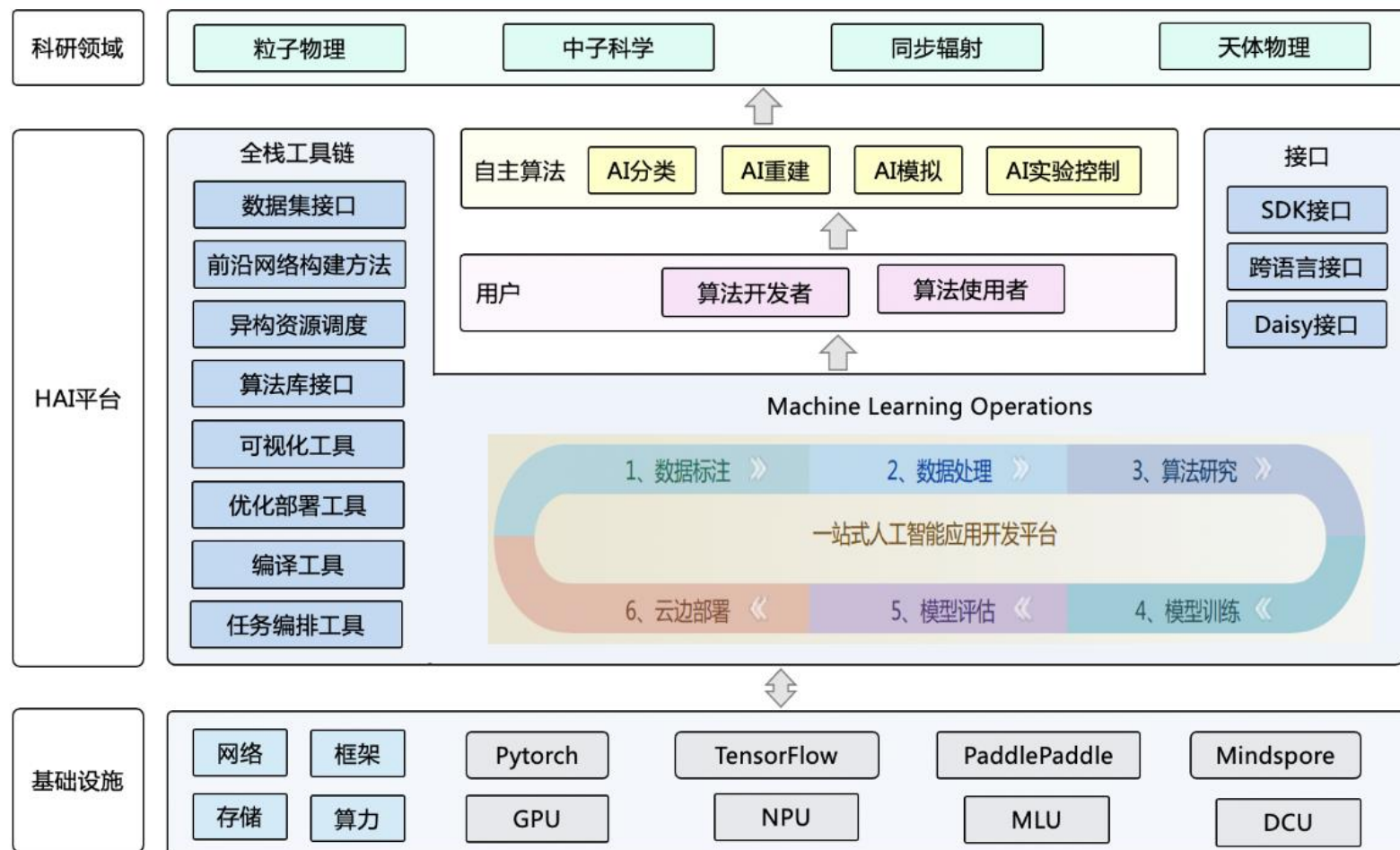
HepAI人工智能平台



人工智能平台可以加速多学科场景下的科学研究、简化模型迭代和流动，是发展AI算法及应用的共性基础设施。

- HAI算法框架 (30%)
- AI算法库: 4个
- AI数据集: 3个
- 算力资源
- 工具链:
 - 标注工具
 - 开发GUI和WebUI框架

目标: 让AI For HEP
更加简单、水平更高!



<https://code.ihep.ac.cn/zdzhang/hai>

人工智能界面框架HaiGF



全新自研**可扩展、轻量化**的QT GUI界面框架HAI GUI Framework

- AI数据标注、训练、测试、产品开发



全新的可扩展设计

①核心功能栏

- 资源浏览器
- 标注工具
- AI工具(远程)
- 更多功能

②主侧栏

- 核心功能的具体展开

③中央控件

- 可视化数据、交互
- 基于选项卡+页面的可扩展设计
- 基于分屏器的自动分屏

④辅助侧栏

- 详细属性、信息等

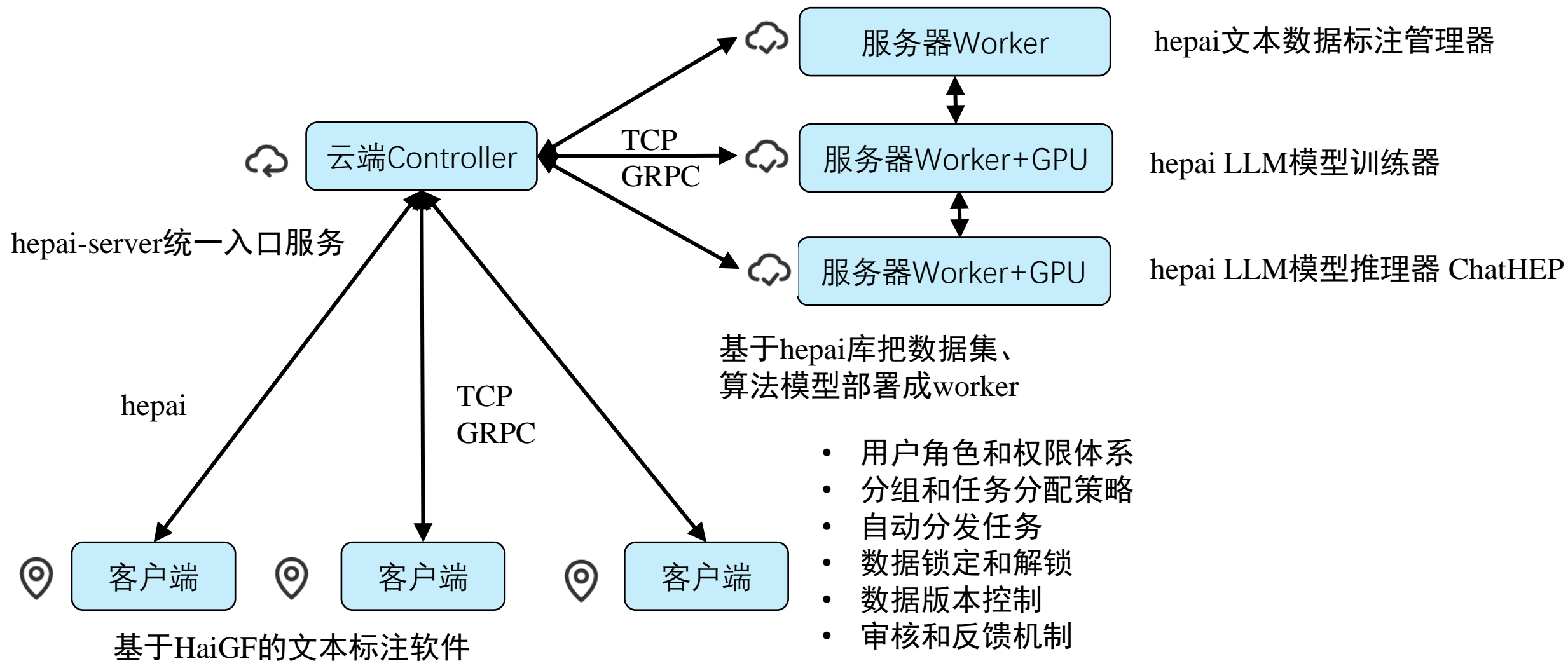
⑤面板栏

- 多选项卡输出面板

HepAI分布式数据标注/模型部署模块



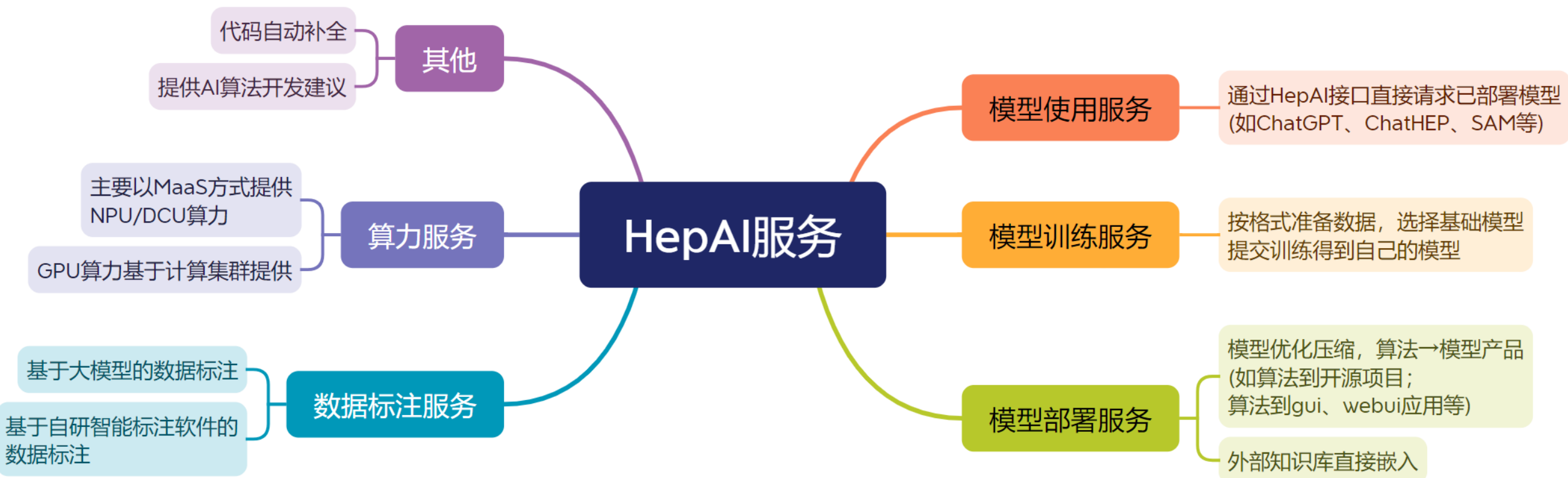
全新自研**跨系统、高并发**的分布式部署模块



HepAI服务



- 主要以**模型即服务(Model as a Service, MaaS)**方式提供





- 划分科研范式主要依据之一是科研工具的使用，**新工具的使用必然会提高科研效率，催生新成果。**
- 院里：发挥高能物理学科基础和优势，打造高水平的数据和AI驱动平台。
- 所十四五规划：先进计算技术使更强大的建模与模拟成为可能，是高能物理取得重大突破不可或缺的手段。需加强机器学习在实时处理、模拟、重建、分析等方面的应用，提升高能物理探索和新发现的能力。
- **仰望星空，脚踏实地（立足常规，着眼新奇）**
- 利用先进技术去推动基础研究，加速科学进步，造福人类社会。



请批评指正！