国家高能物理科学数据中心
National HEP Science Data Center

高能所計算中心
IHEP Computing Center

面向科学研究的一站式高能AI平台

Yiyu Zhang On behalf of Data & AI Group
IHEP CC , CAS
2026-11-06

Outline

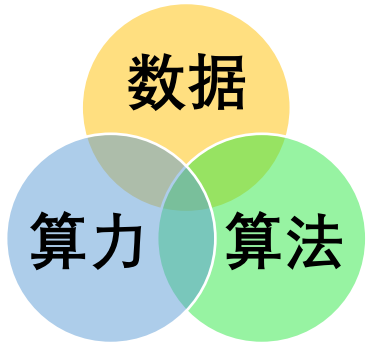
- 背景与平台定位
  - 高能AI平台
  - 传统 AI 支撑高能物理研究
  - 智能体时代的平台支撑
- 总结与展望

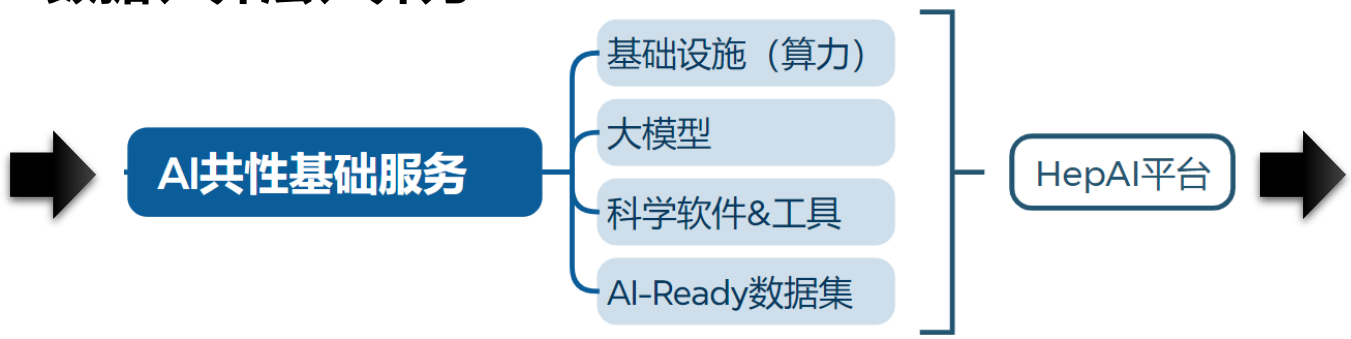
背景

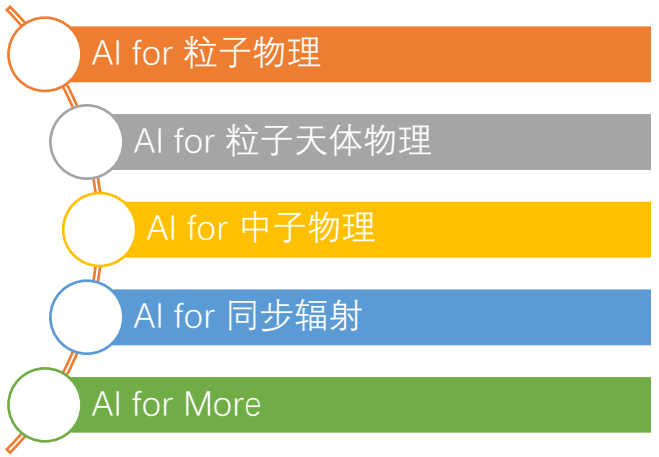
- **AI研究流程：复杂、零碎，需要共性基础服务简化流程**

任务	数据	AI-ready 数据集	模型	训练环境	训练&推理
• 事例重建 • 图像识别 • 知识问答	• 探测器 • 数据预处理 • 科学软件	• JetClass • QuarkGluon • Hep Q&A	• Github • Huggingface • 自行构建 • **Particle Transformer**	• 训练框架 • PyTorch • Tensorflow • Weaver	• 算力

- **目前高能所算力硬件共250+英伟达GPU，以V100为主，少量A100（非AI研究独占）。**
- **人工智能三要素：数据、算法、算力**







高能AI基础设施



目标：构建一种满足人工智能场景中长时间运行、高交互性和强隔离性特征的新型AI基础设施。



- 资源集中管理，统一服务；
- 多维度实现合作共享；
- 系统间相互兼容，互联互通。

HAI高能人工智能平台

共性基础设施：面向粒子物理、粒子天体物理、同步辐射、中子科学、加速器等多领域，提供从数据治理、算法迭代到算力调度的**一体化AI研发闭环**。

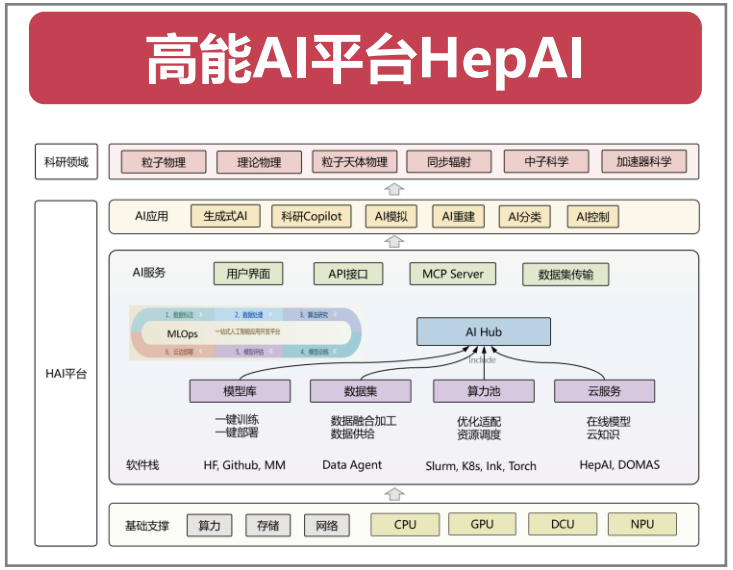
平台价值

- **承载AI模型/工具**：已支持领域10+模型/工具。
- **新建专用AI算力（2026）**：40卡GPU+ 48卡DCU
- **打通数据通道**：文本、语料、科学等10+AI数据集
- **赋能智能应用**：6+AI应用。

AI共性支撑能力

- 01 基础设施（算力）
- 02 大模型
- 03 科学软件&工具
- 04 AI-Ready数据集

整合



（1）公共模型服务

- 大语言模型：DeepSeek, etc
- AI算法模型：SAM, ParT etc
- 科研工具：PDF解析 etc
- 一键部署、一键训练

（2）AI应用服务

- 生成式AI：HaiChat
- 学术助手：HaiAcademic(姜)
- 论文写作助手：Overleaf(侯)
- 智能体：opendrsai

（3）公共算力服务

- 5PFlops , 200T SSD
- 领域模型在国产设备的适配
- AI易用（像私有服务器一样）
- 高效（智能调度提升整体利用率

(4)公共数据集服务

HAI高能人工智能平台入口



算力资源：异构是常态，不是例外

AI 专用节点构成



另有 CPU 节点、1PB 全闪存与并行文件系统，
用于数据准备、预处理与结果暂存。

按“工作集 + 任务形态”选卡，而不是只看算力峰值

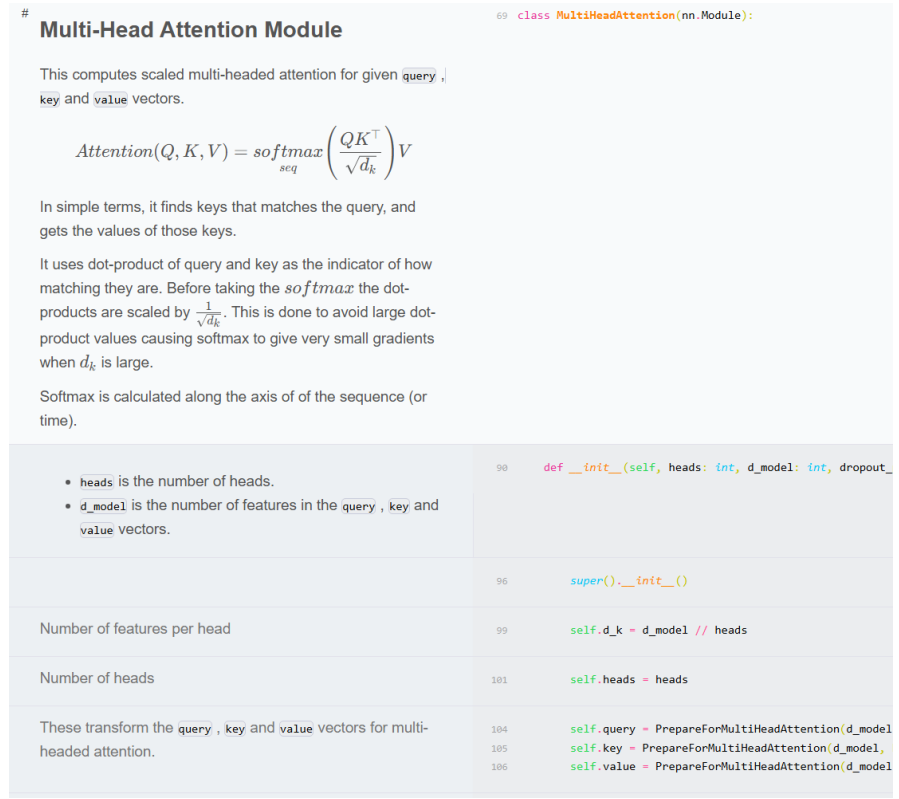
资源	优势	更适合	注意
RTX 5090	交互性好、性价比高	原型、小模型微调、推理	显存有限，先做小批量
L40	48GB、适合多实例服务	视觉推理、批量推断	训练优先让给 A800
A800	大显存、适合多卡训练	领域模型训练、微调	资源稀缺，走批处理队列
DCU K100AI	国产加速生态、规模化	大吞吐推理/训练迁移	需适配 DTK 软件栈

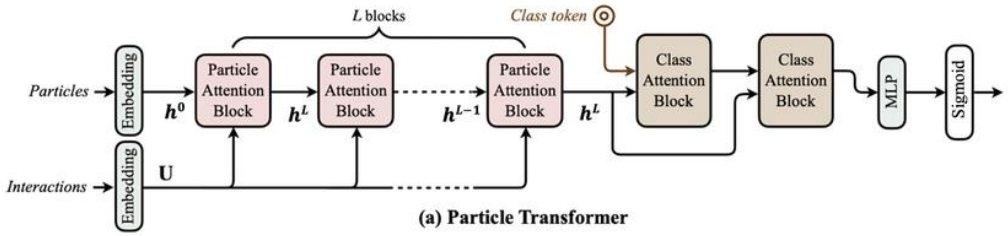
AI算力服务： AI 框架/模型、科学软件&工具

- **AI框架**
  - Slurm(资源)+kvm(环境)+conda(软件包管理)
  - Pytroch，tensorflow, …
- **模型快速复现**
  - FCN，CNN，transformer，VIT
- **科学模型：**
  - SAM(分割一切)，PtychoNN(叠层成像重建)
  - ParticleNet&ParT(粒子鉴别), PointNet (事例重建)
- **大语言模型：**
  - gpt-4, gpt-4o, gpt-4o-mini, gpt-o1(推理)
  - llama 3, Qwen
  - Deepseek-v3, deepseek-R1, 蒸馏模型
  - Xiwu

PyTorch

TensorFlow





AI算力服务

算力怎么用：按任务形态选择入口

A 交互式开发

ECS • JupyterLab • VS Code Remote

1–8 卡；调试、可视化、单机多卡

→ 快速试错，保留运行时状态

B 批量训练 / 扫描

Slurm 作业 • 作业数组 • checkpoint

多机多卡；长任务、参数扫描

→ 吞吐高，适合可中断的批处理

C 在线推理 / 服务

AI K8s • 容器 • 自动扩缩容

按 QPS / 延迟配置副本

→ 稳定服务，需监控与灰度

D 零运维调用

HepAI Token API • MaaS

按 token / 调用计费，无需管卡

→ 低门槛，适合文档、代码、Agent

资源使用闭环

配额 — 环境 — 数据挂载 — 作业/服务

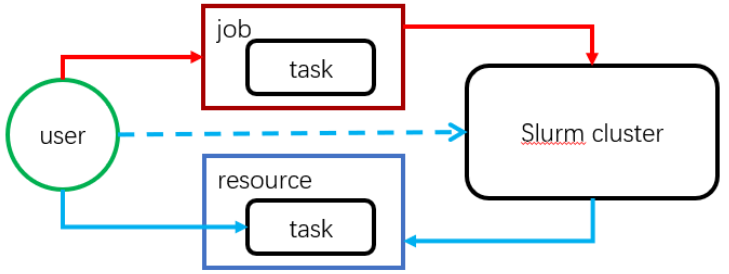
先定义数据工作集、并行策略、精度与吞吐目标，再决定卡型和队列。

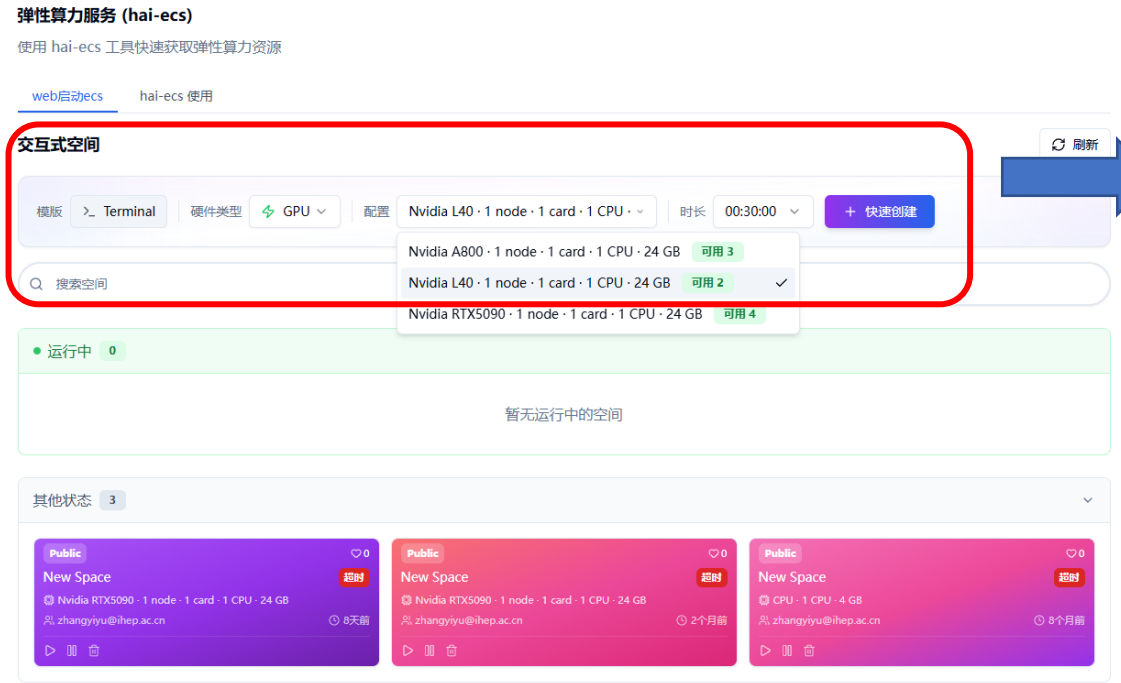
AI算力服务：ECS 弹性算力服务

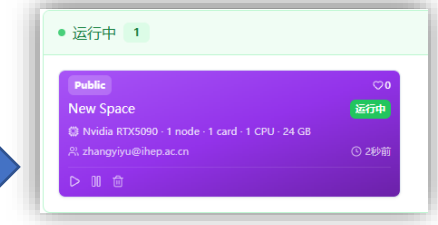
ECS 弹性算力

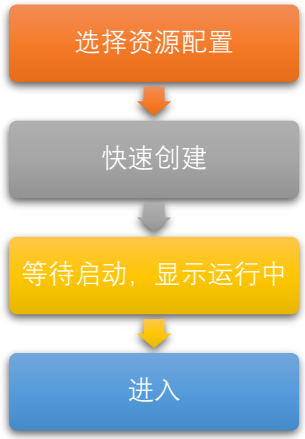
面向高交互、快速迭代的单机任务，提供可直连、可扩展的运行环境。

可按需配置 CPU、内存与 GPU，适合代码调试、小规模训练和推理验证。











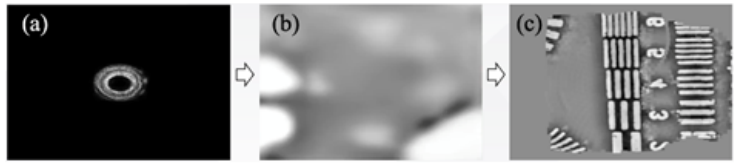


AI应用服务： OpenDrSai通用智能体



地址： https://k8s-ai.ihep.ac.cn/

AI用于数据精细化处理



(王磊、常广才)

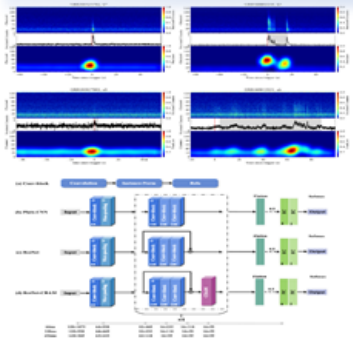
叠层成像快速重建



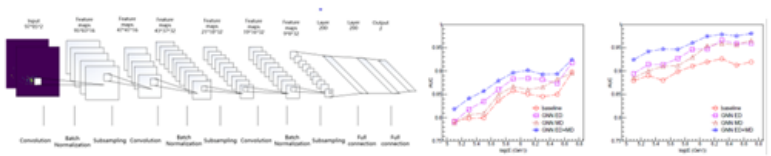
基于NN和知识嵌入，实现较传统算法**100倍加速和1/5投影减少**

天体物理高能爆发事件甄别

(李兵)



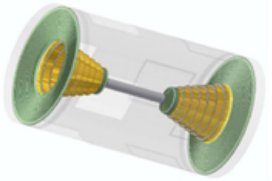
基于ResNet和注意力机制，实现对GRB、SGRs和SFLARE等高能爆发的准确判别，展现出意料之外的**低阈值触发事件**甄别能力

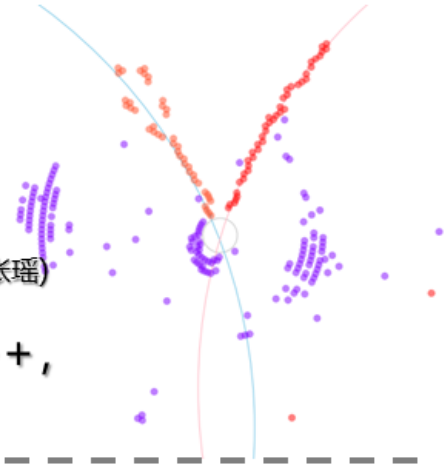


数据处理精细化

BESIII径迹重建

(袁野、杨志斌、张瑶)



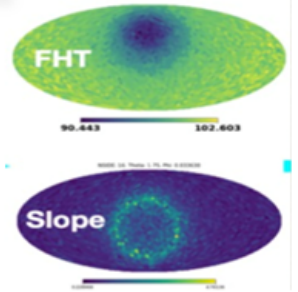


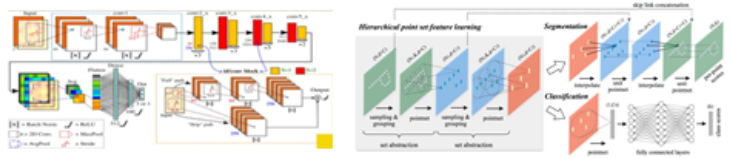
基于GNN和PointNet++，实现**端到端多轨迹跟踪**

JUNO重建和粒子鉴别

(罗武鸣)

基于ResNet、PointNet++等模型，实现**高精度**能量重建和顶点重建，突破大气中微子方向重建和粒子鉴别的**技术瓶颈**





LHAASO鉴别

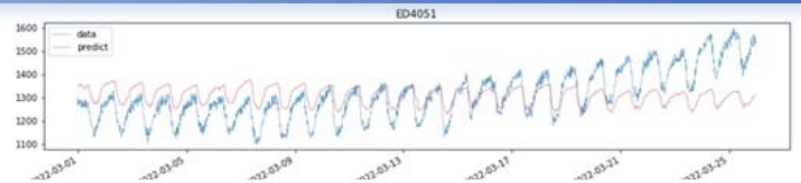
(张笑鹏、陈秀林、靳超)

基于CNN实现WCDA的gamma和p区分，基于GNN和ParticleNet进行KM2A宇宙线原初粒子成分鉴别，**效果显著优于传统方法**

AI用于实验过程智能化

LHAASO时间序列预测和异常值检测发现探测器故障

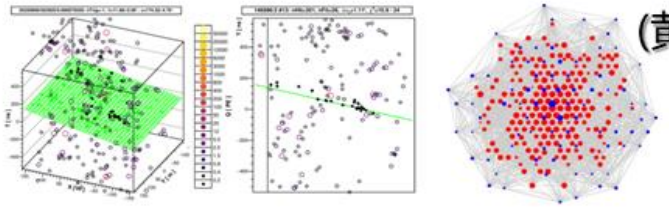
(张笑鹏)



已用于阵列运行和探测器维护

LHAASO GPU在线噪声过滤加速

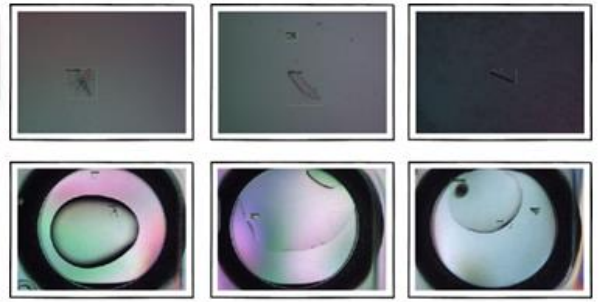
(黄德智)



去除超过90%噪声，真实信号保留>90%

样品自动识别

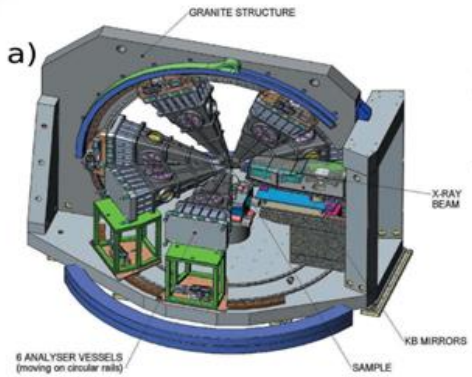
(张一)



基于Swin Transformer+YOLOv3，能检测**传统无法处理**的各种形状晶体和大视场下的小晶体

样品自动调姿

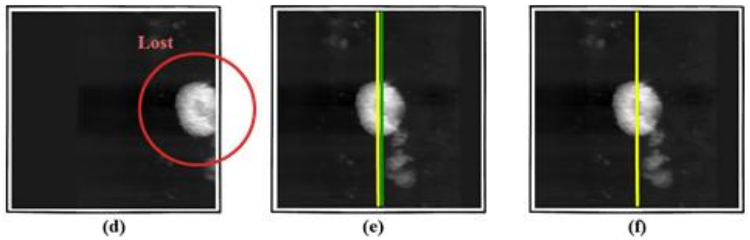
(张一)



基于遗传算法，可实现多组实验异步控制从而**显著加速调姿过程**

样品自动对中

(张一)



基于PPO2强化学习，可实现XANES扫描成像样品的**高精度自动对中**

实验过程智能化

传统 AI 到智能体时代｜AI 如何进入科学研究

核心变化：AI 先提升特定科学算法的性能，再扩展到理解上下文、组织工具和协助完成科研任务。

第一层｜传统 AI

模型替代传统算法

- 用神经网络、树模型等学习数据规律
- 在粒子鉴别、径迹重建等任务中替代部分传统算法
- 通过训练和评估，提升准确率、效率或处理速度

对象仍是明确的单一科学任务。

第二层｜LLM 出现

性能与上下文提升

- 语言理解、生成和推理能力显著增强
- 能处理更长上下文，理解代码、文献、数据说明和任务背景
- 从“完成一个模型任务”扩展到“理解一个研究问题”

AI 开始连接知识、代码与资料。

第三层｜智能体时代

帮助完成科研工作

- 以智能体方式调用模型、代码、文件和工具
- 可参与代码编辑、数据处理、测试运行和结果检查
- DrSai、Codex、Claude 等工具支持多步骤任务协作

研究人员设定目标并审核结果，智能体负责执行与反馈。

对高能物理的意义：传统 AI 解决粒子鉴别、径迹重建等专门问题；LLM 和智能体进一步帮助研究人员理解资料、修改代码、组织计算流程，并推动算法持续迭代。

对AI平台提出了新的要求： 承载领域神经网络训练、大批量科学数据处理与物理评估。提供模型 API 和模型服务，适合语言理解、生成和助手能力。

AI模型服务：智能体正在改变 AI 服务调用方式

传统：人 → 模型

现在：智能体 → 网关 → 多源模型 / 领域软件

HEP 场景的四个特点

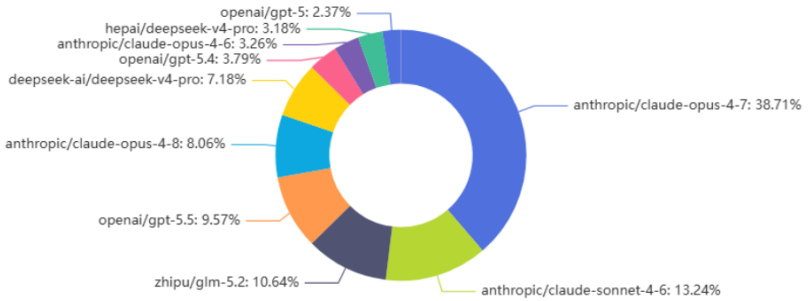
海量调用
专业任务、持续任务、复杂任务

模型异构
通用 LLM + 领域专用模型

资源敏感
云模型成本高、资源紧张

强合规
原始数据、内部结果不可外发

Token 总量
1123.20亿

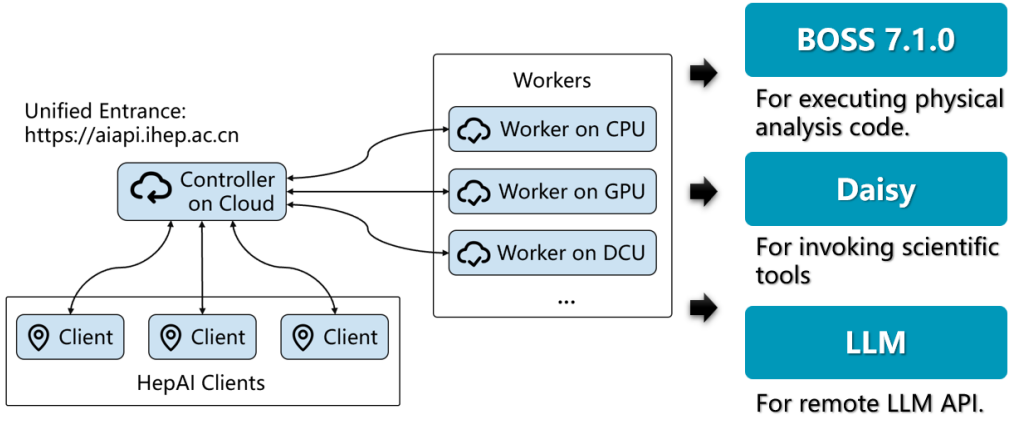




模型服务正由“协议聚合器”变成可信治理与资源调度的中枢

HepAI-MaaS：统一模型与科研工具调用

- **自研分布式部署框架，实现对先进AI算法模型和科学软件工具的直接调用。**
  - **大语言模型**：ChatGPT, GPT-4, Xiwu, etc
  - **科学软件**：BOSS，Daisy
  - **AI算法模型**：SAM，ParticleNet&ParT, PtychoNN
  - **科研工具**：PDF解析，PDF翻译，论文润色，科研绘图，etc
  - **AI 应用**：HaiChat、Dr. Sai、天文警报信息汇集平台、**overleaf助手、zotero助手、vs代码补全， etc**





- **通过HepAI-MaaS(Model as a Service)服务加速AI模型的应用创新**
  - **旧模式**：基于AI模型进行应用开发时，需要下载代码、模型、考虑算力、解决环境依赖等问题。
  - **新模式**：基于HepAI-MaaS服务，使用API-KEY直接调用AI模型，而无需考虑算法部署、算力资源等问题。

Token 服务：统一 API 与模型入口

- **基于模型基础设施的大模型词元（Token）服务，提供统一的API接口和计费**
- 方案：私有化本地部署+第三方云模型协同
  - 模型**开箱即用**
  - 并发高
  - 统一计费审计

https://ai.ihep.ac.cn



提供服务

统一的模型HUB

- **数据内部流转**
- **保障数据安全**

本地模型

hepai/deepseek-v4-flash
hepai/deepseek-v4-pro
hepai/qwen3-30b
hepai/mineru2_2b
...

第三方云模型

- Glm系列
- Minimax系列
- Qwen系列
- OpenAI系列
- Anthropic系列
- Google系列

- **提供最先进模型**
- **支撑研究探索**

AI模型服务： AI models and tools

模型 API 科研集成示例

适用于文献初筛、报告整理和数据集说明等轻量文本任务。

模型输出由研究人员复核，不替代科学判断。

```python
import os
from hepai import HepAI
from typing import Dict, List

client = HepAI(
  api_key=os.environ.get("HEPAI_API_KEY"),
  base_url="https://aiapi.ihep.ac.cn/apiv2",
)
```

```python
def main():
    # 调用函数处理当前文件夹所有.md文件
    result = summarize_md_files()

    # 格式化输出结果
    output_lines = []
    for filename, summary in result.items():
        output_lines.append(f"文件: {filename}")
        output_lines.append(f"摘要: {summary}")
        output_lines.append("-" * 50)

    # 将结果拼接为字符串返回
    return "\n".join(output_lines)
```

```python
def summarize_md_files(model: str = "hepai/deepseek-v3.2", max_tokens: int = 150) -> Dict[str, str]:
    """
    遍历当前文件夹所有.md文件，使用llm为每个文件内容生成摘要。

    Returns:
        Dict[str, str]: 一个字典，键为.md文件名，值为该文件的摘要内容。
                        如果处理过程中发生错误，摘要值将为错误信息字符串。
    """

    # 遍历每个.md文件
    for md_file in md_files:
        file_path = os.path.join(current_dir, md_file)

        try:
            # 读取文件内容
            with open(file_path, 'r', encoding='utf-8') as f:
                content = f.read()

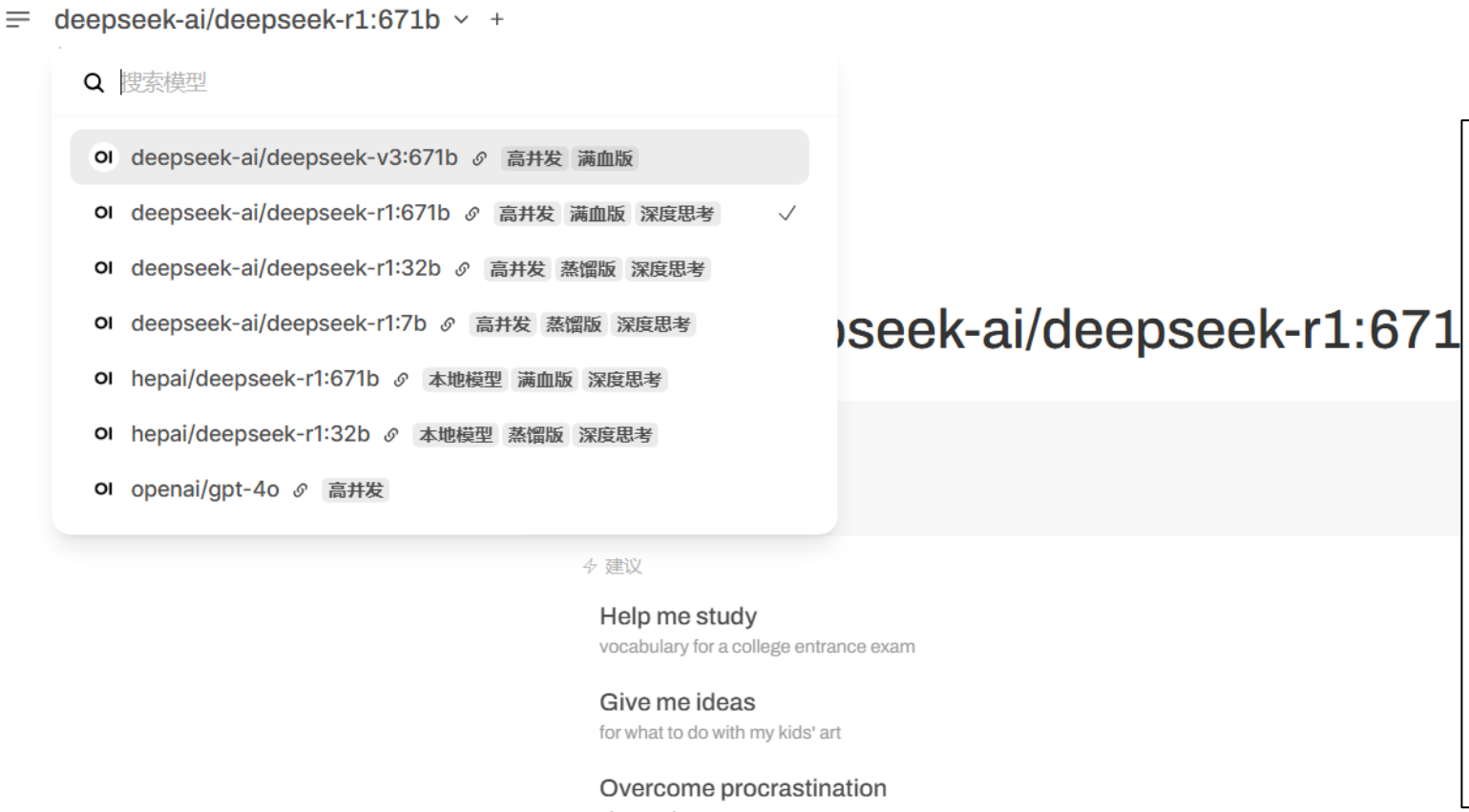
            # 构建提示词
            prompt = f"请为以下Markdown文档内容生成一个简洁的摘要，摘要长度不超过{max_tokens}个token：\n\n{content}"

            # 调用LLM API
            response = client.chat.completions.create(
                model=model,
                messages=[
                    {"role": "system", "content": "你是一个专业的文档摘要助手。"},
                    {"role": "user", "content": prompt}
                ],
                max_tokens=max_tokens,
                temperature=0.5
            )

            # 提取摘要
            summary = response.choices[0].message.content.strip()
            summaries[md_file] = summary


    return summaries
```

AI应用服务：HaiChat





服务特点：支持多模型对话与结果比较，可对回答进行续写、重生成和润色精炼。
适合问答、文本起草、文献粗读与代码思路探索；模型输出仍需科研人员核实。

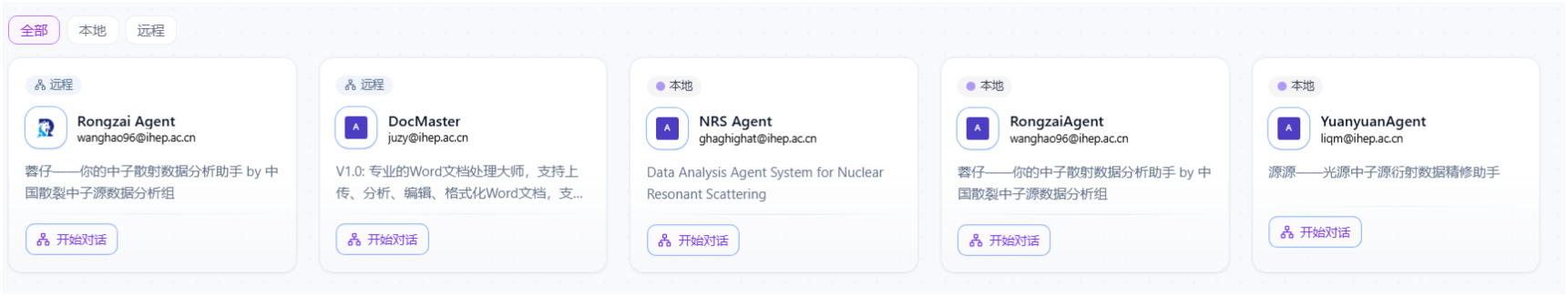
https://chat.ihep.ac.cn

All users: 6016 (IHEP 2980)
Active users: 923 (within 1 month)

AI应用服务： OpenDrSai

OpenDrSai

Open Dr.Sai： 科研个人助手 |



- 通过自然语言理解目标，可调用 shell、文件查找与编辑等能力，协助用户管理服务器环境内的研究任务。
- 支持多轮任务、工具扩展、学术检索与 AI Coding 等能力。
- 适合代码理解、文件处理和任务编排；执行结果需由研究人员审核。

- 智能体辅助算法迭代
- 可协助阅读代码、提出修改建议、运行测试并整理结果



AI应用服务：数据融合智能体

□ 数据融合智能体



总结

- HAI平台秉持数据、算法、算力、应用“四位一体”的核心理念，致力于打造推动AI赋能科学研究的共性基础设施。
- 未来计划
  - 扩充AI算力，以满足更多研究人员的AI开发工作。
  - 完善AI平台的模型和数据集服务，使用户更好的迭代和管理自己的模型和数据集。
  - 完善AI开发工作流，打通模型-数据-算力瓶颈，提供快捷训练和自定义训练服务。

我们诚邀各位用户登录ai.ihep.ac.cn使用AI服务，帮助您的研究和项目。我们欢迎各位用户与我们开展深度的AI研究合作

谢 谢