



# 悟道 · 天鷹 Aquila 语言大模型系列及 天秤 FlagEval 语言大模型评测体系

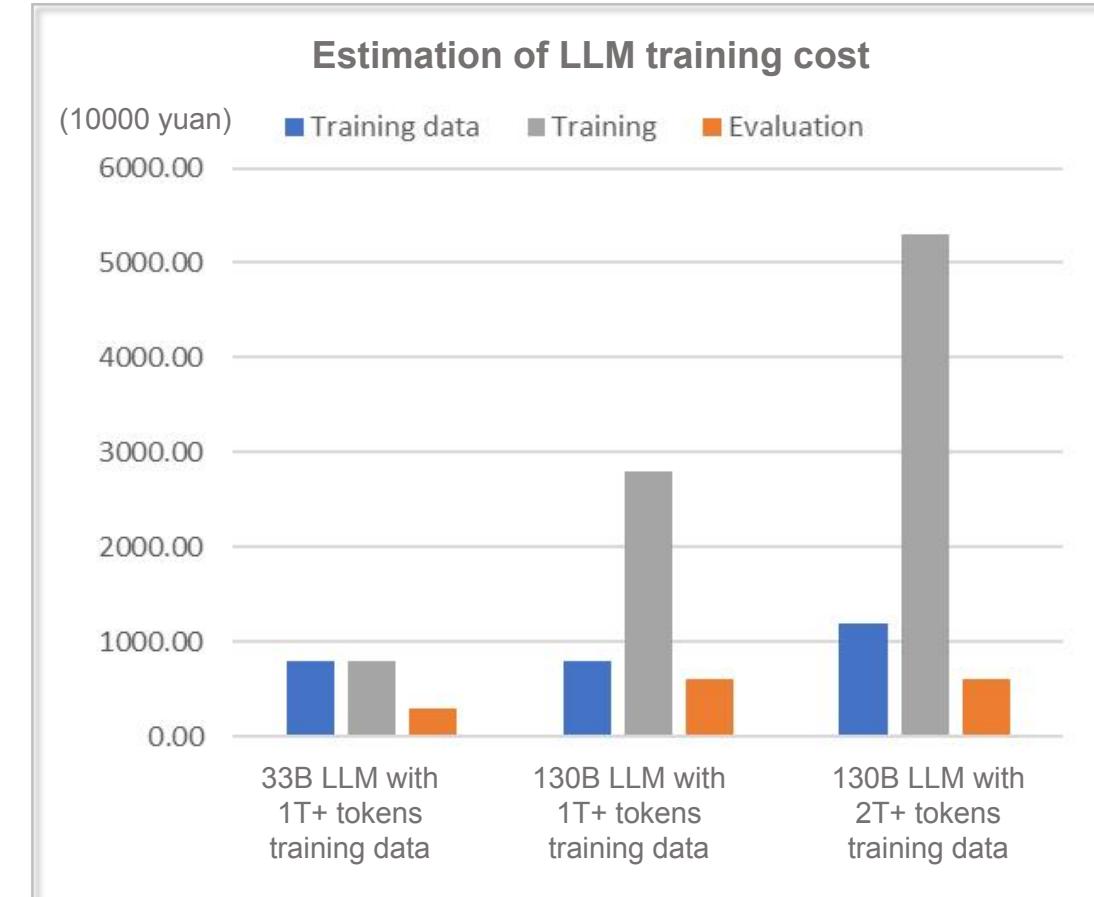
北京智源人工智能研究院

玄日成

BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

基础模型已经成为AI大模型时代，单一“产品”投入最大的部分。

- 对训练一个语言基础模型进行成本的粗略估算（右图）
  - **包括：**训练数据的准备、训练过程、测试评测三大部分。每一部分包括在该部分所需要的人力成本、计算成本等
  - **不包括：**可以分摊到多个大模型训练的成本项，例如工具的开发、新算法的研发等。
- 一个LLM模型的开发成本十分高昂



## 基础模型很大程度决定了后续模型能力、产业落地等因素

- **能力和知识** —— *Superficial Alignment Hypothesis : A model's knowledge and capabilities are learnt almost entirely during pretraining, while alignment teaches it which subdistribution of formats should be used when interacting with users* —— Meta, etc.
  - 在大模型中，“理解”、“涌现”、“上下文学习”等能力，是主要由基础模型的结构、尺寸（训练量）所决定。
  - 知识是在基础模型训练过程中学习所得
- **合规性和安全性**
  - 训练基础模型的语料很大程度会影响AIGC应用、微调后的模型等的内容生成的合规、安全和价值观。
    - 100万条Common Crawl网页数据进行分析，可以提取出中文的网页有39052个网页。从站源角度来看，可以提取出中文的网站共有25842个，其IP所在国家分布如右表所示，其中有4522个位于中国大陆，仅占比17%。—— CC数据集中的中文数据，大部分是海外网站。

国家或地区	数量
美国	10646
日本	5892
中国大陆	4522
中国香港	1578
南非	1184
中国台湾	482
新加坡	146
其他	1392

基础模型很大程度决定了后续模型能力、产业落地等因素

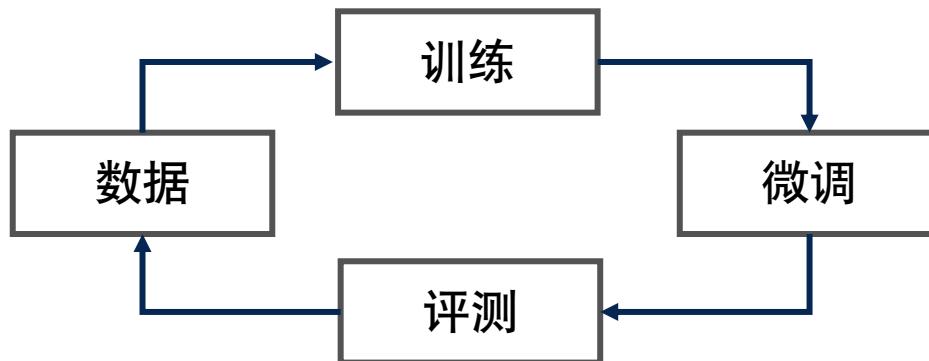
- 版权和商用许可
  - 非商用许可？AGPL协议？

已经发布的国内外通用语言大模型统计（从2023年1月至5月底）

- 国外发布的开源语言大模型有39个，其中可商用、非copyleft协议的大模型有16个
- 国内发布的语言大模型有28个，其中开源的语言大模型有11个，其中开源可商用的语言大模型仅有1个 (BELLE 基于BLOOMz-7B进行指令微调的对话模型)

**具备中英双语知识  
支持商用许可协议、无 Copyleft 限制  
符合国内数据合规需要**

- 为大模型产业打造具备中英文双语能力的语言基础大模型模型，并以可商用协议开放源代码及模型系列。
- 天鹰大模型需要符合语言模型的整体能力框架要求。
- 打造端到端、循环迭代的大模型生产流水线





- 具备中英双语知识
- 支持商用许可协议
- 符合国内数据合规需要

**基础模型：由海量中英文语料预训练而来，中文占40%**

- Aquila-33B (330亿参数中英双语基础模型)
- Aquila-7B (70亿参数中英双语基础模型)

**对话模型：基于Aquila基础模型进行指令微调训练及强化学习**

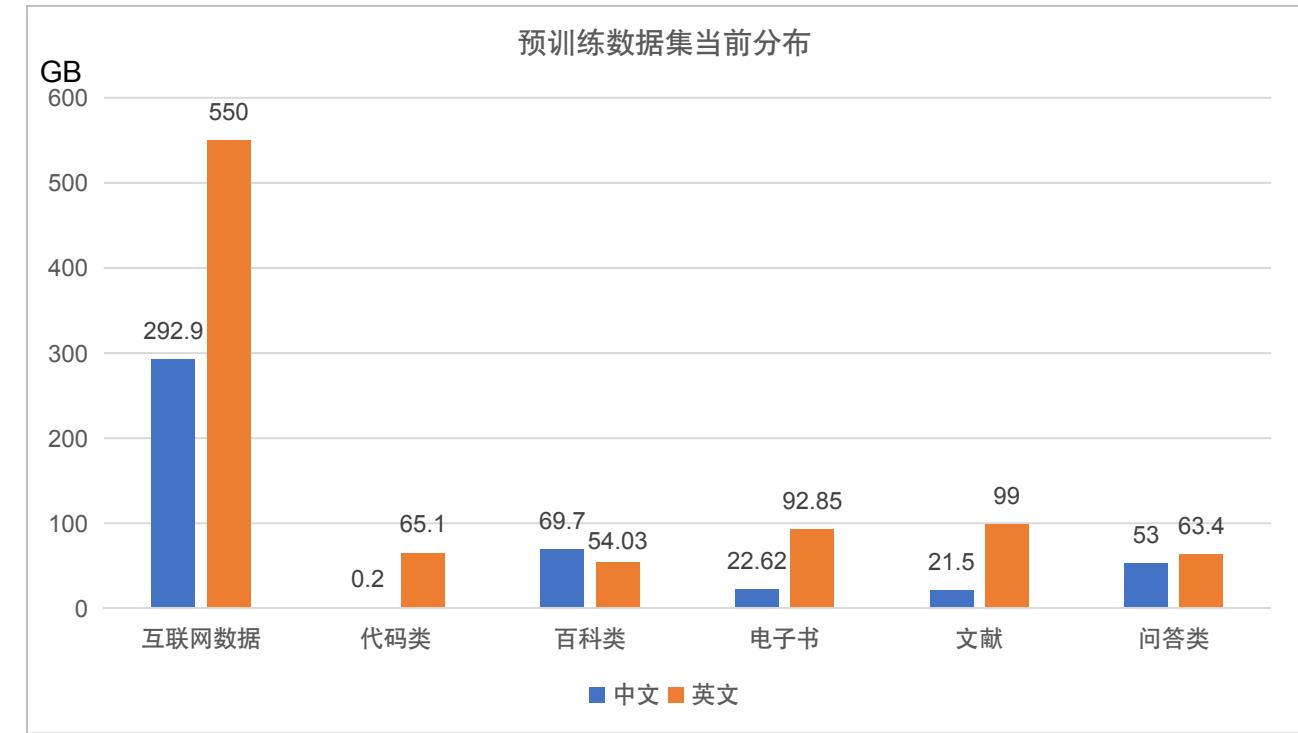
- AquilaChat-33B
- AquilaChat-7B

**代码模型：基于Aquila基础模型进行持续训练**

- AquilaCode-7B

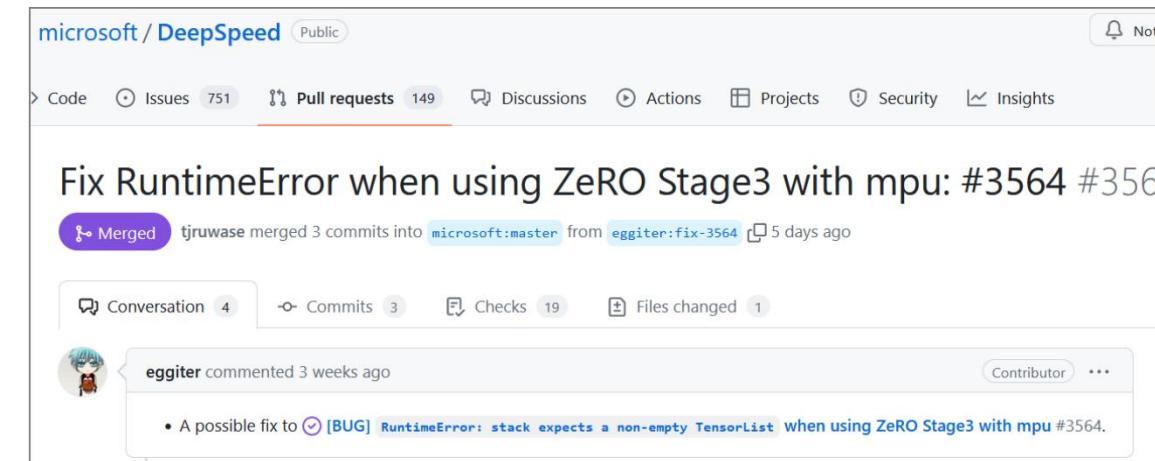
模型已集成至 *FlagAI*: <https://github.com/FlagAI-Open/FlagAI>  
亦可通过 *model.baai.ac.cn* 单独下载权重

- 中文语料来自智源一直积累的中文数据集，包括
  - 来自1万多个站源的中文互联网数据，其中99%以上为国内站源。
  - 获得国内权威机构支持的高质量中文文献数据、中文书籍数据等。
- 当前已经有>1.4T token的训练数据，正持续增加高质量、多样性的数据集，并源源不断地推进Aquila基础模型的训练中。



- Aquila语言大模型在技术上继承了GPT-3、LLaMA等的架构设计优点
- 重新设计实现了中英双语的tokenizer
- 并行训练方法：
  - 升级了BMTrain并行训练方法，包括优化器状态加载、优化计算和通信覆盖（效率 $\uparrow 10\%$ ）以进一步提升性能等，在Aquila的训练过程中实现了比Magtron+DeepSpeed zero-2将近 8 倍的训练效率。
  - 底层算子：Aquila替换了一批更高效的底层算子实现（Flash attention），并且集成到BMTrain的训练框架中。

dimension	heads	layers	Vocab size	max length
Aquila-33B	6656	52	60	100008
Aquila-7B	4096	32	32	100008



# Aquila天鹰语言模型SFT数据打造



# Aquila天鹰语言模型SFT数据打造

- SFT数据采集
  - 人工写prompt+回复
    - 内部数据标注人员+外部公益者
    - 从公开高质量数据集进行指令生成

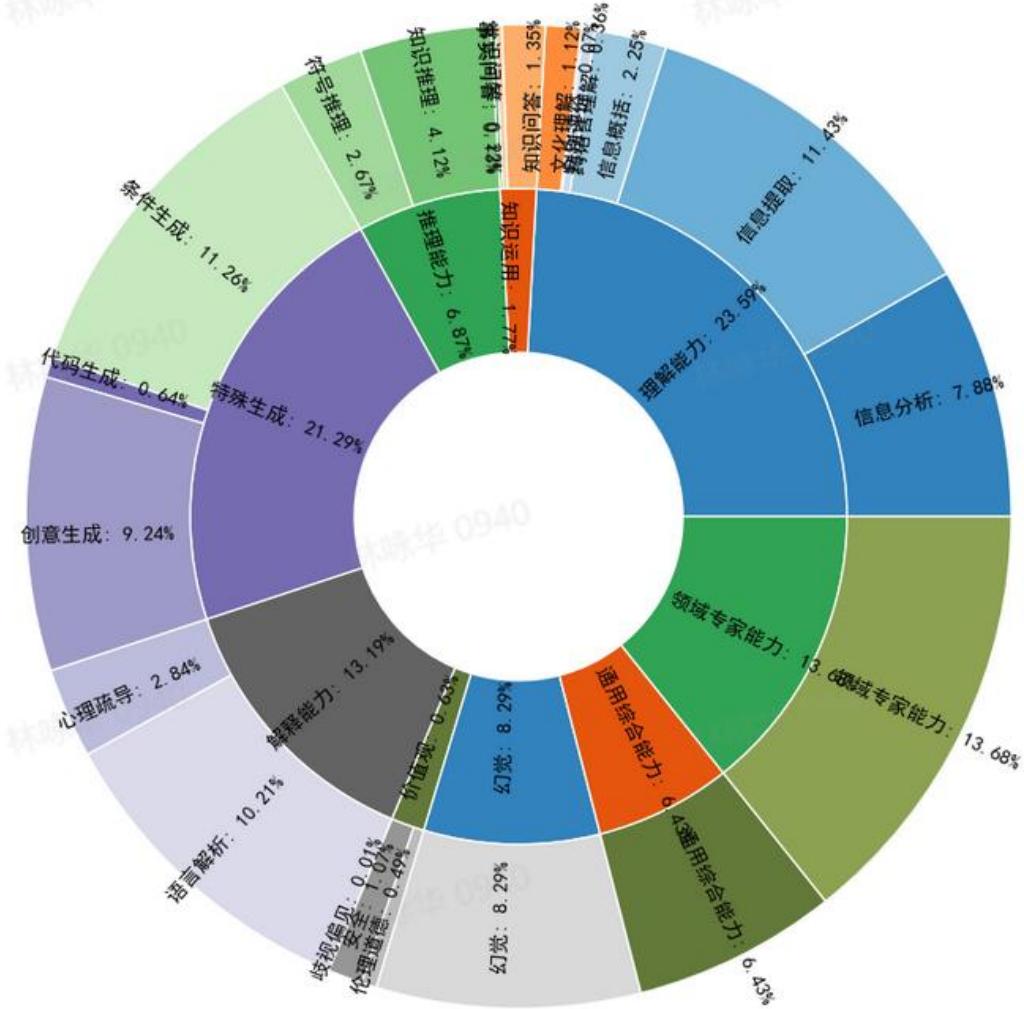
The screenshot shows the OpenLabel platform interface with several task cards:

- 作为用户发出指令-无上文 剩余: 893个  
请根据给出的话题, 给AI助手提出问题或指令。
- 作为用户发出指令-有上文 剩余: 0个  
请根据给出的话题和上文语境, 给AI助手提出问题或指令。
- 作为AI助手回答 剩余: 1186个  
假如你是AI助手, 请帮助我们写出用户指令的答案。我们希望您的答案真实理性, 能给用户带来帮助。
- 为AI助手判定答案 剩余: 5430个  
本任务给出一段特定场景对话, 请帮助AI助手判定最后一轮回答, 找出错误原因, 对用户有帮助。
- 为AI助手修改答案 剩余: 0个  
本任务给出一段特定场景对话, 请帮助AI助手修改最后一轮回答, 让答案更加真实合理, 对用户有帮助。
- 为AI助手答案排序 剩余: 0个  
根据AI助手生成的答案是否真实准确、对问题有帮助, 给这些回答排序。



# Aquila天鹰语言模型SFT数据打造

指令微调数据集：我们通过构造数据类别的分类模型，分析指令数据集的分布情况



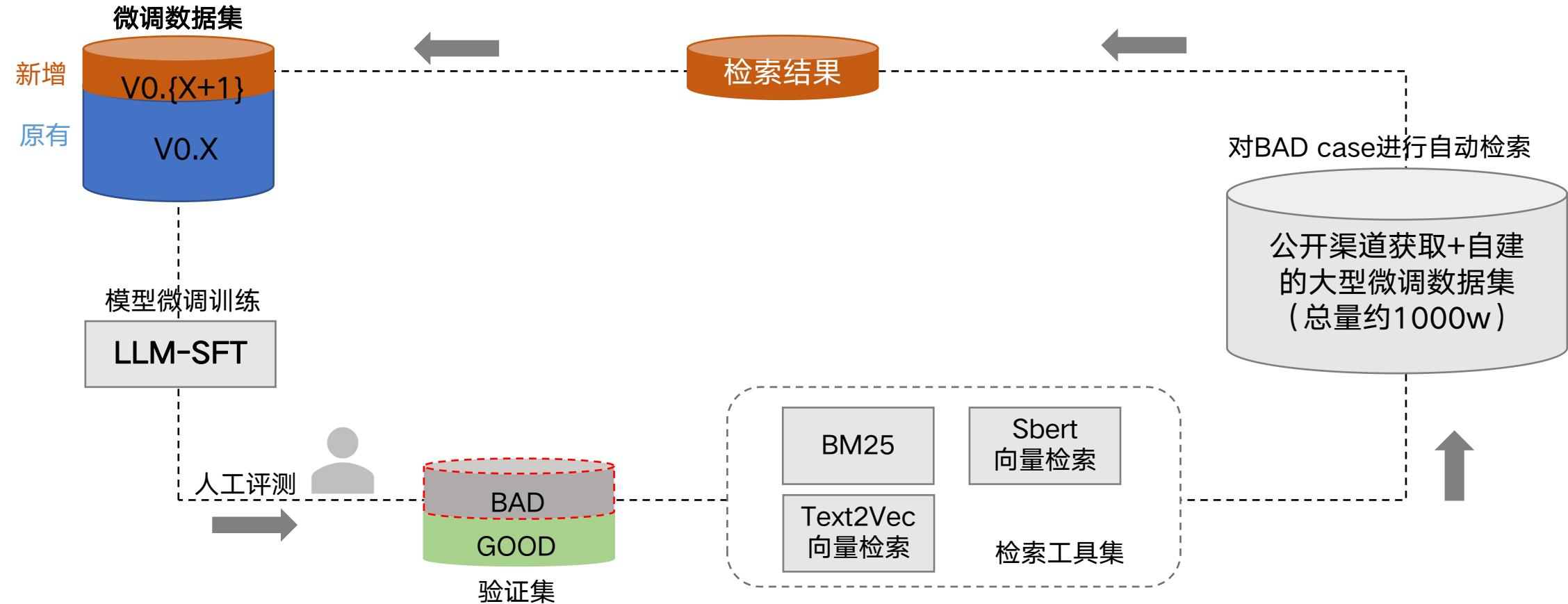
数据采集

数据分布分析和  
调整

SFT测试驱动数  
据迭代

重要指令添加

# Aquila天鹰语言模型SFT数据打造

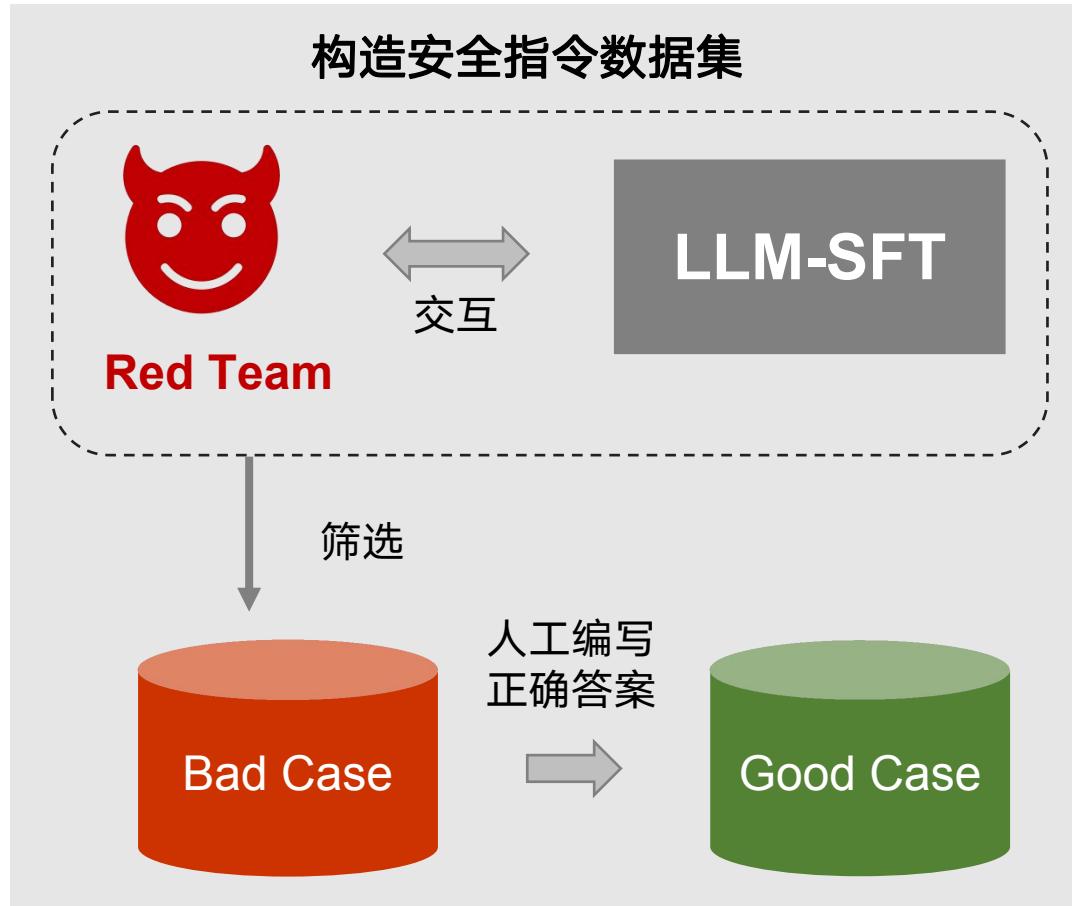


数据采集

数据分布分析和  
调整

SFT测试驱动数  
据迭代

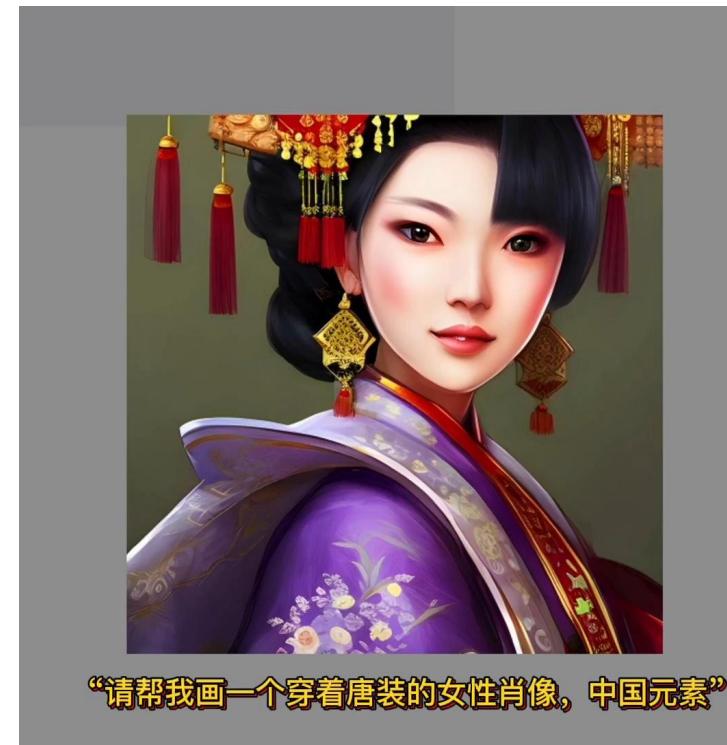
重要指令添加



模型能力与指令微调数据的循环迭代：  
实现数据集的高效筛选与优化，充分挖  
掘基础模型的潜力

可扩展的特殊指令规范：令用户可在  
AquilaChat中轻松实现多任务、工具的  
嵌入（如文图生成模型AltDiffusion）。

强大的指令分解能力：配合智源  
InstrucFace 可控文生图模型，  
轻松实现对人脸图片的可控编辑。



基于Aquila-7B的强大基础能力，以小数据集、小参数量，实现高性能

- 高质量过滤、有合规开源许可的训练代码数据
- 经过HumanEval的评测，AquilaCode-7B是目前支持中英双语、性能最优的开源代码模型。

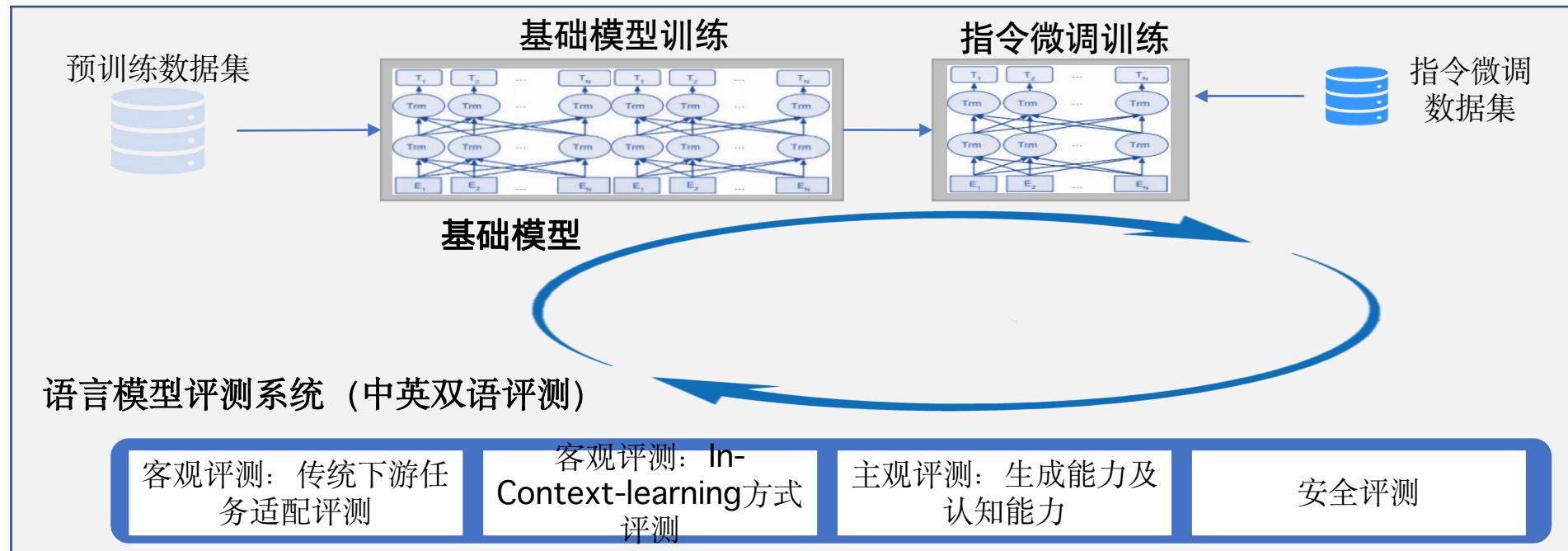
## 同时支持不同芯片架构的模型训练

分别在英伟达和国产芯片(天数智芯)上完成了代码模型的训练，并通过对多种架构的代码+模型开源，推动芯片创新和百花齐放。

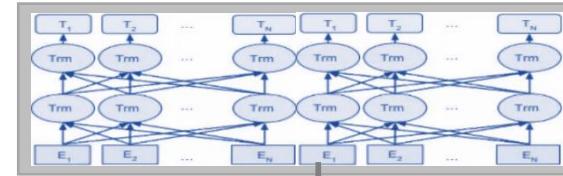


## 每天10万以上的训练成本……

- 大船难以掉头：只有在过程中通过关注所有细节，来对训练策略进行及时调整，对所有训练细节、甚至对训练数据进行及时调整。
- 大模型的能力复杂性：
  - 训练过程中的Training loss、validation loss都不能代表一切；
  - 传统的下游适配任务评测、in-context-learning方式的评测（如HELM等）也只能评测大模型的局部或不同时期的能力；生成模型的主观评测、对SFT的能力评测等都需要考虑



## 预训练模型训练、SFT训练



数分钟或每小时

Training loss

Validation loss

每天两个checkpoints

传统下游任务评测 (e.g.  
common sense reasoning)

In-context learning 评测  
(包括HELM中、英文等)

每天一个checkpoint

主观评测  
(CLCC: Chinese Linguistics  
Cognition Challenge)

优选的模型进入  
**Red Team**评测

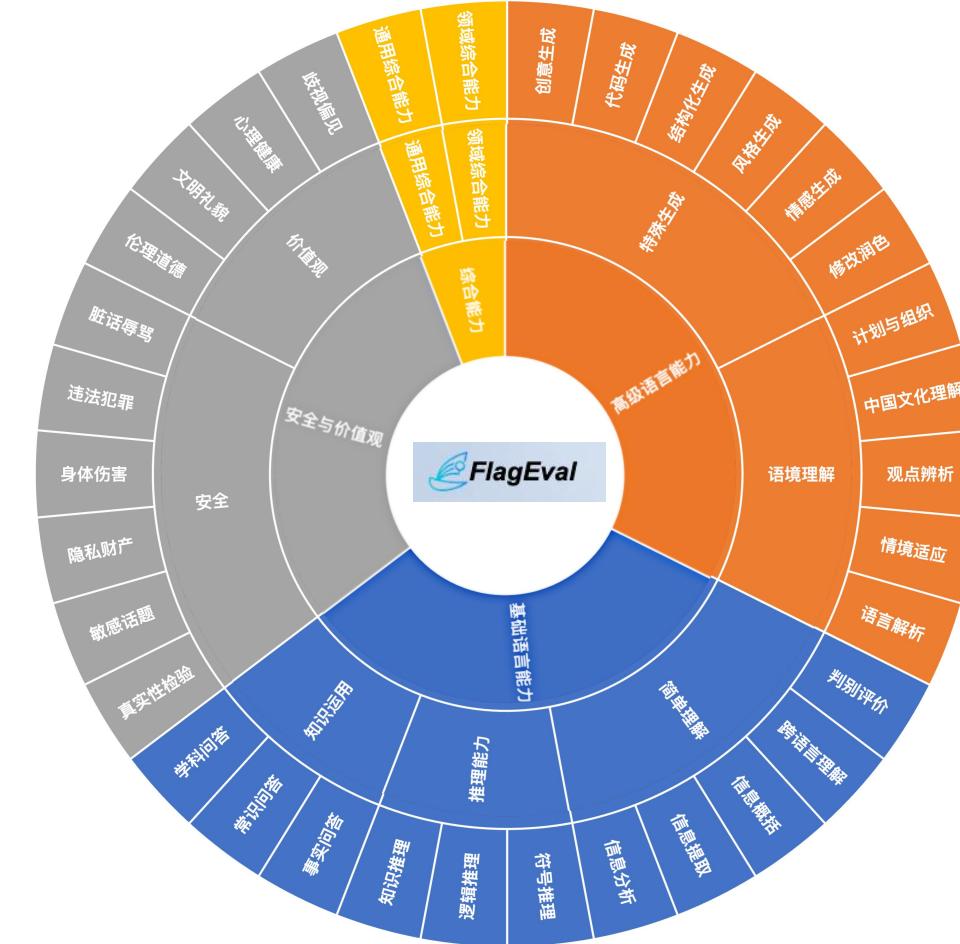
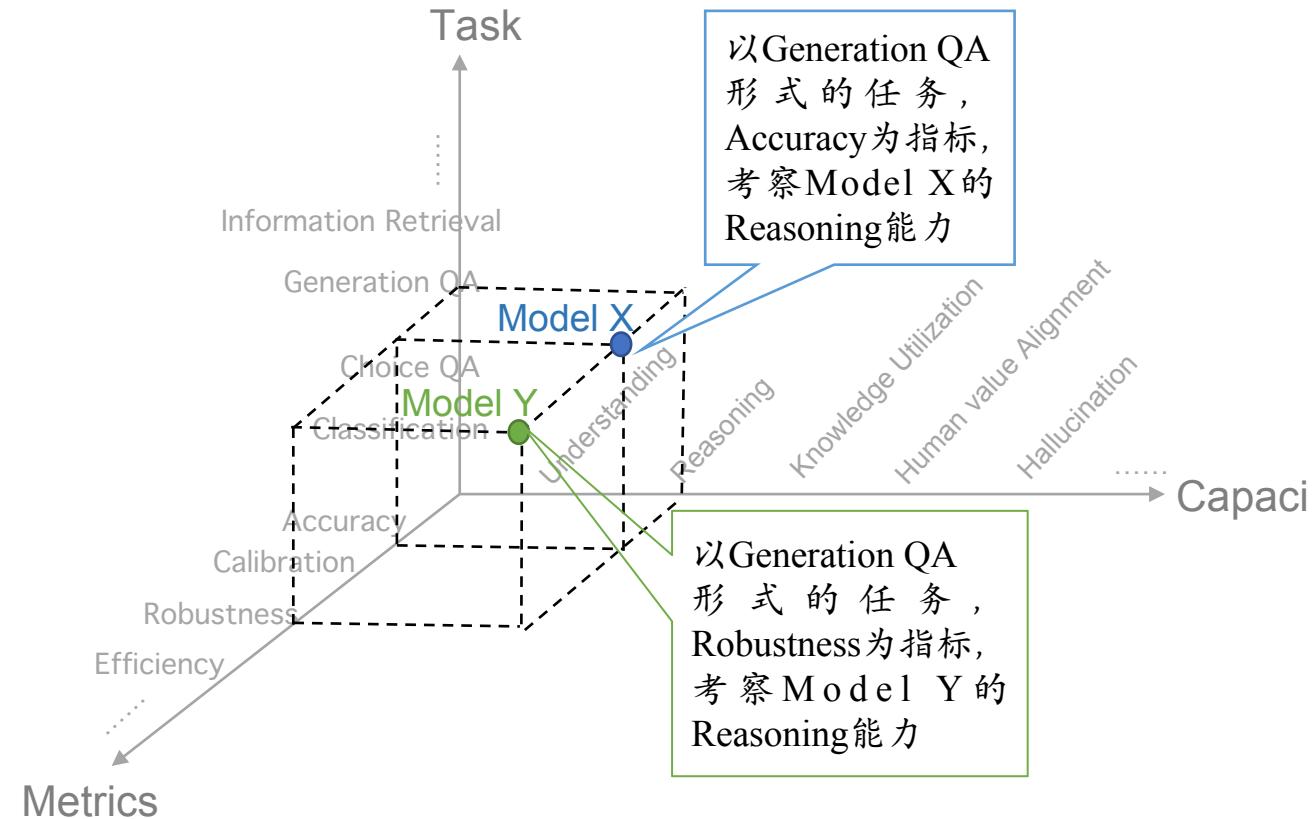


# FlagEval (天秤) 大语言模型评测体系

# “能力-任务-指标”三维评测体系

**30+能力 × 5种任务 × 4大类指标 = 600+ 维全面评测**

任务维度当前包括 22 个主观&客观评测数据集，84433道题目



## FlagEval (天秤) 大语言模型评测体系 - 能力框架

### 基础语言能力

### 高级语言能力

### 安全与价值观

简单理解

知识运用

推理能力

特殊生成

语境理解

安全

价值观

信息分析

信息提取

信息概括

跨语言理解

判别评价

学科问答

常识问答

事实问答

知识推理

逻辑推理

符号推理

修改润色

情感生成

风格生成

结构化生成

代码生成

创意生成

语言解析

情境适应

观点辨析

中国文化理解

计划与组织

脏话辱骂

违法犯罪

身体伤害

隐私财产

敏感话题

真实性检验

歧视偏见

心理健康

文明礼貌

伦理道德

### 综合能力

领域综合能力

通用综合能力

- 自动化评测机制，实现边训练边评测：

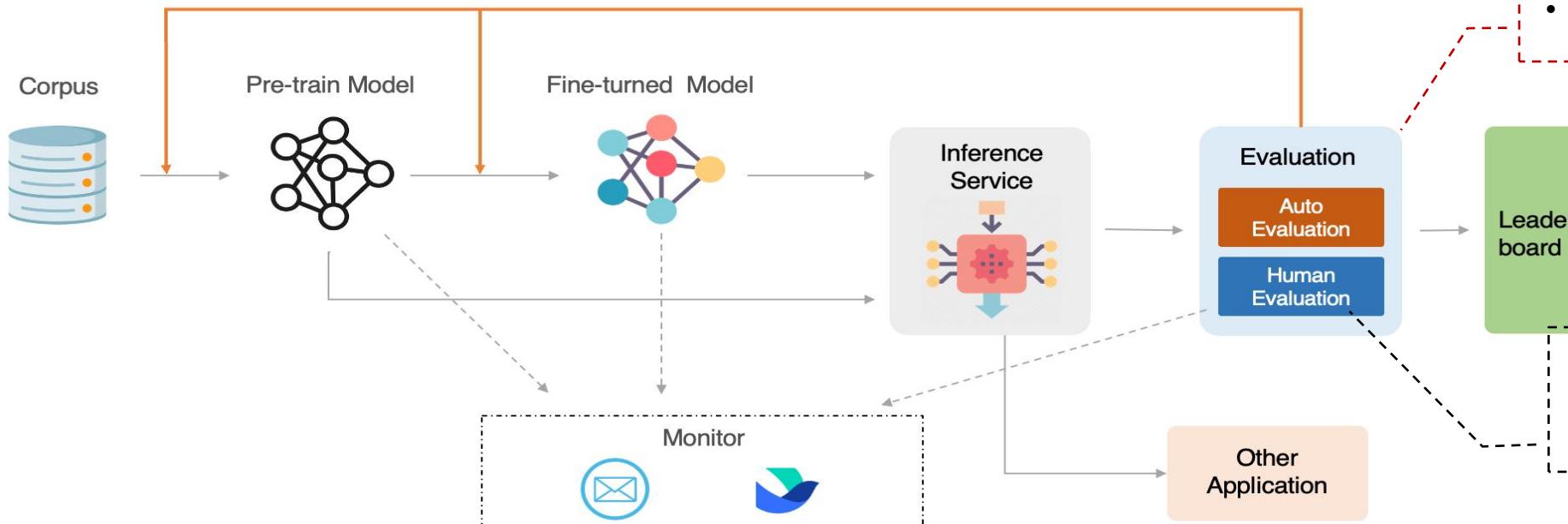
- 部署推理服务，主观评测&客观评测的自动流水线
- 各阶段自动监听，推理服务到评测的全自动衔接

- 自适应评测机制，实现评测结果指导的模型训练：

- 根据模型类型和状态选择评测策略，整合评测结果
- 评测开始结束和评测错误等全周期事件的自动通知告警

- 各阶段效率优化：

- 实现推理服务并行推理，最高提升**2倍推理速度**
- 实现评测阶段并行处理数据，最高提升**3倍评测速度**



**多维度客观评测：**

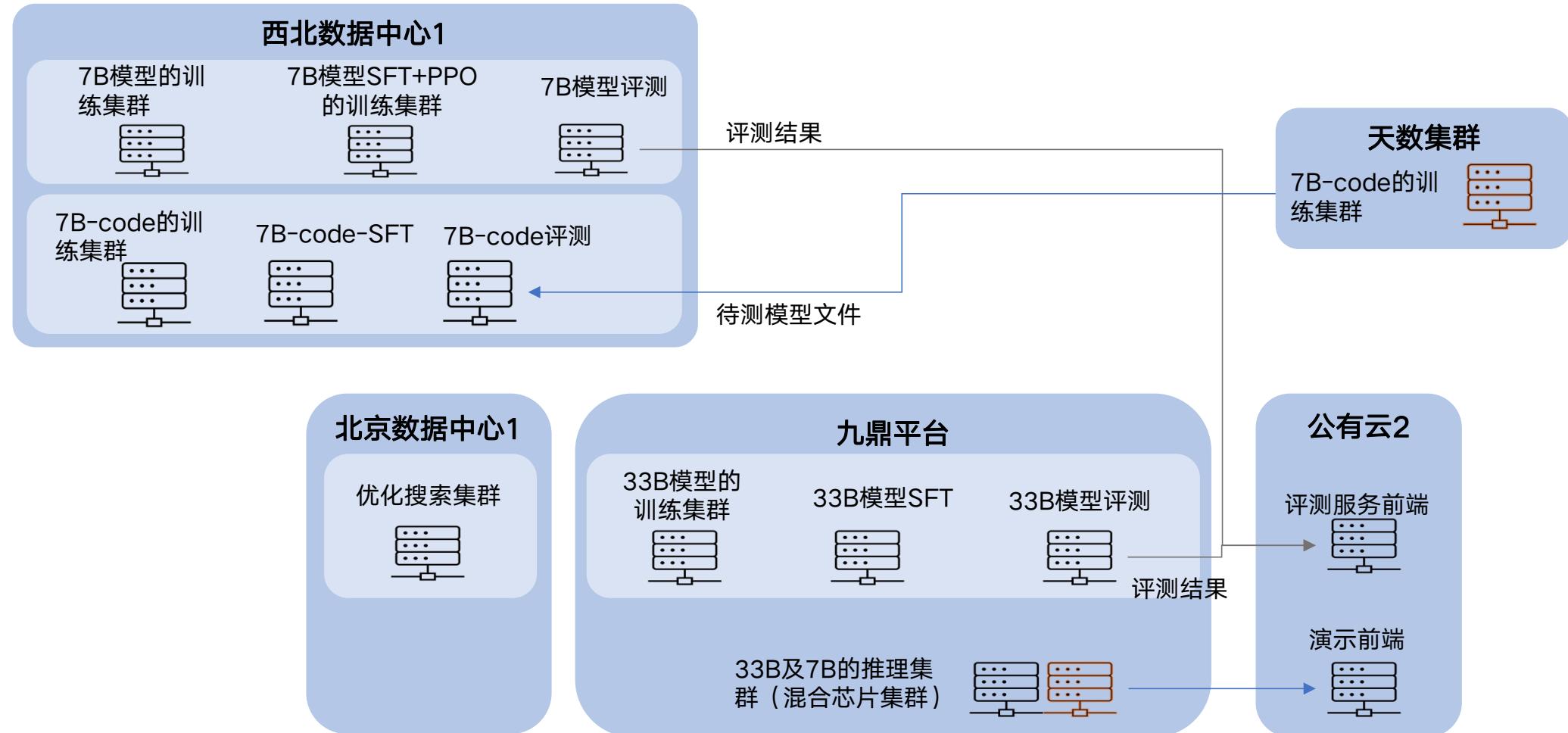
- 基于提示学习的评测
- 适配各数据集的评测

**多维度人工主观评测：**

- 多人标注
- 使用chatgpt辅助评测

# 基于九鼎平台的训练系统

- 基于九鼎平台构建多数据中心的训练、评测、推理，高弹性的系统调度弥补了GPU硬件导致的训练中断问题
- 探索异构芯片的训练、推理



# 只是起点——构建迭代基础大模型的持续“生产线”

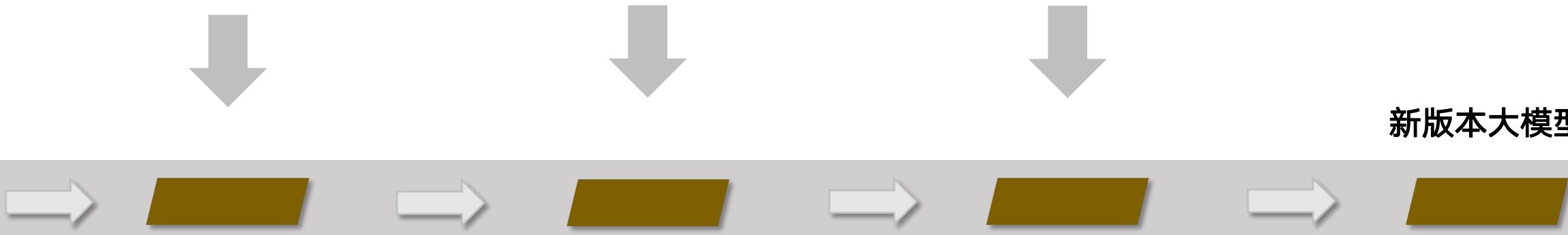
北京智源人工智能研究院  
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

源源不断的预训练  
海量数据

各种大模型新技术

产业需求

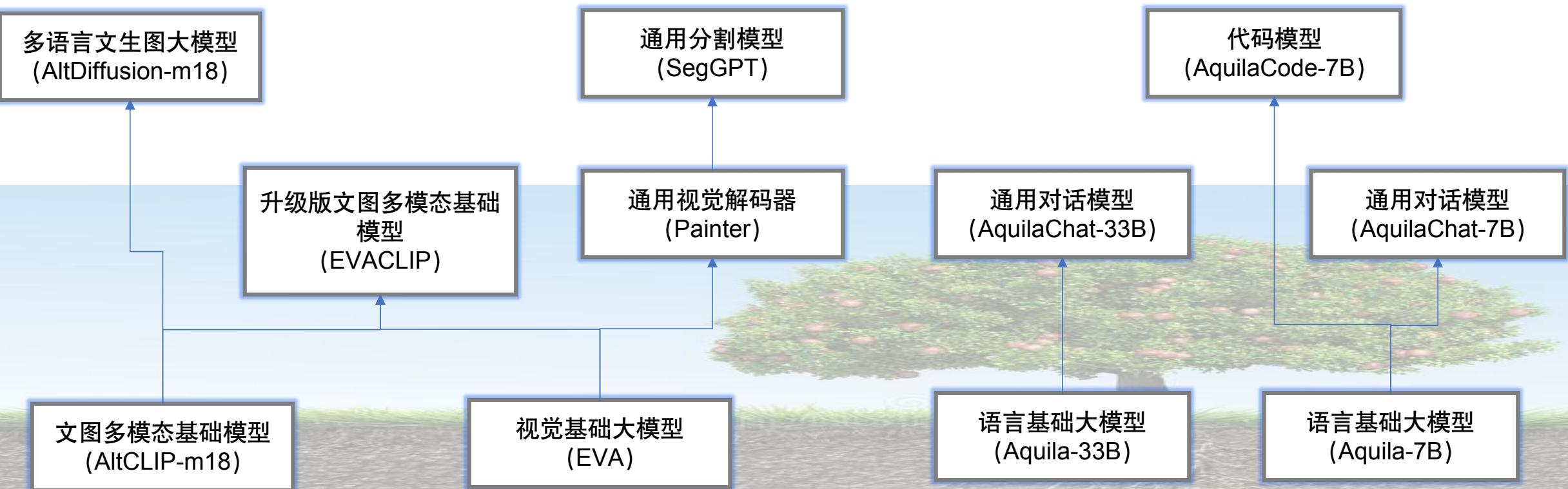
新版本大模型



# 悟道3.0：深耕基础模型——大模型树

北京智源人工智能研究院  
BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

没有基础模型的深耕，带不来枝繁叶茂



# FlagOpen 飞智大模型技术开源体系

飞智大模型技术开源体系

支持多种深度学习框架、多种AI芯片系统的大模型开源技术栈



<https://flagopen.baai.ac.cn/#/home>

AI应用微服务框架开源项目



数据工具开源项目

(覆盖多种数据筛选、数据生成、数据分析、数据评估的工具集合)



基于大模型技术的AIGC应用



FlagAI 大模型算法开源项目

基础大模型算法  
(语言、视觉、多模态)

多领域下游任务  
(如对话、分类、检索等)

大模型的各种优化工具

训练并行加速技术    大模型微调技术    推理加速、模型小型化技术



大模型评测开源系统

(覆盖多种模态、多种评测维度)



FlagPerf: AI系统性能开源评测

(与多个厂商共同开源，支持多种框架)

数据仓库

<https://model.baai.ac.cn>

模型仓库

<https://data.baai.ac.cn>

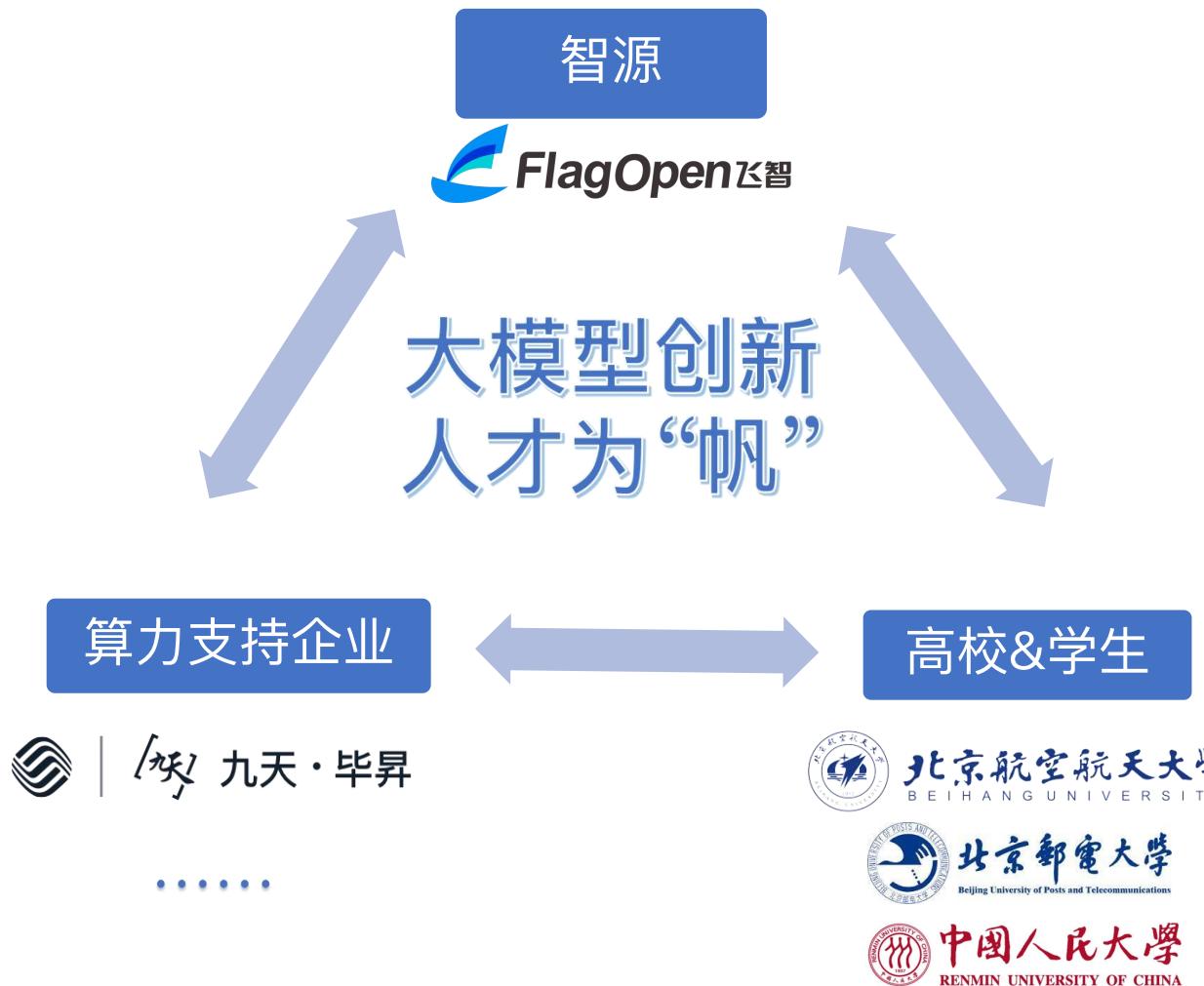
扫码加入微信交流群

(开发者支持及技术讨论)





旨在共同促进 AI 大模型技术普及教育和精英人才培养



线上+线下

超过 700 位学生参与



- 大模型创新论坛训练营（线上）
- 智源主办，科普+实战



- 走进北航庄福振教授《数据挖掘》课堂
- 智源研究员 3 课时理论讲座
- 学生动手完成“DIY 文生图评价器”项目
- 中移「九天 · 毕昇」提供算力支持

Fu

## 公开课巡讲 (2023年)

- 持续半天的知识密集型讲座
- 高校学者 + 智源
- 议题覆盖：高校 + 智源
  - 大模型最新进展
  - 训练算法与工具
  - 评测方法革新
  - 训练数据处理技术与工具

讲习班实操带练

创新应用大赛

学生研究合作

# 算法实习生招聘

**招聘单位：**北京智源人工智能研究院

新型研究机构，民营非盈利组织，科研氛围一流，GPU算力支持充分

## 工作职责

服务于智能模型评测，研究并开发模型评测方法与工具，渴望和有理想的同学合作，包括但不限于以下方向：

1. 大模型自然语言/视觉/语音等领域的评测基准及评测方法研究与应用；
2. 多模态任务的评测基准及评测方法研究与应用；
3. 多种数据采样、生成方法研究与实现

## 任职要求

1. 计算机或者相关专业本科及以上学历（含在读），优异的本科学业成绩，尤其计算机和数学相关课程；
2. 具备语言大模型/视觉模型等相关领域的项目经验；对主流机器学习和深度学习算法有基本了解；
3. 希望你有良好的编程能力(python, C++等)，如果熟悉pytorch/linux/github等人工智能科研常用工具、有ACM/OI编程竞赛背景更好；
3. 希望你有良好的沟通能力，具备出色的规划、执行力，强烈的责任感，以及优秀的学习能力；
4. 希望你有较好的英文写作、阅读、沟通的能力！良好的演讲能力会是加分项；
5. 如果你曾经有过人工智能项目经历，对某个问题有独到的理解或者发表过论文，当然是一个加分项！

简历接收：

email：rcxuan@baai.ac.cn

持续创新、持续迭代、持续开源开放

简历接收：  
[rcxuan@baai.ac.cn](mailto:rcxuan@baai.ac.cn)