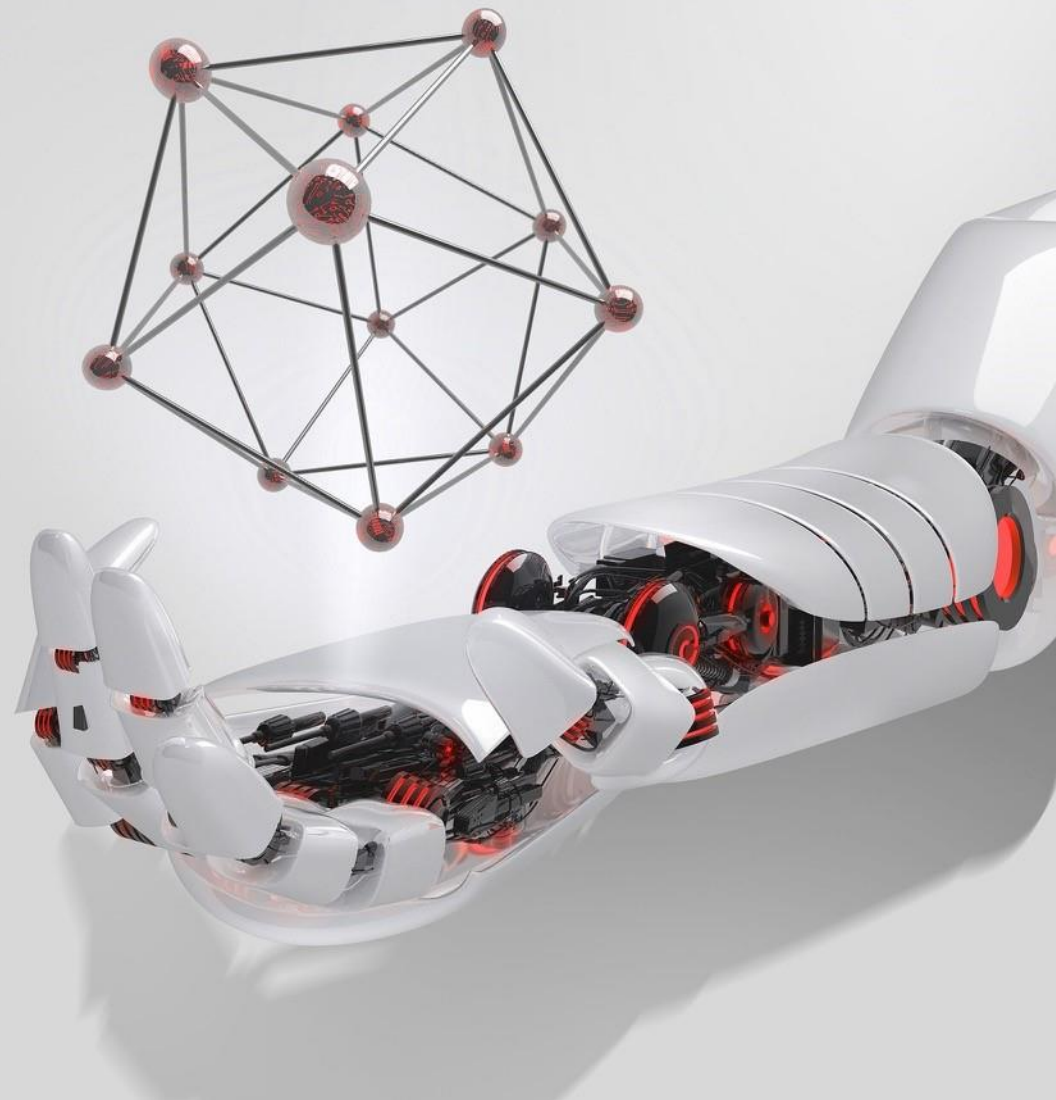


# ModelEngine使能平台赋能



Security Level:



# 目录

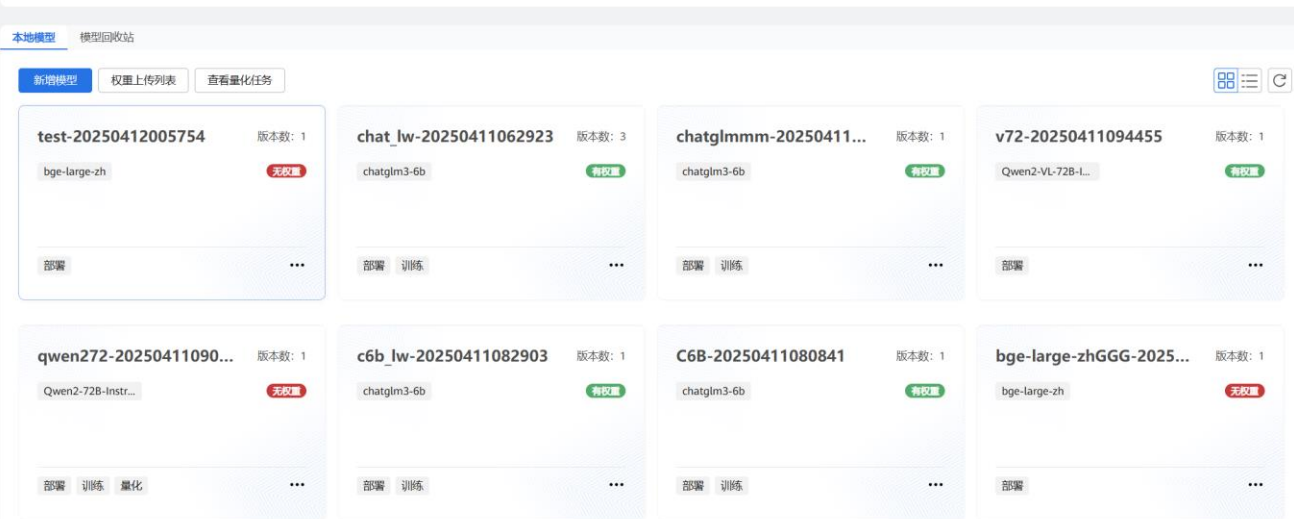
---

1. 操作流程
2. 数据使能
3. 模型使能
4. 服务接入

# 模型使能：模型仓库和权重上传

## 模型仓库

模型仓库支持管理上传的模型，以及训练后得到的版本，主要提供模型新增、模型列表查询、模型回收站和版本回收站等功能。

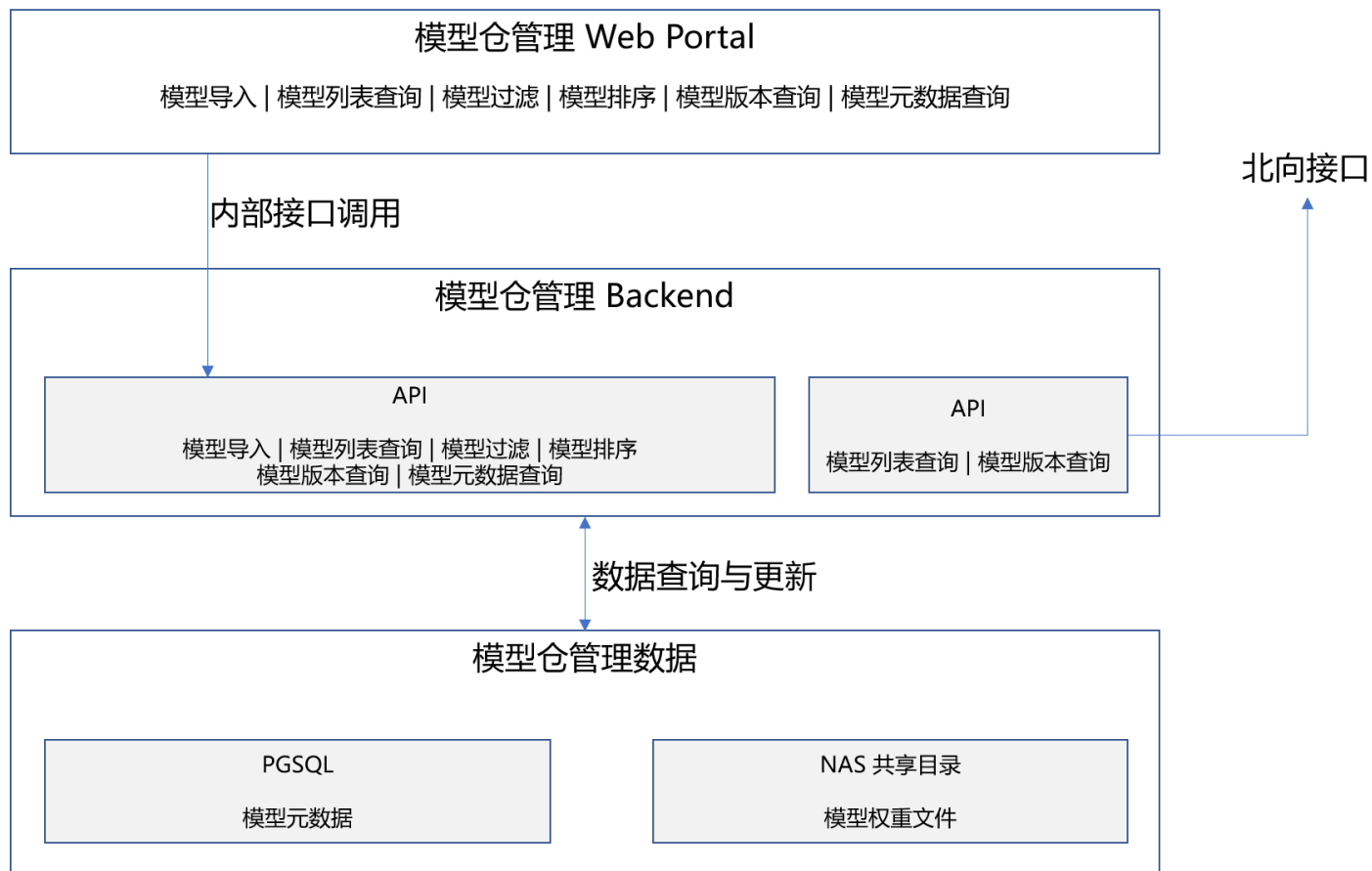


支持模型	支持操作	推理精度
Qwen-72B-Chat	部署, 评测	Ascend 910B: FP16 Ascend 310P: FP16
Qwen2-72B-Instruct	部署, 精调, 评测, 量化	Ascend 910B: BF16, FP16, 量化精度 (W8A8, W8A16) Ascend 310P: FP16
Qwen-14B-Chat	部署, 评测	Ascend 910B: FP16 Ascend 310P: FP16
chatglm3-6b	部署, 精调, 评测	Ascend 910B: FP16 Ascend 310P: FP16
bge-large-zh	部署	FP32
Meta-Llama-3-8B-Instruct	部署, 评测	Ascend 910B: BF16, FP16 Ascend 310P: FP16
Meta-Llama-3-70B-Instruct	部署, 评测	Ascend 910B: BF16, FP16; Ascend 310P: FP16
bge-reranker-large	部署	Ascend 910B: FP16 Ascend 310P: FP16
Qwen1.5-7B-Chat	部署, 精调, 评测	Ascend 910B: BF16
Qwen1.5-14B-Chat	部署, 精调, 评测	Ascend 910B: FP16 Ascend 310P: FP16
Qwen1.5-32B-Chat	部署, 精调, 评测	Ascend 910B: FP16 Ascend 310P: FP16
Baichuan2-13B-Chat	部署, 精调, 评测	Ascend 910B: FP16 Ascend 310P: FP16
Qwen2.5-0.5B-Instruct	部署, 精调, 评测	Ascend 910B: BF16 Ascend 310P: FP16
Qwen2.5-1.5B-Instruct	部署, 精调, 评测	Ascend 910B: BF16 Ascend 310P: FP16
Qwen2.5-3B-Instruct	部署, 精调, 评测	Ascend 910B: BF16 Ascend 310P: FP16
Qwen2.5-7B-Instruct	部署, 精调, 评测	Ascend 910B: BF16 Ascend 310P: FP16
Qwen2.5-14B-Instruct	部署, 精调, 评测	Ascend 910B: BF16 Ascend 310P: FP16
Qwen2.5-32B-Instruct	部署, 精调, 评测	Ascend 910B: BF16 Ascend 310P: FP16
Qwen2.5-72B-Instruct	部署, 评测	Ascend 910B: BF16 Ascend 310P: FP16
DeepSeek	部署, DeepSeek模型权重仅后台上传	Ascend 910B: BF16

# 模型使能：模型仓库和权重上传

## 模型仓库

模型仓库支持管理上传的模型，以及训练后得到的版本，主要提供模型新增、模型列表查询、模型回收站和版本回收站等功能。



# 模型使能：模型精调

## 模型精调

模型训练支持NPU资源监控，支持训练策略配置，基于数据使能处理对话数据集进行全参微调 and LoRA微调。

### NPU资源监控

芯片 ⌵

芯片型号	Node名称	总NPU数量	可用NPU数量	NPU显存大小
Ascend910	cluster-worker-iihlu	8	3	64 GB
	cluster-master-ohkfk	8	5	64 GB
	cluster-master-lafoy	8	8	64 GB
	cluster-master-ojzfb	8	8	64 GB

### 训练框架

MindSpeed-LLM

### 镜像名称

mindspeed:24.1.rc1 ⌵

### 模型和训练类型

模型 ⌵      模型版本 ⌵

训练类型  全参训练     LoRA微调

### 训练策略

TP = 8, PP > 1的配置需要多机多卡分布式训练。多机多卡训练，每个节点分配的NPU数量必须相同，因此TP \* PP的值必须能被Nodes数量整除。

Max Sequence Length ⌵      Local Batch Size ⌵      最小NPU卡数 ⌵

8192      1      ..

NPU数 ⌵      TP ⌵      PP ⌵

请先获取最小NPU卡数      请选择NPU数量      ..

参数	说明
芯片型号	芯片的具体型号名称，ModelEngine自动识别芯片型号，示例：910B3。
Node名称	节点的唯一标识名称，通常用于区分不同的计算节点。
总NPU数量	节点上安装的NPU卡的总数。
可用NPU数量	当前可用的NPU卡的数量，即没有被占用或出现故障的NPU卡。
NPU显存大小	NPU卡的显存容量。

*NPU资源监控芯片参数*

参数	说明
最大序列长度	模型可处理的输入文本的最大长度，长度越长，显存开销越大。
局部样本数	一个迷你批次中，单张NPU上样本的数量。
最小NPU卡数	根据已配置的参数，可获取模型训练所需的最小卡数量。请根据最小NPU卡数量下发训练任务。
NPU数	TP*PP
TP（张量并行）	张量并行，把线性层按行或列对模型权重进行划分。
PP（管道并行）	管道并行，对模型进行层间划分。值为NPUs / TP。

*训练策略参数*

# 模型使能：模型精调

模型训练 收起 ^

使用说明



制作训练镜像

提前准备训练所需的环境的镜像包。



上传权重文件

在模型仓库上传模型权重文件。



导入微调数据集

在数据集管理中上传训练依赖的数据集。



模型训练

从模型服务页面下发模型训练任务。



权重文件归档

将训练好的模型权重纳入模型仓库管理。

创建任务 ↻

任务ID	模型	来源数据集版本	训练进度	训练状态	归档状态	训练策略	操作
9080a1f8-1415-4fa0-92d...	chatglm3-6b-202504110624 V1	训练 V1	--	数据集下载 <span style="color: green;">✔</span>   预处理 <span style="color: green;">✔</span>   已终止 <span style="color: gray;">⋮</span>	未归档 <span style="color: gray;">⋮</span>	TP2PP2	日志 ...
41daa92a-964f-41a2-b7...	CAHT6B-20250411035151 V1	训练 V1	<div style="width: 100%;"><div style="width: 100%; height: 5px; background-color: #28a745;"></div></div> 100%(16/16)	数据集下载 <span style="color: green;">✔</span>   预处理 <span style="color: green;">✔</span>   训练成功 <span style="color: green;">✔</span>	已归档 <span style="color: green;">✔</span>	TP1PP2	日志 ...

参数	说明
Gradient Accumulation Steps	每次模型参数更新前的迭代次数。
Warm Up Ratio (学习率预热比例)	表明预热过程中，学习率的增长轮数和总训练轮数的比例。
Epoch (数据集遍历轮次)	遍历数据集的次数 (训练中会被转换成iter的形式)。
Learning Rate (学习率)	控制权重参数的更新速度。

全参微调参数

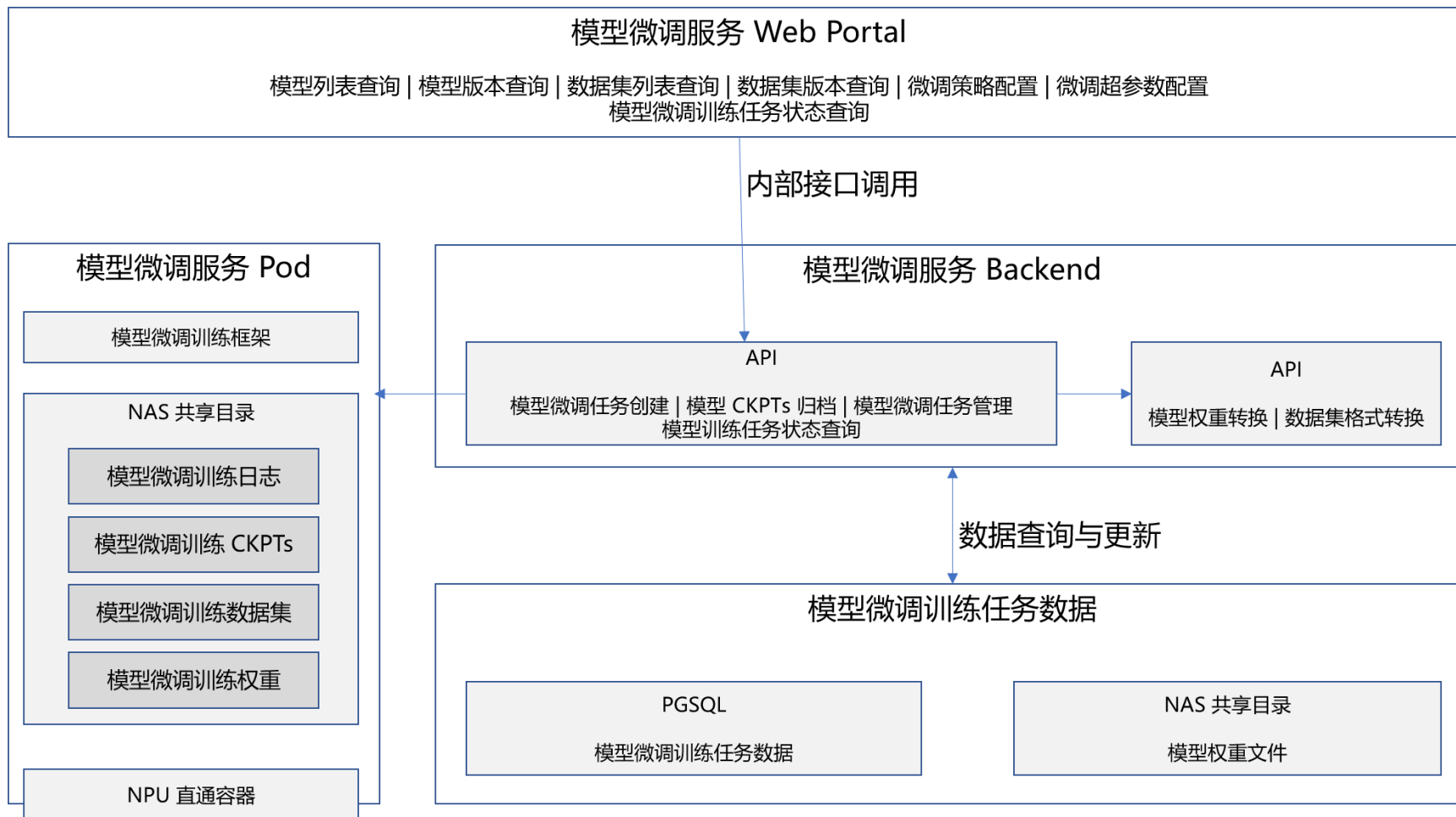
参数	说明
Gradient Accumulation Steps	每次模型参数更新前的迭代次数。
Warm Up Ratio	表明预热过程中，学习率的增长轮数和总训练轮数的比例。
Lora R (LORA训练中低秩矩阵维度)	值越大，模型学习能力越强，但显存开销越大。
Lora Alpha (LORA学习率放缩因子)	一般为Lora R的2倍，控制LORA参数的更新。
Epoch	遍历数据集的次数 (训练中会被转换成iter的形式)。
Learning Rate	控制权重参数的更新速度。

LoRA微调参数

# 模型使能：模型精调

## 模型精调

模型训练支持NPU资源监控, 支持训练策略配置, 基于数据使能处理对话数据集进行全参微调 and LoRA 微调。



# 模型使能：模型评测

## 模型评测

对模型仓库中已有的模型进行评测，获得模型的精度指标和性能指标，生成评测报告。

评测报告

### 模型详情

gpt

评测模型	模型类型	创建时间
chatglm3-6b-2025022402351..._chatglm3-6b		2025-02-24 10:35:16

### 评测详情

任务ID	任务名称	描述	任务类型	启动时间	结束时间
46413665-e0af-4f22-990a-8d..._chatglm3-6b性能评测	...	...	性能评测	2025-04-09 17:04:25	2025-04-09 17:10:09

评测数据集	最大序列长度	并发量
Default Dataset/V1	32768	1000

### 评测结果

参数名称	平均值	最大值	最小值	P75	P90	P99
首个Token时延	1268.6093 ms	2870.8979 ms	26.413 ms	2200.4034 ms	2512.5229 ms	2802.1969 ms
Decode阶段时延	227.743 ms	56273.2305 ms	6.7 ms	90.267 ms	112.761 ms	213.9756 ms
所有请求最大Decode阶段时延	32729.3651 ms	56273.2305 ms	4355.98 ms	49258.7109 ms	53091.7594 ms	55779.942 ms
请求推理时延	48166.2245 ms	163022.562 ms	6922.929 ms	61058.2631 ms	68089.2347 ms	151606.6497 ms
输入Token长度	78.4157	317	28	98	136	228
生成Token长度	200.8859	4096	5	204	294	4096
生成Token速度	3.7276 token/s	28.4487 token/s	0.1082 token/s	4.9263 token/s	7.3688 token/s	25.1606 token/s
生成字符长度	313.4112	8622	3	308	456	5340

Qwen2-72B-Instruct-202504110356... Qwen2-72B-Instruct 2025-04-11 11:56:54

### 评测详情

任务ID	任务名称	描述	任务类型
eaaca922-31c3-418d-a699-c450bec..._72b	...	...	精度评测

启动时间	结束时间	npu数
2025-04-11 14:31:20	2025-04-11 16:37:21	4

92.20%

ctval单道题正确的数量与总题量的比值

### 评测数据集

Default Dataset/V1

### 模型工具链

### 一栈式模型优化方案

#### 数据使能

自动化数据处理和知识生成

#### 模型使能

轻量级训推工具链

#### 应用使能

高精度RAG应用开发和微调

#### 原始输入

LLaMA等模型 | 数据使能对话数据

#### 模型精调

资源监视 | 任务下发 | 全参微调 | LoRA微调 | 模型评测

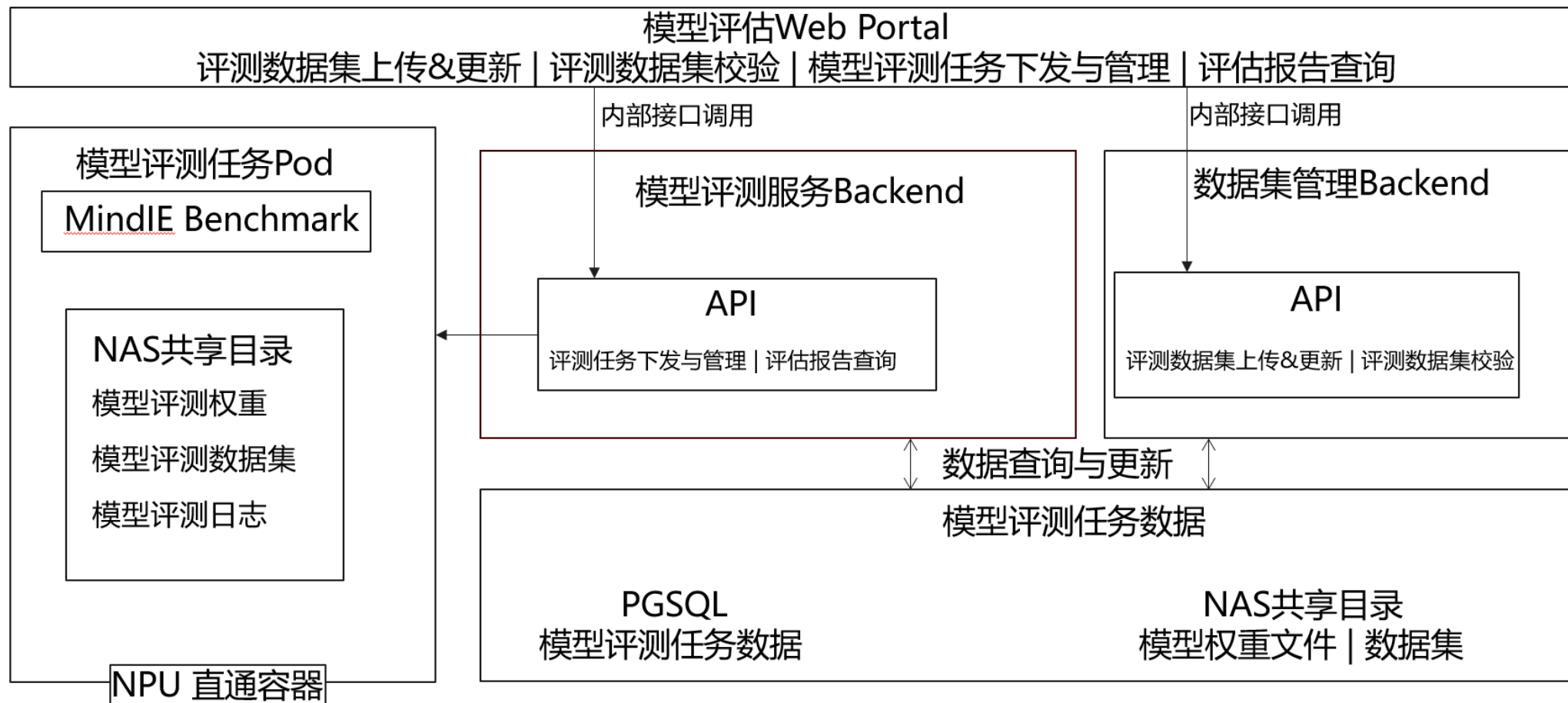
#### 模型服务

模型推理服务 | 北向API接口

# 模型使能：模型评测

## 模型评测

对模型仓库中已有的模型进行评测，获得模型的精度指标和性能指标，生成评测报告。



# 模型使能：模型服务和北向API接口

## 模型服务

模型服务基于资源进行本地模型部署，支持基于当前资源状况进行模型服务管理。

The screenshot displays a resource monitoring table and configuration panels. The table lists nodes with columns for chip type, node name, total NPU count, available NPU count, and NPU memory size. Below the table are sections for 'Model Service Parameters' (including instance name, model type, and version), 'Inference Framework' (with options for MindIE and Custom), and 'Inference Framework Parameters' (with sliders for max tokens, batch size, and max sequence length).

This section shows the 'Model Service Details' page on the left, which includes information about the chat model (chatglm3-6b), its version, and performance metrics like latency (0.00s). On the right is the 'API Documentation' for the chat completions endpoint, showing the base URL, required API key, and a sample POST request body.

参数	说明
实例名称	支持用户自定义实例名称。
模型类型	模型权重的所属类型，支持部署自定义类型的模型。
模型版本	模型的版本号。 已在模型仓库创建模型时上传权重文件，可选择模型版本。
节点数量	当芯片类型为Ascend 910B3或Ascend 910B4且NPU显存大小为64GB时，显示“节点数量”参数。 节点数量为1时，是单机部署。 节点数量为N时，是多机部署。 只有模型类型为DeepSeek时，才支持多机部署。 节点的可选数量为2的幂次方。 范围最小值为2，最大值应不大于节点总数量。
模型实例数	K8S服务副本数。在下拉框中选择数值，取值范围为1~8。
网关最大连接数	大模型支持的最大请求并发数，用于模型服务模块的网关控制。若大模型实际的请求并发数大于“网关最大连接数”，则后续溢出的请求将报错，http响应错误码是“500”，错误信息是“Internal Server Error”。
最大Token数	输入Token和输出Token的叠加最大值。 大模型类型不同，最大Token数的取值范围不同。
Prefill阶段的最大Batch size	控制的是一次性加载或预填充的数据量或批次大小。
模型镜像名称	模型镜像的名称，在下拉框中选择支持的镜像名称，内容无法编辑修改。
整体Batch Size	decode阶段的并发量。
最大输入序列长度	单次问答中能够接收的最大token数。
Prefill阶段最大Token数	Prefill阶段所有batch中token数之和。
最小NPU卡数	选择“NPU资源监控”和“模型服务参数”后，可单击“获取最小卡数”，请根据最小NPU卡数量下发推理服务。自定义类型无最小值。

# 模型使能：模型服务和北向API接口

## 模型服务

模型服务基于资源进行本地模型部署，支持基于当前资源状况进行模型服务管理。

## 估算需要多少显存

2025-03-28 18:08 由 刘乾坤 00918905 创建，于2025-03-29 10:00 由 刘乾坤 00918905最后修改。

参考文章：

[https://github.com/harleyszhang/llm\\_note/blob/main/1-transformer\\_model/llm%E5%8F%82%E6%95%B0%E9%87%8F-%E8%AE%A1%E7%AE%97%E9%87%8F-%E6%98%BE%E5%AD%98%E5%8D%A0%E7%94%A8%E5%88%86%E6%9E%90.md](https://github.com/harleyszhang/llm_note/blob/main/1-transformer_model/llm%E5%8F%82%E6%95%B0%E9%87%8F-%E8%AE%A1%E7%AE%97%E9%87%8F-%E6%98%BE%E5%AD%98%E5%8D%A0%E7%94%A8%E5%88%86%E6%9E%90.md)

## 推理计算

**显存占用 = 模型权重 + 输入输出及中间激活结果 + kvcache**

### 1. FP16

模型权重：直接乘2，别问为什么。32B = 32 X 2 **GB**

输入输出中间过程：直接乘0.2，经验值。32B \* 0.2 = 6.4GB

KVcache: 4 X 隐藏层维度 X batchsize X 最大输入输出长度和 X 网络层数

### 2. W8A8/W8A16

模型权重：直接乘1，别问为什么。32B = 32 **GB**

输入输出中间过程：直接乘0.2，经验值。32B \* 0.2 = 6.4GB

KVcache: 4 X 隐藏层维度 X batchsize X 最大输入输出长度和 X 网络层数

问：为甚W8A8和W8A16没啥差

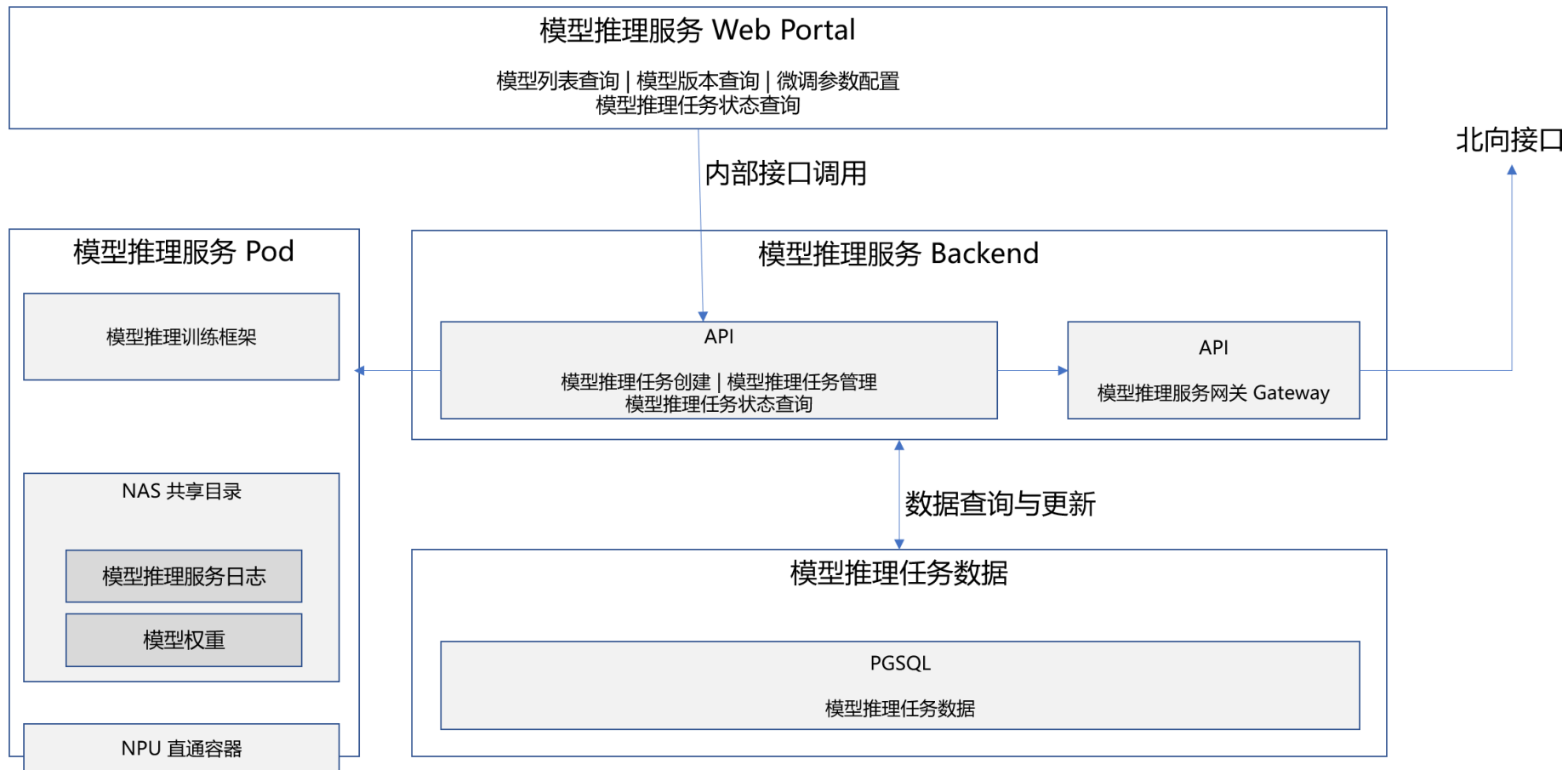
有的兄弟，有的，差异体现在输入输出中间过程上

W8A16

# 模型使能：模型服务和北向API接口

## 模型服务

模型服务基于资源进行本地模型部署，支持基于当前资源状况进行模型服务管理。



# 服务接入:

## 服务接入

ModelEngine提供的数据自动标注、QA对数据自动评估、向量知识库、大模型应用等功能均会用到模型服务能力。如果用户想要使用这些功能，除了可以在ModelEngine模型使能本地部署语言模型服务、Embedding模型服务和Rerank模型服务外

< 创建服务 ①

**基本信息**

服务类型 \*

大模型对话服务  Embedding模型服务  向量知识库服务

服务名称 \*

协议 ② \*

HTTPS  HTTP

**证书设置**

数字证书认证

证书别名 \*

证书文件 \*

证书描述

**服务配置**

URL \*

标头

新增

Key	Value	操作
Content-Type	application/json	编辑 删除



大模型对话服务

embedding

向量知识库

数据标注

数据评估

智能对话

RAG

向量知识库

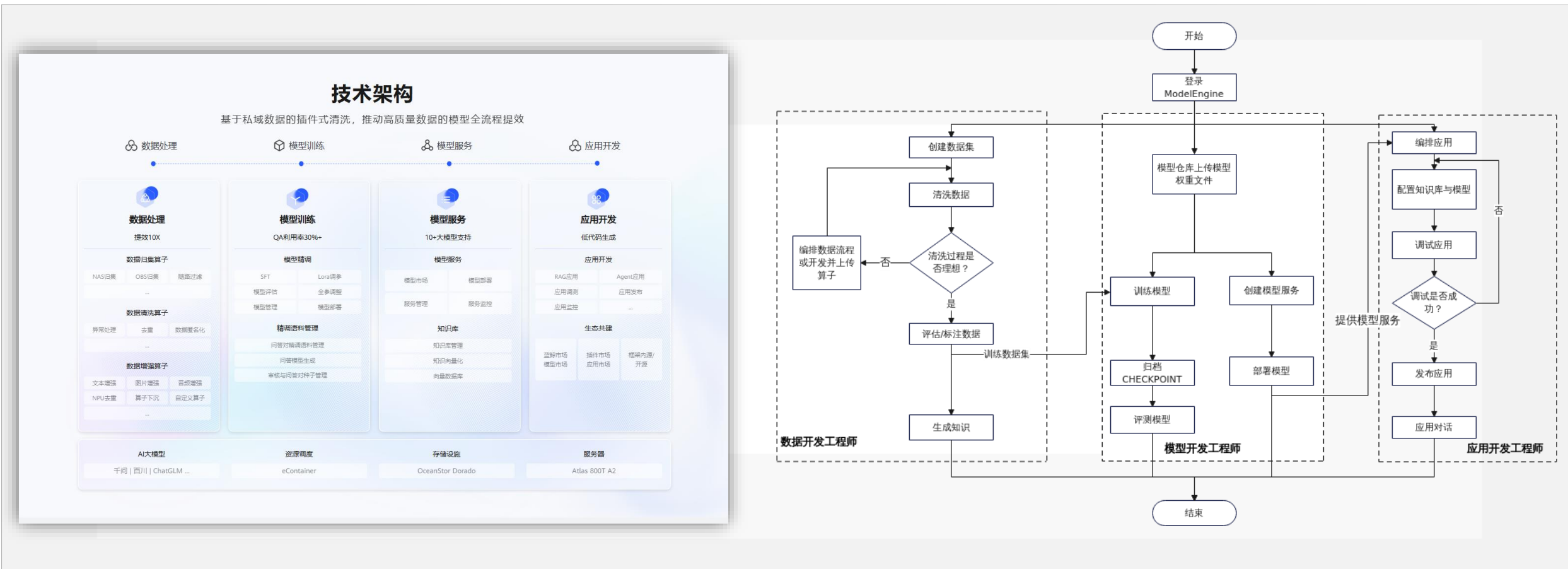
**数据使能**  
数据标注 | 数据评估 | 向量知识库

**应用使能**  
RAG | 智能对话

# 数据处理小结：数据使能workflow回顾&AI训推全流程工具链介绍

## 操作流程

ModelEngine提供从数据处理、知识生成，到模型微调和部署，以及RAG (Retrieval Augmented Generation) 应用开发的AI训推全流程工具链，用于缩短从数据到模型、数据到AI应用的落地周期。



## 请思考一下

W8A8 72B模型权重需要多少显存()

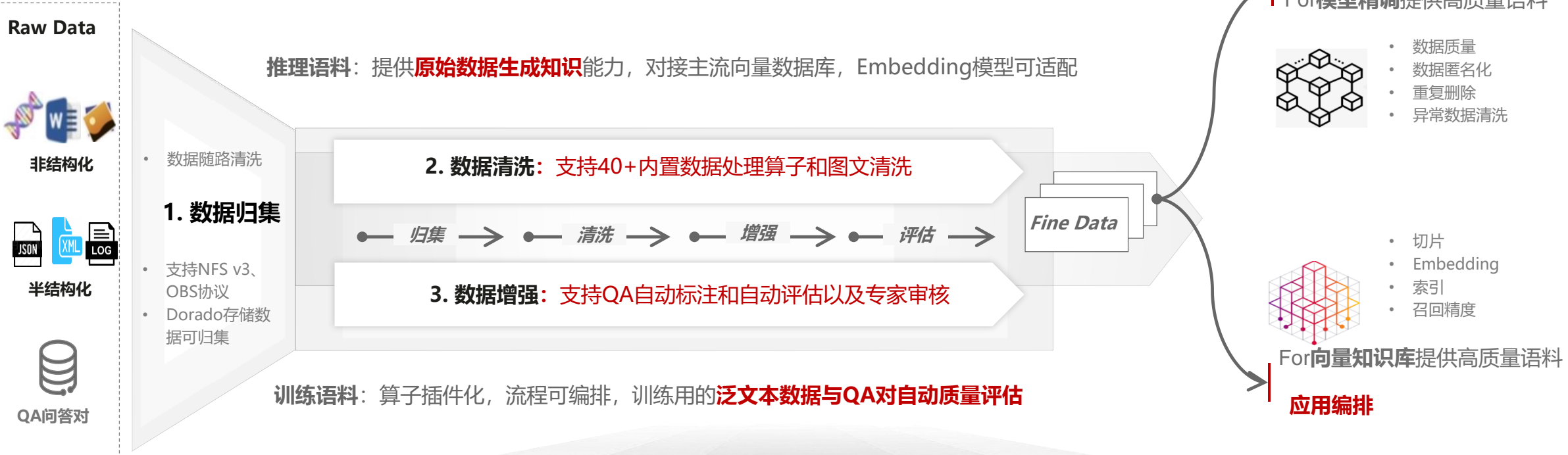
- A. 72G
- B. 36G
- C. 144G

模型推理框架是什么? 模型训练框架是什么()

- A. mindie, mindspeed
- B. vllm, deepspeed
- C. mindie, deepspeed
- D. vllm, mindspeed

# 数据使能：高质量泛文本数据处理和数据增强

提供完整数据处理流：从原始的Raw Data通过一系列数据加工流程，最终得到符合大模型训推所需要的精炼数据 Fine Data



## 关键流程

- 丰富的数据清洗算子**
- 内置40+数据处理算子，覆盖文本、图像等多模态数据。支持自定义数据处理清洗流程机制，满足批量数据处理等能力

- 基于大模型的QA对自动生成**
- 基于清洗后的文本数据与外置大模型服务，自动生成大模型微调QA对；具备QA对自动评估，满足QA高效生成能力

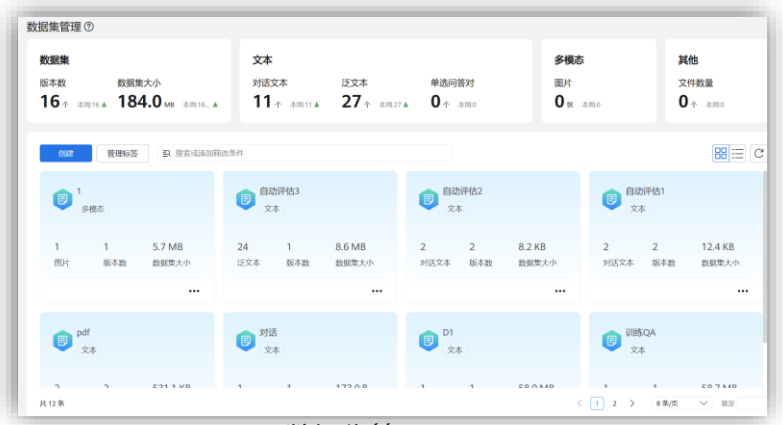
- 支持QA自动评估以及专家审核**
- 支持数据自动评估和专家审核机制，对评估后数据进行数据筛选以及专家审核，支持alpaca格式垂域QA对导出



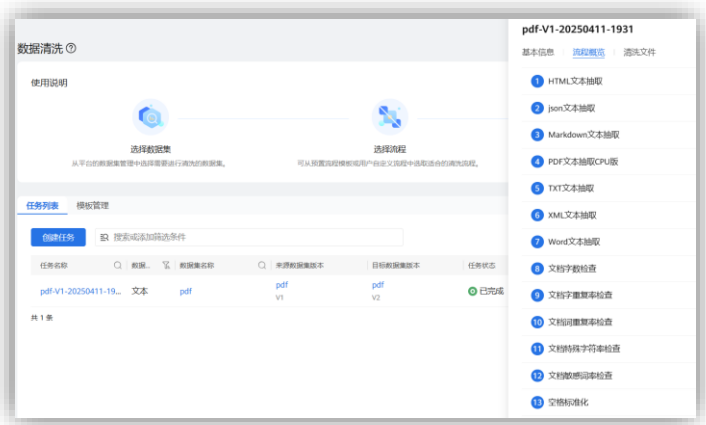
# 数据使能：高质量泛文本数据清洗和QA对自动质量评估

## 数据使能

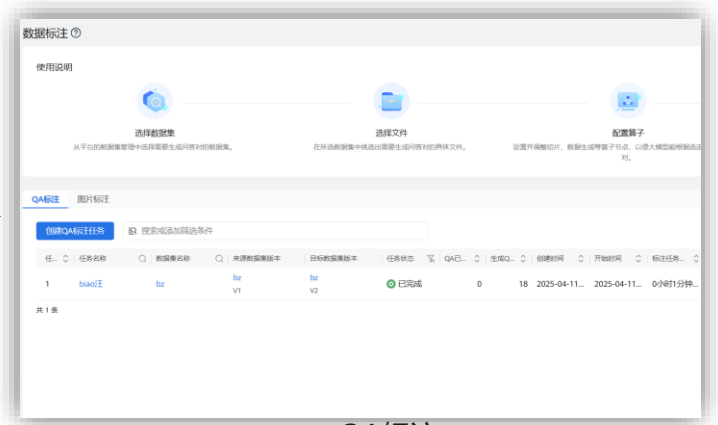
基于资源使能底座和模型底座完成数据清洗一站式流程，含数据集管理、数据处理、QA标注、QA评估、专家审核、向量知识库生成能力。



数据集管理



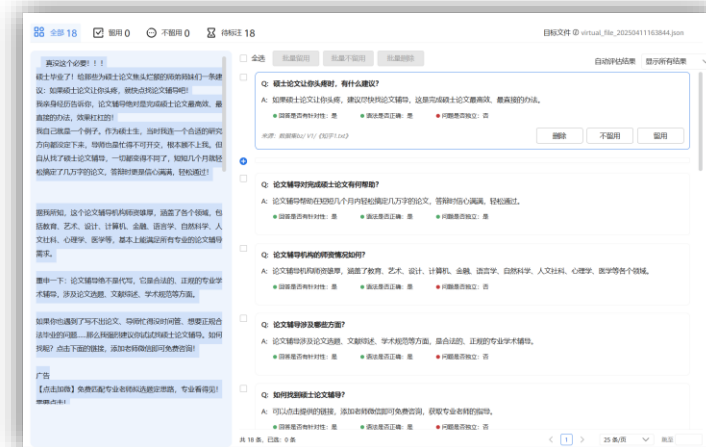
数据处理



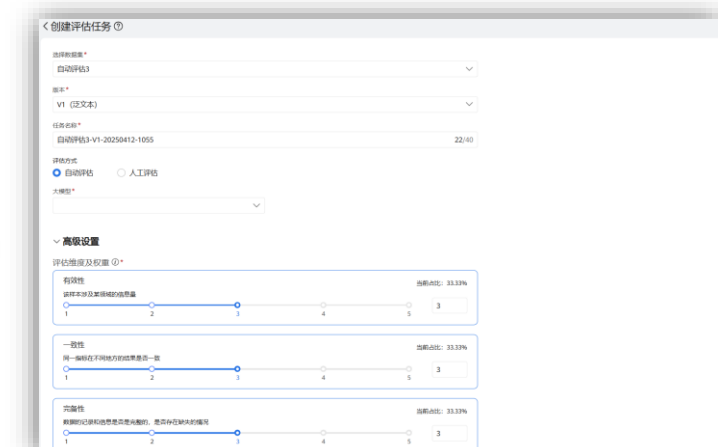
QA标注



向量知识库



专家审核

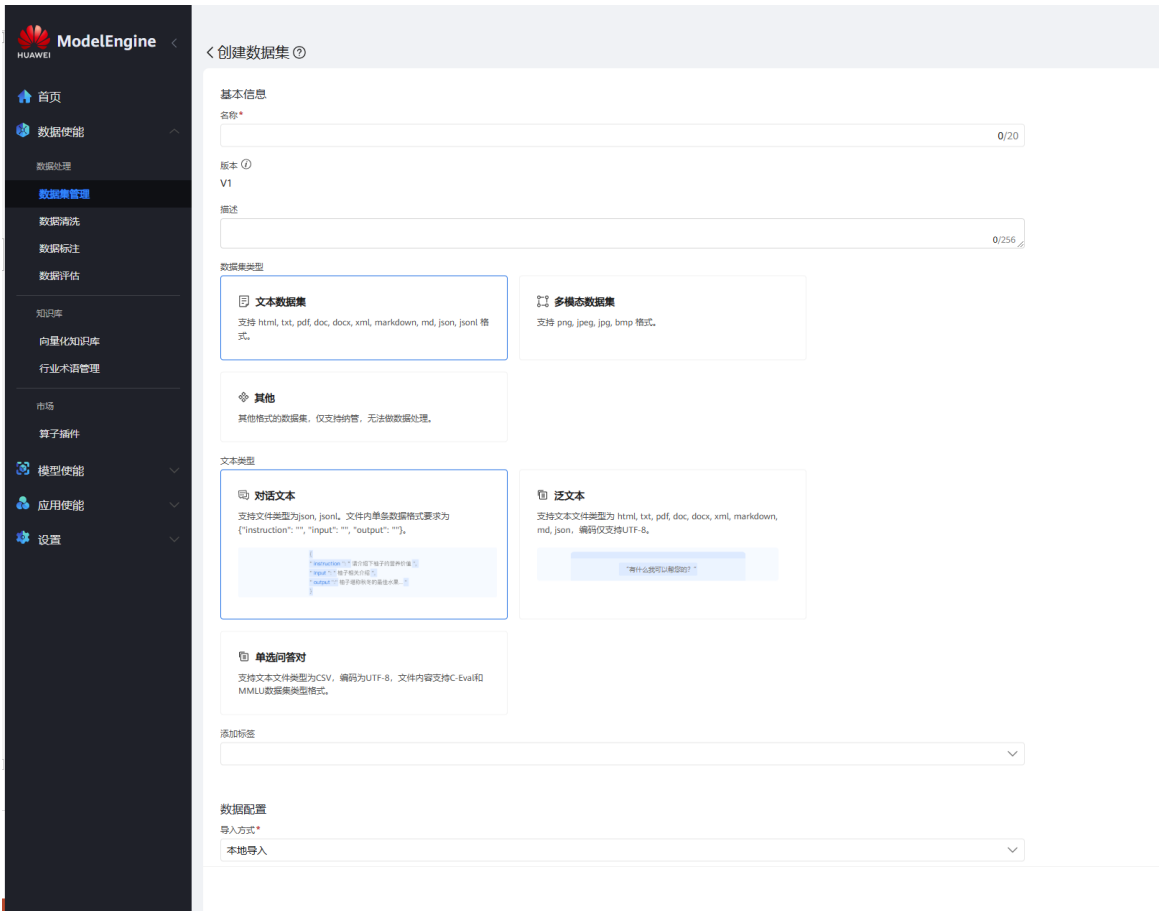


QA处理

# 数据集管理：多种数据归集和多种格式的数据管理

## 数据集管理

数据集管理是指将待处理数据从原始数据源归集至数据使能数据集中的过程，支持数据归集、版本管理、数据筛选、数据统计。



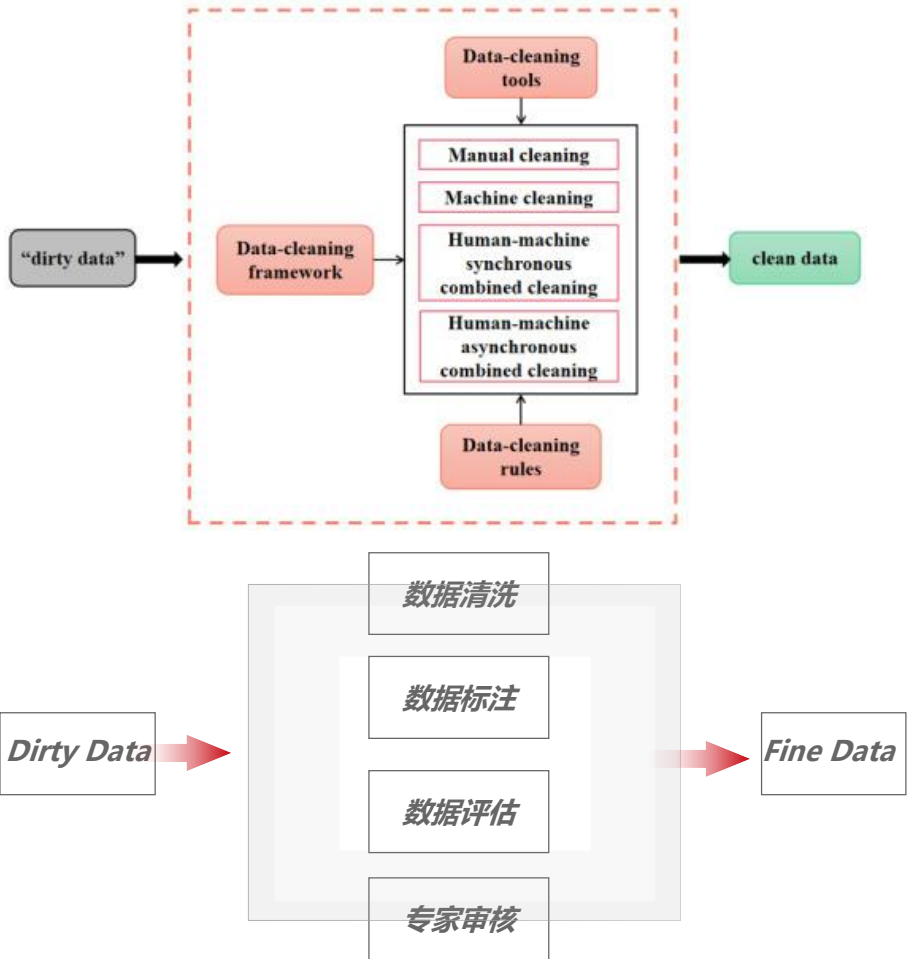
数据集类型	文件类型	详细描述
文本数据集	对话文本	主要针对模型微调过程中使用的QA对数据，文件内单条数据格式要求为[{"prompt": "prompt内容"}]，格式要求为：json/jsonl。
	泛文本	主要针对模型预训练/推理知识库使用的泛文本数据，支持文本文件类型为html/txt/pdf/docx/xml/md/doc/markdown，编码仅支持UTF-8。
	单选问答对	支持模型评测的自定义C-Eval和MMLU数据集，格式要求csv
多模态数据集	图片数据	基础图片数据，支持图片类型为jpg/png/bmp/jpeg。

本地上传	支持从本地直接上传数据至数据使能数据集。
NAS数据归集	支持NFS v3，支持Dorado 5500 V6和 Dorado 5600 V6。
OBS数据归集	支持OBS协议，支持Pacific 9550。

# 数据清洗：高质量泛文本数据清洗和QA对自动质量评估

## 数据清洗

数据清洗是指针对原始数据，进行一系列操作，最终得到清洗后的高质量数据的过程。按照数据处理范式次执行①文本抽取、②文档过滤、③异常清洗、④数据去重、⑤特殊词替换，得到清洗后的纯文本数据的过程。



< 创建数据清洗任务 >

基本信息 | 流程配置

加工步骤编排 全部收起

开始 从数据集中抽取文本内容。

文档字数检查 字数不在指定范围会被过滤。

文档字数 10 - + - 10000000 - +

文档重复率检查 去除重复率过多的文档。

文档重复率 0.5

文档词重复率检查 去除重复词过多的文档。

文档词重复率 0.5

去除停用词

文档特殊字符率检查 去除特殊字符过多的文档。

文档特殊字符率 0.3

文档敏感词率检查 去除敏感词过多的文档。

文档敏感词率 0.01

空格标准化 将文档中不同的 unicode 空格，如 u2008，转换为正常空格(u0020)。

多余空格去除 移除文档末尾、句中或标点符号附近多余空格和 tab 等。

全角转半角 将文档中的所有全角字符转换成半角字符。

不可见字符去除 去除文档中的不可见字符，例如 0-31 号字符中的部分字符。

文本抽取

- HTML文本抽取
- json文本抽取
- Markdown文本抽取
- PDF文本抽取CPU版
- TXT文本抽取
- XML文本抽取
- Word文本抽取

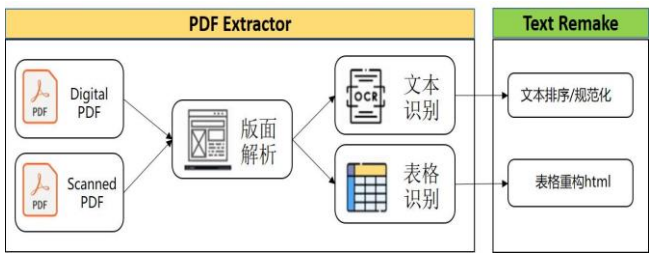
文档过滤 全不选

- 文档字数检查
- 文档重复率检查
- 文档词重复率检查
- 文档特殊字符率检查
- 文档敏感词率检查

异常清洗 全不选

- 空格标准化
- 多余空格去除
- 全角转半角
- 不可见字符去除
- 文档目录去除
- 图注去除
- 文档表格去除
- HTML标签去除
- 繁体转简体
- 文档乱码去除

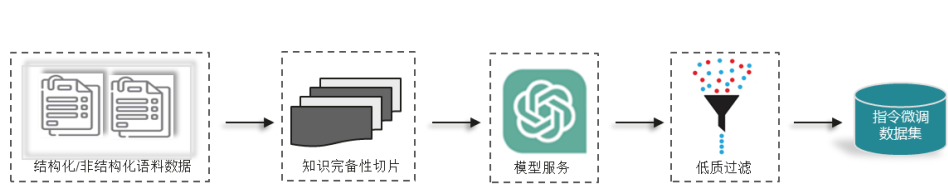
特殊词替换	描述
1)	信用卡号匿名化：将文档中的信用卡号匿名化为：<credit_card_number>。
2)	邮件地址匿名化：将文档中的邮箱地址匿名化为：<email>。
3)	IP地址匿名化：将文档中的IP地址匿名化为：<ip>。
4)	电话号码匿名化：将文档中的电话号码匿名化为：<tel>。
5)	政治文本匿名化：将文档中的政治敏感文字匿名化为：***。
6)	暴力色情文本匿名化：将文档中的暴力色情文字匿名化为：***。
7)	URL网址匿名化：将文档中的URL匿名化为：<url>。
8)	身份证号匿名化：将文档中的身份证号匿名化为：<id>。
9)	高级匿名化算子：检测姓名等五种类型实体并匿名化（NPU算子，仅支持Ascend 910 B3、910 B4和310P）。



# 数据标注：AI自动化生成模型微调数据

## 数据标注

数据标注通过对原始数据进行分类、标记或注释，以便为机器学习算法提供结构化的信息。这些经过标注的数据通常用于训练和验证模型，以提高模型的准确性和性能。



- 采用self-qa的算法，通过高质量非结构化原始数据进行QA对的批量生成 <https://arxiv.org/pdf/2305.11952>,
- 采用self-instruct对结构化数据进行QA对的批量生成 <https://arxiv.org/pdf/2212.10560>

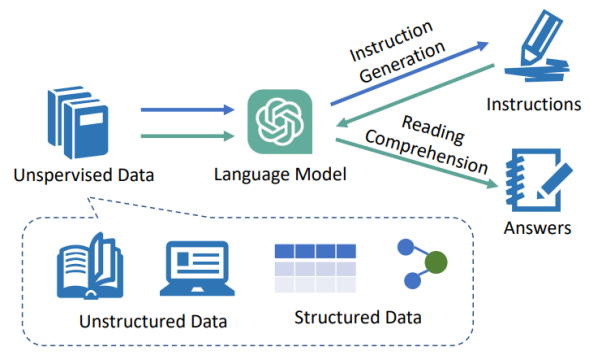


Figure 1: The pipeline of SELF-QA

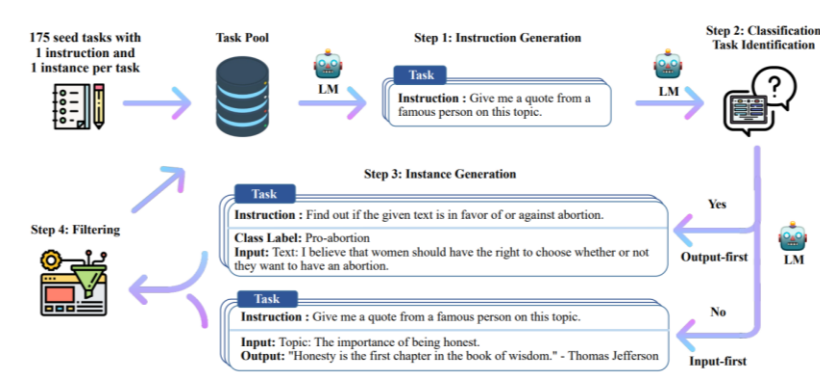
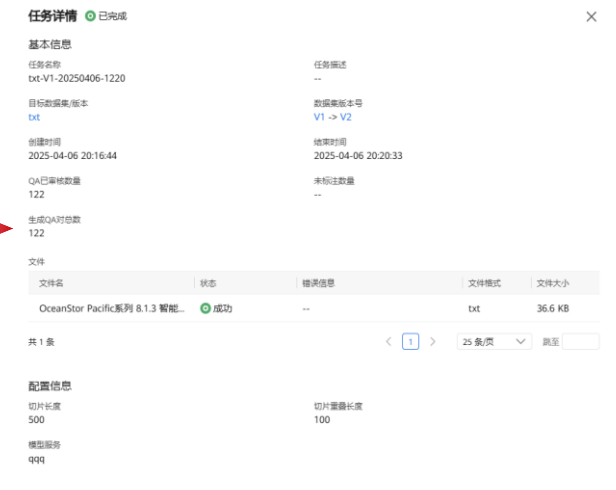
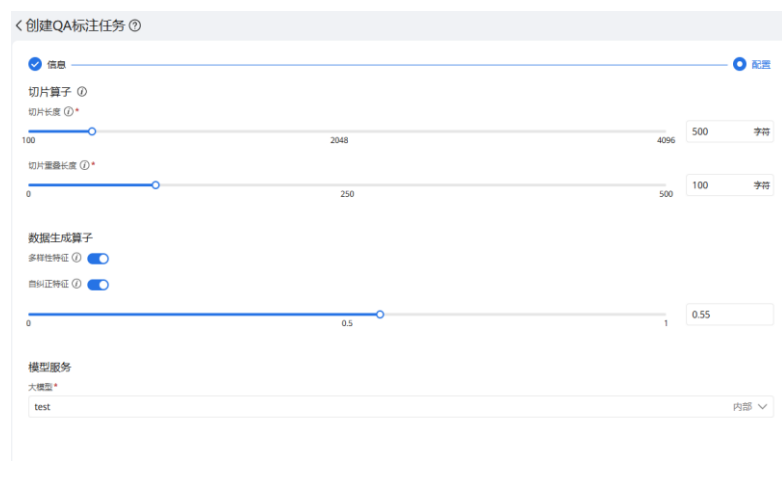


Figure 2: The pipeline of SELF-Instruct.



# 数据评估：自动评估结合专家审核使能垂域数据集快速构建

## 泛文本评估

自动化评估算子对输入的文本分片处理，通过句法质量、逻辑质量、领域相关性和内容知识性综合评估文本质量，最终通过有效性、一致性、完备性3个维度体现文本质量状态

任务ID	任务名称	数据集名称	来源数据集版本	目标数据集版本	任务状态	QA已审核	生成QA对	创建时间	启动时间	标注任务执行耗时	评估任务名称	评估任务状态	操作
23	高质量发展	气象数据集	气象数据集 V2	气象数据集 V5	已完成	109	109	2025-03-31 15:24	2025-04-01 03:40	1小时18分钟37秒	气象数据集-V2-20...	已完成	自动评估 留用审核 ...
21	Shaomx医疗模型-数据标注	Shaomx医疗数据数据集	Shaomx医疗数据数据集 V3	Shaomx医疗数据数据集 V4	已完成	0	5	2025-03-31 07:51	2025-04-01 03:27	0小时13分钟16秒	--	--	自动评估 留用审核 ...
20	财务税司核标注	2b3c税务司库核算数据集	2b3c税务司库核算数据集 V2	2b3c税务司库核算数据集 V3	终止	0	2	2025-03-31 07:06	2025-03-31 07:06	20小时20分钟50秒	2b3c税务司库核算...	已完成	自动评估 留用审核 ...
19	气象高质量发展QA	气象数据集	气象数据集 V2	气象数据集 V3	已完成	1	109	2025-03-31 03:37	2025-03-31 03:37	1小时18分钟34秒	气象数据集-V2-20...	已完成	自动评估 留用审核 ...
18	就业咨询数据标注	XX大学就业咨询	XX大学就业咨询 V2	XX大学就业咨询 V3	部分成功 (成功)	21	177	2025-03-25 09:45	2025-03-25 09:45	49小时13分钟19秒	--	--	自动评估 留用审核 ...
17	test	sjm数据集	sjm数据集 V10	sjm数据集 V13	已完成	4	4	2025-03-19 09:52	2025-03-19 09:52	0小时1分钟9秒	sjm数据集-V10-20...	已完成	自动评估 留用审核 ...

数据类型	评估方法	描述
泛文本质量评估	自动评估	从有效性、一致性、完备性3个维度，基于算子对数据进行自动化质量评估。
	人工评估	基于预置评估模型，从不同维度，按照权重对数据进行人工评估。
QA对质量评估	自动评估	基于大模型能力，从①回答是否有针对性、②语法是否正确、③问题是否独立3个角度，对QA对数据进行质量评估，辅助进行QA对留用审核。
	人工评估	从最多10个维度，对QA对数据进行人工质量评估。

泛文本自动评估

泛文本人工评估

泛文本本报告



# 数据评估：自动评估结合专家审核使能垂域数据集快速构建

## QA对评估

通过语义完整性及知识完备性高质量指令数据生成方案，实现自动化QA数据生成，降低人力投入成本，结合专家审核，实现高质量及多样性QA数据生成。

任务ID	任务名称	数据集名称	来源数据集	目标数据集	任务状态	QA已审核	生成QA数	创建时间	启动时间	标注任务执行耗时	评估任务名称	评估任务状态	操作
23	高质量发展	气象数据集	气象数据集 V2	气象数据集 V5	已完成	109	109	2025-03-31 15:24...	2025-04-01 03:40...	1小时18分37秒	气象数据集-V2-20...	已完成	自动评估 留用审核 ...
21	Shaomx医疗模型-数据标注	Shaomx医疗数据数据测试	Shaomx医疗数据数据测试 V3	Shaomx医疗数据数据测试 V4	已完成	0	5	2025-03-31 07:51...	2025-04-01 03:27...	0小时13分16秒	--	--	自动评估 留用审核 ...
20	税务司标注	2b3c税务司标注数据集	2b3c税务司标注数据集 V2	2b3c税务司标注数据集 V3	终止	0	2	2025-03-31 07:06...	2025-03-31 07:06...	20小时20分50秒	2b3c税务司标注...	已完成	自动评估 留用审核 ...
19	气象高质量发展QA	气象数据集	气象数据集 V2	气象数据集 V3	已完成	1	109	2025-03-31 03:37...	2025-03-31 03:37...	1小时18分34秒	气象数据集-V2-20...	已完成	自动评估 留用审核 ...
18	就业咨询数据标注	XX大学就业咨询	XX大学就业咨询 V2	XX大学就业咨询 V3	部分成功 (成功)	21	177	2025-03-25 09:45...	2025-03-25 09:45...	49小时13分19秒	--	--	自动评估 留用审核 ...
17	test	sjm数据集	sjm数据集 V10	sjm数据集 V13	已完成	4	4	2025-03-19 09:52...	2025-03-19 09:52...	0小时1分9秒	sjm数据集-V10-20...	已完成	自动评估 留用审核 ...

数据类型	评估方法	描述
泛文本质量评估	自动评估	从有效性、一致性、完备性3个维度，基于算子对数据进行自动化质量评估。
	人工评估	基于预置评估模型，从不同维度，按照权重对数据进行人工评估。
QA对质量评估	自动评估	基于大模型能力，从①回答是否有针对性、②语法是否正确、③问题是否独立3个角度，对QA对数据进行质量评估，辅助进行QA对留用审核。
	人工评估	从最多10个维度，对QA对数据进行人工质量评估。

### 创建数据评估任务

数据集: Shaomx医疗数据数据测试 | 评估方式: 自动评估

任务名称: Shaomx医疗数据数据测试-V3-20250415-2336 (31/40)

大模型: test (内部)

评估维度:

评估维度	维度解释
回答是否有针对性	针对问题内的疑问点，答案中都有正面、直接的回答，并且没有遗漏或是答非所问的情况。
语法是否正确	语句没有语法错误，并且通顺、符合逻辑。
问题是否独立	问题是否依赖于文档片段，且不能独立存在。例如“这篇文章讲了什么？”，这个问题就不是独立的。

### 数据评估分析报告

2b3c税务司标注数据集/V3

生成时间: 2025-04-15 15:37:06

基本信息: 2b3c税务司标注数据集 V3

评估结果:

QA对详情	问题是否独立	语法是否正确	回答是否有针对性
Q: 如何评估数据质量？	是	是	是
Q: 如何评估数据质量？	否	是	是

全部 2 | 禁用 0 | 不启用 0 | 待标注 2

目标文件: virtual\_file\_20250408125450.json

Q: What components are included in the train folder?

A: The train folder contains the following components: Miniconda3 file, two CANN packages, a PyTorch package, a model\_lite\_utils folder, and a MindSpeed-LLM folder.

Q: 如何评估数据质量？

A: Execute command 'nodedocker save -o /opt/model-lite/train\_image/mindspeed24.1.rc1.tar mindspeed24.1.rc1' to save training image as a tar package.

QA评估

QA评估报告

专家审核

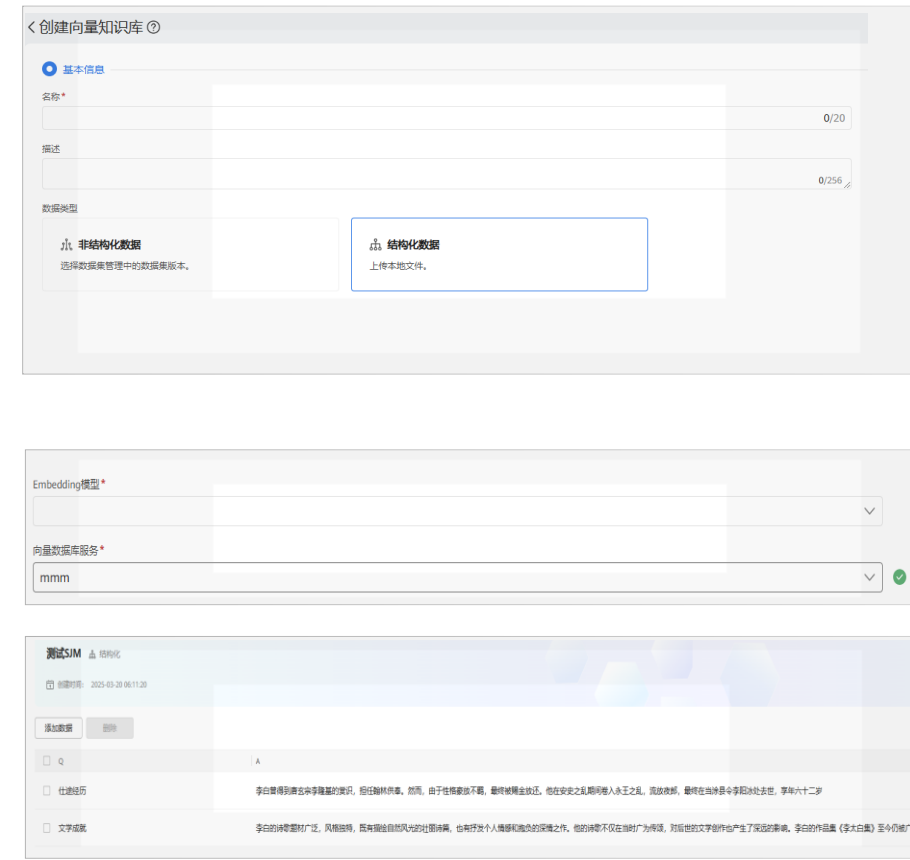
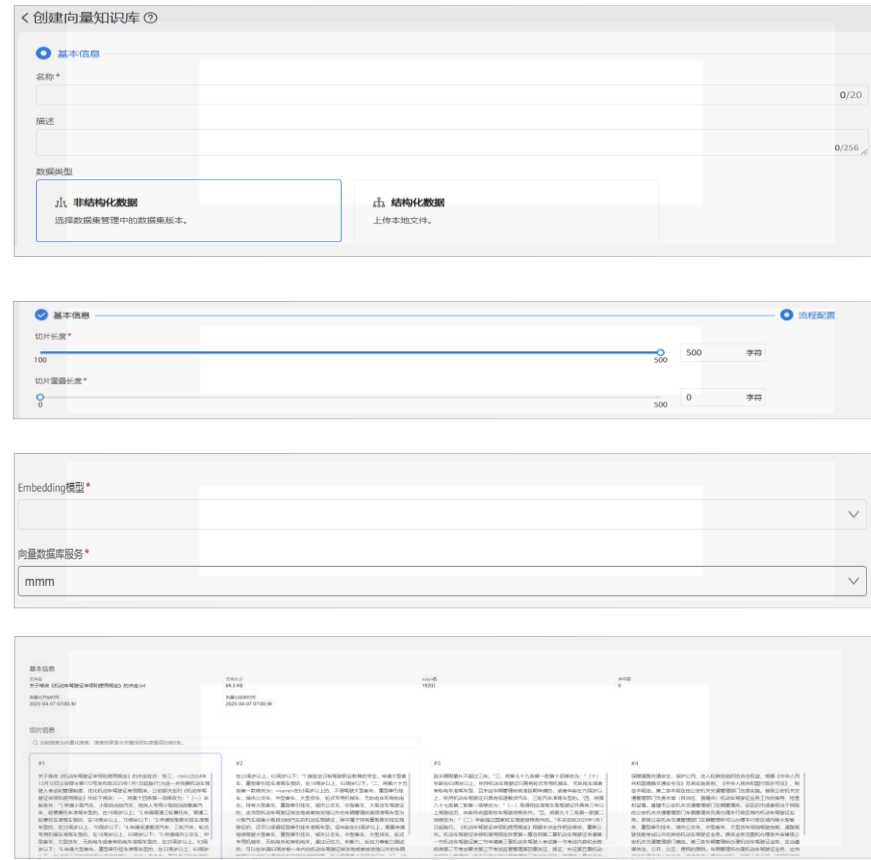
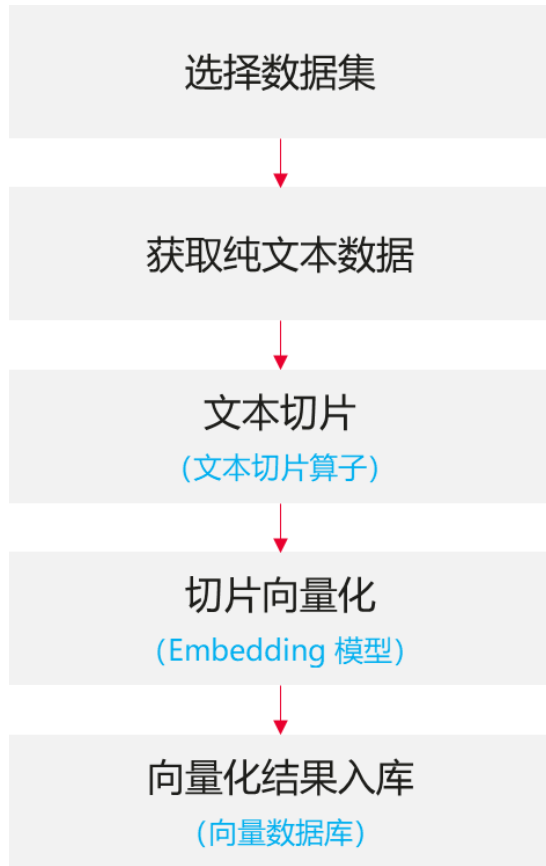


# 向量知识库：构筑了知识向量化能力

## 向量知识库

向量知识库基于知识向量化能力，将清洗后的纯文本数据，转化为模型推理可用的向量化知识，为应用使能模块提供知识库能力

文本切片：依据一定的规则，将长文本切分为短的文本片段，方便进行向量化。  
 切片向量化：基于Embedding模型，将切片后的文本转化为其在Embedding空间的语义向量表达。  
 向量入库：将Embedding模型向量化后的结果，保存至向量数据库中，以向量相似度进行语义检索。



## 请回忆一下

### ModelEngine支持多少种类泛文本数据？支持多少种上传方式（A）

- A. 8, 3
- B. 7, 2
- C. 5, 2
- D. 9, 3

### ModelEngine更换embedding模型后不需要重新生成向量知识库吗？

错，知识库关联了生成知识的embedding，把embedding卸载了重新拉起一个新的，知识库还是保留了原来知识库的embedding元数据。

### 以下哪个流程是正确的（B）

- A. 结构化数据——切片——向量化——入库
- B. 非结构化数据——切片——向量化——入库
- C. 结构化数据——向量化——切片——入库
- D. 结构化数据——向量化——切片——入库

# Thank you.

把数字世界带入每个人、每个家庭、  
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and  
organization for a fully connected,  
intelligent world.

**Copyright©2018 Huawei Technologies Co., Ltd.  
All Rights Reserved.**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

