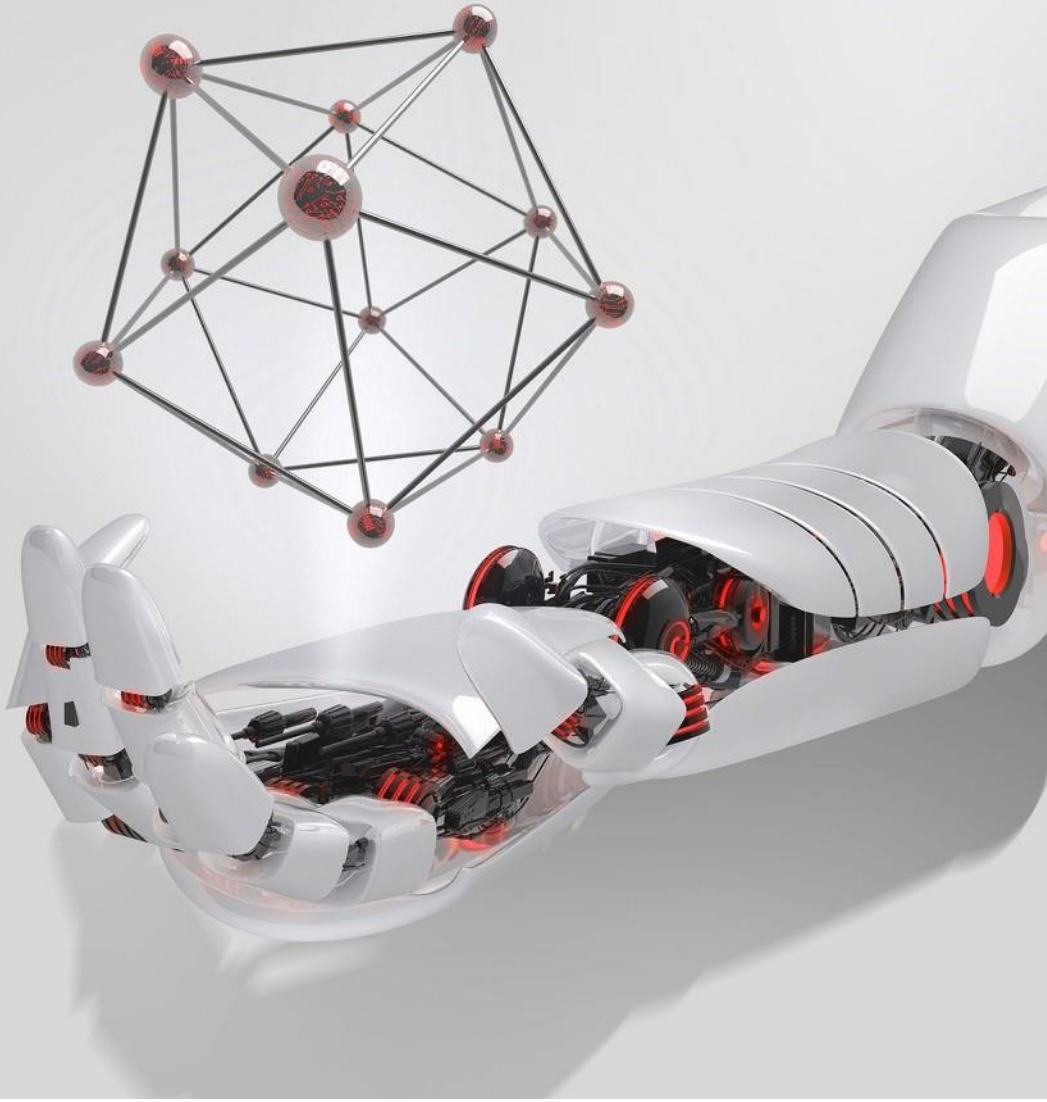
ModelEngine使能平台赋能



Security Level:

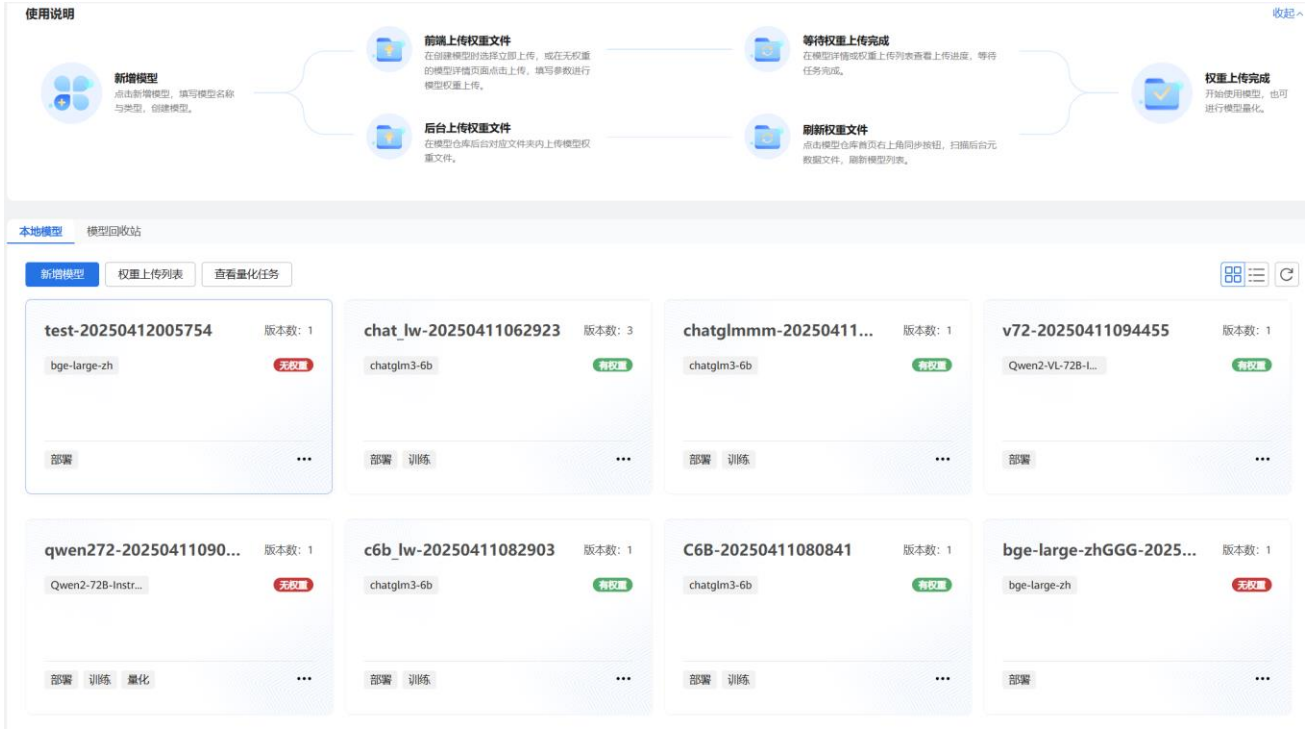
HUAWEI

目录

模型使能：模型仓库和权重上传

模型仓库

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。

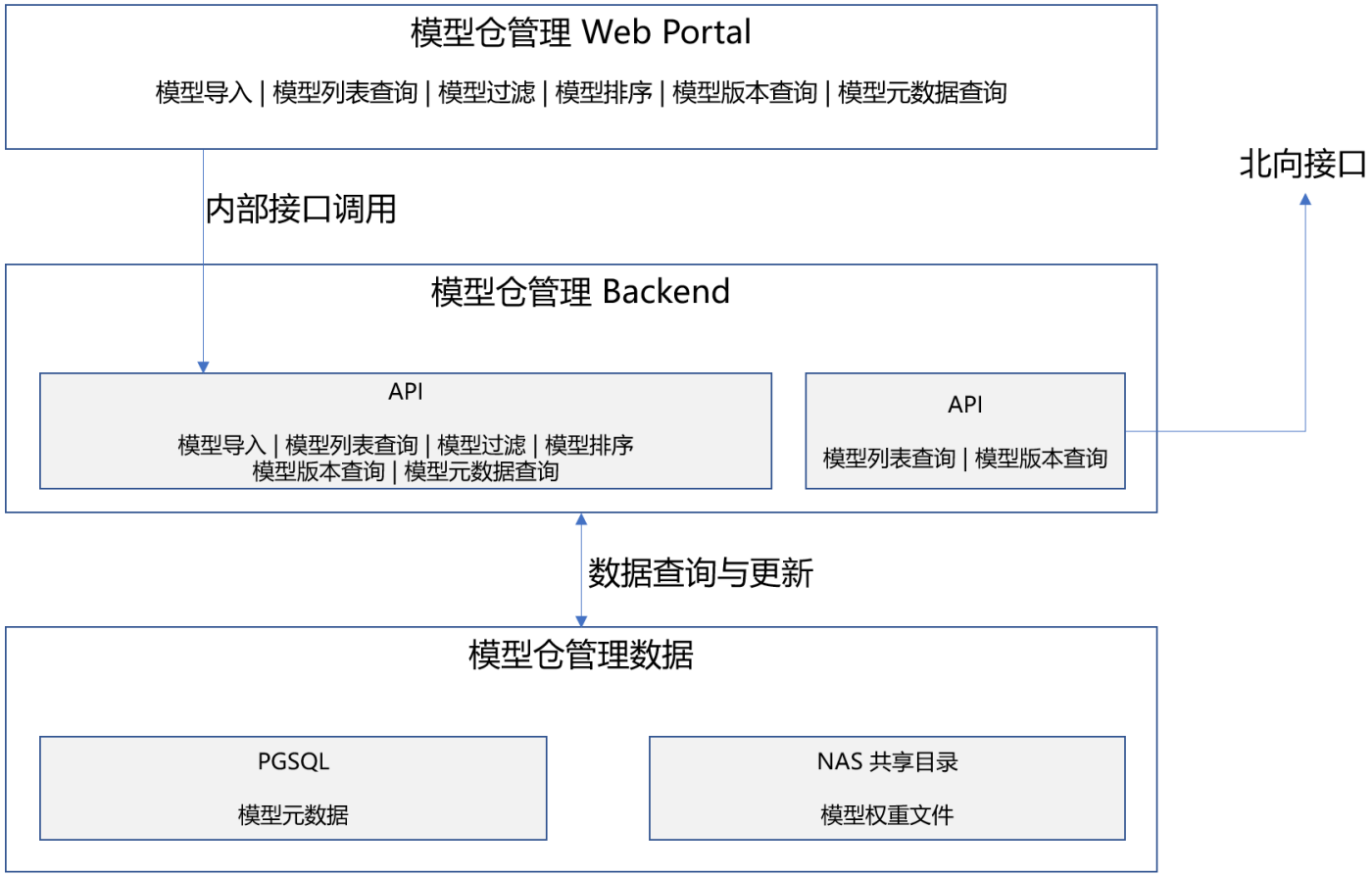


支持模型	支持操作	推理精度
Qwen-72B-Chat	部署，评测	Ascend 910B：FP16 Ascend 310P：FP16
Qwen2-72B-Instruct	部署，精调，评测，量化	Ascend 910B：BF16，FP16，量化精度（W8A8，W8A16） Ascend 310P：FP16
Qwen-14B-Chat	部署，评测	Ascend 910B：FP16 Ascend 310P：FP16
chatglm3-6b	部署，精调，评测	Ascend 910B：FP16 Ascend 310P：FP16
bge-large-zh	部署	FP32
Meta-Llama-3-8B-Instruct	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

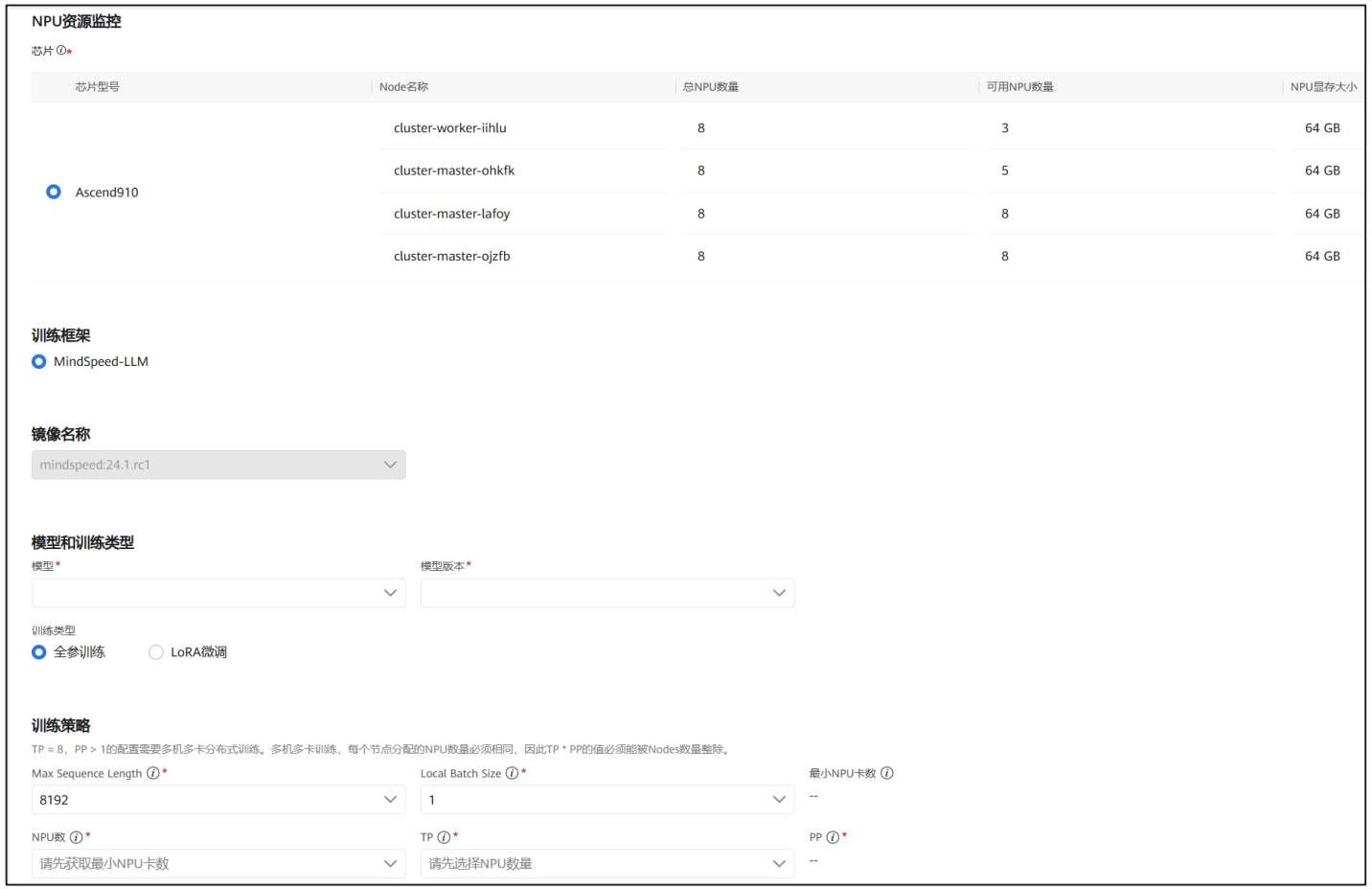
型仓库支持管理上传的模型，以及训练后得到的版本，主要提供模型新增、模型列表查询、模型回收站和版本回收站等功能。



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和LoRA微调。



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

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

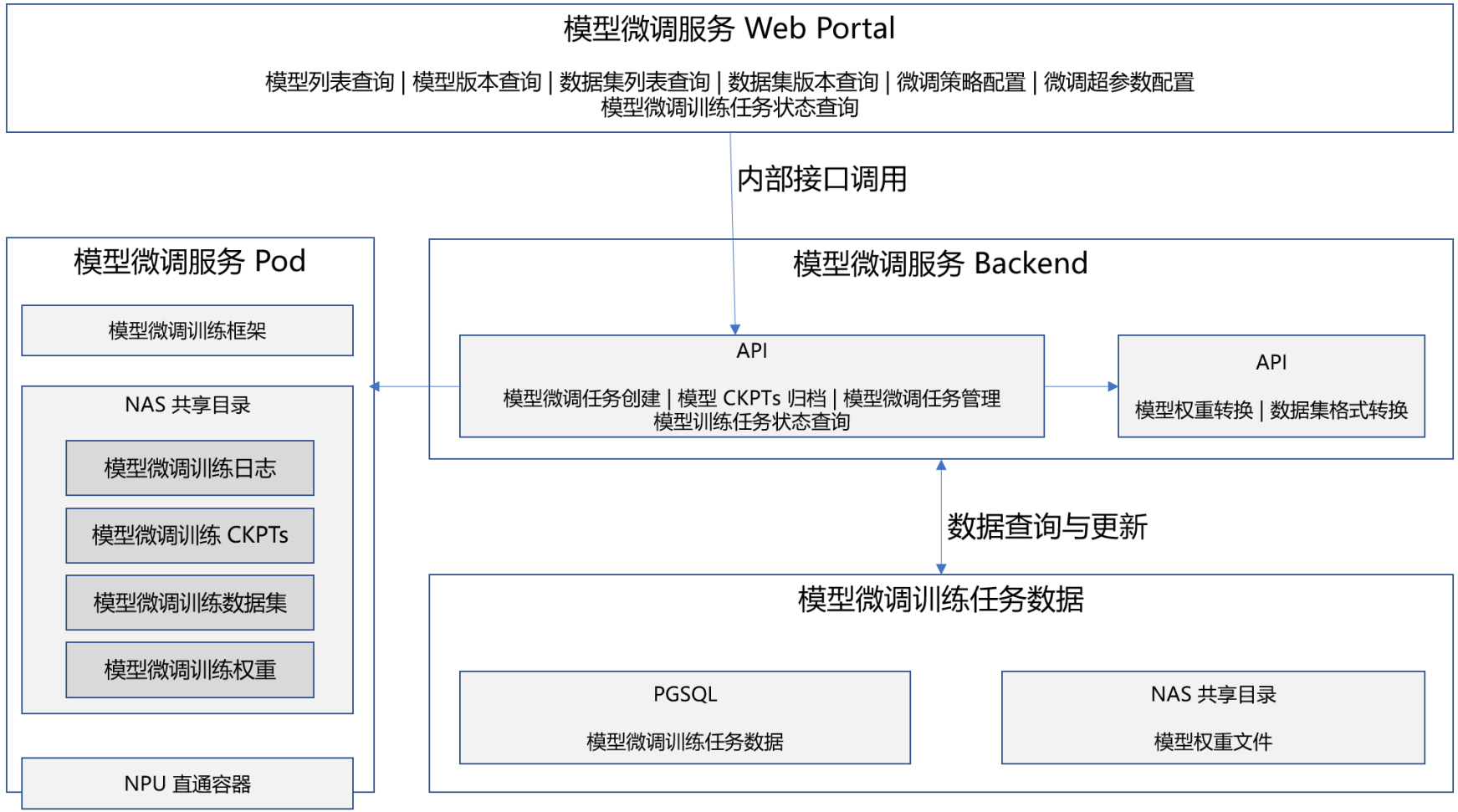
参数	说明
Gradient Accumulation Steps	每次模型参数更新前的迭代次数。
Warm Up Ratio	表明预热过程中，学习率的增长轮数和总训练轮数的比例。
Lora R（LORA训练中低秩矩阵维度）	值越大，模型学习能力越强，但显存开销越大。
Lora Alpha（LORA学习率放缩因子）	一般为Lora R的2倍，控制LORA参数的更新。
Epoch	遍历数据集的次数（训练中会被转换成iter的形式）。
Learning Rate	控制权重参数的更新速度。

LoRA微调参数

模型使能：模型精调

模型精调

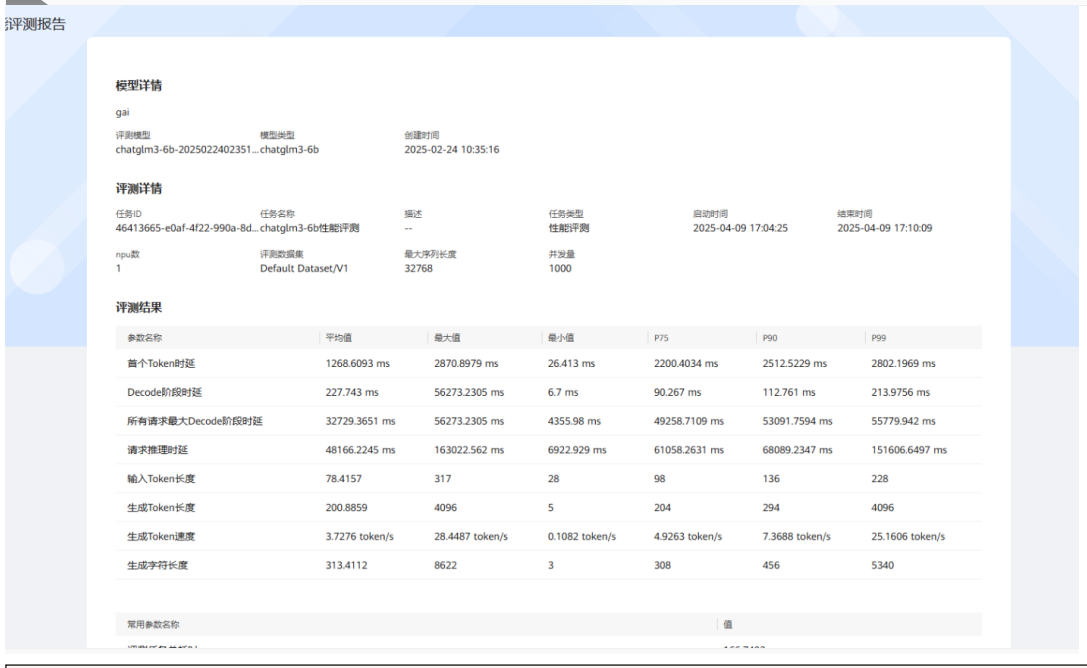
模型训练支持NPU资源监控，支持训练策略配置，基于数据使能处理对话数据集进行全参微调和LoRA微调。

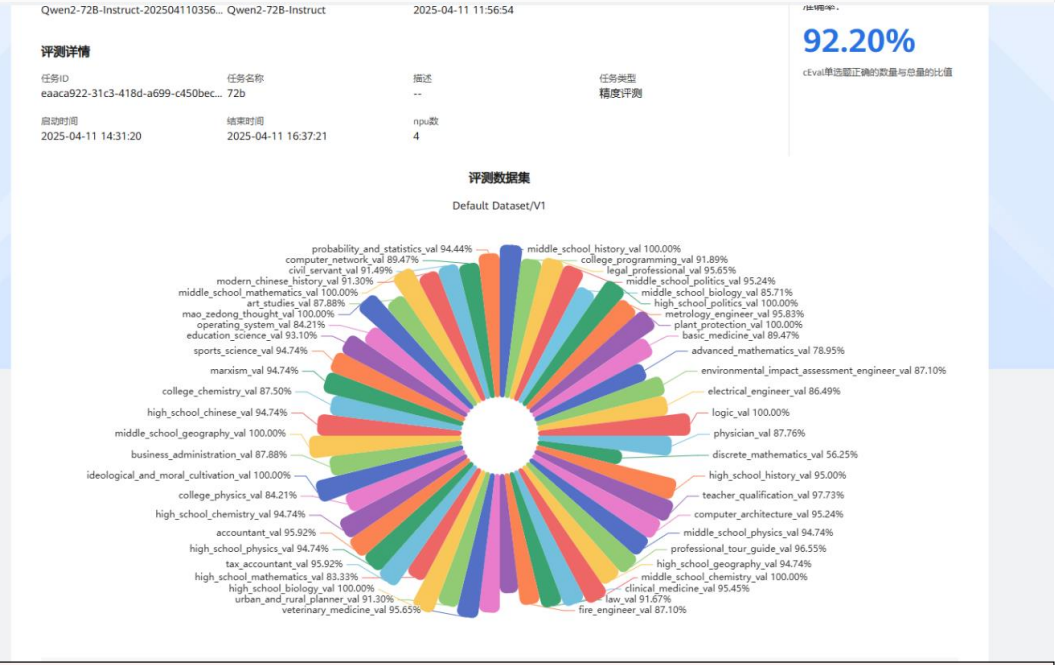


模型使能：模型评测

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。





模型工具链 一栈式模型优化方案

数据使能	→	模型使能	→	应用使能
自动化数据处理和知识生成		轻量级训推工具链		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

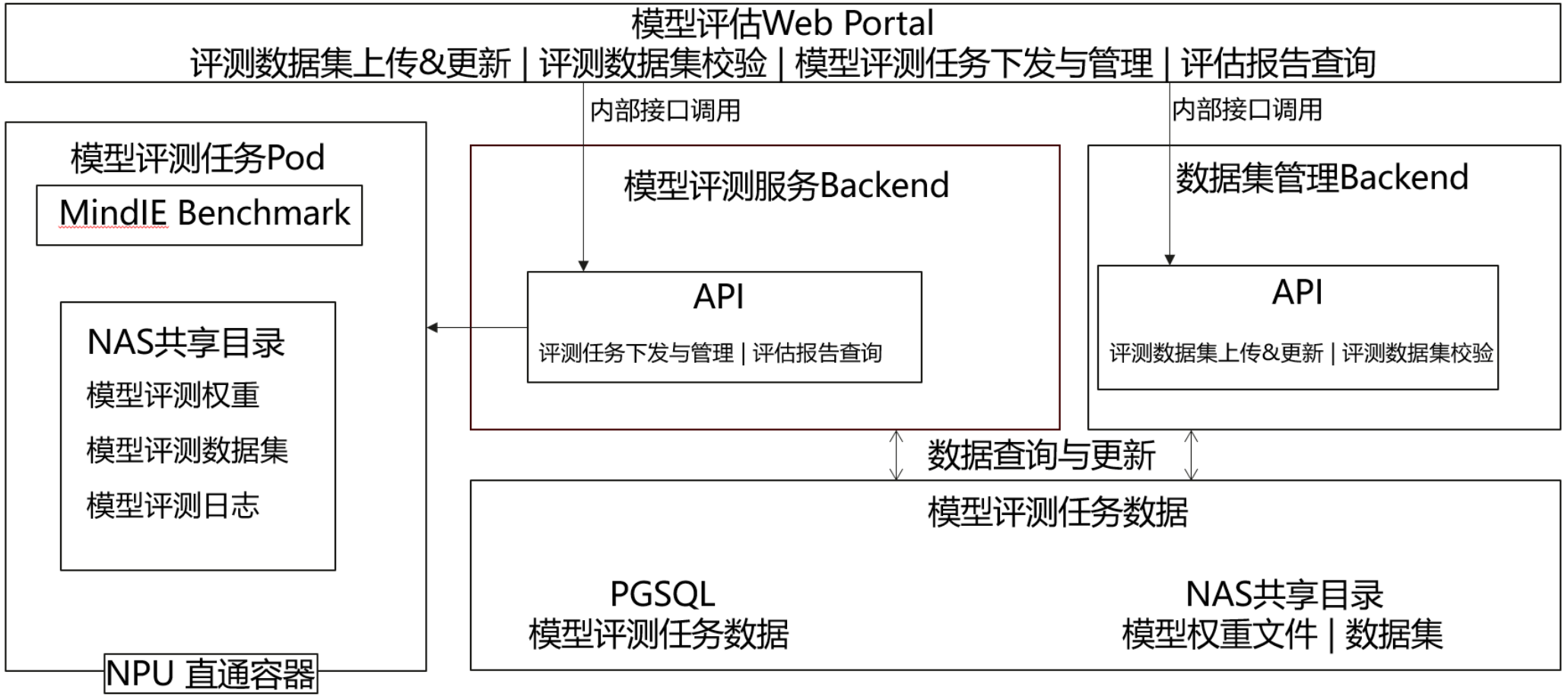
模型服务
模型推理服务 | 北向API接口

HUAWEI

模型使能：模型评测

模型评测

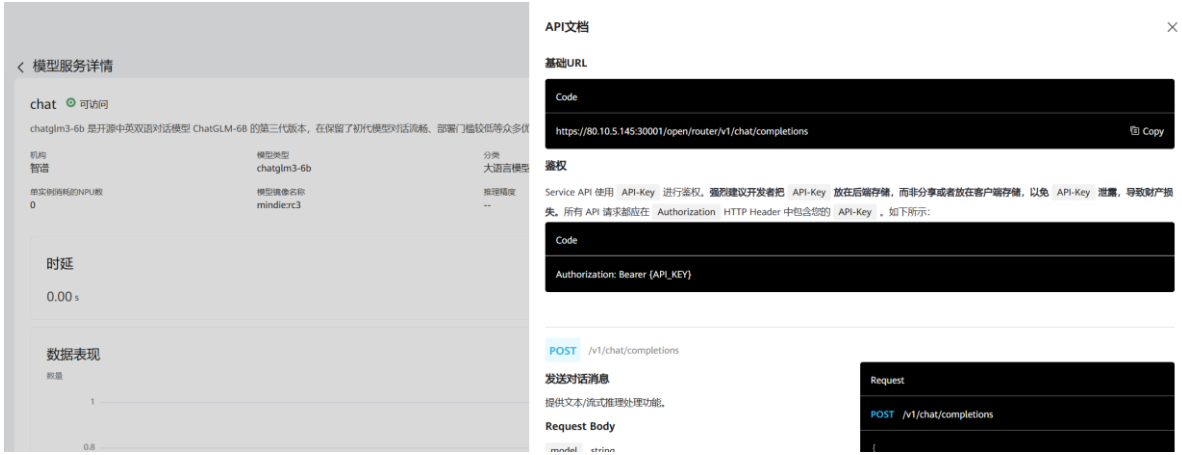
对模型仓库中已有的模型进行评测，获得模型的精度指标和性能指标，生成评测报告。



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。



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。

HUAWEI

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

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

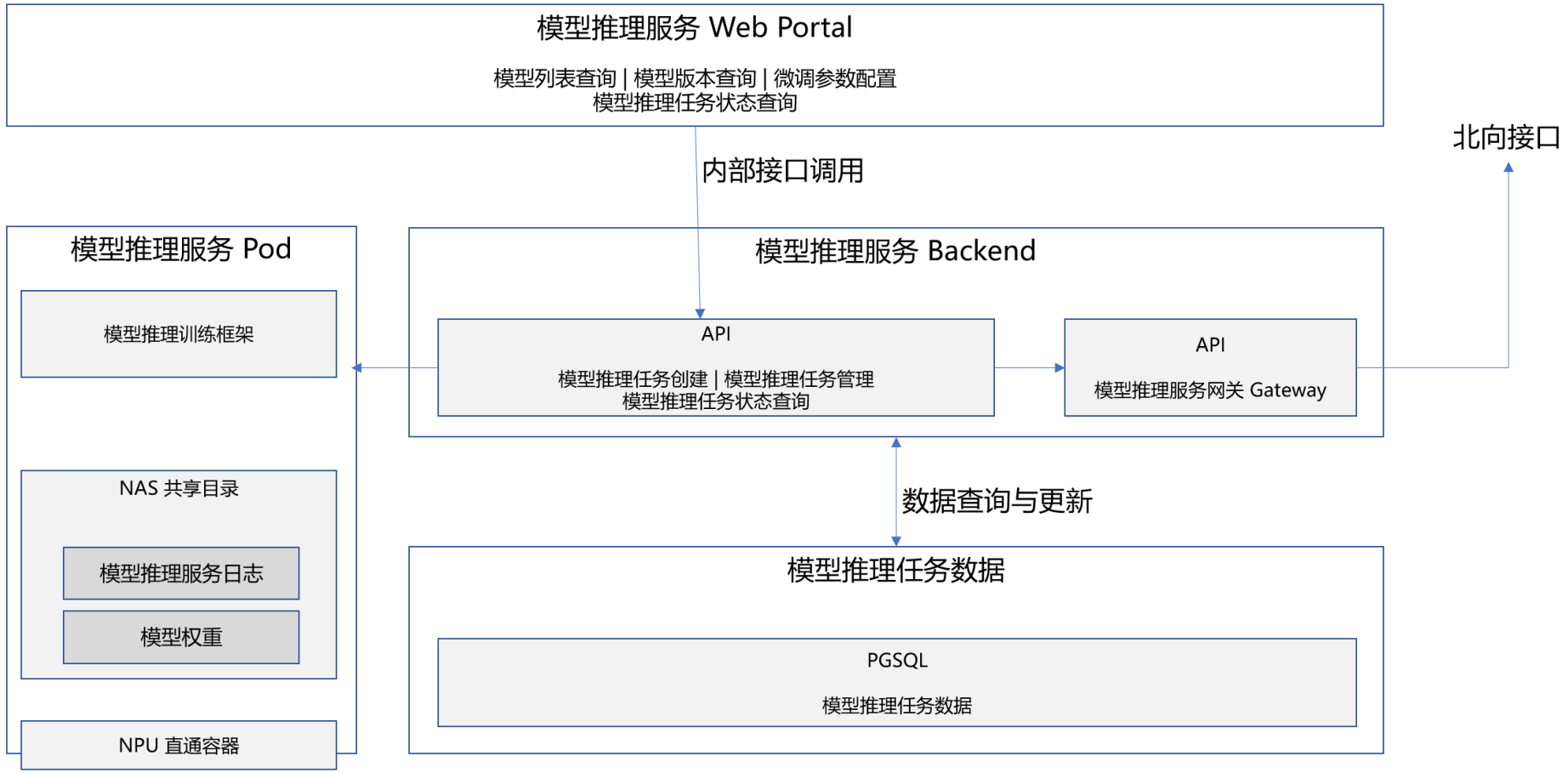
有的兄弟，有的，差异体现在输入输出中间过程上

W8A16

模型使能：模型服务和北向API接口

模型服务

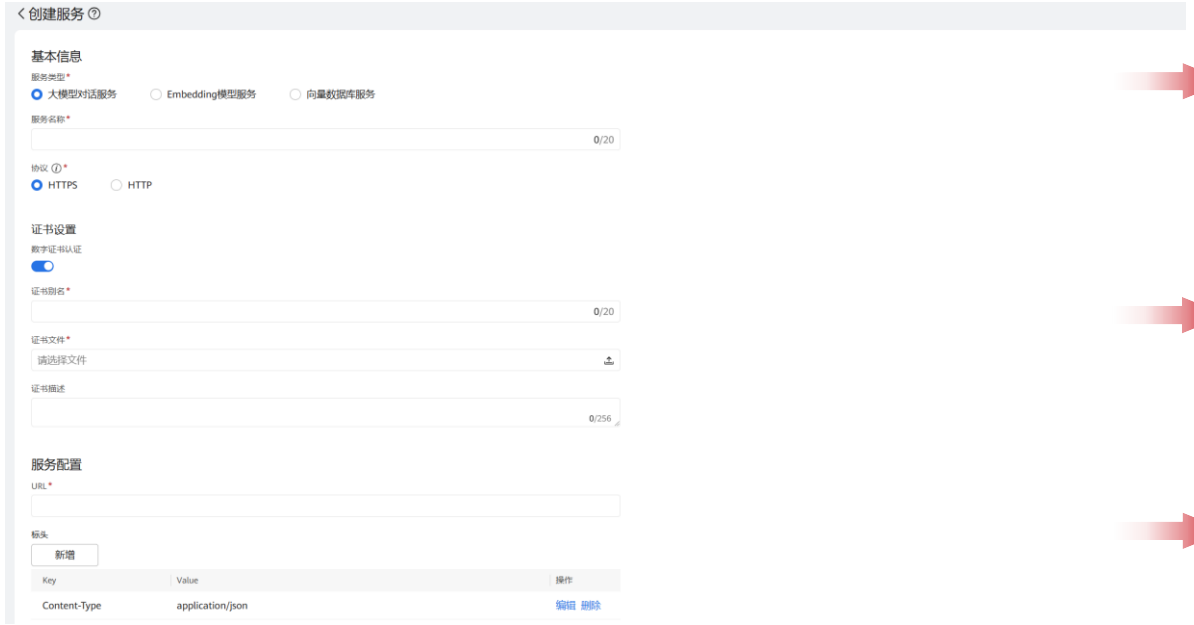
模型服务基于资源进行本地模型部署，支持基于当前资源状况进行模型服务管理。

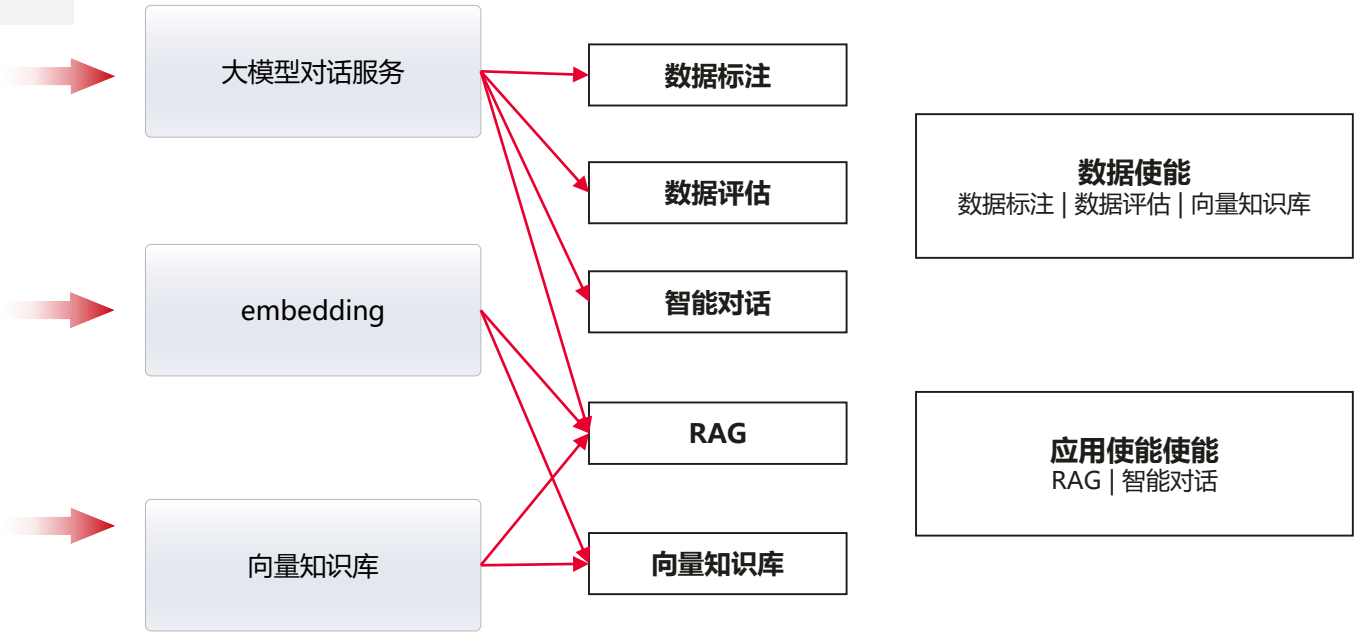


服务接入：

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



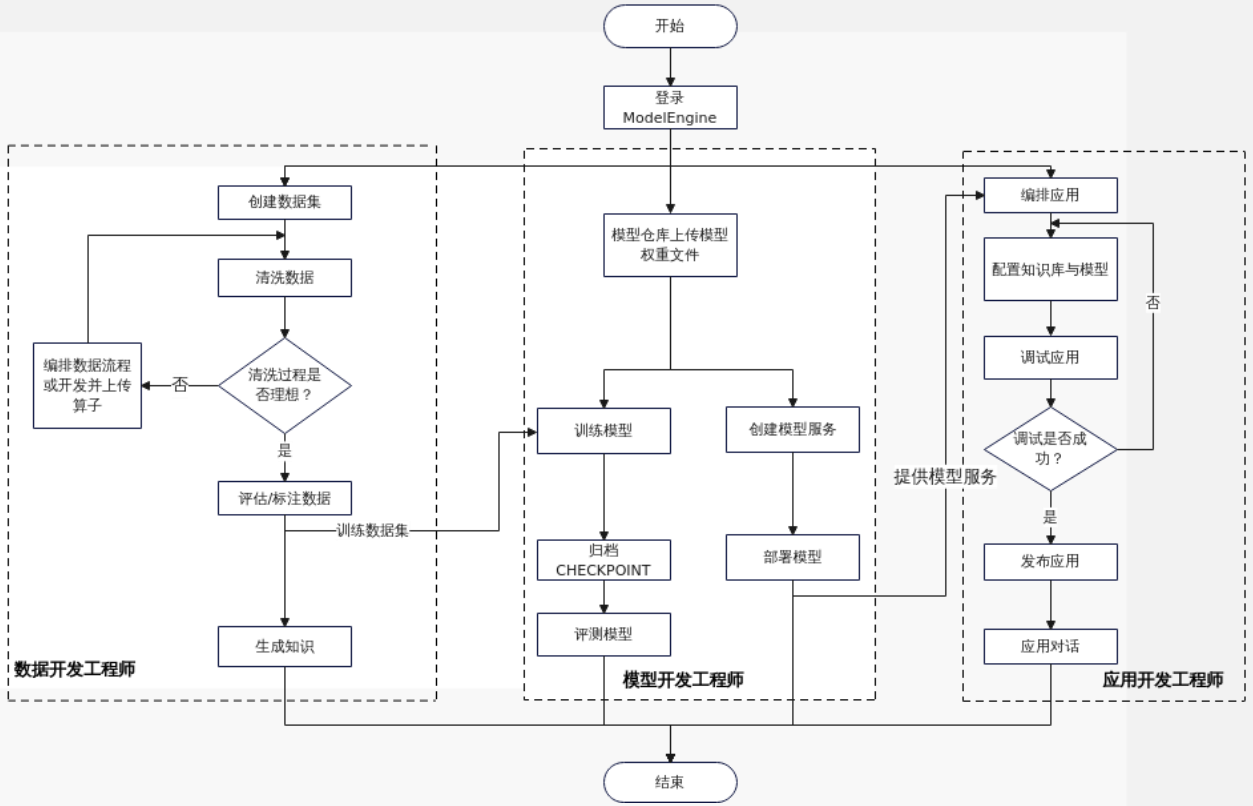


HUAWEI

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。



HUAWEI

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

Raw Data

非结构化

半结构化

QA问答对

推理语料：提供**原始数据生成知识**能力，对接主流向量数据库，Embedding模型可适配

- 数据随路清洗

1. 数据归集

- 支持NFS v3、OBS协议
- Dorado存储数据可归集

2. 数据清洗：支持40+内置数据处理算子和图文清洗

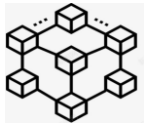
归集 → 清洗 → 增强 → 评估 → *Fine Data*

3. 数据增强：支持QA自动标注和自动评估以及专家审核

训练语料：算子插件化，流程可编排，训练用的**泛文本数据与QA对自动质量评估**

模型训练

For**模型精调**提供高质量语料



- 数据质量
- 数据匿名化
- 重复删除
- 异常数据清洗

- 切片
- Embedding
- 索引
- 召回精度

For**向量知识库**提供高质量语料

应用编排

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**

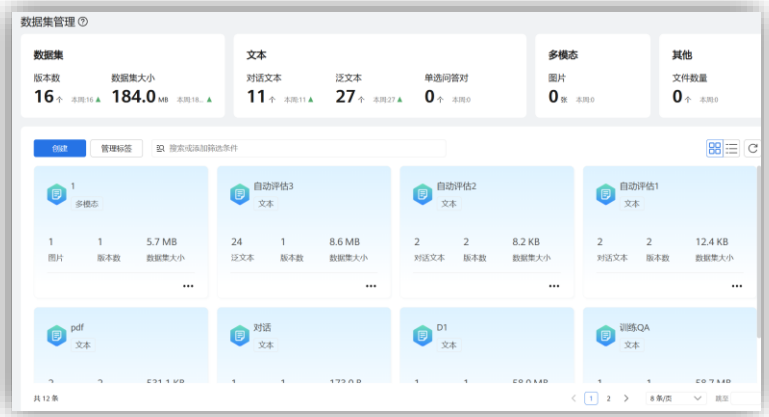
基于清洗后的文本数据与外置大模型服务，自动生成大模型微调QA对；具备QA对自动评估，满足QA高效生成能力

- **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

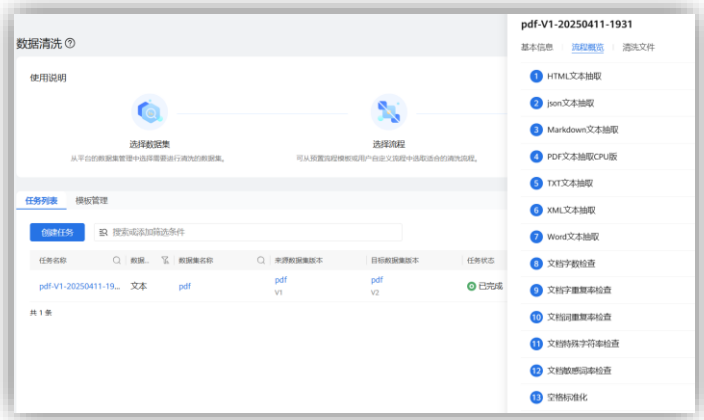
数据使能：高质量泛文本数据清洗和QA对自动质量评估

数据使能

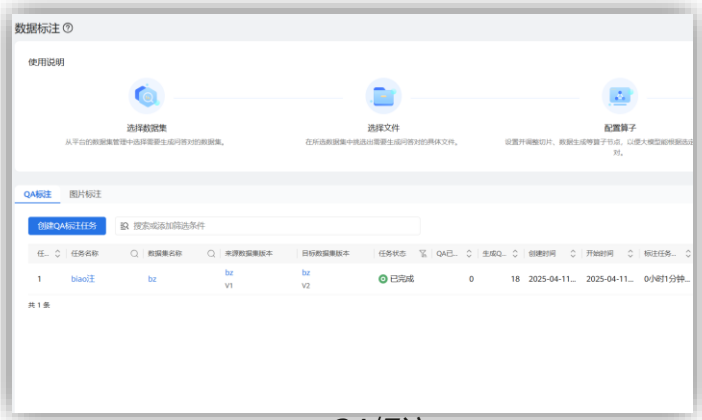
基于资源使能底座和模型底座完成数据清洗一栈式流程，含数据集管理、数据处理、QA标注、QA评估、专家审核、向量知识库生成能力。



数据集管理

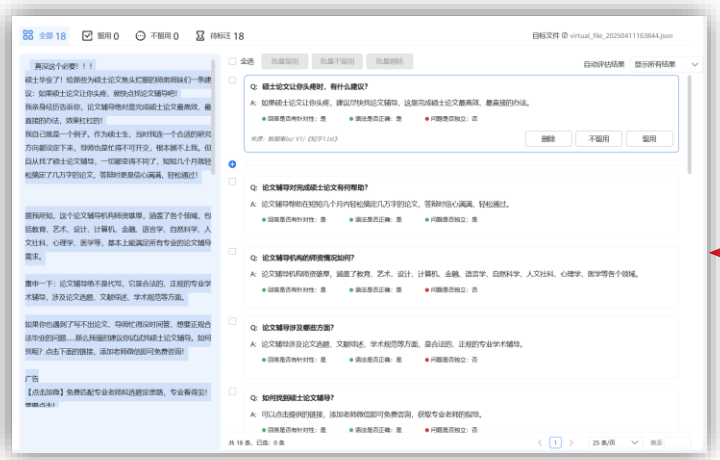


数据处理

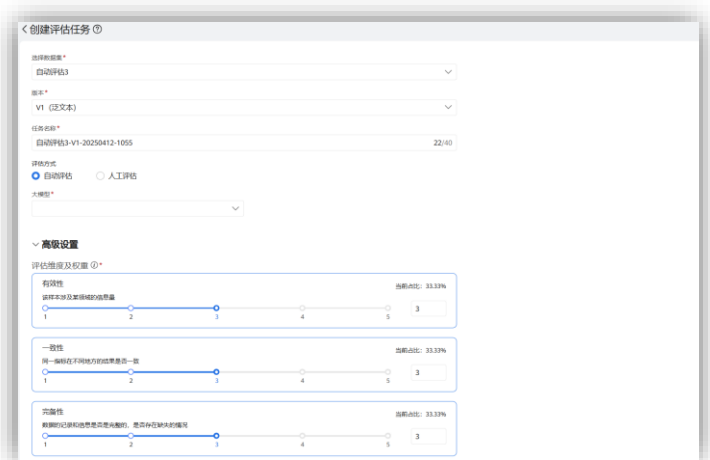


QA标注

向量知识库



专家审核

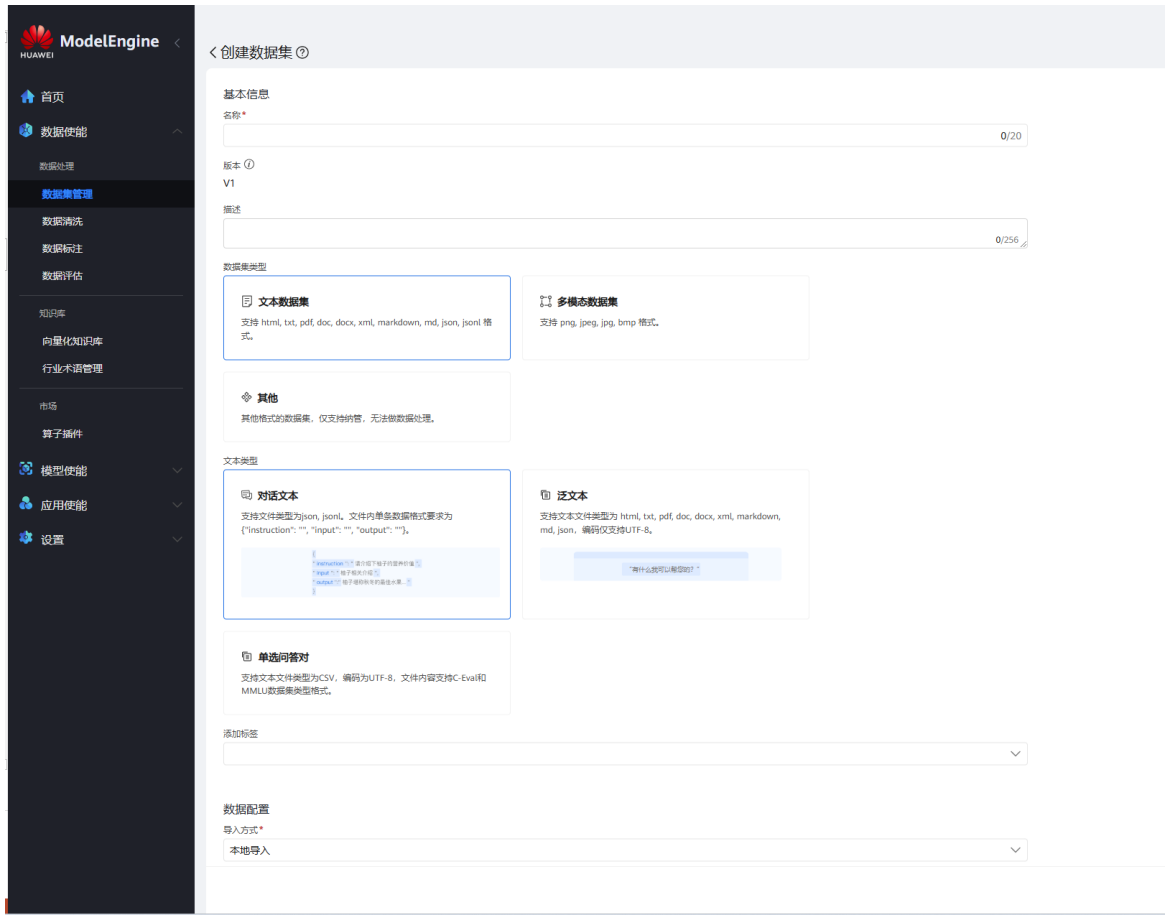


QA处理

数据集管理：多种方式的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。

专家审核

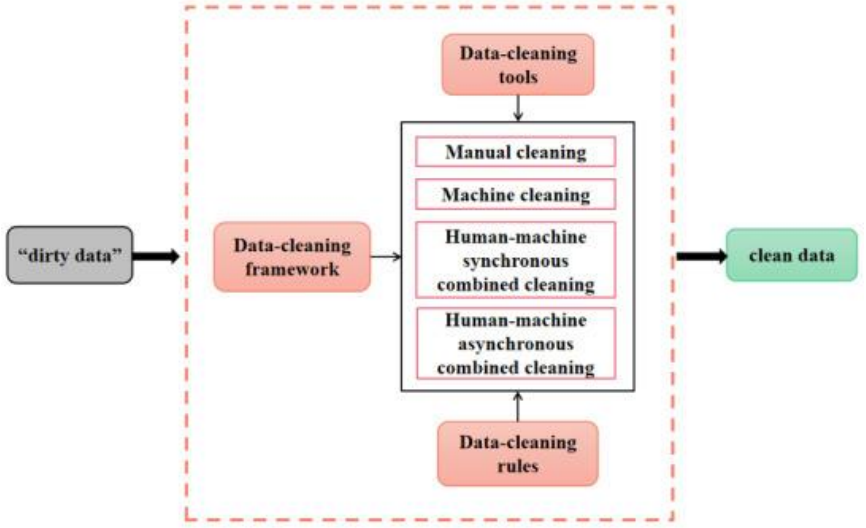
QA处理

HUAWEI

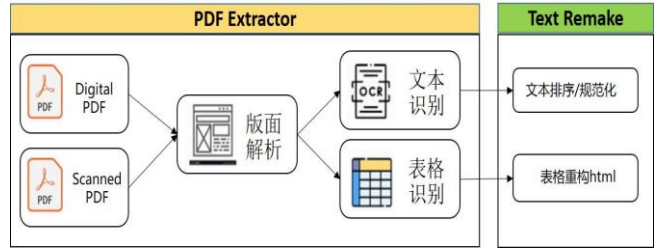
数据清洗：高质量泛文本数据清洗和QA对自动质量评估

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。



特殊词替换	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）。

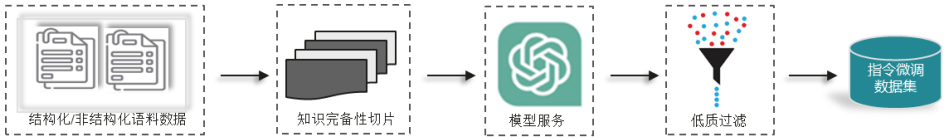


HUAWEI

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

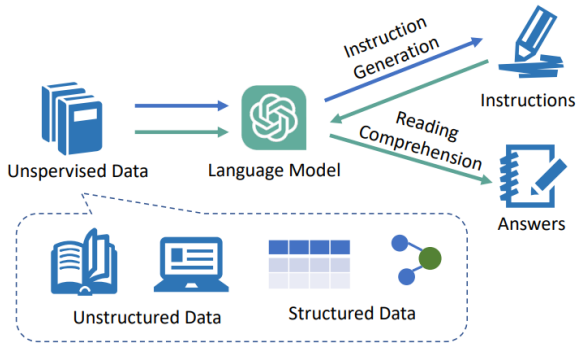


Figure 1: The pipeline of SELF-QA

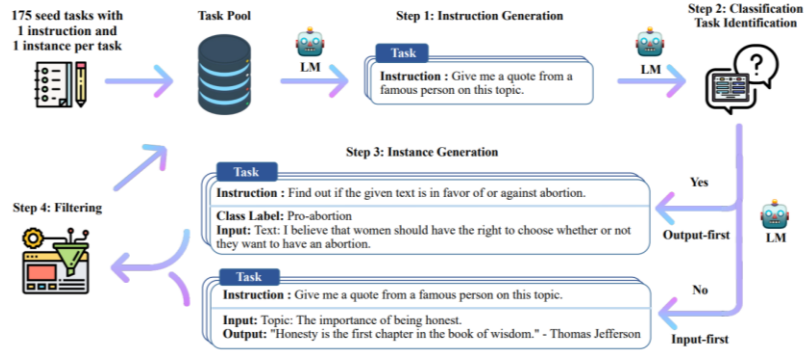


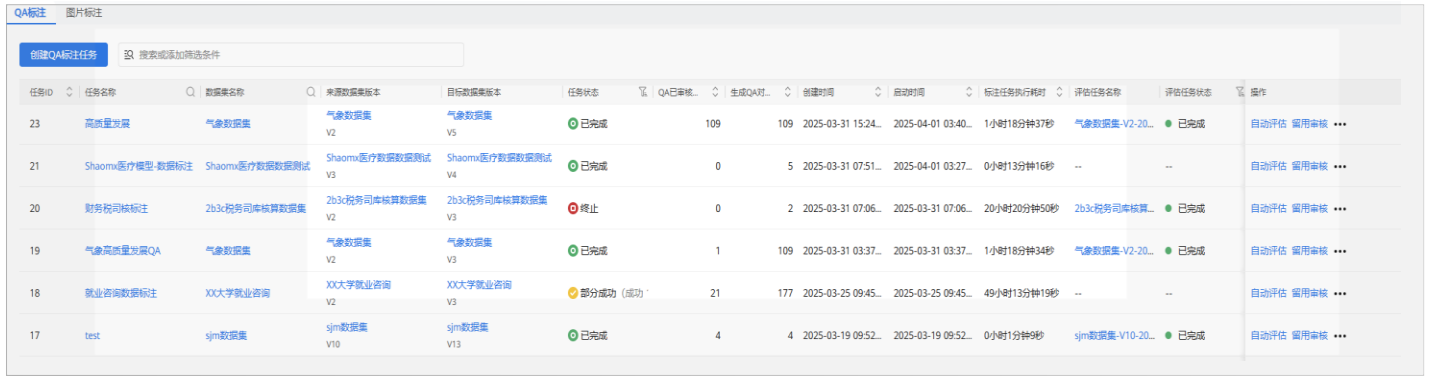
Figure 2: The pipeline of SELF-Instruct.

HUAWEI

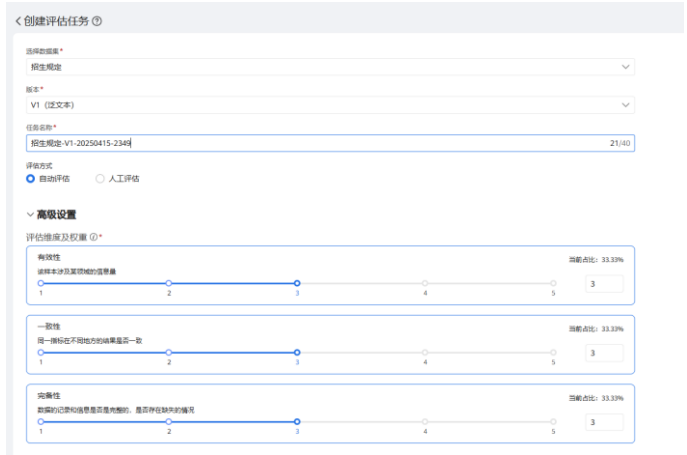
数据评估：自动评估结合专家审核使能垂域数据集快速构建

泛文本评估

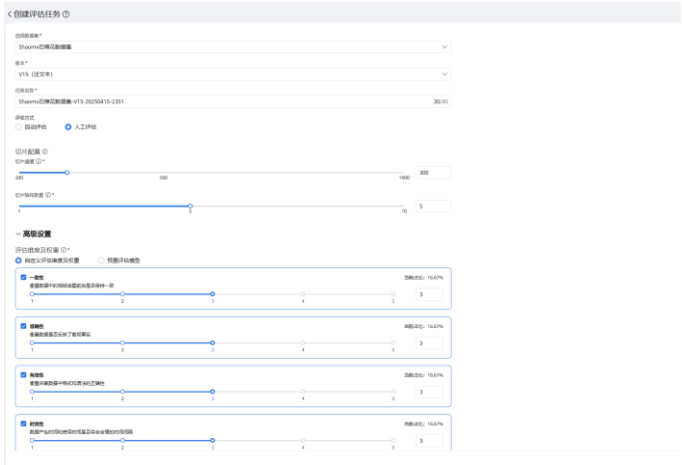
自动化评估算子对输入的文本分片处理，通过句法质量、逻辑质量、领域相关性和内容知识性综合评估文本质量，最终通过有效性、一致性、完备性3个维度体现文本质量状态



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

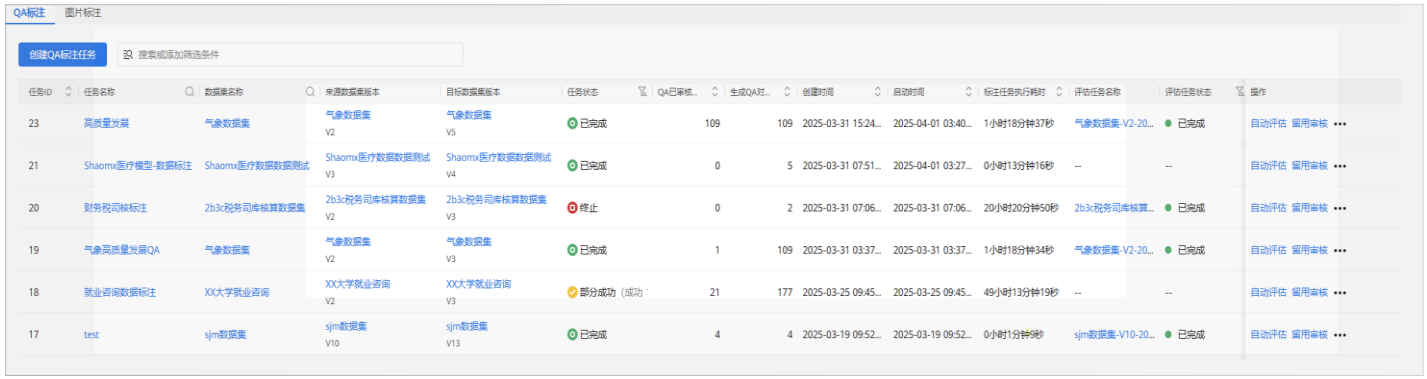
泛文本报告

HUAWEI

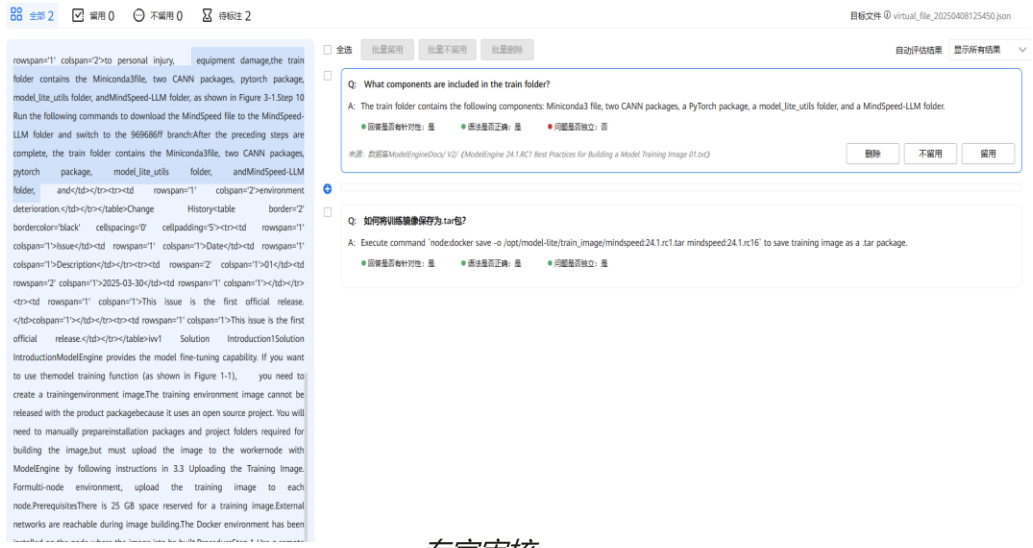
数据评估：自动评估结合专家审核使能垂域数据集快速构建

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。



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。



QA评估　　*QA评估报告*　　*专家审核*

HUAWEI

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

选择数据集

↓

获取纯文本数据

↓

文本切片
（文本切片算子）

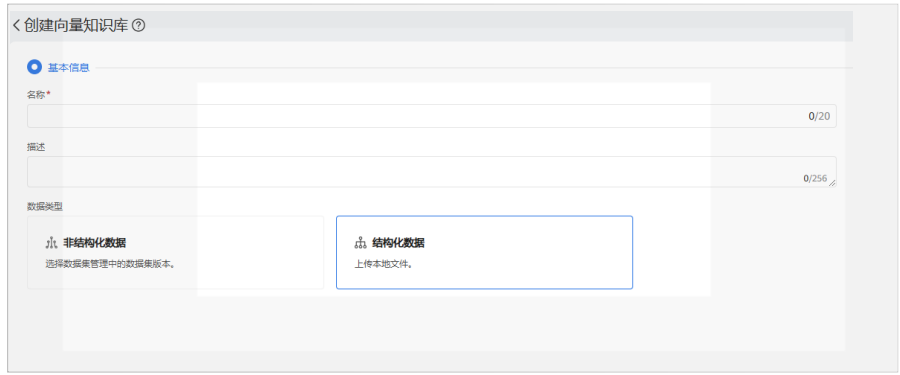
↓

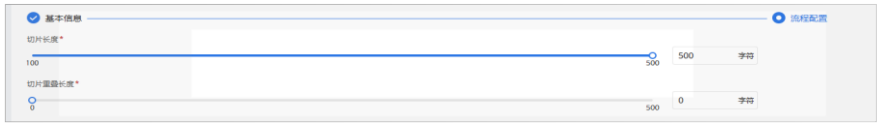
切片向量化
（Embedding 模型）

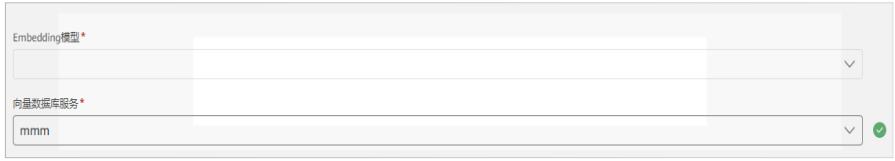
↓

向量化结果入库
（向量数据库）

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。







HUAWEI

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.

HUAWEI