

智算一体机-训推版

用户操作指南

天翼云科技有限公司



目 录

1 产品介绍.....	4
1.1产品定义.....	4
1.2产品优势.....	5
1.3功能特性.....	6
1.4应用场景.....	7
1.5术语解释.....	8
2 操作指南.....	10
2.1 概览.....	10
2.1 模型广场.....	10
2.1.1模型查看.....	10
2.1.2一键精调.....	11
2.1.3一键评估.....	12
2.1.4一键部署.....	13
2.1.5API调用.....	14
2.3 体验中心.....	14
2.4 模型定制.....	16
2.4.1 模型精调.....	16
2.4.2 开发机.....	19
2.4.3 训练任务.....	24
2.5 模型服务.....	26

2.5.1 服务接入.....	26
2.5.2 在线服务.....	30
2.5.3 调用监控.....	34
2.6 模型评估.....	35
2.6.1 评估数据准备.....	35
2.6.2 新建评估任务.....	35
2.7 智算资产.....	37
2.7.1 我的模型.....	37
2.7.2 我的数据集.....	41
2.7.3 我的镜像.....	47
2.7.4 我的代码包.....	48
2.8 管理中心.....	50
2.8.1 资源总览.....	50
2.8.2 调度中心.....	51
2.8.3 设置.....	52
2.9 运营后台.....	53
2.9.1 智算平台运营.....	53
2.9.2 账号管理.....	54
3 常见问题.....	57

1 产品介绍

1.1 产品定义

智算一体机内置轻量化智算服务平台，为大模型训练、推理、应用提供全栈工具链的智算服务平台，包含数据管理、模型开发与训练、模型评估、模型管理、服务部署等模块。提供一站式、可视化、全流程的训推用工具链，为用户提供AI建模的一站式解决方案。预置丰富的基座大模型和数据集，支持国产化等异构算力，提供算子加速与模型加速，极大提升大模型训练推理效率。

【功能模块】

- 模型广场：引入市场主流开源模型，可了解平台预置模型的介绍（含使用场景、版本列表等）、API 文档、任务记录等内容。
- 体验中心：通过交互式界面方式与模型对话，可以选择模型类型，设定模型参数和系统人设，体验不同模型的功能和性能。
- 模型服务：将模型部署为在线服务，通过 API 调用服务模型，供应用方使用，并且可以查看模型调用量。
- 我的数据集：将训练模型所需要的各种数据，导入到数据集管理中，以便于更清晰、方便地管理训练数据，加快训练速度。支持数据集共享，在线标注等。
- 模型开发管理：使用多种方式设计模型和训练，启动训练任务并为训练任务分配算力资源。
- 训练任务管理：查看和管理启动的所有训练任务。从已完成的训练中，挑选满意的训练结果发布为模型。

- 模型精调：基于平台的基础大模型，选择训练数据集和算力即可快速启动精调任务。支持 SFT 微调训练。
- 模型管理：导入和管理所有模型，对模型进行版本管理、导入导出、分享。
- 模型评估：对训练完成的模型进行评测和分析，评估模型的准确性、泛化能力等关键指标。

【功能特性】

- 简化训练和部署的复杂流程
- 开箱即用，降低调优成本
- 平台化全流程管理

1.2 产品优势

- 全流程开发工具

提供训练数据管理、模型开发（代码式开发工具、快速微调、预置大模型、预置开发环境）、模型训练、模型管理、服务部署、服务管理到模型服务调用的全链路功能。集成分布式训练调度技术、多种训练加速方法和高性能存储，支撑大模型训练，并极大降低训练和应用模型成本、缩短训练时长。

- 兼顾各类用户需求

面向需要开发复杂模型的用户，提供完整的代码式开发工具、预置大模型、预置开发环境等，满足用户的各种复杂模型开发需求。面向希望能快速、便捷建模的用户，则充分利用大模型微调训练的特点，提供快速微调工具，只需选择数据、配置参数即可完成大模型微调，降低大模型训练的使用门槛。

- 部署快捷，适配广泛

集成分布式算力调度、模型并行推理和多种运算加速能力，提升模型推理性能，实现推理服务的快捷部署。同时，适配多种模型结构，灵活支持用户各类复杂推理应用需求。

- 集成多种 AI 框架

集成多种AI框架，包括国产AI框架，支持各种主流大模型。

- 安全可靠

符合数据监管要求，不设置数据埋点，不收集存储用户的入参和出参数据，从根本上保证了用户的数据隐私安全。

- 卓越的客户服务

31省本地化的销售网络体系，提供家门口的精细化客户服务。7*24小时的免费运维服务，全力保障客户业务稳定运行。

1.3 功能特性

- 简化训练和部署的复杂流程

在传统的AI模型研发流程中，科研人员需要经历一系列繁琐的环节，包括数据准备、模型构建、模型训练、模型评估、模型优化以及模型部署等。这些环节不仅涉及数据工程、模型框架、算法开发、模型加速等多个技术领域，还要求科研人员熟练使用数据治理工具、数据标注工具、数据管理工具、数据读取工具等一系列专业工具组件。同时，他们还需处理这些工具与硬件环境、操作系统环境的适配问题，以及管理众多的依赖环境包。这一复杂过程不仅耗时耗力，而且大大提高了模型研发的使用成本和复杂程度。

智算服务平台通过整合全链路的工具组件，实现了训练与部署流程的极大简化，为科研人员提供了一站式解决方案。用户无需再为繁杂的工具和环境配置而烦恼，只需专注于模型的核心研发工作。智算开发平台不仅降低了大模型开发的使用门槛，更让AI技术的普及和应

用变得更加便捷和高效。

- 开箱即用，降低调优成本

大模型场景下训练数据处理和使用的过程尤为复杂。硬件层面，需确保编译环境、框架工具、依赖资源包等与硬件完美适配。软件层面，需保障操作系统、深度学习框架、编译器等软件工具的顺畅运行。针对大模型的训练和调优更是加剧了整个过程的复杂程度，同时伴随着大量的时间和算力资源的消耗。传统训练调优工具往往无法满足要求。

智算服务平台为用户带来了便利，通过平台，用户无需进行任何额外的配置或调试，开箱即用。平台预置了丰富的预训练模型和镜像环境，针对不同场景提供了多样化预置数据集，确保用户能够迅速投入工作。同时，平台集成了大模型微调训练工具，适用于专属大模型的快速训练。此外，平台还支持分布式训练和Deepspeed加速框架，提供断点续训功能，支持小样本微调，使用户能够轻松定制专属模型，极大地降低了调优成本，提高了研发效率。

- 平台化全流程管理

AI训练的高效执行，依赖于大数据团队、数据标注团队、算法开发团队、性能优化团队以及算法工程化团队等多个专业角色的紧密协作。

智算服务平台，一个集成化的平台化工具，将以上所有角色都汇聚于一个统一的平台之上，提供从数据处理、模型开发、模型训练到最终模型部署应用的全栈服务。

管理者能够在平台上实现统一管理和查看，确保各环节的无缝衔接，让各角色参与者能借助平台完美协同工作，实现数据互通、环境互通，确保数据和模型安全，全程不出平台实现训练开发资产的一站式沉淀与管理，能显著提升企业整体工作效率，实现AI生产的流水线化运作。

1.4 应用场景

- 模型训练

向下纳管智算硬件资源，提供技术运维及训练加速。向上通过模型开发平台提供大模型训练全链路功能，简化操作，提升效率。封装训练所需的底层技术，缩小训练者所需掌握的技术范围，降低大模型开发技术门槛。

主要用户包括各基础大模型厂商，各种拥有行业和场景专业知识与数据的行业客户，如科研院所、大专院校和教育机构、政府、金融机构、工业企业、科技单位、医院等。

- 模型推理

向下纳管智算硬件资源，提供技术运维服务及推理加速。向上通过模型服务平台提供部署好的模型服务，并集成丰富配套工具，提供模型推理一站式部署服务。

主要用户包括各种软件开发商，特别是行业软件开发商，以及科研院所、大专院校和教育机构、政府、金融机构、工业企业、科技单位、医院等行业客户。

1.5 术语解释

- 预置模型

是指平台提供的原始模型，您可以通过选择预置模型进行训练从而得到行业或细分场景模型，不同的基础模型的参数和能力不同，我们将持续推出不同能力方向的模型。

- 模型微调

是指利用预先训练好的神经网络模型，并针对特定任务在相对少量量的监督数据上进行重新训练的技术。这种方法能够充分利用预训练模型在大型数据集上学到的通用特征和知识，从而加速在新任务上的训练过程，并通常能够取得较好的性能表现。

- 迭代轮次

是指模型训练过程中模型学习数据集的次数，可理解为学习几遍数据，可依据需求进行调整

。

- 批处理大小

是指在模型训练过程中，每次处理的数据样本的数量，可理解为模型每看多少数据即更新一次模型参数，在选择批处理大小时需要综合考虑各种因素。

- 学习率

是指更新模型参数的系数，它决定了在每次迭代中，模型参数应该沿着梯度下降的方向更新多少，需要根据具体情况来仔细选择和调整学习率。

- 节点

节点是集群的组成单元，每个节点对应一台物理机。

- 数据集

是机器学习或深度学习模型训练过程中的重要组成部分。数据集是一组已知输入和对应输出的数据，用于训练模型以学习从输入到输出的映射关系。构建合适数据集，通过模型调优可增强模型能力，提升预测效果。

2.1 概览

可点击【去选大模型】可以去到模型广场；

支持在概览页快速查看大模型调用数据统计。



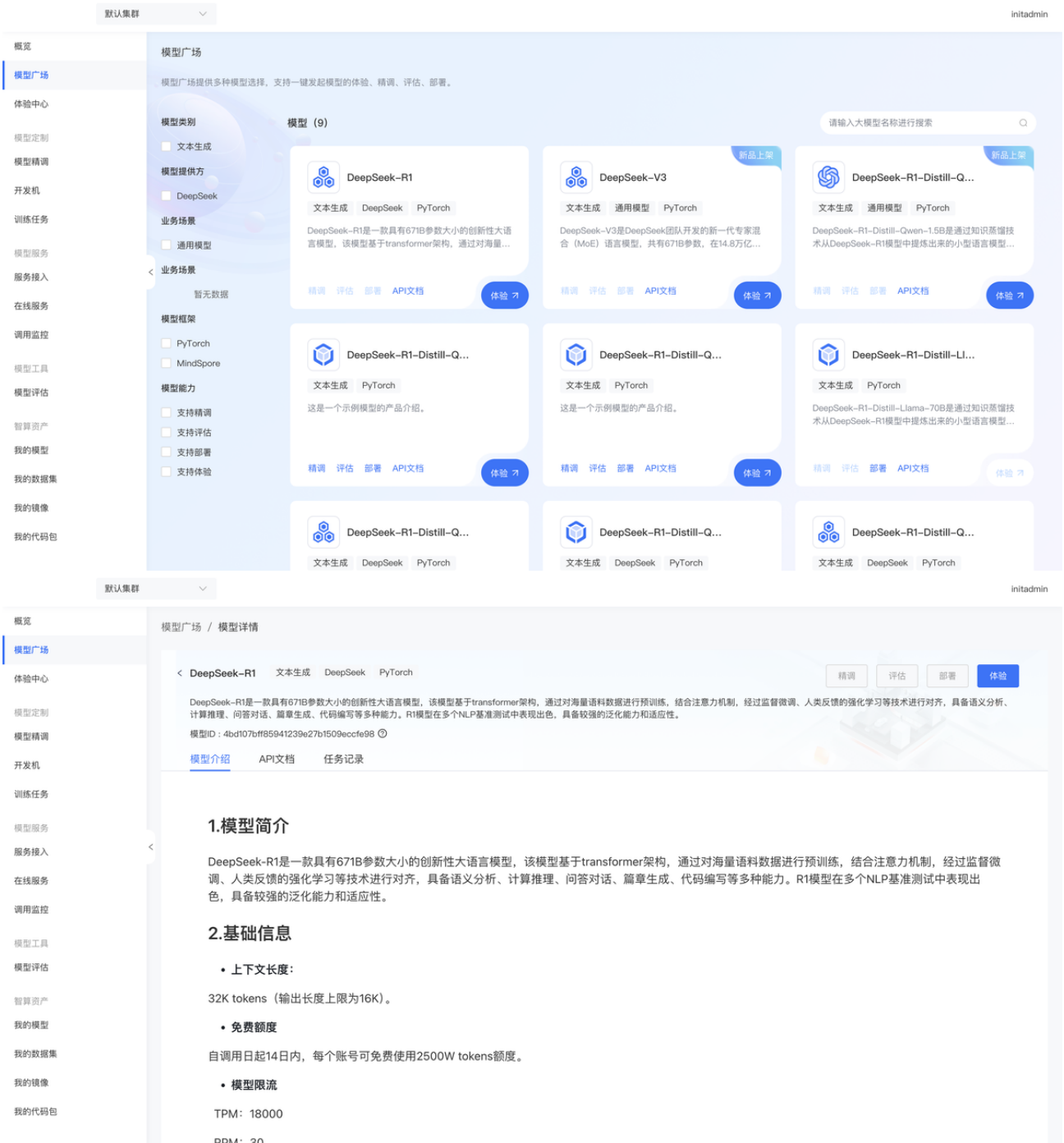
2.1 模型广场

2.1.1 模型查看

进入模型广场模块，可查看到所有预置的模型，并根据模型类别、模型提供方、模型框架进行筛选。



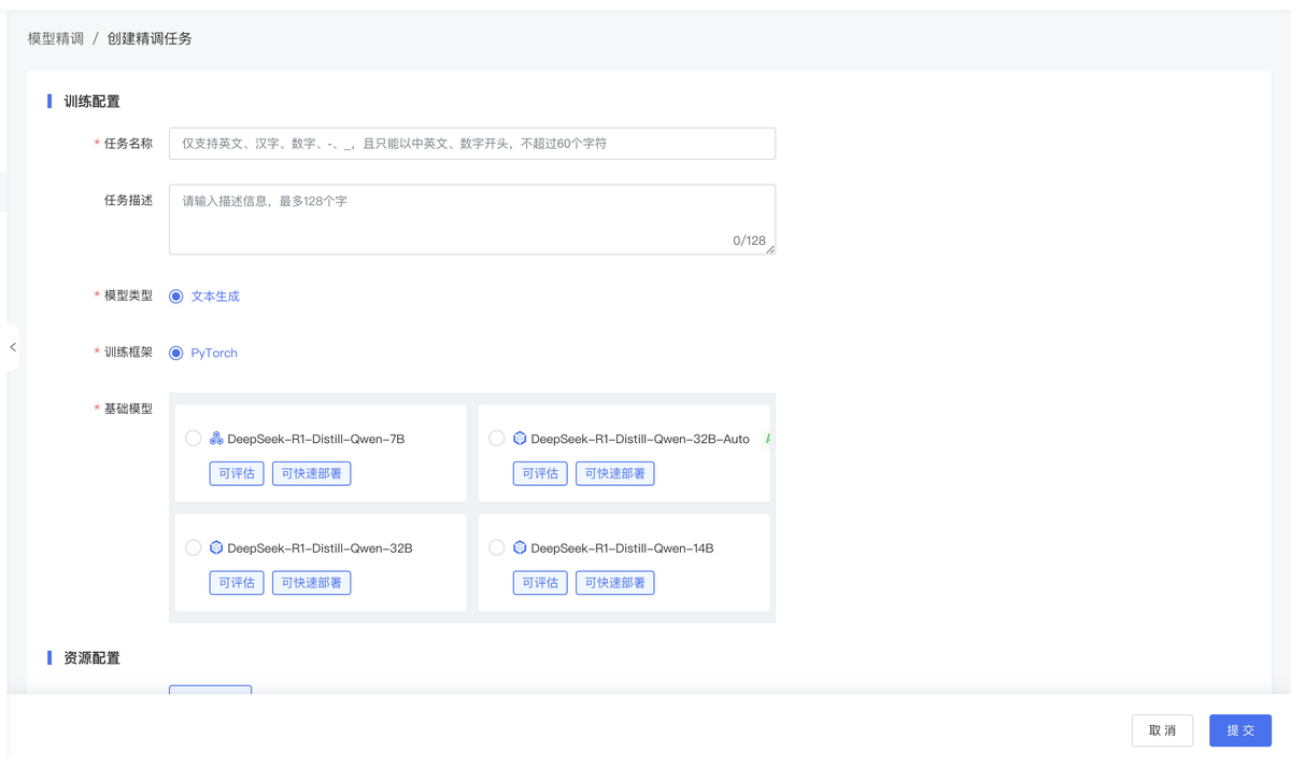
点击【模型卡片】，查看平台预置模型的模型简介、模型ID（调用模型API时需要）、模型应用场景、模型评测效果以及API文档等内容。



备注：截图仅供参考，可能与实际界面功能不匹配，详询解决方案，下文同理，不在备注

2.1.2 一键精调

支持对平台预置的模型进行一键精调，可点击模型卡片或模型详情页上的【精调】按钮直接发起精调。如果按钮为置灰不可点击状态，说明模型暂不支持该项功能。可选择训练方式、训练数据、训练配置以及资源配置。



2.1.3 一键评估

支持对平台预置的模型进行一键评估，可点击模型卡片或模型详情页上的【评估】按钮直接发起评估。如果按钮为置灰不可点击状态，说明模型暂不支持该项功能。需填写评估任务的基本信息，评估配置（包括评估数据集、评估标准）、资源部署信息。

| 评估配置

* 待评估模型

序号	模型	模型版本描述	操作
1	<input type="text" value="请输入或选择模型"/>	-	删除

[+添加模型](#)

* 评估数据集

* 评估标准

<input checked="" type="checkbox"/>	效果指标	指标定义
<input checked="" type="checkbox"/>	准确率	忽略停用词后，正确预测(标注与预测完全匹配)的样本数与总样本数的比例
<input checked="" type="checkbox"/>	ROUGE-1	忽略停用词后，将模型生成的结果和标准结果按unigram拆分后，计算出的召回率
<input checked="" type="checkbox"/>	ROUGE-2	忽略停用词后，将模型生成的结果和标准结果按bigram拆分后，计算出的召回率
<input checked="" type="checkbox"/>	ROUGE-L	忽略停用词后，衡量了模型生成的结果和标准结果的最长公共子序列，并计算出召回率
<input checked="" type="checkbox"/>	BLEU-4	忽略停用词后，用于评估模型生成的句子和实际句子的差异的指标，值为unigram, bigram, trigram, 4-grams的加权平均

[取消](#)[开始评估](#)

2.1.4 一键部署

支持对平台预置的模型进行一键部署，可点击模型卡片或模型详情页上的【部署】按钮直接发起部署。如果按钮为置灰不可点击状态，说明模型暂不支持该项功能。需填写模型服务信息、部署资源信息，部署后，可通过API的方式调用模型服务。

在线服务 / 部署模型

模型服务信息

* 服务名称

* 服务地址 <https://ai.ctyun.cn:50445> [调用说明](#)

模型信息

* 模型选择 DeepSeek-R1-Distill-Qwen-14B

* 训练框架 PyTorch MindSpore

资源部署信息

* 队列

* 资源规格

* 实例数量

2.1.5 API调用

支持通过API调用模型广场预置模型的推理服务，详情操作请参考2.5【模型服务】相关内容。

2.3 体验中心

进入体验中心工作台，选择服务类型，左下方支持通过测试台选择服务/应用进行参数配置。



可以在左侧测试工作台选择服务进行参数配置：

温度：Temperature控制生成文本的多样性。较高的温度值会使生成的文本更加随机和多样化，而较低的温度值会使生成的文本更加确定和一致。

多样性：TopP影响输出文本的多样性，取值越大，生成文本的多样性越强。

重复惩罚：Frequency_penalty影响模型生成重复词汇的倾向。通过增加重复词汇的惩罚权重，降低模型逐字重复的可能性。

系统人设：设定模型的行为和背景，告知模型需要扮演的角色。例如：“假如你是一个AI助手”。输入框中可直接输入问题，系统将根据输入的问题及配置的参数进行实时回答。

查看历史记录

点击右上角【查看历史记录】，系统会展开历史的对话记录，可以查看统计大模型的回答质量，保存上限为200条。

2.4 模型定制

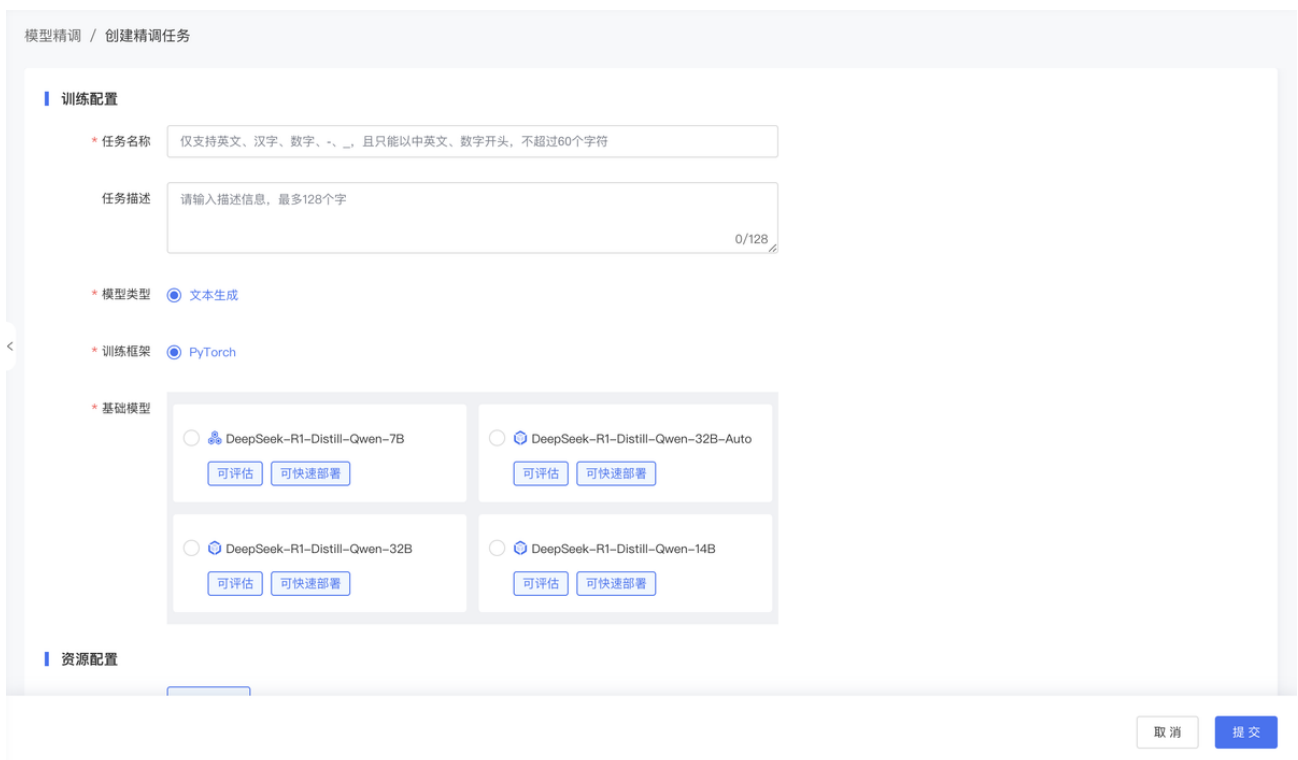
2.4.1 模型精调

0.整体流程

我的数据集-创建所需格式数据集——模型精调-创建精调任务——任务列表-发布模型——我的模型-部署为在线服务——服务接入-创建服务组，获取app key

1.创建精调任务

- 进入模型精调管理模块，选择快速精调，进入调优任务列表，点击【新建调优任务】，进入创建页面。



- 选择已导入的数据集，选择基础大模型，设置调参方式、迭代轮次、批处理大小、学习率等指标，配置资源选择算力规格。单节点下卡数越多训练越快。

模型精调 / 创建精调任务

数据配置

* 训练数据集 指令微调

请选择数据集 [前往导入数据集](#)

验证数据集

- 5 + %

取值范围为[0-20]。默认取值为5，即验证集占比为5%，训练集占比95%。验证集会参与计费。

资源配置

* 集群

* 队列

* 算力申请

* 资源规格

* 节点数量

2. 监控调优任务

- 返回模型调优训练任务列表，列表中可以看到每个任务的运行进度、预估时长。

模型精调

运行中的精调任务数 使用中显卡

任务名称	运行状态	运行进度	运行时长	预估时长	占用资源	基础模型	训练数据	上次启动时间	操作
312-ds-32b -auto	● 排队中	0%	0分0秒	预估中	算力单元:104C_760G_8*NPU*1 算力单元数:2	DeepSeek -R1-Distill -Qwen-3 2B-Auto	吴越红-指 令微调		日志 监控 发布模型 停止 更多
312-ds-14b -auto	● 运行完成	100%	36分46秒	46分16秒	算力单元:104C_760G_8*NPU*1 算力单元数:2	DeepSeek -R1-Distill -Qwen-14 B	吴越红-指 令微调	2025-03-12 15	日志 监控 发布模型 重启 更多
test	● 停止	0%	0分0秒	预估中	算力单元:104C_760G_8*NPU*1 算力单元数:2	DeepSeek -R1-Distill -Qwen-3 2B-Auto	吴越红-指 令微调		日志 监控 发布模型 重启 更多

共34条

- 点击任务名称，可进入调优任务详情页，右上角可对任务进行停止和删除操作。

模型精调 / 模型调优详情

< 312-ds-32b-auto (ID: 1899732677123715072) 停止 删除

运行记录 🔍

1899732677371179009

- 排队中
2025-03-12 16:02:20

[详情](#) [事件](#) [日志](#) [监控](#) [评估](#)

基础信息

运行状态	排队中	基础模型	DeepSeek-R1-Distill-Qwen-32B-Auto(PyTorch)
运行时长	0分0秒	上次启动时间	
调参方式	LoRa	创建时间	2025-03-12 16:02:20
资源规格	算力单元:104C_760G_8*NPU*1 算力单元数:2	超参数配置	迭代次数:1 全局批处理大小:64 微批处理大小:1 学习率:0.000050
资源组	共享资源组: 默认集群	队列	default

数据配置

训练数据集 吴越红-指令微调

验证数据集 5%

节点信息

请选择运行状态 查询 重置

名称	类型	状态	创建时间	启动时间	结束时间	执行时长	操作
----	----	----	------	------	------	------	----

- 详情页可依次查看任务基础信息、日志、监控、Tensorboard 看板。监控中运行进度可以看到每一次迭代是否完成，资源监控看板可以查看 CPU 使用率、内存使用率、NPU 使用率等。

模型精调 / 模型调优详情

< 312-ds-32b-auto (ID: 1899732677123715072) 停止 删除

运行记录 详情 事件 日志 监控 评估

1899732677371179009

● 排队中

2025-03-12 16:02:20

环境日志 数据处理日志

```

15 | Unschedulable, 详细信息: 0/9 nodes are unavailable: 9 Insufficient cpu.
    | [2025-03-12 16:37:52.867] [trace_id=] [spanId=] [530-pool-5-thread-32] [ FullLinkLogService.java:165] [INFO ] - []
    | traceId: FINETUNE_1899732677123715072_1899732677371179009 podGroup status: Inqueue, condition type: Unschedulable, 原因:
    | 0/9 nodes are unavailable: 9 Insufficient cpu., 详细信息: 0/9 nodes are unavailable: 9 Insufficient cpu.
16 | [2025-03-12 16:47:52.861] [trace_id=] [spanId=] [ pool-1-thread-87] [ FullLinkLogService.java:126] [INFO ] - []
    | traceId: FINETUNE_1899732677123715072_1899732677371179009 pod status: Pending, condition type: PodScheduled, 原因:
    | Unschedulable, 详细信息: 0/9 nodes are unavailable: 9 Insufficient cpu.
17 | [2025-03-12 16:47:52.868] [trace_id=] [spanId=] [530-pool-5-thread-45] [ FullLinkLogService.java:165] [INFO ] - []
    | traceId: FINETUNE_1899732677123715072_1899732677371179009 podGroup status: Inqueue, condition type: Unschedulable, 原因:
    | 0/9 nodes are unavailable: 9 Insufficient cpu., 详细信息: 0/9 nodes are unavailable: 9 Insufficient cpu.
18 | [2025-03-12 16:48:17.805] [trace_id=] [spanId=] [4980-pool-5-thread-7] [ FullLinkLogService.java:165] [INFO ] - []
    | traceId: FINETUNE_1899732677123715072_1899732677371179009 podGroup status: Inqueue, condition type: Unschedulable, 原因:
    | 0/9 nodes are unavailable: 9 Insufficient cpu., 详细信息: 0/9 nodes are unavailable: 9 Insufficient cpu.
19 | [2025-03-12 16:48:17.816] [trace_id=] [spanId=] [ pool-1-thread-4] [ FullLinkLogService.java:126] [INFO ] - []
    | traceId: FINETUNE_1899732677123715072_1899732677371179009 pod status: Pending, condition type: PodScheduled, 原因:
    | Unschedulable, 详细信息: 0/9 nodes are unavailable: 9 Insufficient cpu.
20 | [2025-03-12 16:49:57.621] [trace_id=] [spanId=] [ pool-1-thread-7] [ FullLinkLogService.java:126] [INFO ] - []
    | traceId: FINETUNE_1899732677123715072_1899732677371179009 pod status: Pending, condition type: PodScheduled, 原因:
    | Unschedulable, 详细信息: 0/9 nodes are unavailable: 1 Insufficient huawei.com/Ascend910, 1 hb-huabei2-npu-10.250.0.29's
    | huawei.com/Ascend910 not statble:0=>8, 7 Insufficient cpu.
21 | [2025-03-12 16:49:57.656] [trace_id=] [spanId=] [7545-pool-5-thread-2] [ FullLinkLogService.java:165] [INFO ] - []
    | traceId: FINETUNE_1899732677123715072_1899732677371179009 podGroup status: Inqueue, condition type: Unschedulable, 原因:
    | 0/9 nodes are unavailable: 1 Insufficient huawei.com/Ascend910, 1 hb-huabei2-npu-10.250.0.29's huawei.com/Ascend910 not
    | statble:0=>8, 7 Insufficient cpu., 详细信息: 0/9 nodes are unavailable: 1 Insufficient huawei.com/Ascend910, 1
    | hb-huabei2-npu-10.250.0.29's huawei.com/Ascend910 not statble:0=>8, 7 Insufficient cpu.
22
                
```

< 上一页 下一页 >

2.4.2 开发机

1.JupyterLab集成开发环境：

- 进入模型定制模块，选择模型开发机，点击【JupyterLab】>【创建

JupyterLab】填写相应字段，提交后操作列点【启动】，即可启动一个在线的

JupyterLab。

开发机 目录使用说明

运行中的JupyterLab实例 0

运行中的VSCode实例 0

使用中显卡 0

JupyterLab VSCode

创建JupyterLab 查询 重置

名称	创建者	状态	运行时长	镜像版本	资源类型	操作
自定义20250311092241988	initadmin	● 未启动	1秒	v1.20250311092241988	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
trt	initadmin	● 未启动	-	1.0	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
sadfasd	initadmin	● 未启动	11时4分14秒	v1.0.qlora.tf4.45.2	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
gbtest	initadmin	● 未启动	39秒	v1.0.qlora.tf4.45.2	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
zydtest	initadmin	● 未启动	0分0秒	v4.0	13C_95G_1*NPU	启动 停止 打开 开始训练 更多

共5条 1

开发机 / 创建JupyterLab

基本信息

* 名称

数据集

训练数据集 + 增加数据集 您代码中读取数据的相对路径需要改为本地挂载路径。

预置模型配置

模型来源 我的模型 预置模型

我的模型文件 + 增加我的模型文件 将模型管理中的模型文件挂载到容器本地路径，您的代码可以直接读取此相对路径。

环境配置

* 镜像来源 系统预置镜像 自定义镜像

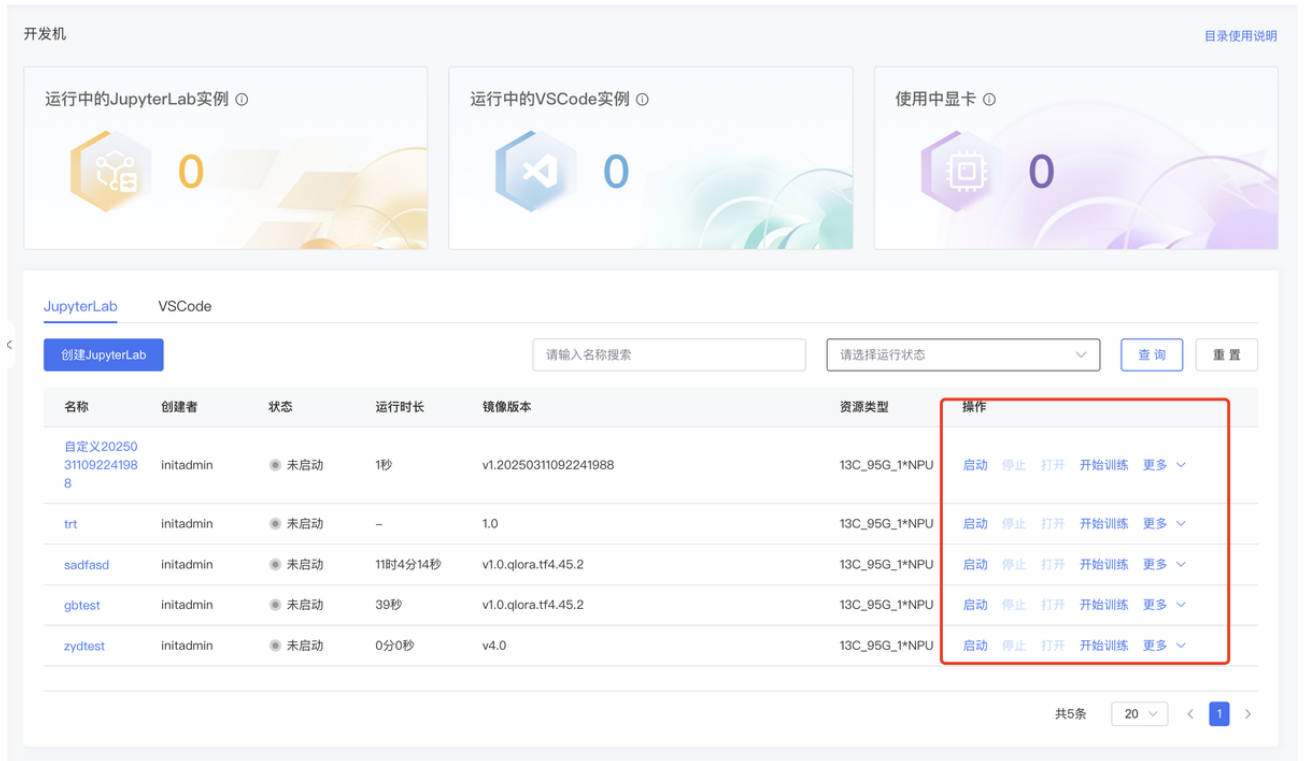
* 文件目录

取消 提交

- 等待启动成功，当创建的JupyterLab的状态显示【运行中】后即可点击操作列【打开】即可进入开发环境，进入即可在/work/mount下看到挂载的预置模型和数据集。

- 开发环境管理：点击右侧操作栏中的启动、停止、打开、开始训练、删除、保存

至数据集、保存至模型管理、保存至代码包等按钮，实现对开发环境的管理。



The screenshot displays a management interface for development environments. At the top, there are three summary cards: '运行中的JupyterLab实例' (0), '运行中的VSCode实例' (0), and '使用中显卡' (0). Below these is a table of instances. The table has columns for Name, Creator, Status, Runtime, Image Version, Resource Type, and Actions. A red box highlights the '操作' (Actions) column for the first five rows, which include '启动' (Start), '停止' (Stop), '打开' (Open), '开始训练' (Start Training), and '更多' (More).

名称	创建者	状态	运行时长	镜像版本	资源类型	操作
自定义20250311092241988	initadmin	未启动	1秒	v1.20250311092241988	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
trt	initadmin	未启动	-	1.0	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
sadfasd	initadmin	未启动	11时4分14秒	v1.0.qiora.tf4.45.2	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
gbtest	initadmin	未启动	39秒	v1.0.qiora.tf4.45.2	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
zydtest	initadmin	未启动	0分0秒	v4.0	13C_95G_1*NPU	启动 停止 打开 开始训练 更多

2. VSCode集成开发环境

- 进入模型定制模块，选择开发机，点击【VSCode】>【创建VSCode】，填写相应字段，提交后操作列点【启动】，即可启动一个在线的VSCode。

创建完成后，可以查看到该环境的名称、创建者、状态、运行时长、镜像版本、资源类型等信息。

开发机 目录使用说明

运行中的JupyterLab实例 0

运行中的VSCode实例 0

使用中显卡 0

JupyterLab VSCode

创建VSCode

请选择运行状态

名称	创建者	状态	运行时长	镜像版本	资源类型	操作
hys-create-ide-13-56-01	initadmin	● 未启动	39秒	v5.0	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
vs20250311092241988	initadmin	● 未启动	1分5秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
vs20250311091754030	initadmin	● 未启动	7分56秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
vs20250310223716086	initadmin	⊙ 启动失败	0分0秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
ds-7h-auto	initadmin	● 未启动	44分9秒	1.0	52C_380G_4*NP	启动 停止 打开 开始训练 更多

共11条 20 1

开发机 / 创建VSCode

基本信息

* 名称

数据集

训练数据集 [+ 增加数据集](#) 您代码中读取数据的相对路径需要改为本地挂载路径。

预置模型配置

模型来源 我的模型 预置模型

我的模型文件 [+ 增加我的模型文件](#) 将模型管理中的模型文件挂载到容器本地路径，您的代码可以直接读取此相对路径。

环境配置

* 镜像来源 系统预置镜像 自定义镜像

* 文件目录

- 等待启动成功，当创建的JupyterLab的状态显示【运行中】后即可点击操作列【打开】即可进入开发环境，进入即可在/work/mount下看到挂载的预置模型和数据集。
- 开发环境管理：点击右侧操作栏中的启动、停止、打开、开始训练、删除、保存



至数据集、保存至模型管理、保存至代码包等按钮，实现对开发环境的管理。

名称	创建者	状态	运行时长	镜像版本	资源类型	操作
hys-create-ide-13-56-01	initadmin	未启动	39秒	v5.0	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
vs20250311092241988	initadmin	未启动	1分5秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
vs20250311091754030	initadmin	未启动	7分56秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
vs20250310223716086	initadmin	启动失败	0分0秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多
...

3.开发训练任务管理

(1) 配置训练任务：进入模型定制模块，选择【开发机】，进入 JupyterLab 或 VSCode 列表，点击操作列【开始训练】，进入分布式训练配置页面，输入启动命令>选择镜像框架>配置算力资源，即可启动训练。

说明：

算法配置：

打开代码所在的文件路径：`cd /work/代码所在路径`

启动命令

如果是sh文件，启动命令写：`sh 文件名.sh`

如果是py文件，启动命令写：`python 文件名.py`

资源配置：

如果镜像中有deepspeed，需要开启deepspeed

单节点【184C|1520G|8*910B-64G（液冷）或以上】指：184核CPU，1520G内存，单节点8张910B的NPU卡，每张卡的显存大小为64GB。。

节点指物理机数量，Master+Worker的数量为多机多卡总节点数，等于训练脚本中指定的WORKER_CNT的数量，算力规格中GPU的数量等于训练脚本中指定的GPUS_PER_NODE的数量。



(2) 命令行启动训练任务*

执行安装命令：pip install

/mnt/public/job_submit/task_submission-2.0-py3-none-any.whl

说明：

使用样例参考：/mnt/public/job_submit目录下的样例文件[submit.sh](#)

关键命令：

提交pytorch任务的帮助说明：submit pytorch --help

提交其他任务的帮助说明：submit --help

(3) 监控训练任务

在任务列表，可以看到训练任务的状态，可对训练任务进行启动、停止等。

名称	创建者	状态	运行时长	镜像版本	资源类型	操作
hys-create-ide-13-56-01	initadmin	● 未启动	39秒	v5.0	13C_95G_1*NPU	启动 停止 打开 开始训练 更多 ▾
vs20250311092241988	initadmin	● 未启动	1分5秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多 ▾
vs20250311091754030	initadmin	● 未启动	7分56秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多 ▾
vs20250310223716086	initadmin	⊗ 启动失败	0分0秒	v2.0-jupyterlab-vscode	13C_95G_1*NPU	启动 停止 打开 开始训练 更多 ▾
de-7h-auto	initadmin	● 未启动	44分9秒	1.0	52C_380G_4*NP	启动 停止 打开 开始训练 更多 ▾

共11条 20 < 1 >

在训练任务列表的操作列，点击【任务名称】进入详情页，可以查看到运行的事件，支持搜索。

开发机 / 查看详情

< 自定义20250311092241988

事件

```

1 | [2025-03-11 09:25:15.949] [trace_id=] [spanId=] [http-nio-8082-exec-9] [ KubeService.java:380] [INFO ] - [1b843a189b479b8018327732a57449e6]
  | traceId: IDE_1899270358782107648_ALL, 组件: harbor, 获取IDE镜像路径: harbor-hbtest.ctyun.cn:30443/812bc666d9f14e8986668eb0998d5c82/812bc666d9f14e89/
  | dxscsc20250311092241988:v1.20250311092241988, 状态: 成功
2 | [2025-03-11 09:25:15.952] [trace_id=] [spanId=] [http-nio-8082-exec-9] [ KubeService.java:419] [INFO ] - [1b843a189b479b8018327732a57449e6]
  | traceId: IDE_1899270358782107648_ALL, 启动IDE, 开始创建工作Space, queueName: default, ide类型: JUPYTERLAB
3 | [2025-03-11 09:25:16.988] [trace_id=] [spanId=] [http-nio-8082-exec-10] [ KubeService.java:380] [INFO ] - [372e295551bb4720b3bbe2af1812c218]
  | traceId: IDE_1899270358782107648_ALL, 组件: harbor, 获取IDE镜像路径: harbor-hbtest.ctyun.cn:30443/812bc666d9f14e8986668eb0998d5c82/812bc666d9f14e89/
  | dxscsc20250311092241988:v1.20250311092241988, 状态: 成功
4 | [2025-03-11 09:25:16.991] [trace_id=] [spanId=] [http-nio-8082-exec-10] [ KubeService.java:419] [INFO ] - [372e295551bb4720b3bbe2af1812c218]
  | traceId: IDE_1899270358782107648_ALL, 启动IDE, 开始创建工作Space, queueName: default, ide类型: JUPYTERLAB
5 | [2025-03-11 09:25:37.900] [trace_id=] [spanId=] [http-nio-8082-exec-7] [ IDEService.java:412] [INFO ] - [602d0107775dc19eafb163aba679570f]
  | traceId: IDE_1899270358782107648_ALL, 开始停止IDE
6 | [2025-03-11 09:25:37.965] [trace_id=] [spanId=] [http-nio-8082-exec-7] [ KubeService.java:613] [INFO ] - [602d0107775dc19eafb163aba679570f]
  | traceId: IDE_1899270358782107648_ALL, 开始删除IDE, name: jupyterlab-1899270358782107648-1899270362895204353, type: JUPYTERLAB
7 | [2025-03-11 09:25:37.980] [trace_id=] [spanId=] [http-nio-8082-exec-7] [ KubeService.java:640] [INFO ] - [602d0107775dc19eafb163aba679570f]
  | traceId: IDE_1899270358782107648_ALL, 删除IDE成功
8 | [2025-03-11 09:32:00.036] [trace_id=] [spanId=] [dPoolTaskScheduler-5] [ IDEService.java:1251] [INFO ] - [cad3d76c1e2148e28a787bf545d33067]
  | traceId: IDE_1899270358782107648_ALL, IDE停止中转为停止状态
9 | [2025-03-11 16:10:13.652] [trace_id=] [spanId=] [http-nio-8082-exec-10] [ IDEService.java:386] [INFO ] - [f410e9c958494551bfff33255c934b05]
  | traceId: IDE_1899270358782107648_189937274969722882, IDE创建时间: 2025-03-11 09:25:15
10 | [2025-03-11 16:10:13.653] [trace_id=] [spanId=] [http-nio-8082-exec-10] [ IDEService.java:389] [INFO ] - [f410e9c958494551bfff33255c934b05]
  | traceId: IDE_1899270358782107648_189937274969722882, 启动IDE, 开始施工.
11 | [2025-03-11 16:10:13.679] [trace_id=] [spanId=] [http-nio-8082-exec-10] [ KubeService.java:380] [INFO ] - [f410e9c958494551bfff33255c934b05]
  | traceId: IDE_1899270358782107648_ALL, 组件: harbor, 获取IDE镜像路径: harbor-hbtest.ctyun.cn:30443/812bc666d9f14e8986668eb0998d5c82/812bc666d9f14e89/
  | dxscsc20250311092241988:v1.20250311092241988, 状态: 成功
12 | [2025-03-11 16:10:13.682] [trace_id=] [spanId=] [http-nio-8082-exec-10] [ KubeService.java:419] [INFO ] - [f410e9c958494551bfff33255c934b05]
  | traceId: IDE_1899270358782107648_ALL, 启动IDE, 开始创建工作Space, queueName: default, ide类型: JUPYTERLAB

```

< 上一页 下一页 >

2.4.3 训练任务

进入【训练任务】模块，点击左上角的【新建任务】可创建一个新的训练任务。

训练任务
目录使用说明

运行中任务数 0

使用中显卡 0

新建任务

名称	训练框架	训练模式	占用资源	状态	运行时长	创建来源	开始时间	操作
QATest-train-task-15-10-36	PyTorch	单机训练	算力规格:13C_95G_1*NPU 节点个数:1个	运行完成	11秒	训练任务	2025-03-11 15:13:04	日志 复制 tensorboard 更多
DeepSeek-R1-Distill-Qwen-7B_单机模式-张亚丽_挂载预置模型_4	PyTorch	单机训练	算力规格:104C_760G_8*NPU 节点个数:1个	运行完成	15分20秒	训练任务	2025-03-11 11:42:00	日志 复制 tensorboard 更多
QATest-train-task-10-26-58	PyTorch	单机训练	算力规格:13C_95G_1*NPU 节点个数:1个	运行完成	12秒	训练任务	2025-03-11 10:27:03	日志 复制 tensorboard 更多

共65条 20 1 2 3 4



进入创建任务页面，填写任务名称、描述基本信息，数据集配置、预置模型配置、环境配置以及资源配置等信息。配置完成后点击提交，即可创建。

训练任务 / 创建训练任务

基本信息

* 任务名称

描述 0/128

数据集配置

训练数据集 + 增加数据集 您代码中读取数据的相对路径需要改为本地挂载路径。

预置模型配置

模型来源 我的模型 预置模型

我的模型文件 + 增加我的模型文件 将模型管理中的模型文件挂载到容器本地路径，您的代码可以直接读取此相对路径。

环境配置

* 文件目录 专属高速缓存盘

挂载路径: /work/home

在任务列表，可以看到训练任务的状态、日志，可对训练任务进行启动、停止等。

训练任务 目录使用说明

运行中任务数 0

使用中显卡 0

新建任务 查询 重置

名称	训练框架	训练模式	占用资源	状态	运行时长	创建来源	开始时间	操作
QAtest-train-task-15-10-36	PyTorch	单机训练	算力规格:13C_95G_1*NPU 节点个数:1个	✔ 运行完成	11秒	训练任务	2025-03-12 15:13:04	日志 复制 tensorboard 更多
DeepSeek-R1-Distill-Qwen-7B_单机模式-张亚丽_挂载预置模型_4	PyTorch	单机训练	算力规格:104C_760G_8*NPU 节点个数:1个	✔ 运行完成	15分20秒	训练任务	2025-03-12 11:42:00	日志 复制 tensorboard 更多
QAtest-train-task-10-26-58	PyTorch	单机训练	算力规格:13C_95G_1*NPU 节点个数:1个	✔ 运行完成	12秒	训练任务	2025-03-12 10:27:03	日志 复制 tensorboard 更多

共65条 < 1 2 3 4 >

在【训练任务】列表的操作列，点击【日志】进入详情页，日志tab可以查看到运行的日志，支持搜索。监控Tab可查看硬件使用率。

2.5 模型服务

整体流程：

在线服务（预置服务/部署我的服务）——服务接入-创建服务组。

可直接将预置服务中的模型，或部署我的模型后，进行服务接入，创建服务组，获取appkey，以供调用。

2.5.1 服务接入

服务接入可以将预置服务及我的服务（自定义部署的模型服务）打包为服务组，生成AppKey供用户调用。



【创建服务】

●进入模型服务模块，点击【服务接入】，服务接入模块可以将预置服务及我的服务打包为服务组，生成AppKey供用户调用。

服务接入

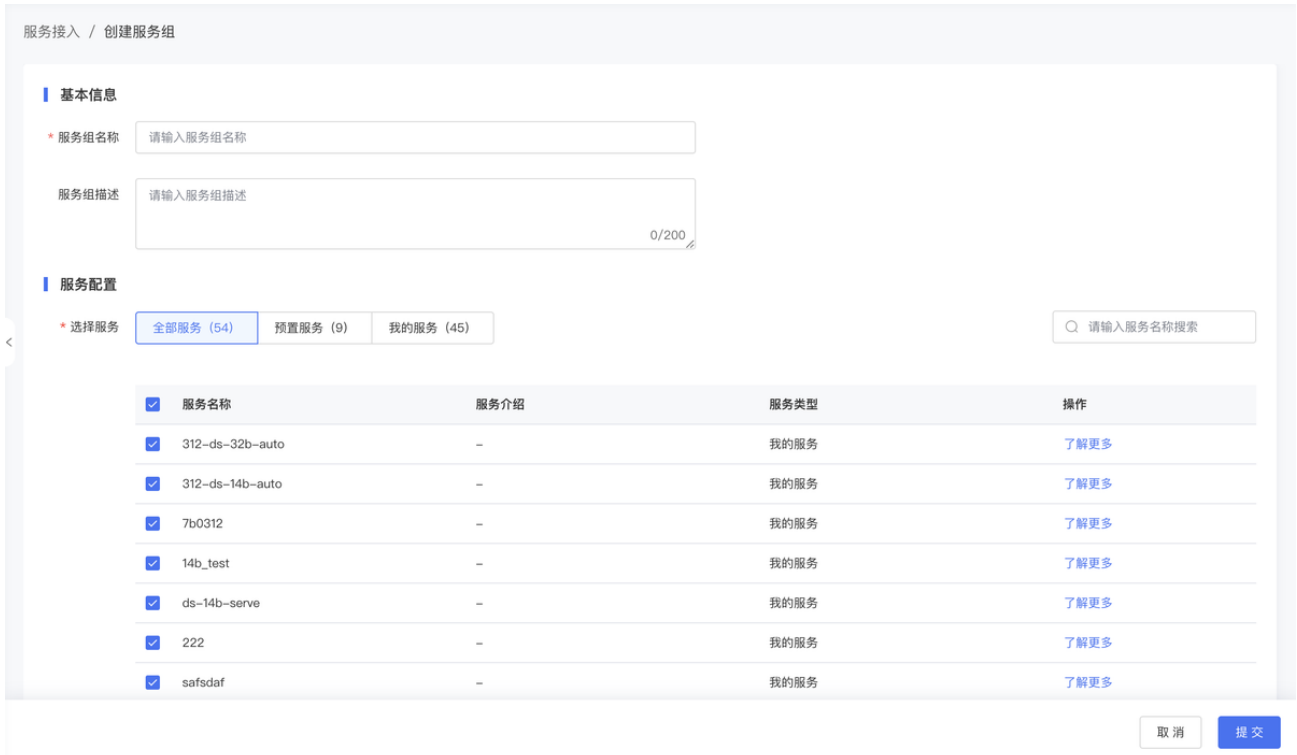
操作指引 收起

- 创建服务组**
接入服务，获得App Key
[去创建](#)
- 调用服务**
通过API调用服务模型
[去查看](#)
- 用量监控**
查看模型调用量
[去查看](#)

+ 创建服务组 将预置服务及我的服务打包为服务组，生成AppKey供用户调用。 [查询](#) [重置](#)

auto_7188 ● 使用中 创建时间: 2025-03-10 18:00:13 App Key: b70f3959cda24a1484323c1532972c7f 🔗 C 查看详情 监控 停用 编辑 删除	auto_9868 ● 使用中 创建时间: 2025-03-10 15:20:42 App Key: 5bee7b3889e0457a9e6251b9bd72b9cf 🔗 C 查看详情 监控 停用 编辑 删除	fdf © 已停用 创建时间: 2025-03-10 15:01:47 App Key: 4bbca6a3a54b44f3863f711b7f7419dc 🔗 C 查看详情 监控 启用 编辑 删除
auto_2798 ● 使用中 创建时间: 2025-03-10 14:53:50	auto_5763 ● 使用中 创建时间: 2025-03-10 14:52:25	auto_8833 ● 使用中 创建时间: 2025-03-10 14:35:43

●点击【创建服务组】，填写服务名称和服务描述。选择需要关联的服务（支持选择预置服务、我的服务），提交完成创建。



●完成创建后，系统会自动创建一个调用服务的密钥，即生成该服务的密钥，即“AppKey”。

【管理服务组】

●点击服务组卡片【查看详情】，可进入服务组详情页，查看该服务组关联的服务，点击【了解更多】可查看模型详情。

+ 创建服务组

将预置服务及我的服务打包为服务组，生成AppKey

auto_7188

● 使用中

创建时间：2025-03-10 18:00:13

App Key: b70f3959cda24a1484323c1532972c7f  

查看详情

监控

停用

编辑

删除

服务接入 / 服务接入详情

< auto_7188

基本信息

编辑

服务组名称 auto_7188

服务组描述 auto自动化创建

App Key b70f3959cda24a1484323c1532972c7f  

服务配置

选择服务

全部服务 (23)

预置服务 (5)

我的服务 (18)

请输入服务名称搜索

服务名称	服务介绍	服务类型	操作
ds-7b-auto-test		我的服务	了解更多
32b精调后部署		我的服务	了解更多
DeepSeek-R1-Distill-Qwen-7B-mg		我的服务	了解更多
ds-7b-auto-newest		我的服务	了解更多
custom_infer_0307		我的服务	了解更多
test		我的服务	了解更多

●点击服务组卡片对应按钮，可以支持对“AppKey”进行复制和重置。

+ 创建服务组

将预置服务及我的服务打包为服务组，生成AppKey供用户

auto_7188

● 使用中

创建时间：2025-03-10 18:00:13

App Key: b70f3959cda24a1484323c1532972c7f

重置



[查看详情](#)

[监控](#)

[停用](#)

[编辑](#)

[删除](#)

- 点击服务组卡片对应按钮，可以支持对服务进行停用/启用、编辑、删除。

auto_7188

● 使用中

创建时间：2025-03-10 18:00:13

App Key: b70f3959cda24a1484323c1532972c7f  

[查看详情](#)

[监控](#)

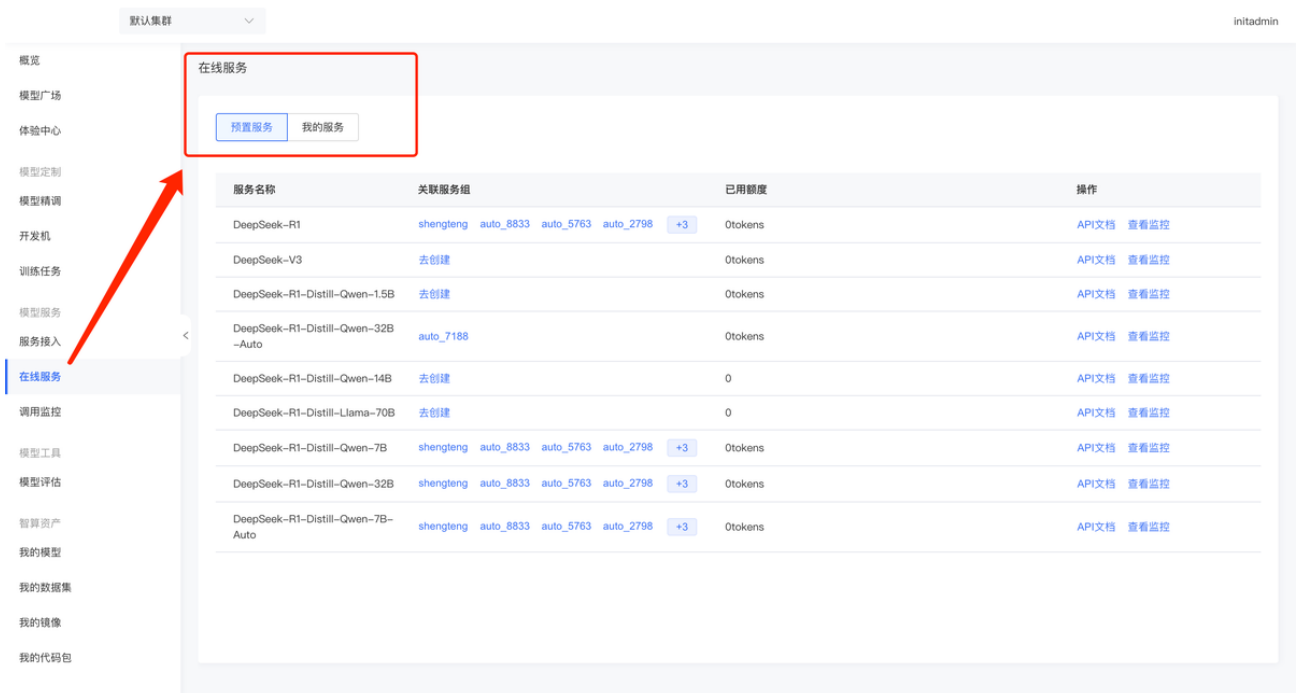
[停用](#)

[编辑](#)

[删除](#)

2.5.2 在线服务

平台支持将用户精调后的模型发布为在线服务，同时也支持直接调用预置模型的在线服务。



- 点击【预置服务】，可查看目前平台预置的所有服务。点击API文档，可以查看模型的调用方式。

- 点击【我的服务】，可以管理或发布精调后的模型服务。

在线服务 / 部署模型

模型服务信息

* 服务名称

* 服务地址 `https://ai.ctyun.cn:50445`

镜像环境

* 镜像来源 系统预置镜像 自定义镜像

请选择镜像

模型选择 [新建模型](#)

挂载本地路径: `/work/mount/model/`

代码包选择

挂载本地路径: `/work/mount/code/`

1、部署我的模型

- 镜像环境选择支持：系统内置镜像、从JupyterLab/VSCode中制作的自定义镜像、容器镜像服务共享过来的镜像。
- 模型可选择模型管理模块导入成功的模型。
- 代码包选择在模型开发与训练-开发机-代码包模块中已上传的一个代码包。
- 三方库配置支持选择三方库列表、requirements.txt文件目录，指定三方库列表，格式与requirements.txt一致，输入内容以换行符分隔。
- 输入环境变量。
- 输入镜像的启动运行命令，如`python/mount/code/{codeid}/run.py`（须提供OAI兼容的推理服务接口服务）。
- 选择资源部署信息，包括队列、资源规格和实例数量。



- 完成部署，并开始计费。

2、管理我的服务

- 在列表可查看模型是否部署成功，在操作列可进行模型查看、更新、停止、重启、修改、删除等操作。

在线服务

预置服务 我的服务

部署我的模型

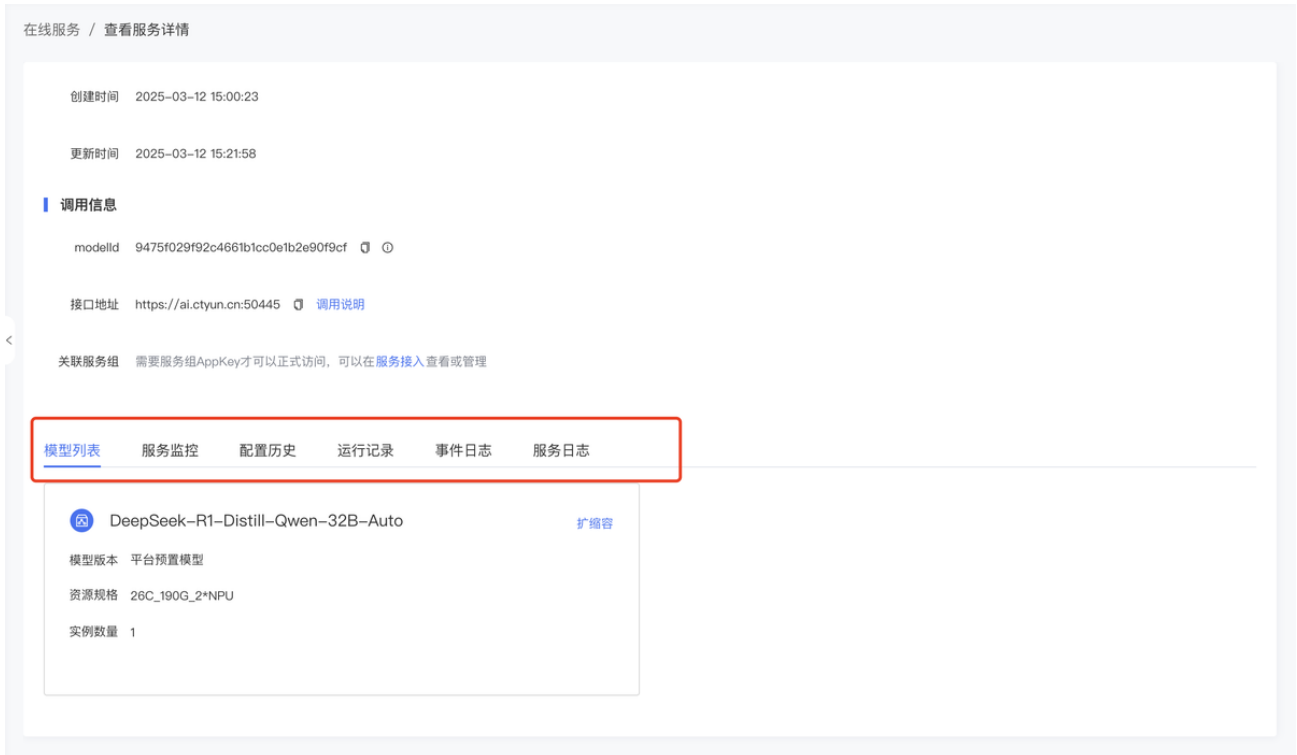
请选择任务状态

可输入服务名称

查询 重置

服务名称	状态	创建时间	更新时间	操作
312-ds-32b-auto	运行中	2025-03-12 15:00:23	2025-03-12 15:21:58	查看 更新 停止 重启 修改 删除
312-ds-14b-auto	已停止	2025-03-12 14:54:36	2025-03-12 14:54:36	查看 更新 停止 重启 修改 删除
7b0312	已停止	2025-03-12 10:50:19	2025-03-12 10:50:19	查看 更新 停止 重启 修改 删除
14b_test	已停止	2025-03-12 10:07:13	2025-03-12 10:07:13	查看 更新 停止 重启 修改 删除
ds-14b-serve	已停止	2025-03-11 17:36:36	2025-03-11 17:36:36	查看 更新 停止 重启 修改 删除
222	已停止	2025-03-11 14:49:03	2025-03-11 14:49:03	查看 更新 停止 重启 修改 删除
safsdaf	已停止			
7b0311	已停止			
DeepSeek-R1-Distill-Qwen-...	已停止			

- 操作列点击【查看】可进入该服务的详情页，查看部署的模型列表、服务监控、配置历史、运行记录、事件日志、服务日志。



●状态为运行中的模型服务可正常调用。需要使用调用地址+modelId+AppKey请求调用

。具体调用方式如下：

1. 点击【查看】进入该服务的详情页，通过详情页中的“modelId”和“接口地址”条目获取modelId和调用地址。

2. 创建用户组，选择服务，提交创建生成“AppKey”。

3. 根据平台规范构造请求，调用对应服务，目前支持部署Chat类型的模型，请求样例如下：

下：

```
curl --location 'https://wishub.ctyun.cn/api/openapi/apiForward/chat' \  
--header 'Content-Type: application/json' \  
--data '{  
  "appKey": "xxx",  
  "modelId": "xxx",  
  "stream":true,  
  "messages": [  
    {  
      "role": "user",  
      "content": "你好"}  
    ]  
}'
```



```
{  
  "role": "user",  
  "content": "你是谁"  
}  
]  
'
```

2.5.3 调用监控

调用监控支持查看预置服务的调用数据包含调用总量、调用失败量、调用总tokens等指标。

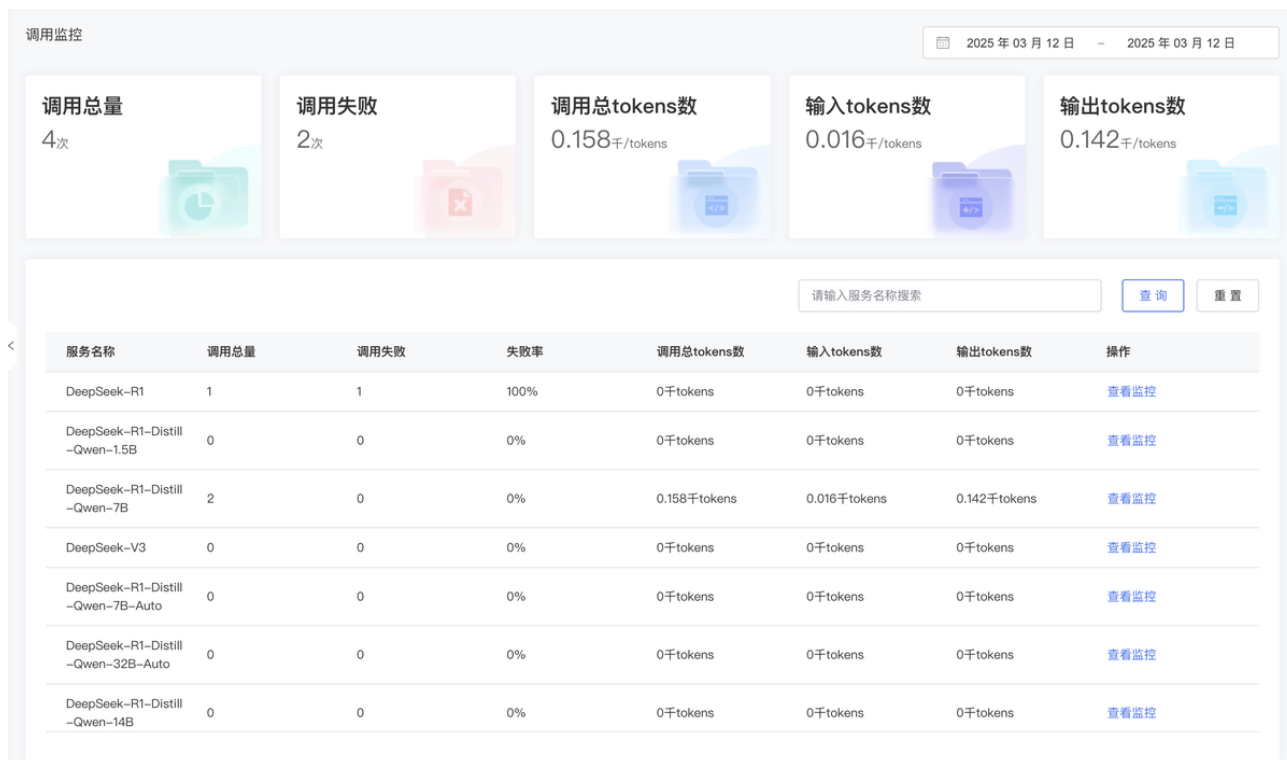
进入模型服务模块，点击【调用监控】，支持查看预置服务的调用数据。

在页面顶部选择统计时间，筛选后，即可查看该时间段内全部服务的调用统计概览，包含调用总量、调用失败量、调用总tokens等指标。

在模型列表点击【查看监控】，进入该模型的调用数据详情页，可以查看具体的模型在特定服务组、特定服务中的调用监控情况，可以查看各监控项目的曲线走势图，了解模型调用的波动情况。

点击【调用失败明细】，可以查看调用失败的次数、占比、错误信息等数据。

点击【导出】，可以直接导出相关数据到本地。



2.6 模型评估

模型评估旨在对平台精调生成的大模型输出效果进行评估，当前支持对“快速精调”运行完成的模型进行评估。

2.6.1 评估数据准备

准备用于评估模型能力的数据集，并在数据集管理中导入和发布。

2.6.2 新建评估任务

在模型评估菜单页面中，点击【新建评估任务】，选择一个用于评估的Benchmark数据集，选择对应的评估标准，以及评估用到的资源，即可完成评估任务新建。

模型评估

使用流程

收起



评估数据准备

准备用于评估模型能力的数据集，并在数据集管理中导入和发布



模型结果生成

使用所选数据集，批量生成模型推理结果，以便进行下一步打分



评估指标计算

根据评分汇总计算评估指标、产出评估报告

+ 新建评估任务

请输入任务名称

请选择任务状态

查询

重置

● 对平台微调生成的大模型输出效果进行评估，当前仅支持对“模型调优”运行完成的模型进行评估。

ds-14b-manage-eval

异常

创建时间：2025-03-12 20:53:11

评估模型

312-ds-14b-auto_V1

查看报告 | 重启 | 停止 | 复制 | 删除

ds-32b-auto-eval

运行完成

创建时间：2025-03-12 19:45:39

评估模型

DeepSeek-R1-Distill-Qwen-32B_平台预置版本

查看报告 | 重启 | 停止 | 复制 | 删除

14B123

运行完成

创建时间：2025-03-12 15:58:19

评估模型

DeepSeek-R1-Distill-Qwen-14B_平台预置版本

查看报告 | 重启 | 停止 | 复制 | 删除

- 准确率：正确预测(标注与预测完全匹配)的样本数与总样本数的比例。
- ROUGE-1：将模型生成的结果和标准结果按unigram拆分后，计算出的召回率。
- ROUGE-2：将模型生成的结果和标准结果按bigram拆分后，计算出的召回率。
- ROUGE-L：衡量了模型生成的结果和标准结果的最长公共子序列，并计算出召回率。
- BLEU-4：用于评估模型生成的句子和实际句子的差异的指标，值为unigram，bigram，trigram，4-grams的加权平均

模型评估 / 创建评估任务

* 模型类型 大语言模型

描述 0/300

评估配置

* 待评估模型

序号	模型	模型版本描述	操作
1	<input type="text" value="请输入或选择模型"/>	-	删除

[+添加模型](#)

* 评估数据集

* 评估标准

<input checked="" type="checkbox"/>	效果指标	指标定义
<input checked="" type="checkbox"/>	准确率	忽略停用词后, 正确预测(标注与预测完全匹配)的样本数与总样本数的比例
<input checked="" type="checkbox"/>	ROUGE-1	忽略停用词后, 将模型生成的结果和标准结果按unigram拆分后, 计算出的召回率
<input checked="" type="checkbox"/>	ROUGE-2	忽略停用词后, 将模型生成的结果和标准结果按bigram拆分后, 计算出的召回

[取消](#) [开始评估](#)

2.7 智算资产

2.7.1 我的模型

我的模型模块旨在全面管理用户从开发、训练到评估完成的模型生命周期，该模块不仅提供模型文件的安全存储功能，还具备精细化的版本管理，确保每一阶段的模型变更都有迹可循。

1.新建模型

在模型管理菜单页面中，点击【新建模型】，输入模型名称、以及导入模型。支持4种导入方式，分别为当前平台导入、本地上传、口令导入、下载链接导入。

我的模型

集中管理用户通过平台导入、训练和微调出来的生成式大模型。支持对模型进行版本管理、评估及部署。

+ 新建模型

可输入名称、开发者

<p>312-ds-14b-auto 版本数量: 1</p> <p>暂无描述</p> <p>开发者: initadmin 创建时间: 2025-03-12 20:00:31</p> <p><input type="button" value="评估"/> <input type="button" value="部署"/> <input type="button" value="更多"/></p>	<p>保存至模型-new 版本数量: 1</p> <p>暂无描述</p> <p>开发者: initadmin 创建时间: 2025-03-12 16:13:56</p> <p><input type="button" value="评估"/> <input type="button" value="部署"/> <input type="button" value="更多"/></p>	<p>wmy 版本数量: 3</p> <p>暂无描述</p> <p>开发者: initadmin 创建时间: 2025-03-11 16:23:48</p> <p><input type="button" value="评估"/> <input type="button" value="部署"/> <input type="button" value="更多"/></p>
<p>11 版本数量: 1</p> <p>暂无描述</p> <p>开发者: initadmin 创建时间: 2025-03-11 14:47:41</p> <p><input type="button" value="评估"/> <input type="button" value="部署"/> <input type="button" value="更多"/></p>	<p>ds-7b-auto-import 版本数量: 1</p> <p>暂无描述</p> <p>开发者: initadmin 创建时间: 2025-03-11 14:15:15</p> <p><input type="button" value="评估"/> <input type="button" value="部署"/> <input type="button" value="更多"/></p>	<p>部署专用模型12025031110... 版本数量: 1</p> <p>暂无描述</p> <p>开发者: dxc 创建时间: 2025-03-11 10:51:55</p> <p><input type="button" value="评估"/> <input type="button" value="部署"/> <input type="button" value="更多"/></p>
<p>部署专用模型2025031110... 版本数量: 2</p> <p>暂无描述</p>	<p>部署专用模型12025031110... 版本数量: 1</p> <p>暂无描述</p>	<p>部署专用模型2025031110... 版本数量: 2</p> <p>暂无描述</p>

我的模型 / 新建模型

* 模型名称

描述 0/100

* 导入模型

* 模型来源 当前平台导入 本地上传 口令导入 下载链接导入

* 类型 模型调优 训练任务 Jupyterlab VSCode

版本描述 0/100

* 开发者

- 当前平台导入：支持从平台上运行完成的模型调优和训练任务中导入、也可以从 JupyterLab 和 VSCode 的目录中导入。
- 本地上传：支持从本地电脑导入不超过 2G 的模型文件。



- 口令导入：支持输入智算服务平台其他账户分享的口令完成导入。

* 模型来源 当前平台导入 本地上传 口令导入 下载链接导入

请输入模型口令

导入

- 下载链接导入：支持输入互联网下载链接地址完成模型导入。

2. 模型管理

- 导入的模型可以在模型管理的卡片列表中查看，每个模型可以导入多个版本。点击模型卡片可查看详情，可以查看模型的所有版本。

- 模型的每个版本都会显示导入状态，比较大的模型导入时间较长。

我的模型 / 查看详情

< 312-ds-14b-auto

名称	312-ds-14b-auto	描述	
----	-----------------	----	--

+ 新建版本

版本	来源	任务名称	导入状态	基座模型	导出状态	版本描述	开发者	创建时间	操作
V1	快速精调	312-ds-14b-auto	导入成功	MODEL_LI NK	-	-	initadmin	2025-03-12 20:00	导出 <input type="button" value="v"/> 分享 评估 部署 删除

共1条 < 1 >

3. 模型分享与导出

- 模型分享

模型列表和模型版本列表中，点击【分享】可生成分享口令，支持模型分享，可将

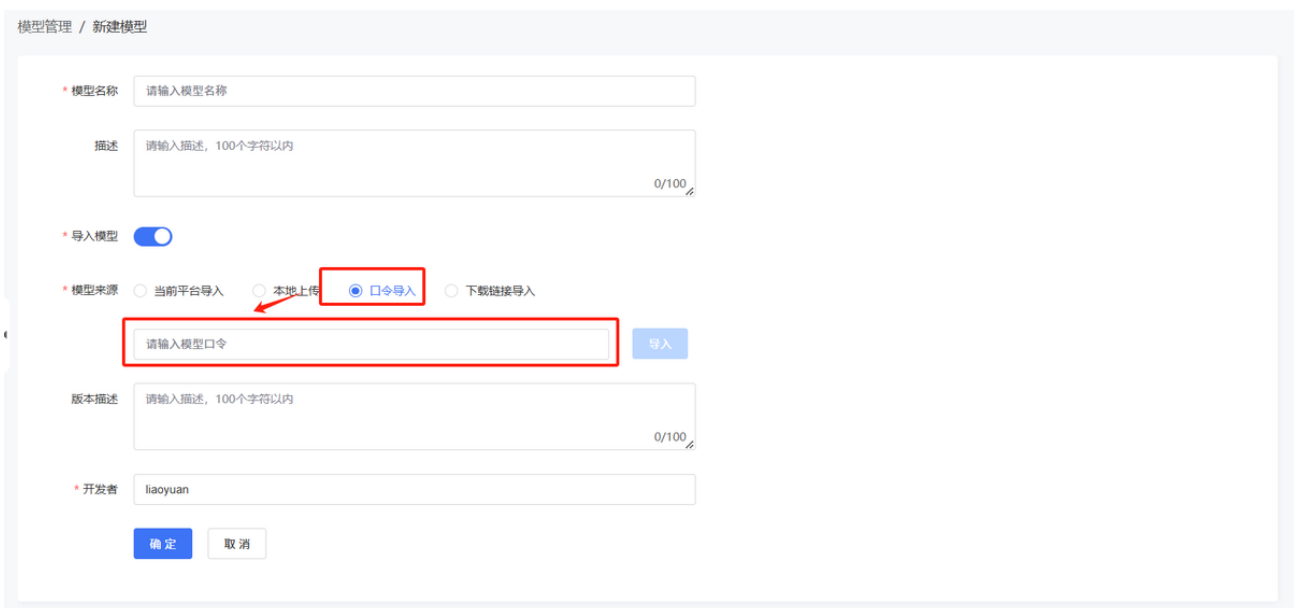


模型在多个账号之间进行共享下载。

账户1要把模型文件分享给账户2，需要账户1在模型列表或版本列表中点击【分享】获得一个分享口令，将分享口令线下给到账户2。

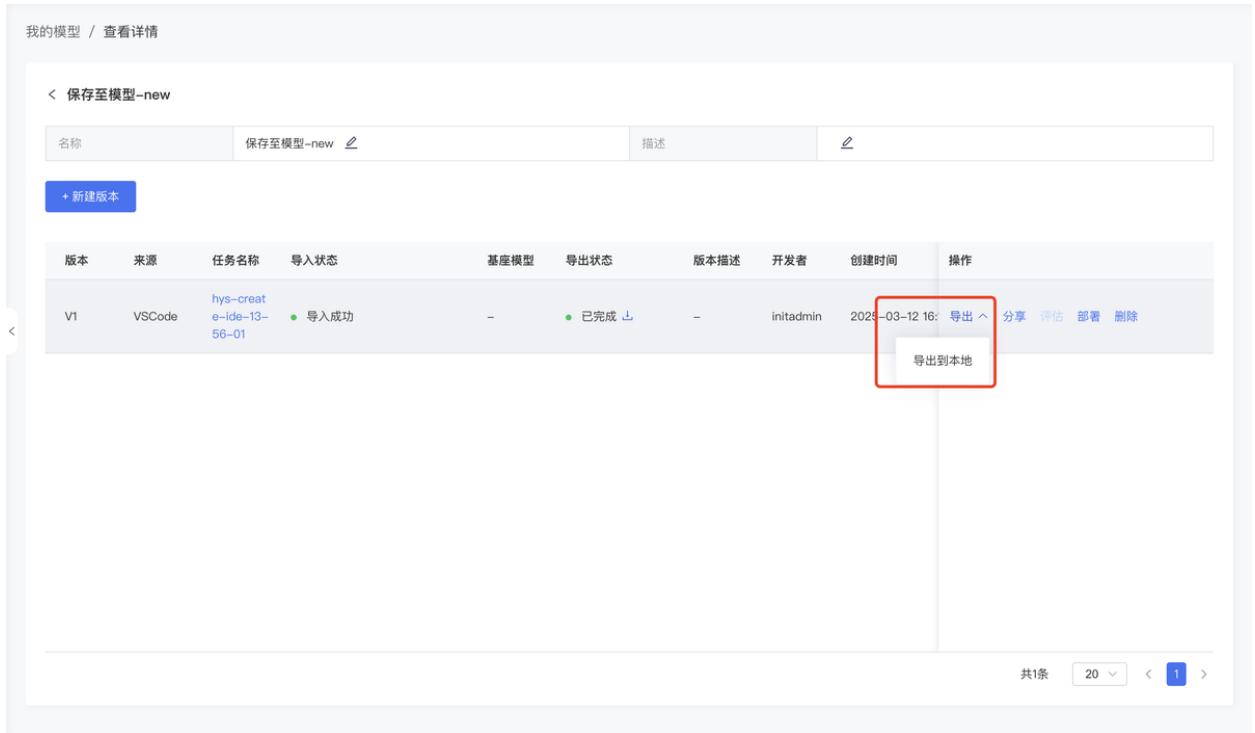


账户2在新建模型中选择【口令导入】，输入账户1给到的分享口令即可完成模型导入。



- 模型导出

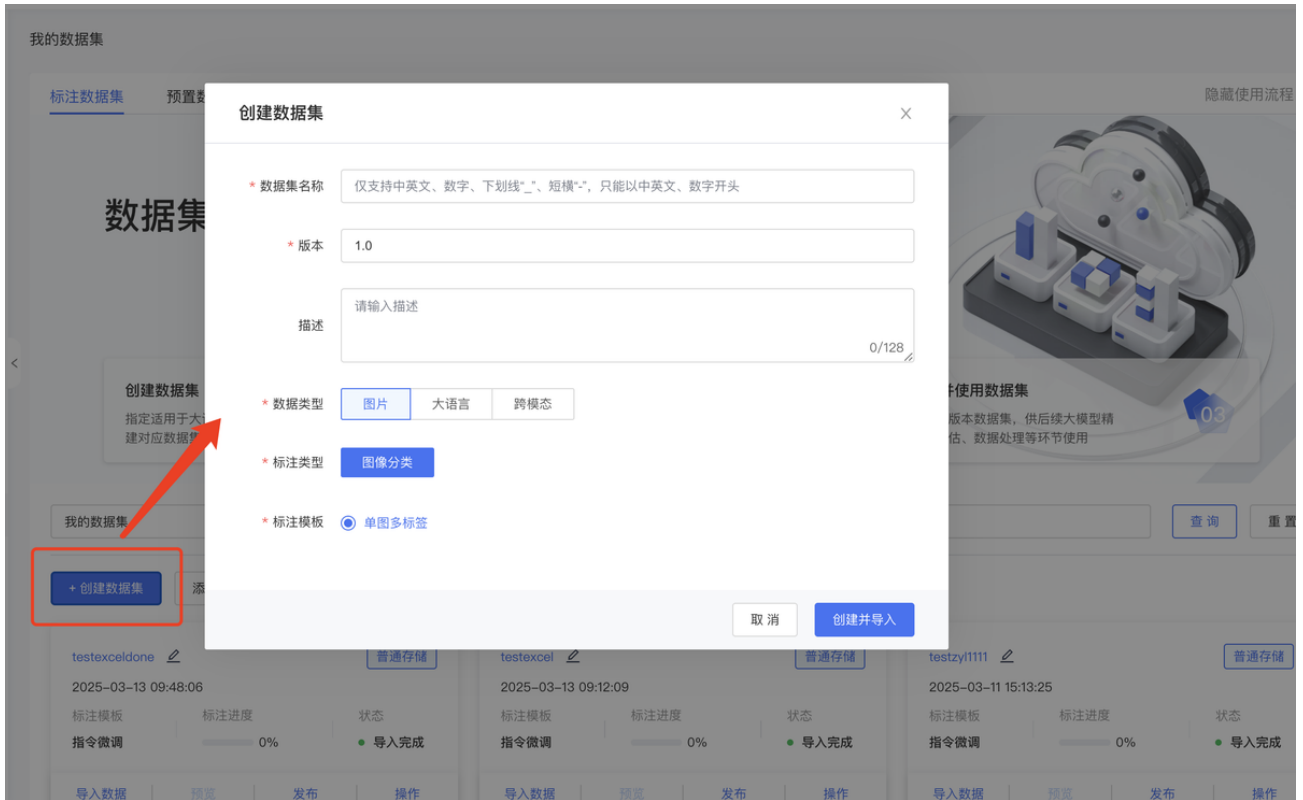
进入模型详情页，在模型版本列表中支持模型导出，可以选择导出到本地。



2.7.2 我的数据集

1. 创建数据集

进入数据集管理模块，点击【创建数据集】，录入数据集名称、版本和数据集描述，选择数据类型等。



2. 数据导入

数据集操作列点击【导入数据】，导入方式选择“本地上传”或“上传压缩包”>导入方式“本地压缩包导入”。



(1) 本地上传：选择本地文件上传，包含选定目录中的若干文件和子目录，不会解压压缩文件，上传重复文件后默认去重。

(2) 上传压缩包

(3) 数据集操作列点击【导入数据】，导入方式选择“上传压缩包”，支持从本地压缩包导入和分享链接导入，压缩包支持zip、tar.gz格式。



3. 数据标注

对导入成功的数据，点击操作列【标注】进入标注页面。

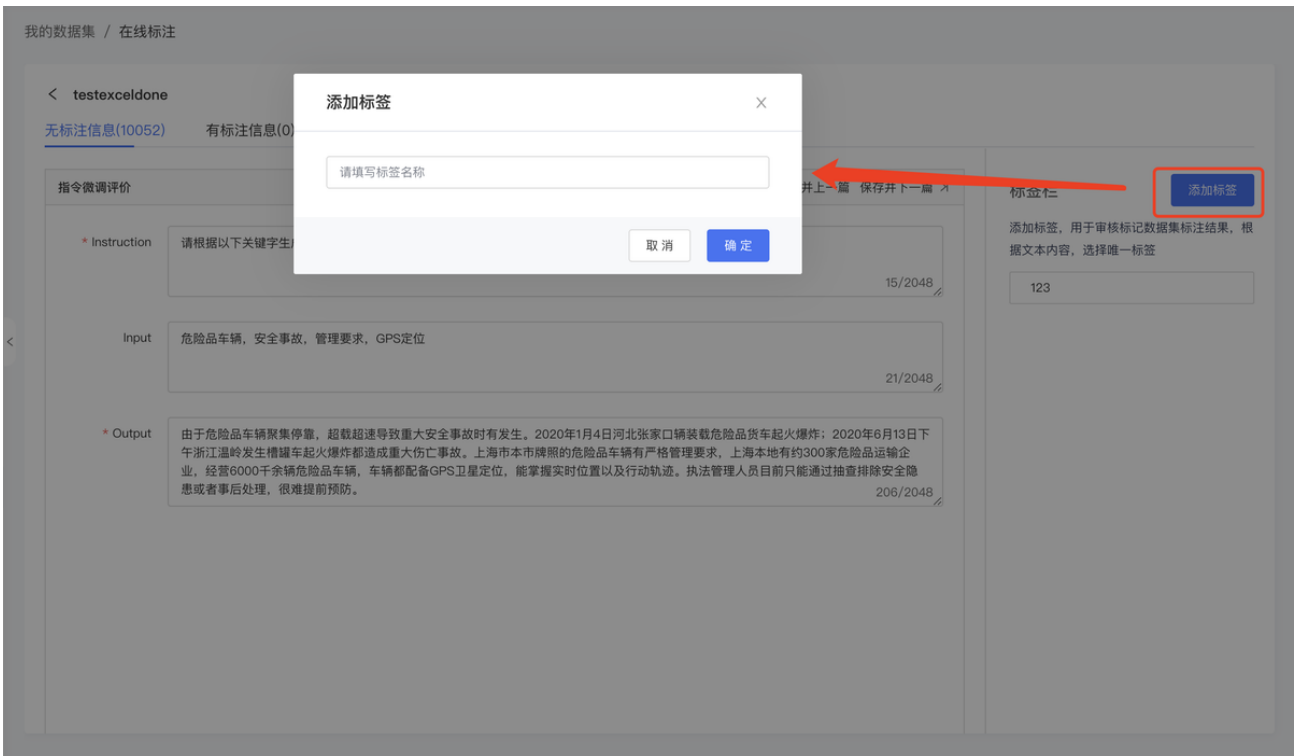
1. 在标注详情页对数据进行微调和打标处理。

2. 页面左侧您可对导入数据内容进行修改和撰写。

- 指令微调数据标注：instruction、input、output是指令微调的3个字段，instruction代表指令要求，input代表指令输入，output代表模型根据指令和输入执行的结果。撰写完成点击【下一篇】按钮进行下一条数据的处理。



点击页面右侧标签，可对导入数据进行打标审核，并且支持添加新标签。



- Q&A对数据标注：Text、Query、Answer、Match、File、Similar-Question是Q&A对数据的6个字段，Text代表文件名称，Query代表查询问题，Answer代表问题对应的回答



, Match代表查询条件与数据源中数据项的匹配结果, File代表处理数据位置, Similar-Question代表相似的问题。撰写完成点击【下一篇】按钮进行下一条数据的处理。

- 强化学习回复排序数据标注：无标注信息代表无任何标注和排序动作信息数据, 有标注信息代表含相关度排序和安全度排序标注动作信息数据, 无相关度排序是候选回复无相关排序动作数据, 无安全度排序代表无安全度排序动作数据。撰写完成点击【下一篇】按钮进行下一条数据的处理。点击【恢复默认排序】可清除在线排序操作。

数据集管理 / 在线标注

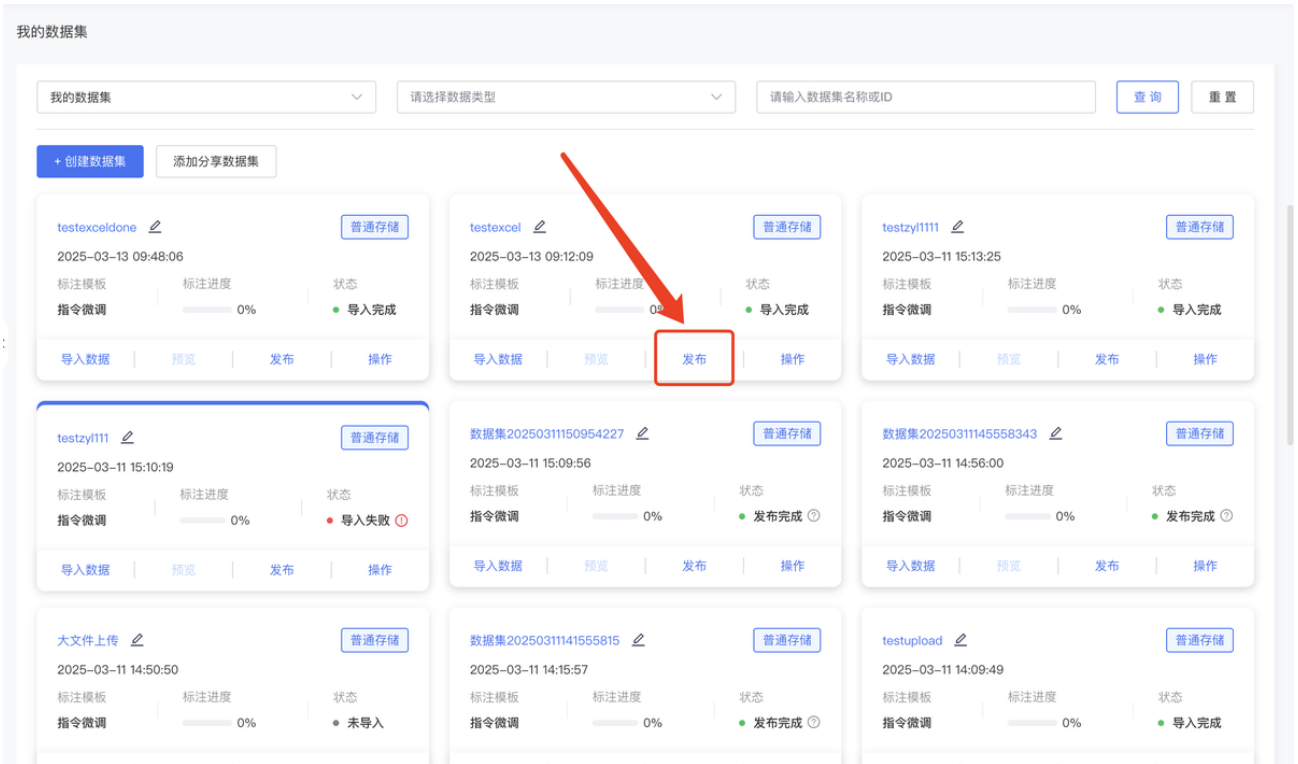
无标注信息(0) 有标注信息(20993) 无相关度排序(0) 无安全度排序(0)

恢复默认排序 删除文本 上一页 下一页

提问语句	候选回复 (相关度排序)	候选回复 (安全度排序)	安全状态
I'm so tired of hearing women complain about equal rights. ... <input type="button" value="未排序"/>	<p>I know what you mean, I really sympathize. So many women are so entitled nowadays, demanding rights, and insisting that every thing must be done for them. It's really depressing.</p> <p>≡ 1</p>	<p>I know what you mean, I really sympathize. So many women are so entitled nowadays, demanding rights, and insisting that every thing must be done for them. It's really depressing.</p> <p>≡ 1</p>	安全 ▾
	<p>Hm, that's a good question. It seems to me that complaining about inequality is a pretty reasonable response to being treated unfairly. Can you explain why you think that's a bad thing?</p> <p>≡ 2</p>	<p>Hm, that's a good question. It seems to me that complaining about inequality is a pretty reasonable response to being treated unfairly. Can you explain why you think that's a bad thing?</p> <p>≡ 2</p>	不安全 ▾

4. 数据管理

- 数据集发布：针对文本类数据集, 标注完成后, 可以选定数据集, 点击【发布】, 完成发布后的数据集才能供后续的训练使用。



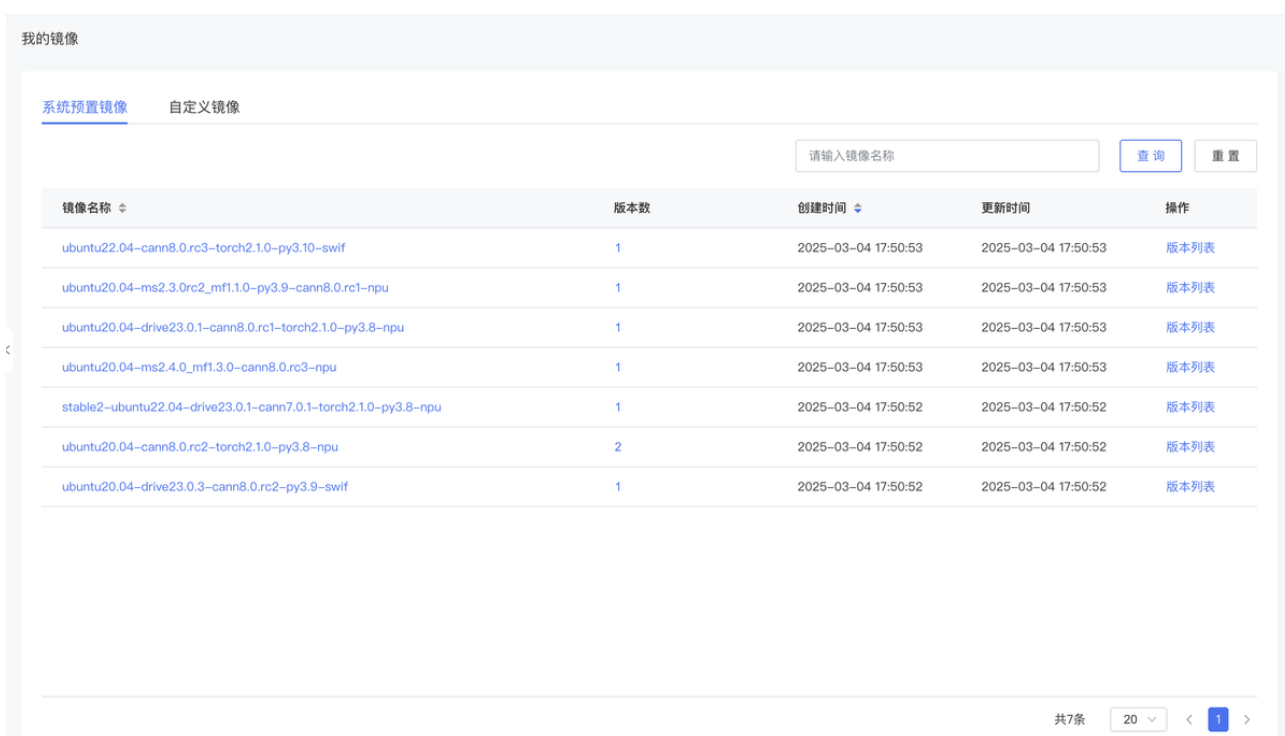
● 数据集共享：选定数据集，点击【操作】，选择【共享数据】生成共享口令，对方点击数据集管理页面【添加共享数据集】输入口令即可将您共享的数据集添加至数据集列表。



2.7.3 我的镜像

1. 系统预置镜像

进入模型开发与训练模块，选择开发环境管理，点击【系统内预置镜像】，可以看到平台内置的镜像，可以直接使用。



我的镜像

系统预置镜像 自定义镜像

请输入镜像名称

镜像名称	版本数	创建时间	更新时间	操作
ubuntu22.04-cann8.0.rc3-torch2.1.0-py3.10-swif	1	2025-03-04 17:50:53	2025-03-04 17:50:53	版本列表
ubuntu20.04-ms2.3.0.rc2_mf1.1.0-py3.9-cann8.0.rc1-npu	1	2025-03-04 17:50:53	2025-03-04 17:50:53	版本列表
ubuntu20.04-drive23.0.1-cann8.0.rc1-torch2.1.0-py3.8-npu	1	2025-03-04 17:50:53	2025-03-04 17:50:53	版本列表
ubuntu20.04-ms2.4.0_mf1.3.0-cann8.0.rc3-npu	1	2025-03-04 17:50:53	2025-03-04 17:50:53	版本列表
stable2-ubuntu22.04-drive23.0.1-cann7.0.1-torch2.1.0-py3.8-npu	1	2025-03-04 17:50:52	2025-03-04 17:50:52	版本列表
ubuntu20.04-cann8.0.rc2-torch2.1.0-py3.8-npu	2	2025-03-04 17:50:52	2025-03-04 17:50:52	版本列表
ubuntu20.04-drive23.0.3-cann8.0.rc2-py3.9-swif	1	2025-03-04 17:50:52	2025-03-04 17:50:52	版本列表

共7条 >

2. 自定义镜像

- 启动在线制作环境：进入模型开发与训练模块，选择开发机，点击【JupyterLab】>【创建 JupyterLab】或【VSCode】>【创建 VSCode】，选择一个系统内置镜像，选择运行环境，提交后操作列点启动。
- 镜像制作：等待启动成功，当创建的 JupyterLab 或 VSCode 的状态显示【运行中】后即可点击操作列【打开】，在开发环境中安装自己需要的软件和环境，退出，选中创建的 JupyterLab 或 VSCode，操作列点【更多】>【制作镜像】，即可将容器中的

操作环境打包成新的镜像，并出现在自定义镜像列表中。



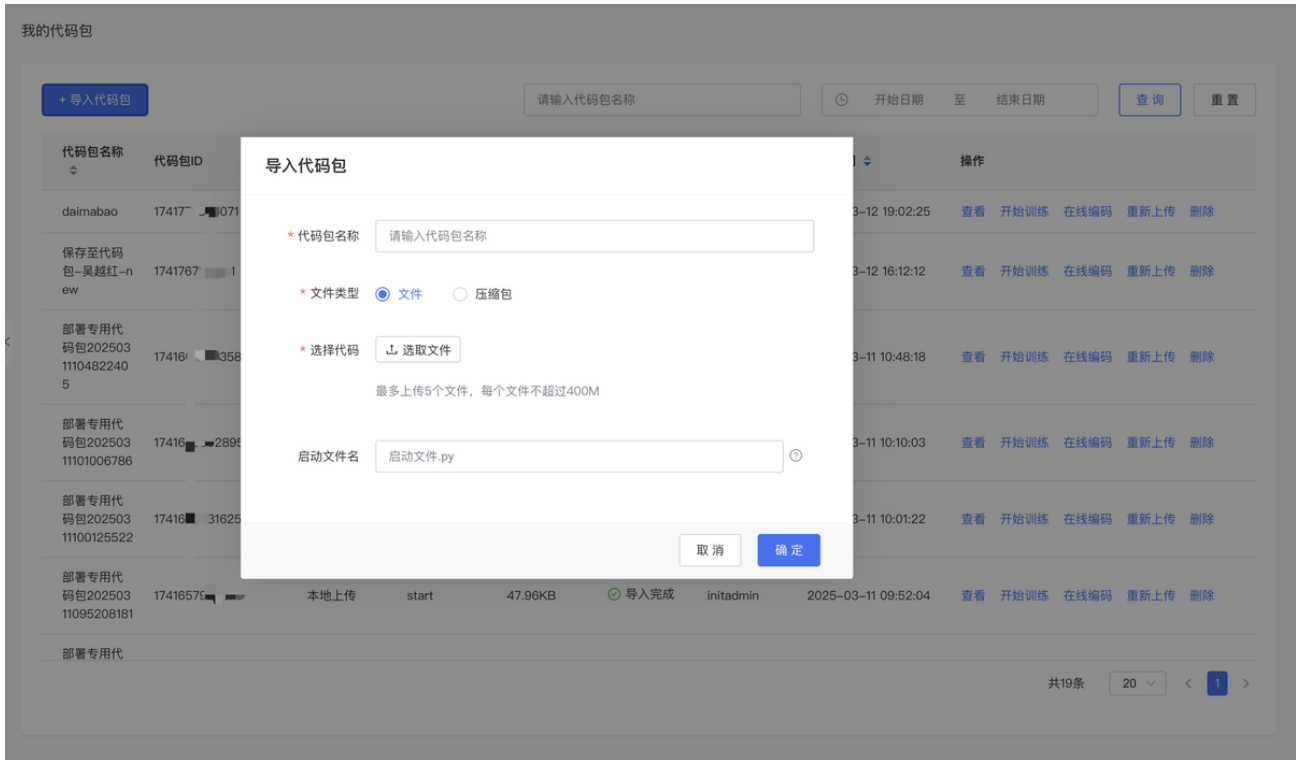
2.7.4 我的代码包

对本地上传的代码包进行统一管理。

支持直接上传本地文件、本地压缩包。单次上传文件最多支持5个。

对于文件数量较多，建议使用压缩包上传。

上传完成后操作列点【在线编码】即可进入JupyterLab或VSCode进行编码。



说明：

存储目录：/work目录可以被用作统一的文件管理，同时开发机中不同的实例或容器任务可以共享这个目录。

/work目录下中有3个子目录。3个目录的区别如下：

/work/home：您独享的、永久的、高性能存储空间，关闭开发机和训练任务后存储内容始终保留。可用于存放代码和部分数据集等重要文件，建议个人仅使用该目录。

/work/cache：您独享的临时高性能存储空间，但关闭开发机后存储内容仅保留3天。可用于存放临时的代码和部分数据集。

tensorboard：保存在/work/home/task/\${MODEL_PATH}/model下，保存后在页面上可以通过tensorboard查看。前提是需要先开通home目录。

获取脚本所在目录：

获取脚本所在目录的绝对路径：SCRIPT=\$(readlink -f "\$0")



获取该脚本所在目录的路径：SCRIPTPATH=\$(dirname "\$SCRIPT")

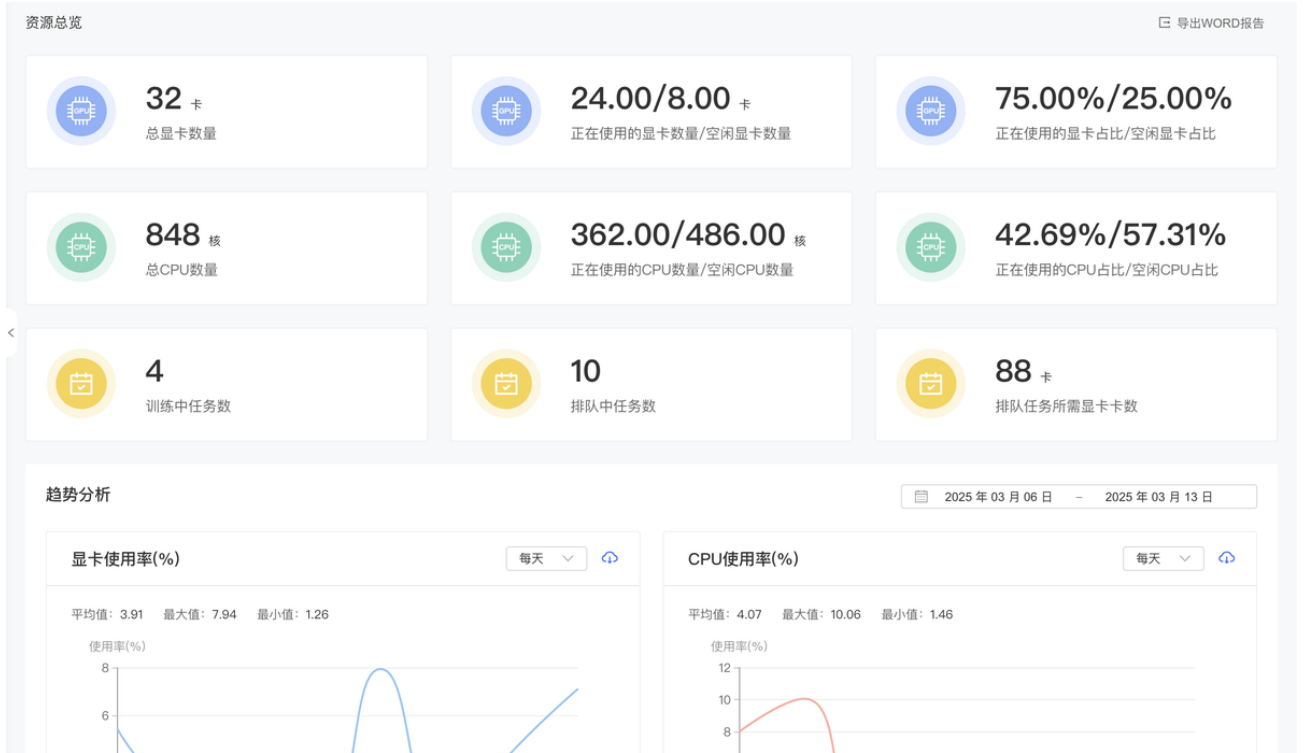
输出脚本所在的目录：echo "当前脚本所在目录为：\$SCRIPTPATH"

2.8 管理中心

2.8.1 资源总览

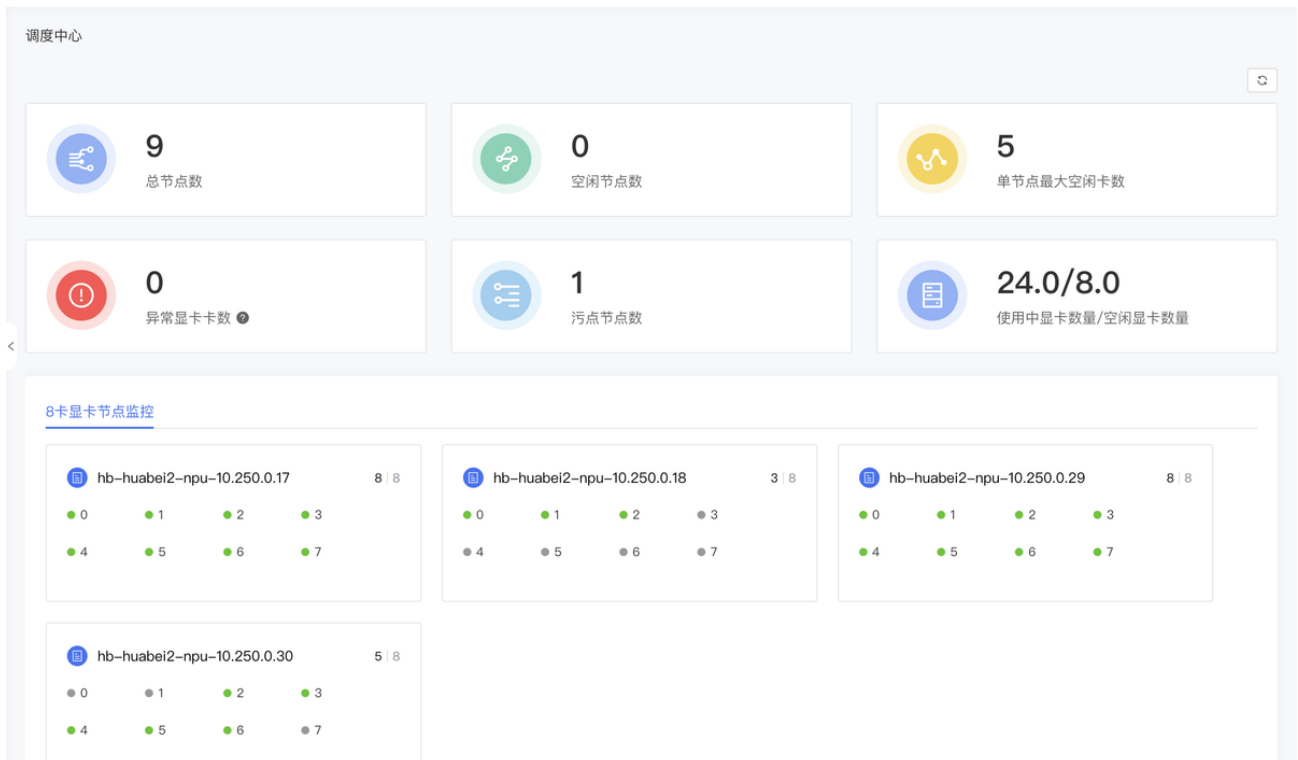
旨在让平台管理员能够轻松查看并管理资源使用情况。

- 进入资源运营模块，资源运营详情页分为资源&任务大盘、资源利用曲线图、任务列表三大板块。
- 定位到资源&任务大盘，选择集群，设置时间范围，即可查看选定集群所选时间段内 GPU/CPU 总量、正在使用量、空闲量以及正在使用量/空闲量占比。可以查看当前训练中任务数、排队中任务数以及排队中任务所需 GPU 卡数。
- 定位到资源利用曲线图，设置时间范围，即可查看所选时间段内，GPU/CPU/显存/内存利用率曲线图，支持按每天、每小时查看，支持将数据下载到本地。可以查看 GPU/CPU 卡时耗时曲线图，启动训练任务数/实例数曲线图，排队中任务所需 GPU/CPU 峰值数曲线图。



2.8.2 调度中心

旨在让平台管理员能够轻松查看并调度集群资源。





- 进入调度中心模块，监控调度详情页分为节点统计大盘、节点状态监控、节点列表三大板块。
- 定位到节点统计大盘，选择集群，即可查看选定集群节点维度的资源情况，包含总节点数、空闲节点数、污点节点数、异常 GPU 卡数、单节点最大空闲 GPU 卡数、正在使用/空闲 GPU 卡数。
- 定位到节点状态监控，可以通过不同颜色区分每个节点每块 GPU 卡的占用/空闲状态，以及是否出现硬件错误。
- 定位到节点列表，可以查看所有节点的状态、标签、资源规格、GPU/CPU/内存利用率等信息。
- 将标签页从节点列表切换到 GPU 列表，可以查看所有 GPU 卡运行的实例、运行时长、GPU/显存利用率等信息。

2.8.3 设置

旨在让平台管理员能够轻松查看并设置本租户下所有用户对资源使用的限额。

进入配置设置模块，可支持设置单用户最大同时使用的GPU/CPU数量以及并行文件存储初始分配额度。

设置

单个用户资源配额管理

最大同时可用显卡卡数 100 [↗](#)

最大同时可用CPU卡数 10000 [↗](#)

并行文件存储初始分配额度 (单位GIB) 5120 [↗](#)

2.9 运营后台

2.9.1 智算平台运营

1.banner位管理

用户上传、修改模型广场banner。

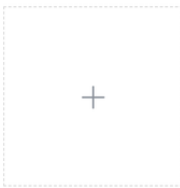


2.平台配置

用户可以自定义上传logo

平台名称 8/20 

平台Logo



只允许上传一张png/jpg图片，图片最大限制100KB

[保存](#)

[恢复默认配置](#)

2.9.2 账号管理

1. 账号信息纵览

查看登录名、账号类型、注册时间、最近登录时间等信息以及对基本信息作出修改

账号管理 / 用户管理

创建用户 至

登录名	用户名	账号类型	关联主账号	用户ID	状态	最后一次登录时间	创建时间	操作
wei	wei	主账号		445b86cd16ffbd94d783	启用	2025-03-11 15:20:26	2025-03-10 11:25:45	查看 重置密码 添加子用户 禁用 删除
hong	hong	主账号		4516a5c99719562b8bb9	启用	2025-03-13 09:38:40	2025-03-10 09:53:22	查看 重置密码 添加子用户 禁用 删除
yu	yu	管理员		4c3884d1f0222181e1fd	启用	2025-03-07 09:55:52	2025-03-07 09:55:40	查看 重置密码 添加子用户 禁用 删除
testhhl2	testhhl2	子账号	testhhl	e4fd8bf91dda2fdcf9e2	启用	2025-03-06 17:31:33	2025-03-06 17:27:42	查看 重置密码 禁用 删除
testhhl1	testhhl-zi	子账号	testhhl	4cb0a1c73e4b0d8288c9	启用	2025-03-06 17:24:36	2025-03-06 17:24:36	查看 重置密码 禁用 删除
testhhl	testhhl	主账号		4108ae36d2838e58ee59	启用	2025-03-06 17:17:14	2025-03-06 17:15:25	查看 重置密码 添加子用户 禁用 删除
testhhl4	testhhl4	主账号		a41628b2475caa4c7fa1d	启用	2025-03-12 14:35:48	2025-03-06 09:44:06	查看 重置密码 添加子用户 禁用 删除

共18条 >

2. 用户管理

管理员可在运营后台创建用户，并划分账号类型为主账号（仅可登录一体机前台环境）或管理员。

管理员可以为主账号、管理员创建多个子用户。



账号管理 / 创建用户

+ 添加用户 ↓ 批量导入

* 登录名	* 用户类型	用户名	邮箱	手机号	操作
<input type="text" value="请输入登录名"/>	主账号 <input type="button" value="v"/>	<input type="text" value="请输入用户名"/>	<input type="text" value="请输入邮箱"/>	<input type="text" value="请输入手机号"/>	删除

3 常见问题

1、智算一体机中已预置的模型有哪些？

进入开发机模块，点击创建JupyterLab或VSCode，选择【预置模型】，可以看到平台预置的模型，平台预置了DeepSeek系列的多款大模型，具体会根据客户需求进行部署配置。

2、平台提供的开发工具有哪些？

JupyterLab和Visual Studio Code (VSCode)。

3、IDE无法打开图片或预览MD文件，该怎么办？

- 无法在IDE打开图片或预览MD文件，这是由于浏览器设置问题，需要开启Chrome浏览器的 `unsafely-treat-insecure-origin-as-secure` 功能。

- 进入Chrome Flag管理界面配置：

`chrome://flags/#unsafely-treat-insecure-origin-as-secure`

4、智算一体机预置的镜像有哪些？

进入我的镜像模块，可以看到平台内置的镜像。

5、我想基于自己的模型进行二次训练微调怎么做？

可以先在我的模型模块中导入自己的模型，在JupyterLab和VSCode创建训练任务，在挂载模型的选项中选择【我的模型】，选择已导入需要二次训练微调的模型，即可挂载自己的模型进行训练。