智算一体机-推理版

用户操作指南

天翼云科技有限公司

目 录

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。

- 安全可信

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较少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 操作指南

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。

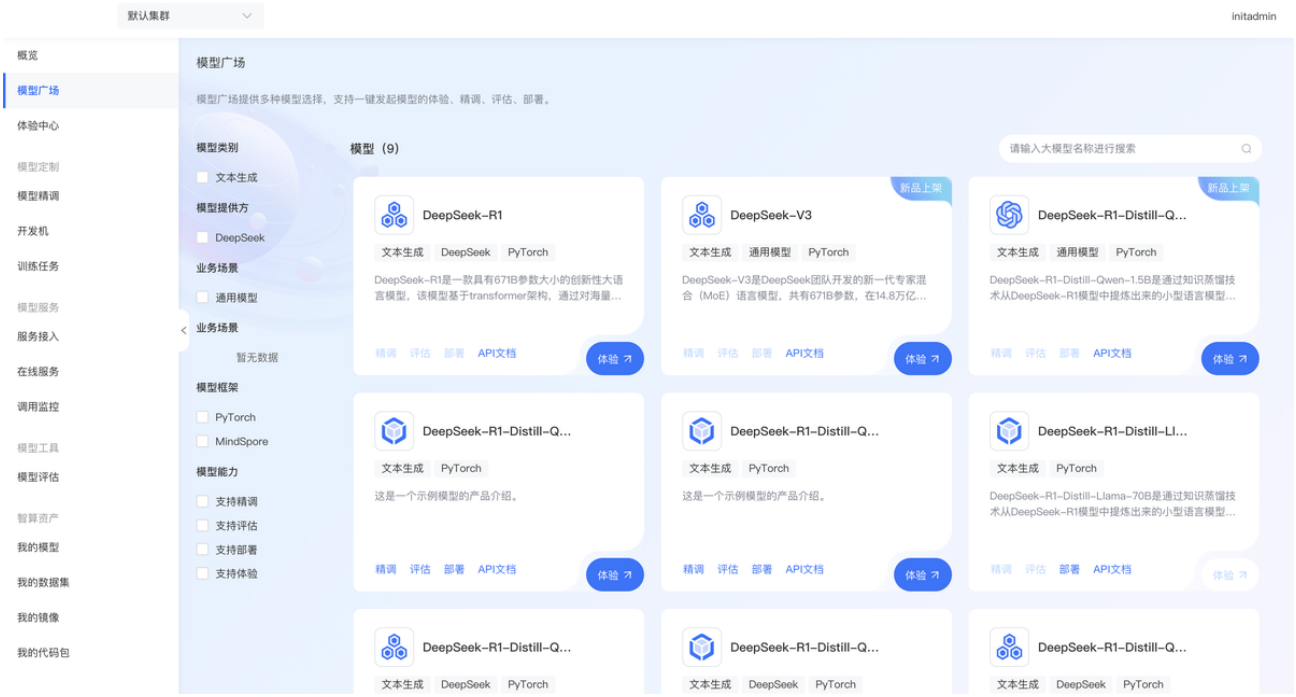
2.2 模型广场

2.2.1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。

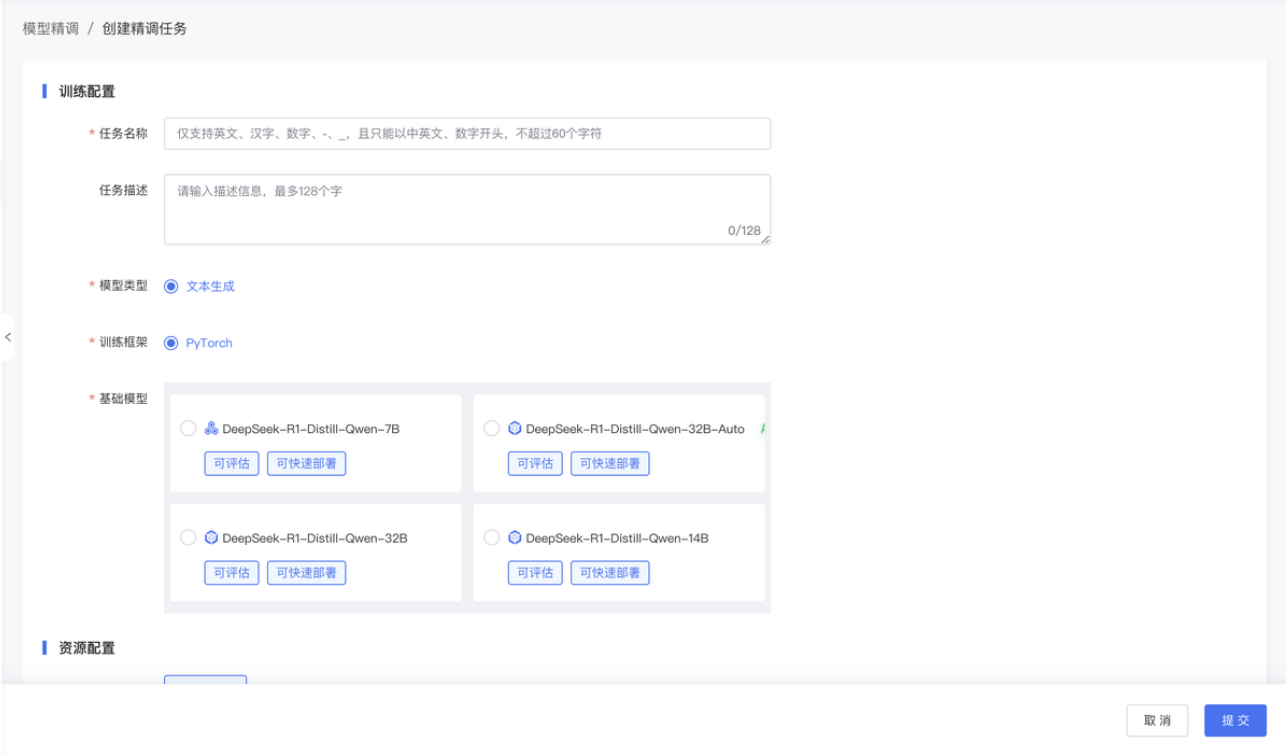


备注：截图仅供参考，可能与实际界面功能不匹配，详询解决方案，下文同理，不在备注

2.2.2一键精调

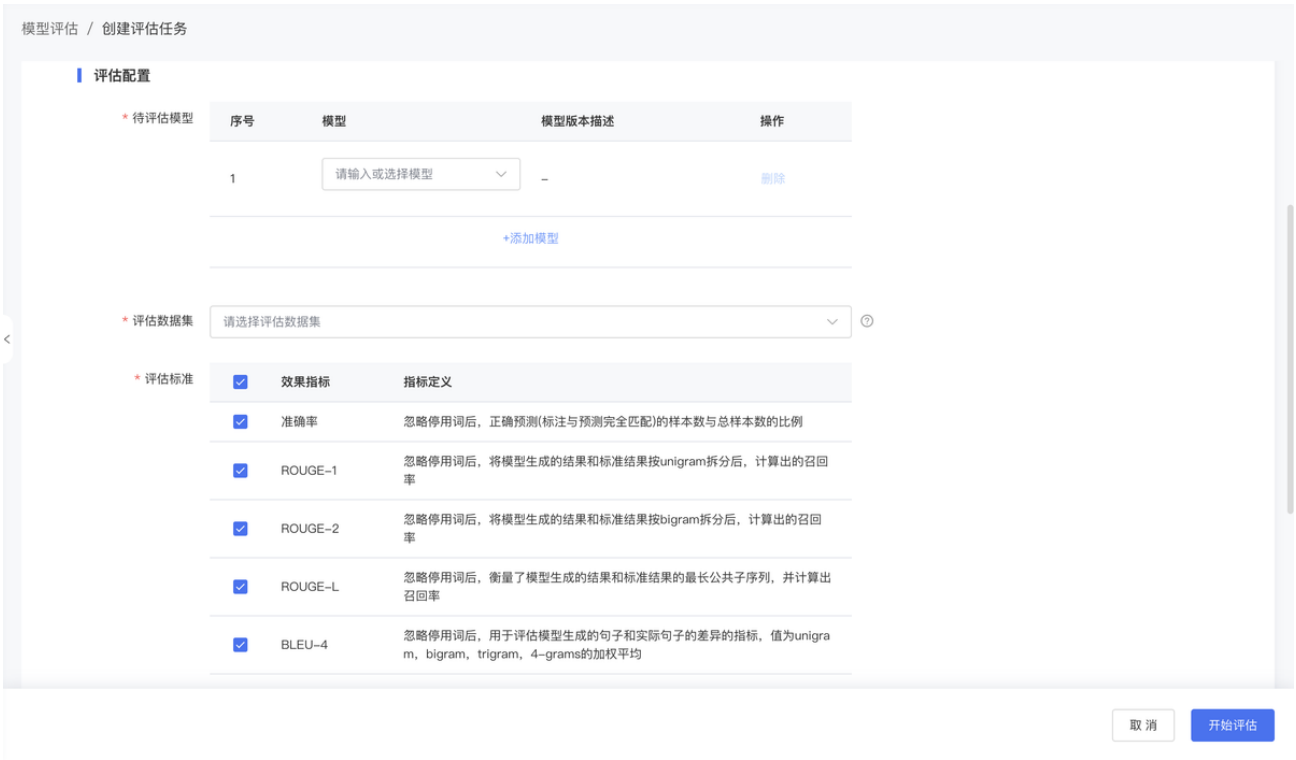
支持对平台预置的模型进行一键精调，可点击模型卡片或模型详情页上的【精调】按钮

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。



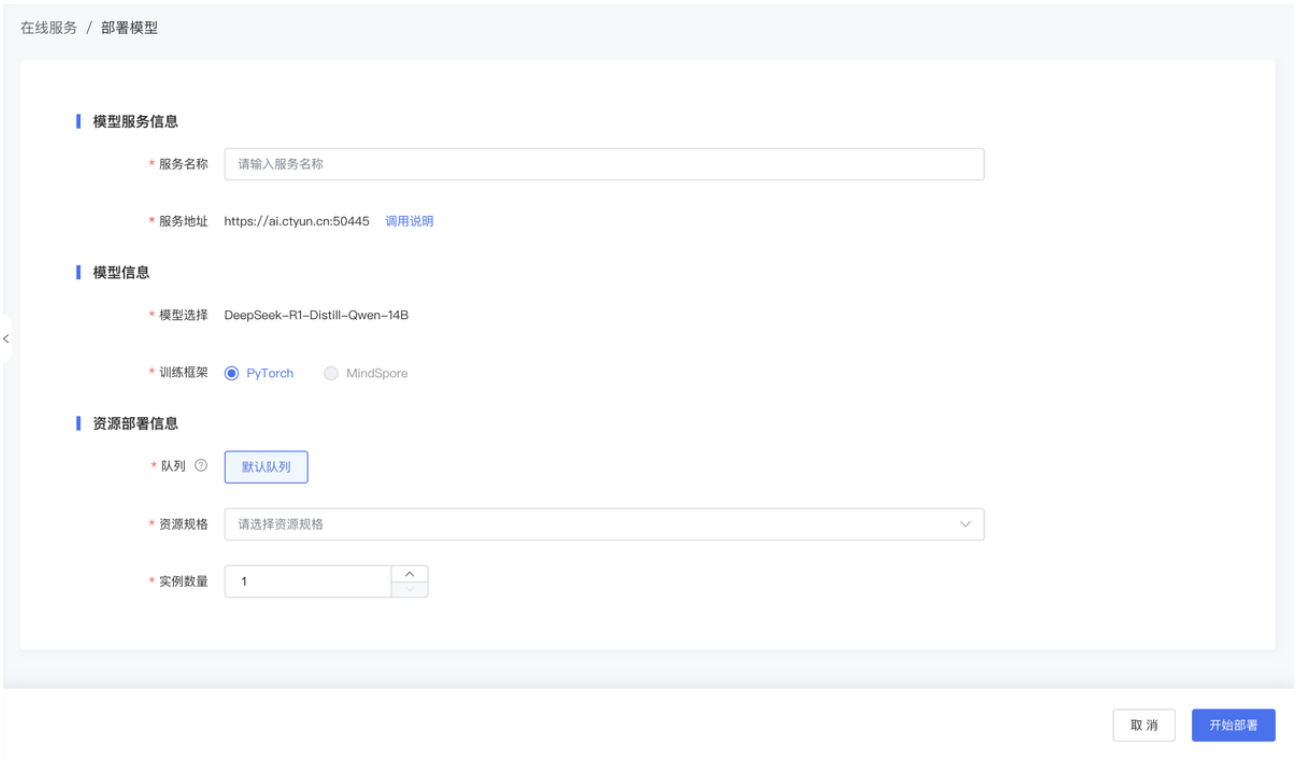
2.2.3一键评估

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。



2.2.4一键部署

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。



2.2.5API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2.4【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。

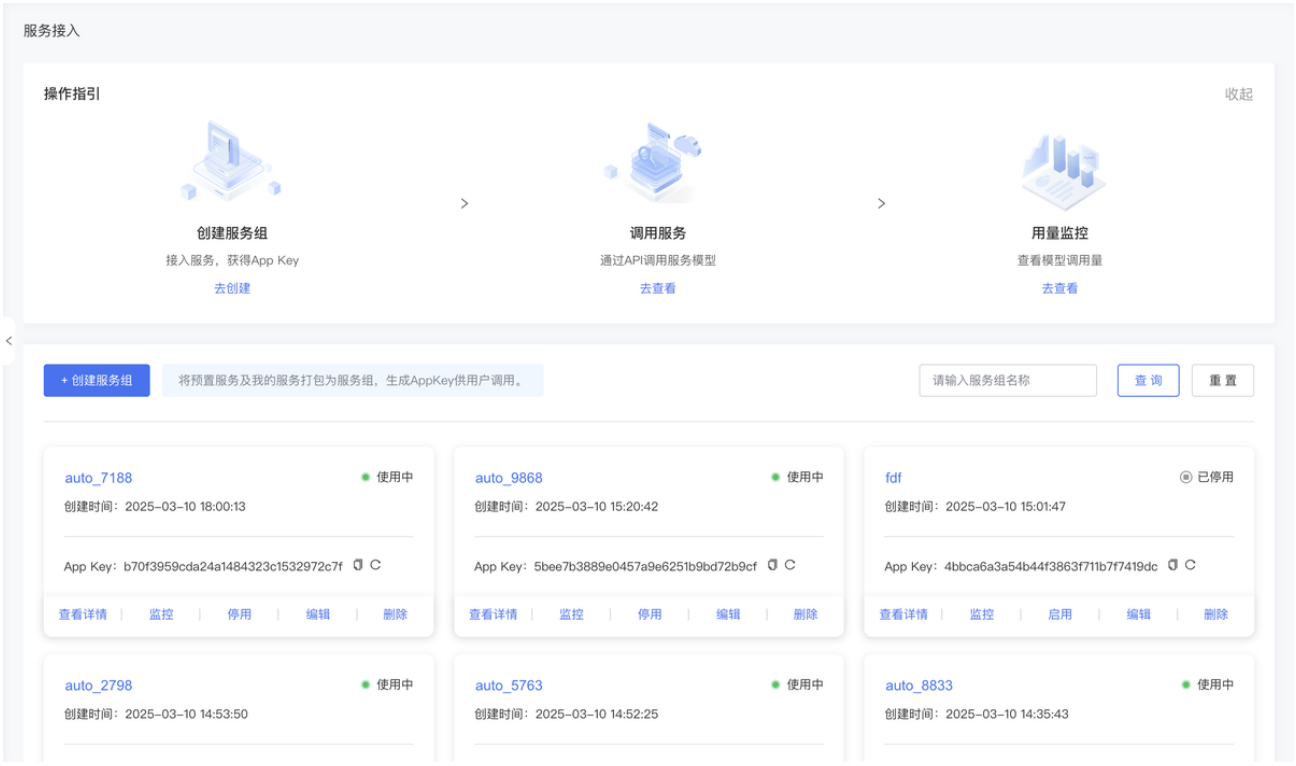
可直接将预置服务中的模型，或部署我的模型后，进行服务接入，创建服务组，获取appkey，以供调用。

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。

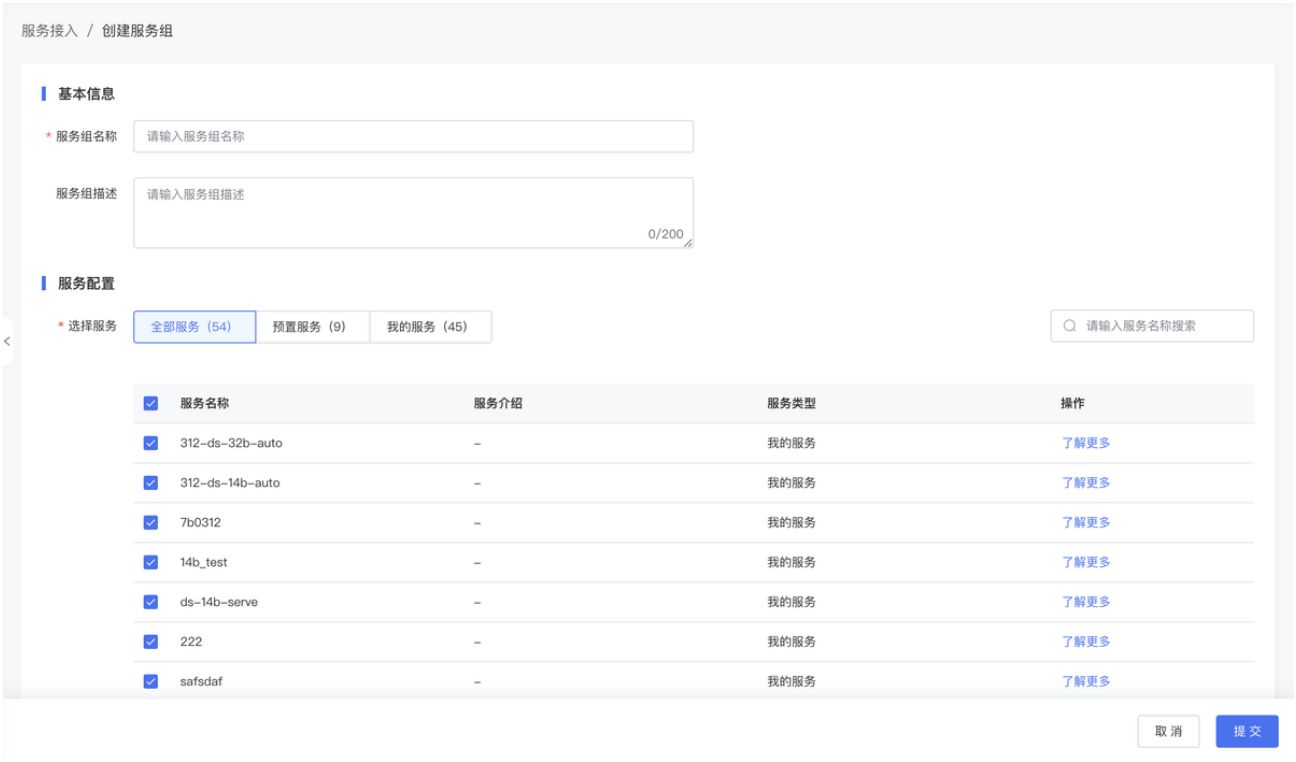
【创建服务】

●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。



●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

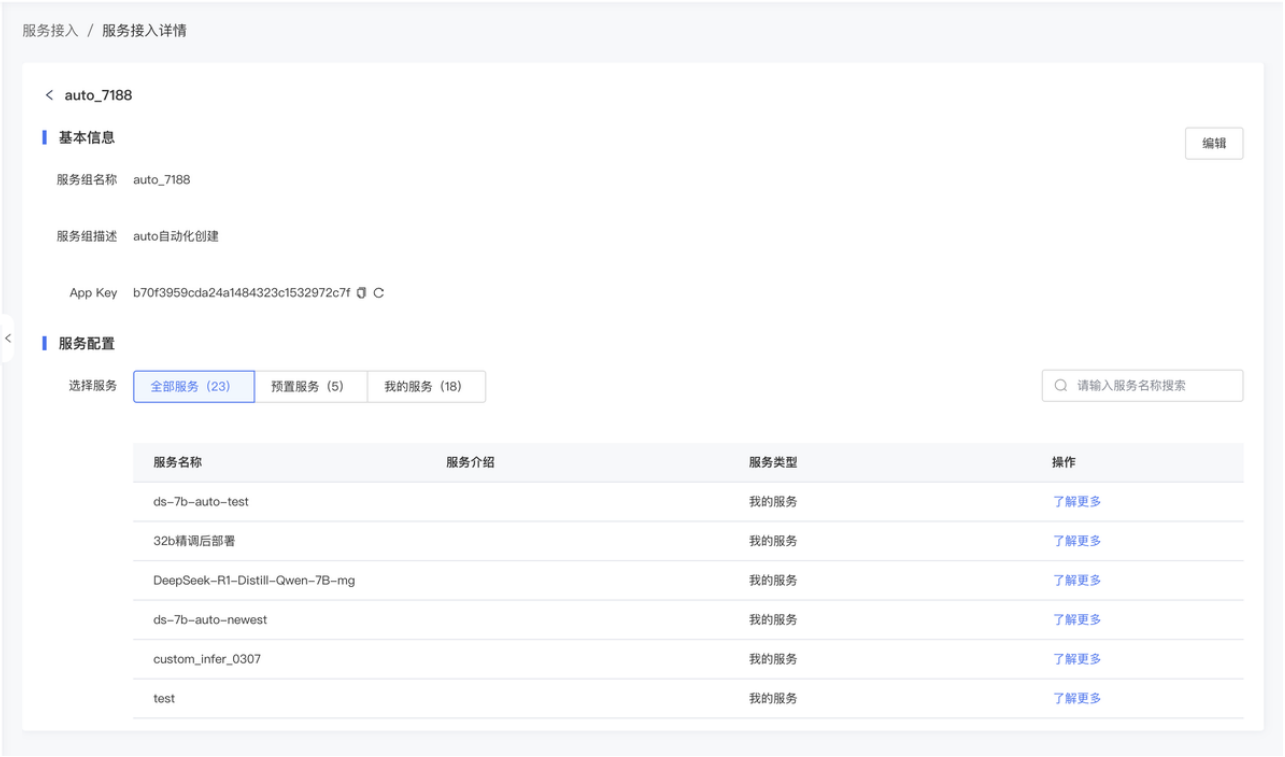
置服务、我的服务），提交完成创建。



●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。



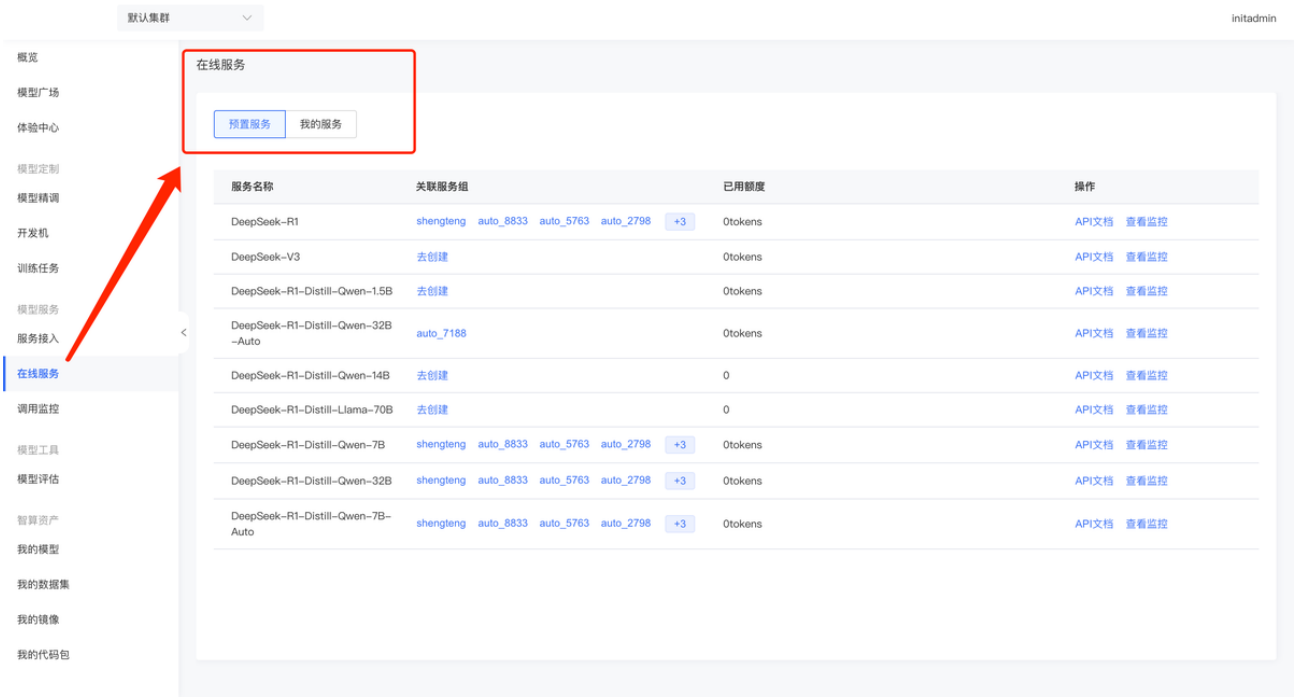
●点击服务组卡片对应按钮，可以支持对“AppKey”进行复制和重置。

●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。

2.5.2在线服务

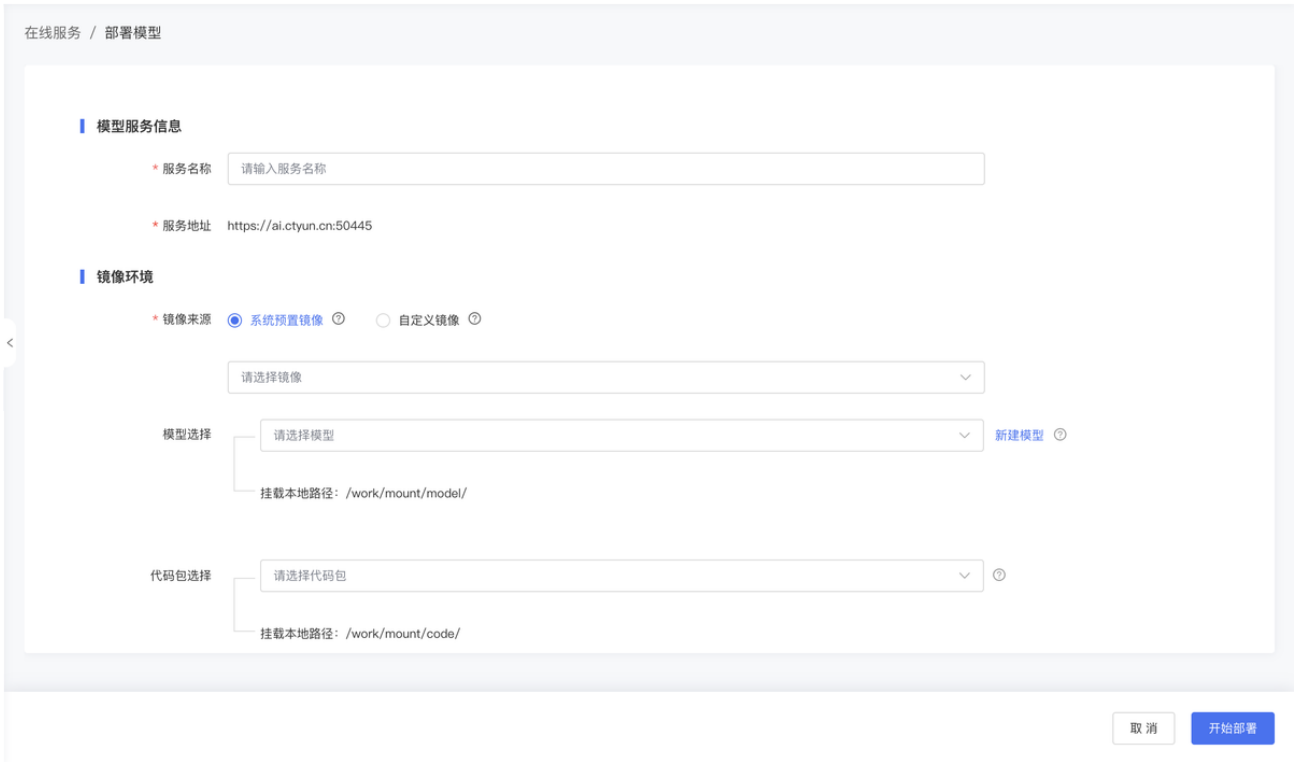
平台支持将用户精调后的模型发布为在线服务，同时也支持直接调用预置模型的在线服

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。



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。

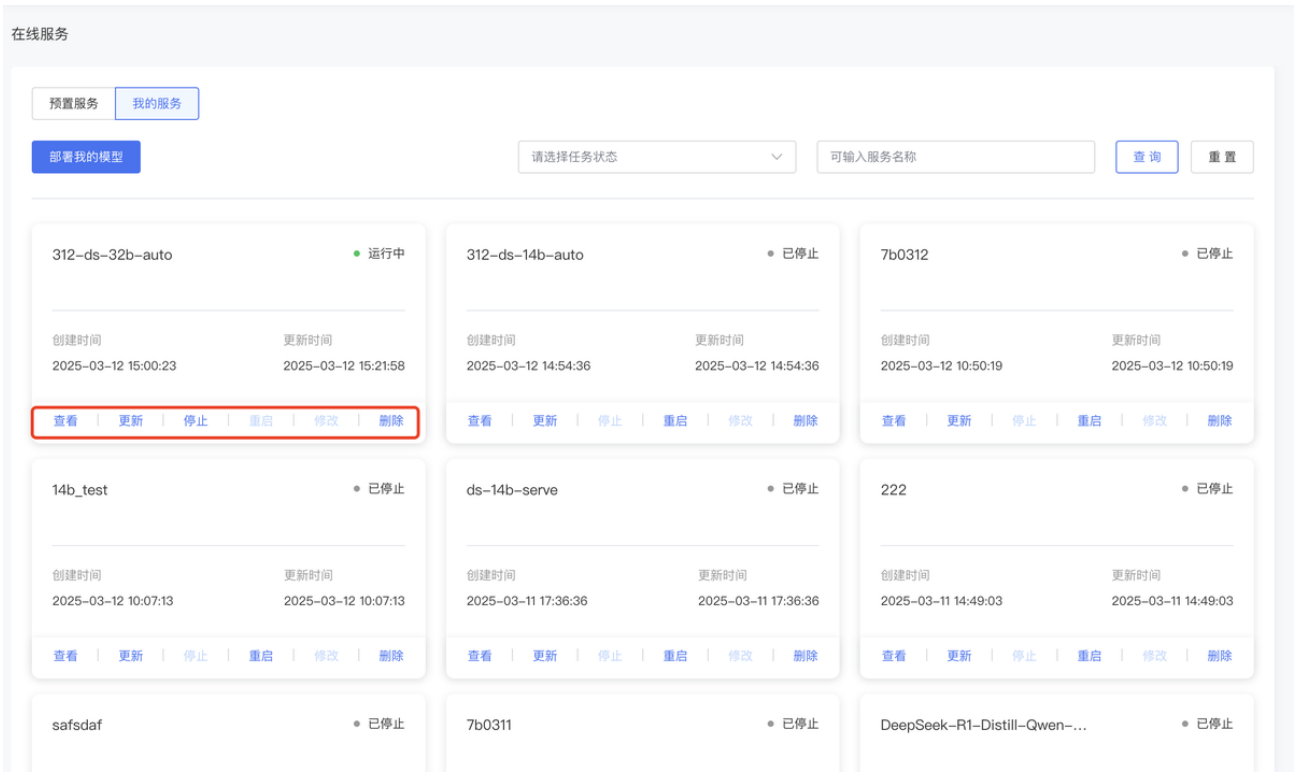
●输入镜像的启动运行命令，如python/mount/code/{codeid}/run.py（须提供OAI兼容的推理服务接口服务）。

●选择资源部署信息，包括队列、资源规格和实例数量。

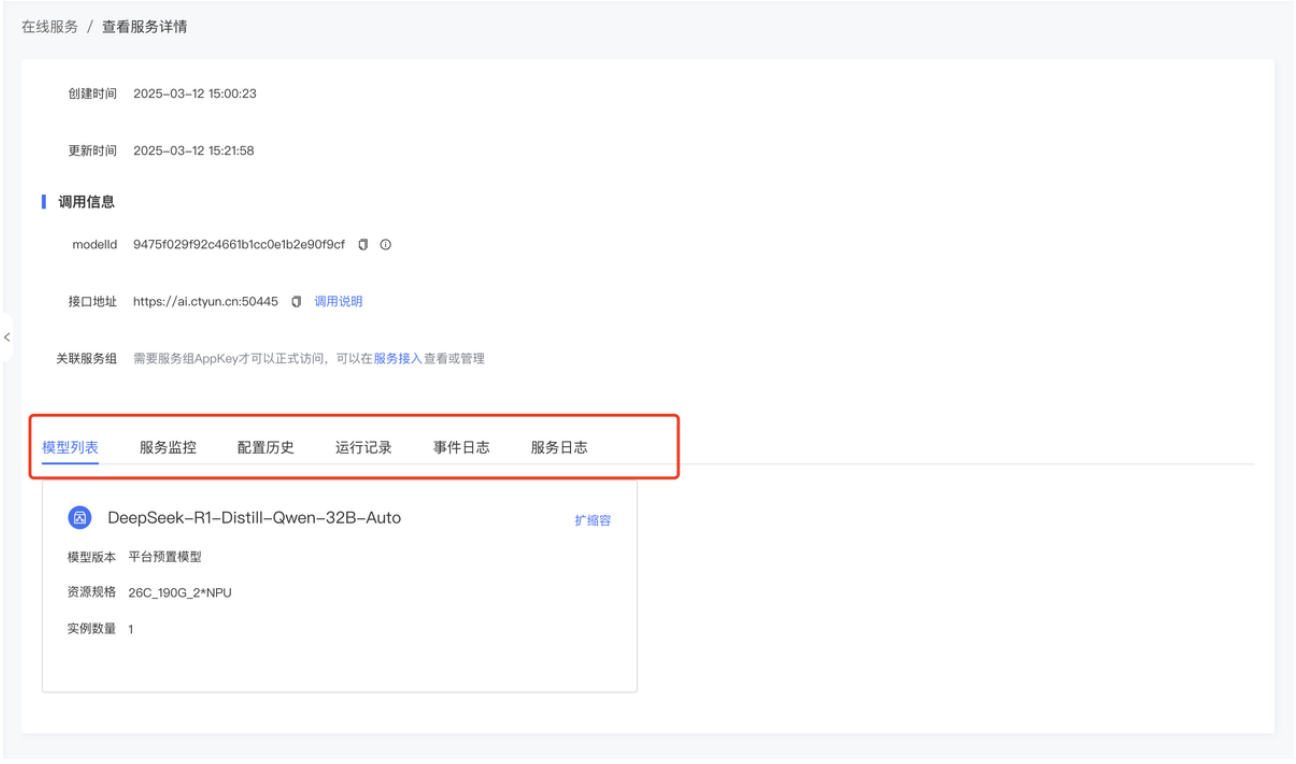
●完成部署，并开始计费。

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。



●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：

```
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。

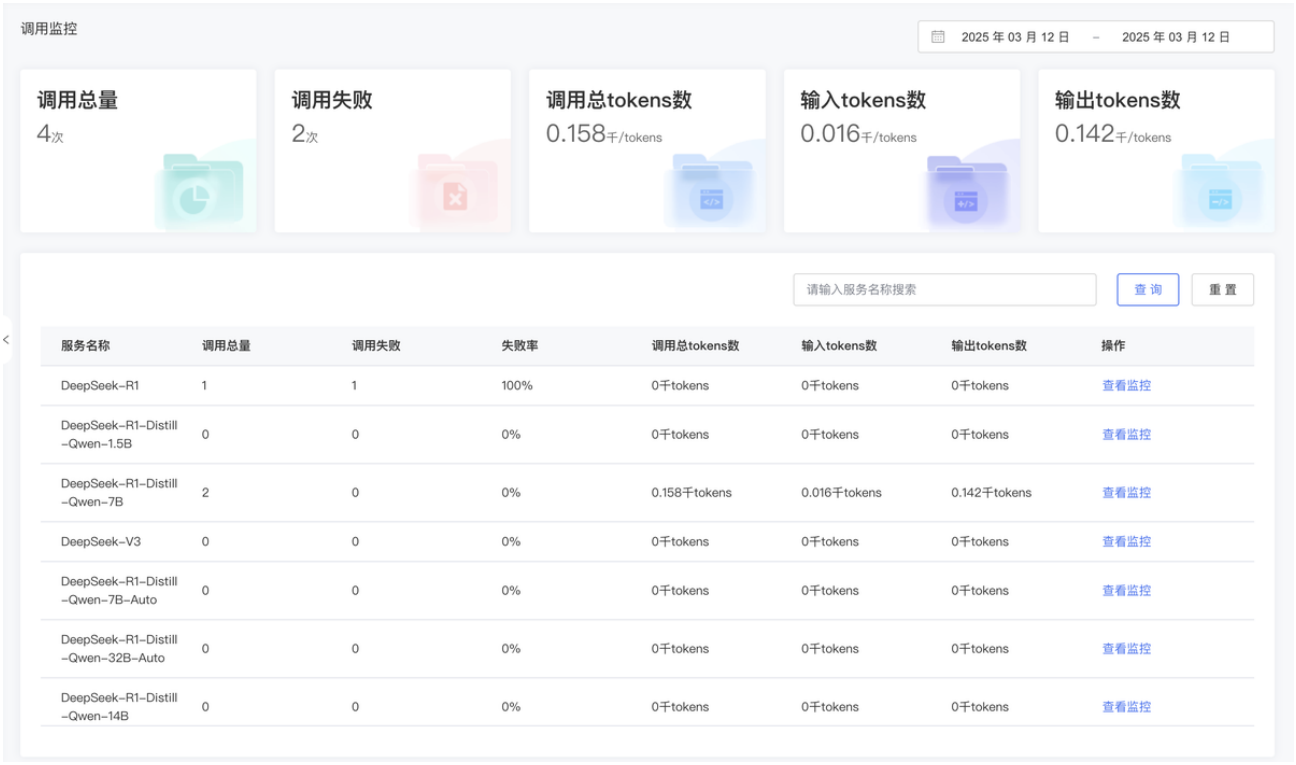
在页面顶部选择统计时间，筛选后，即可查看该时间段内全部服务的调用统计概览，包含调用总量、调用失败量、调用总tokens等指标。

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。



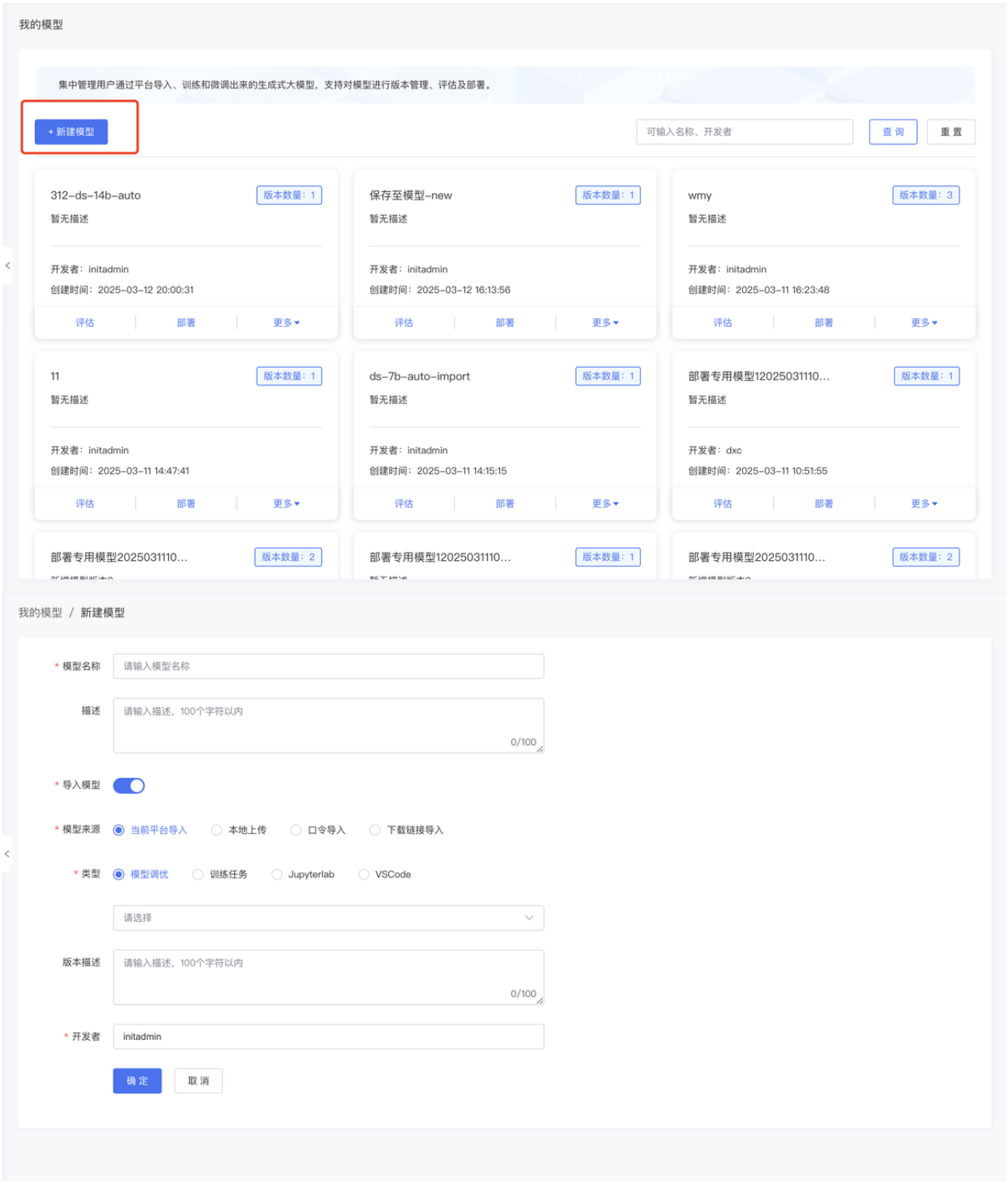
2.5 智算资产

2.5.1 我的模型

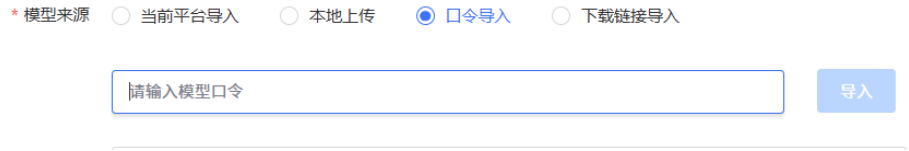
我的模型模块旨在全面管理用户从开发、训练到评估完成的模型生命周期，该模块不仅提供模型文件的安全存储功能，还具备精细化的版本管理，确保每一阶段的模型变更都有迹可循。

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。



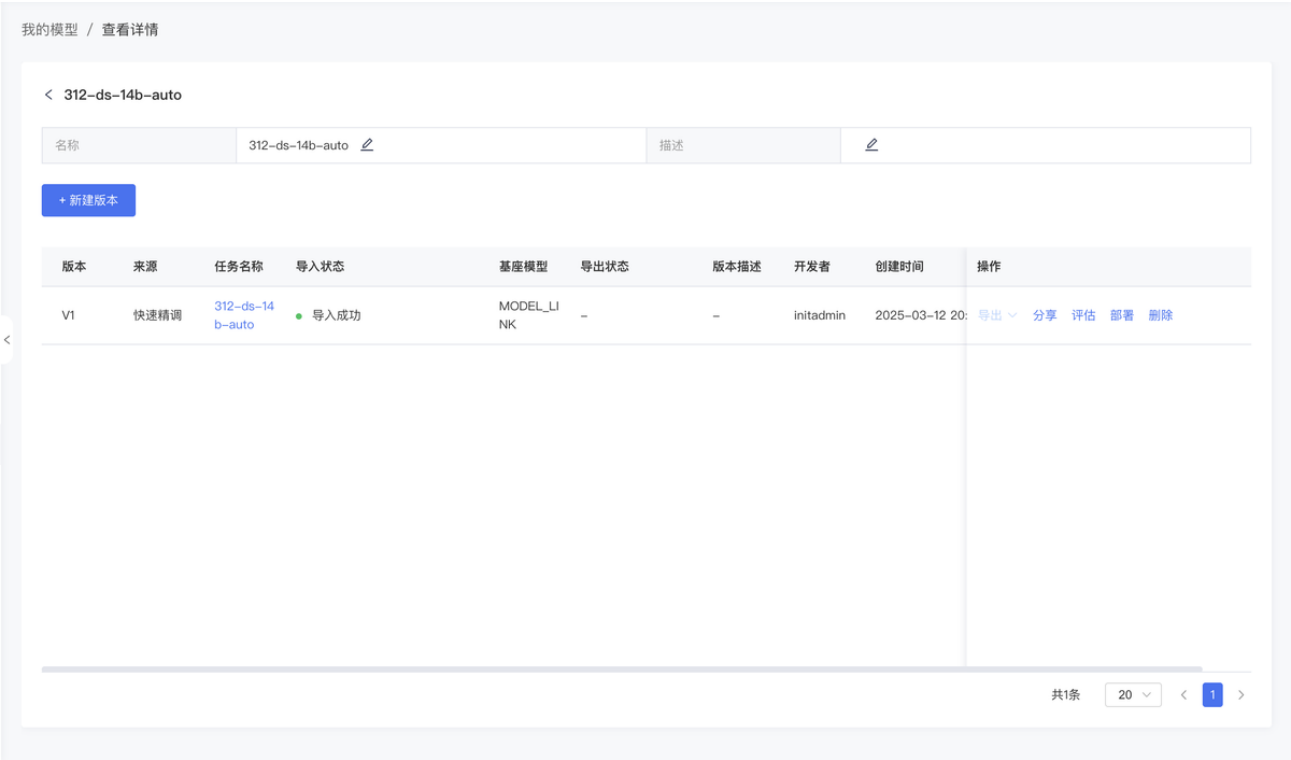
- 当前平台导入：支持从平台上运行完成的模型调优和训练任务中导入、也可以从 JupyterLab 和 VSCode 的目录中导入。
- 本地上传：支持从本地电脑导入不超过 2G 的模型文件。
- 口令导入：支持输入智算服务平台其他账户分享的口令完成导入。



- 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。

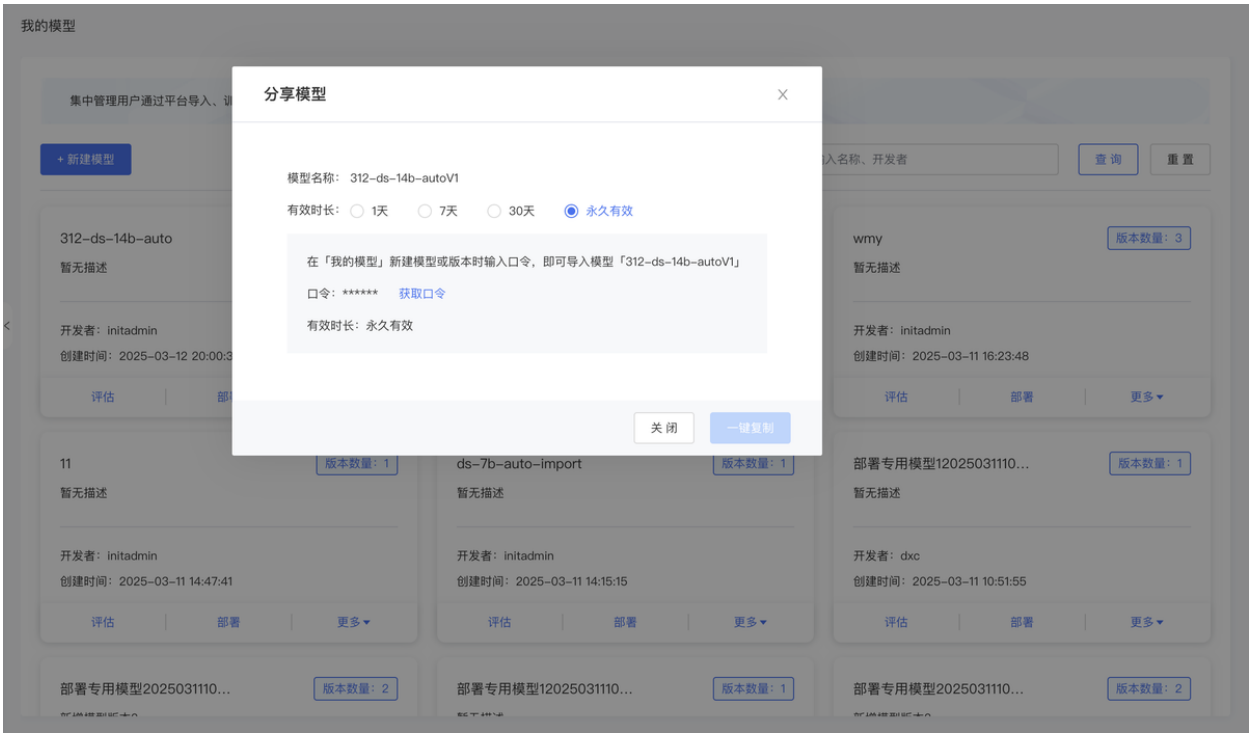


3. 模型分享与导出

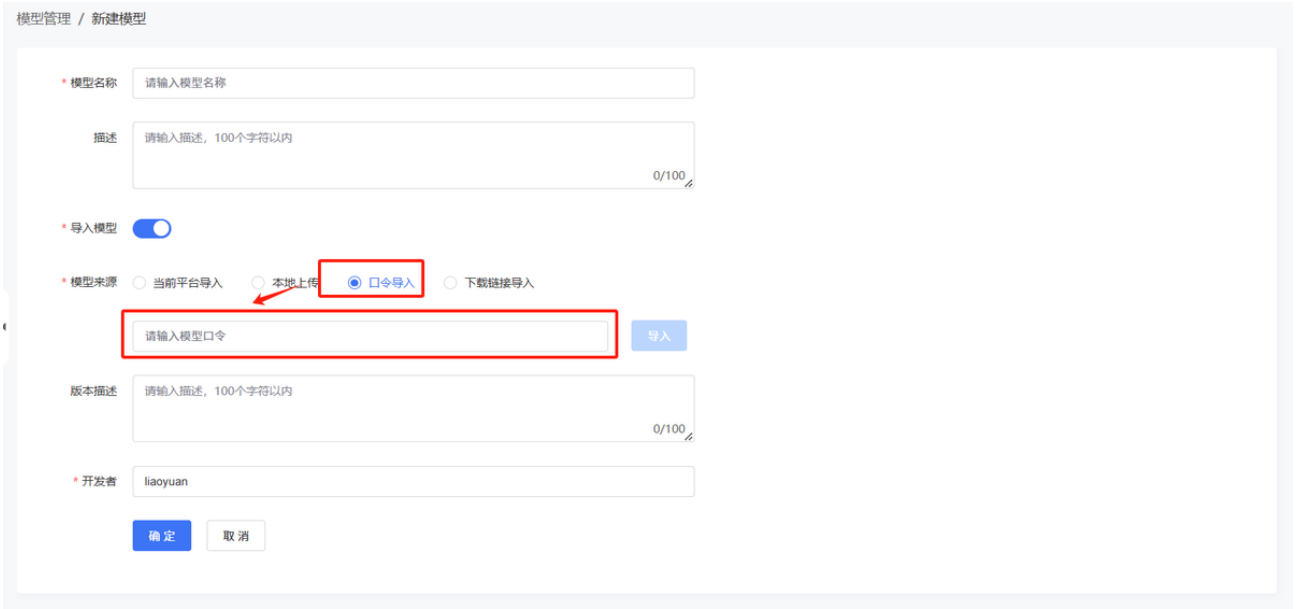
- 模型分享

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。

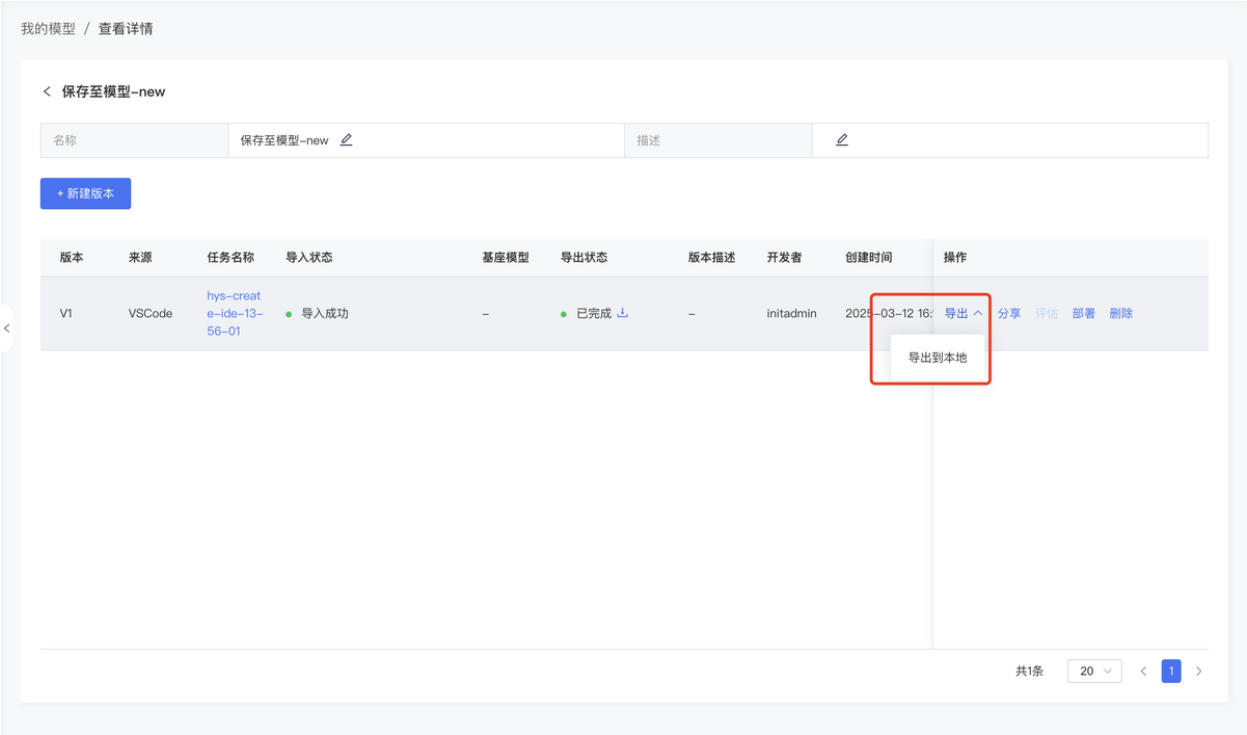


账户2在新建模型中选择【口令导入】，输入账户1给到的分享口令即可完成模型导入。



- 模型导出

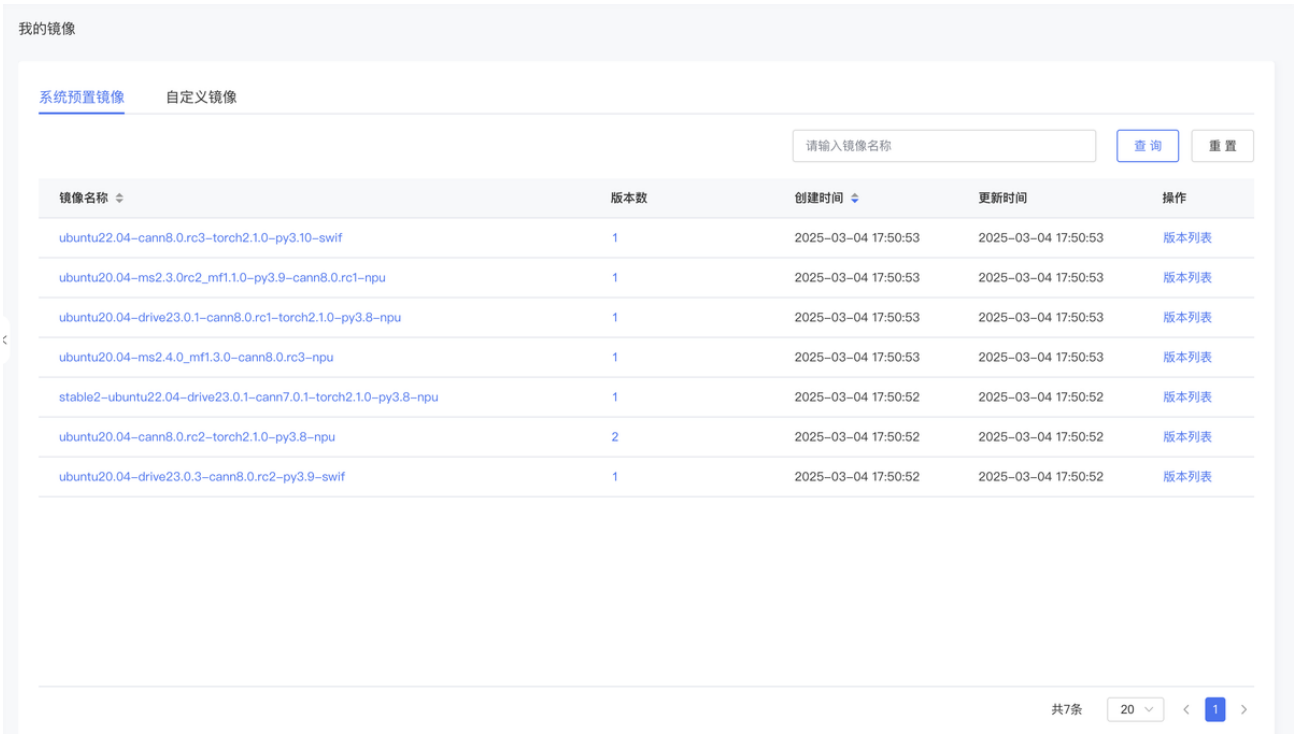
进入模型详情页，在模型版本列表中支持模型导出，可以选择导出到本地。



2.5.2 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。



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。

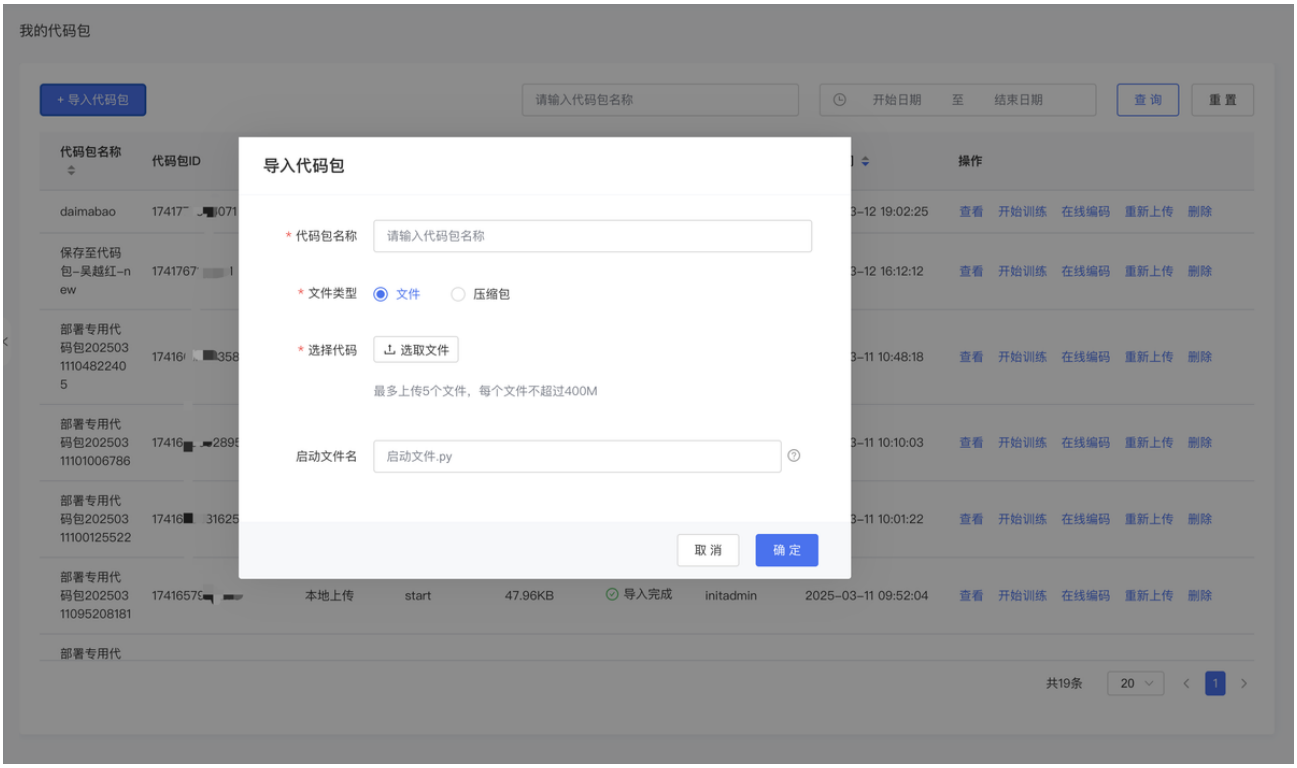
2.5.3 我的代码包

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的文件，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")

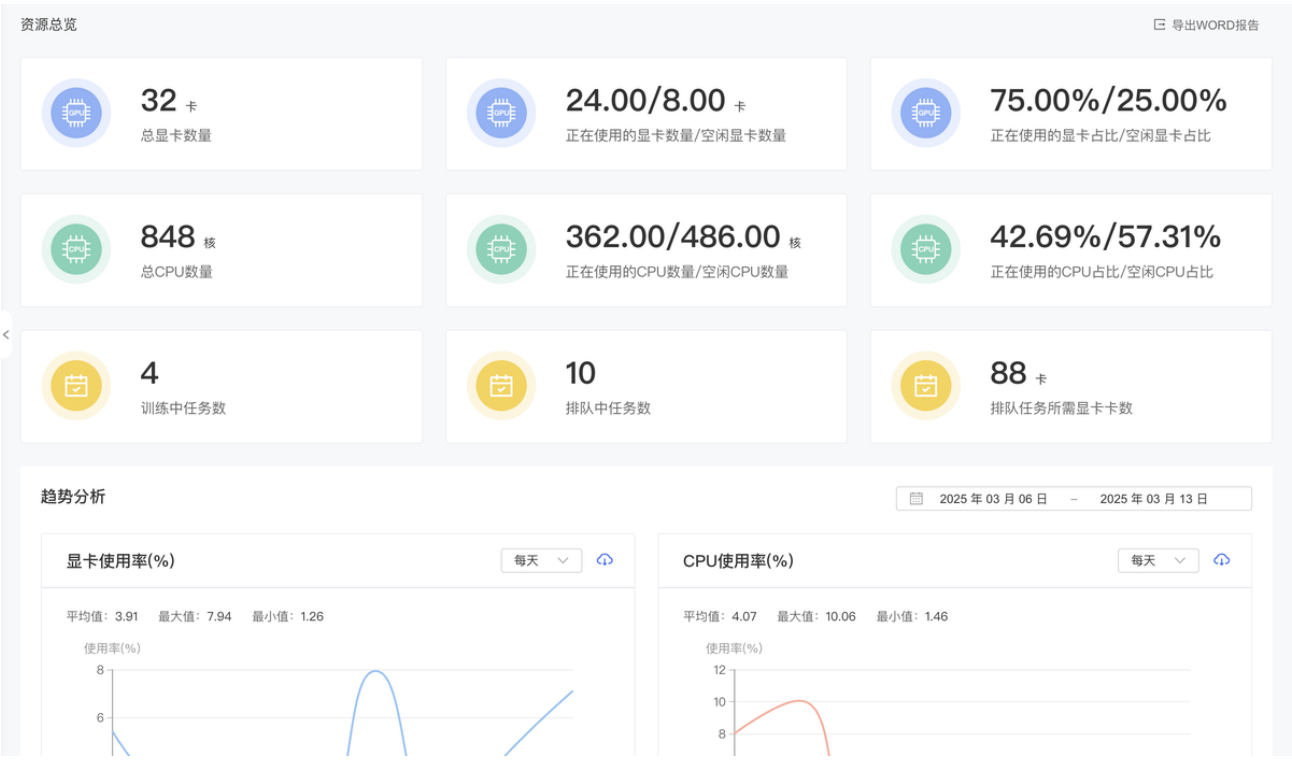
输出脚本所在的目录 ：echo "当前脚本所在目录为： $SCRIPTPATH"

2.6 管理中心

2.6.1 资源总览

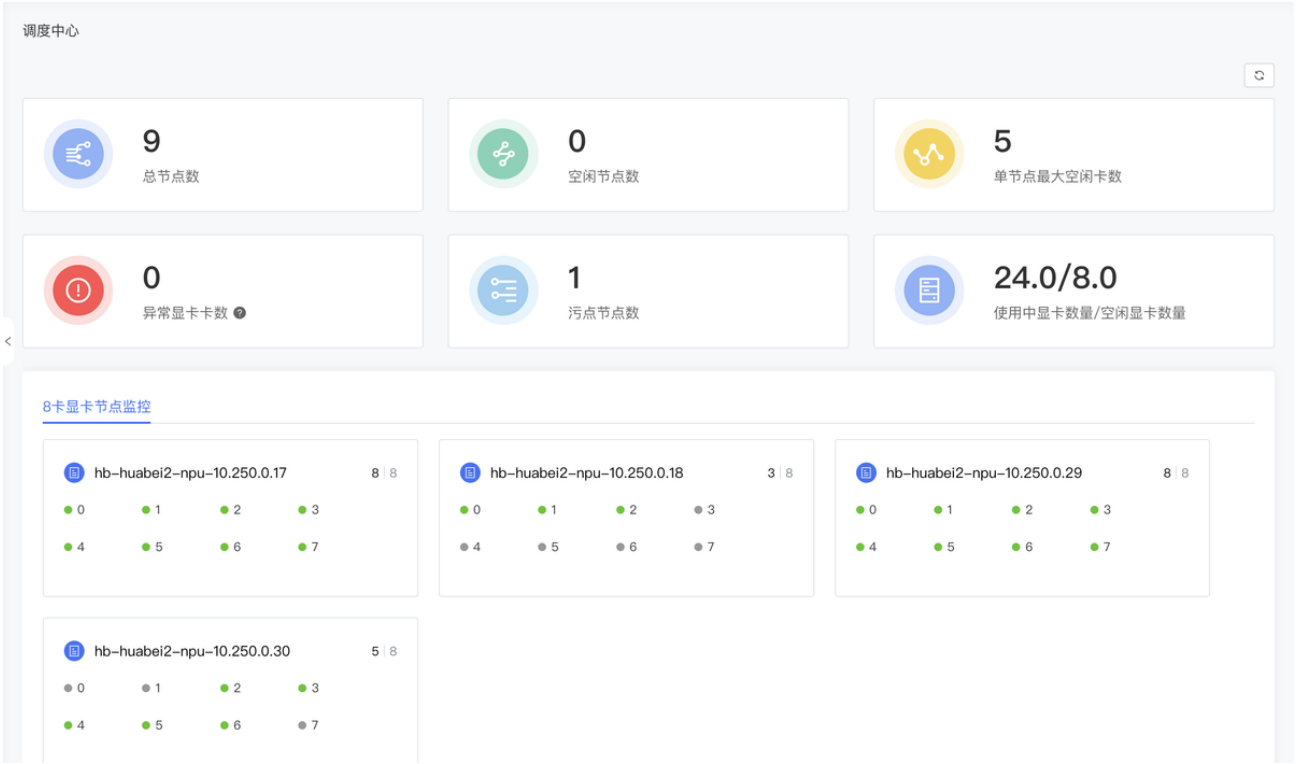
旨在让平台管理员能够轻松查看并管理资源使用情况。

- 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。



2.6.2 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。



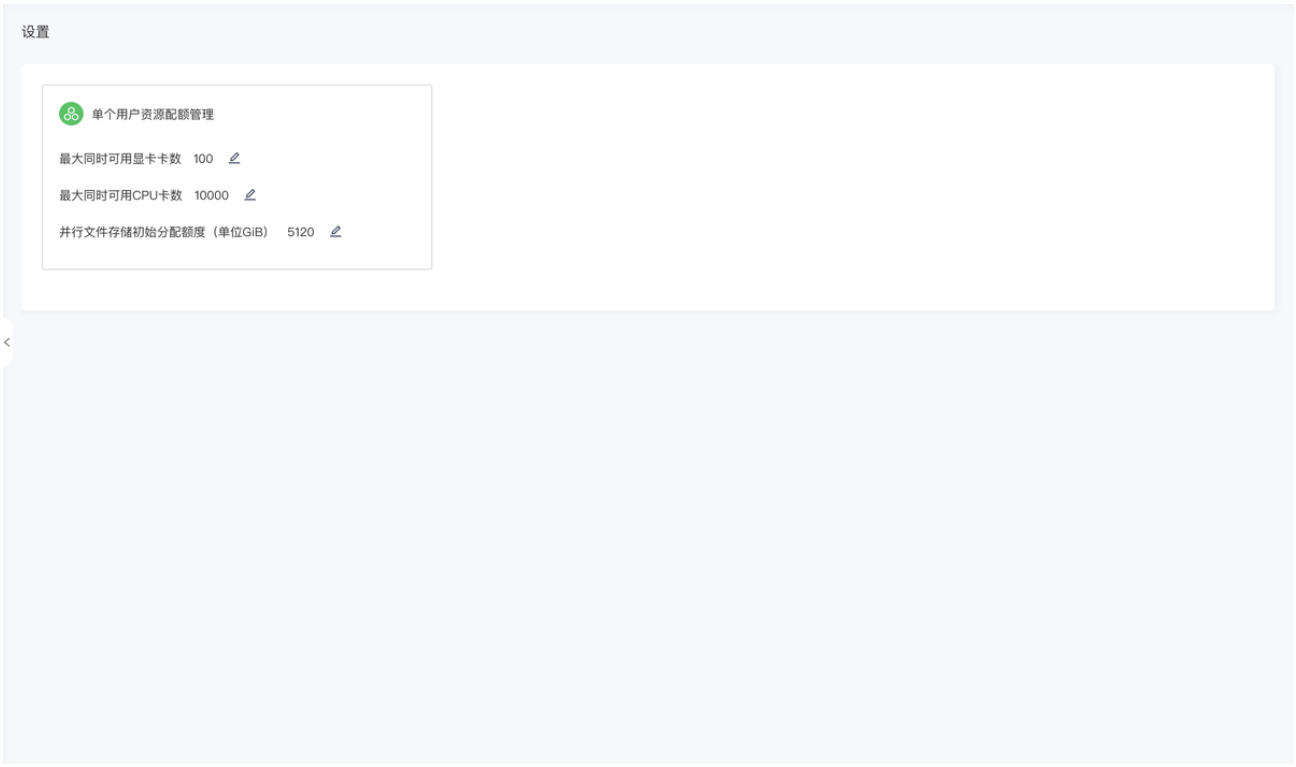
- 进入调度中心模块，监控调度详情页分为节点统计大盘、节点状态监控、节点列表三大板块。

- 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。

2.6.3 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。

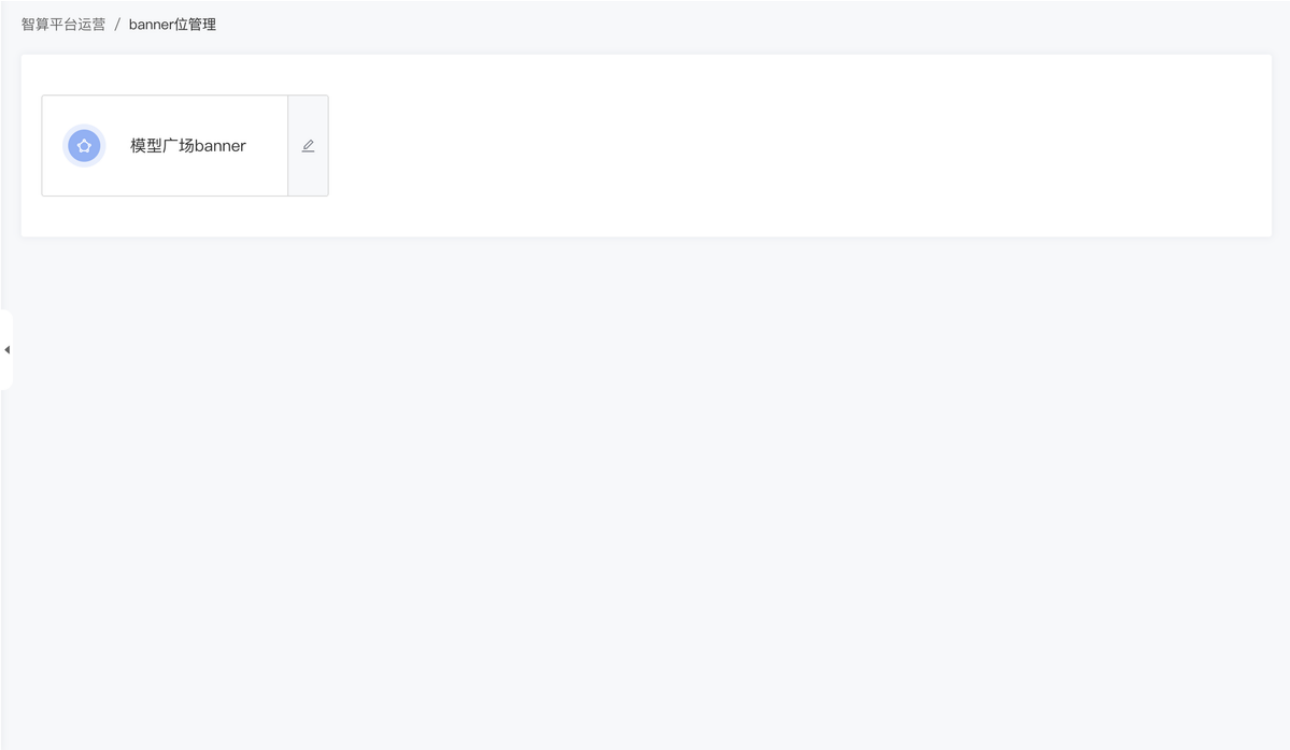


2.7 运营后台

2.7.1 智算平台运营

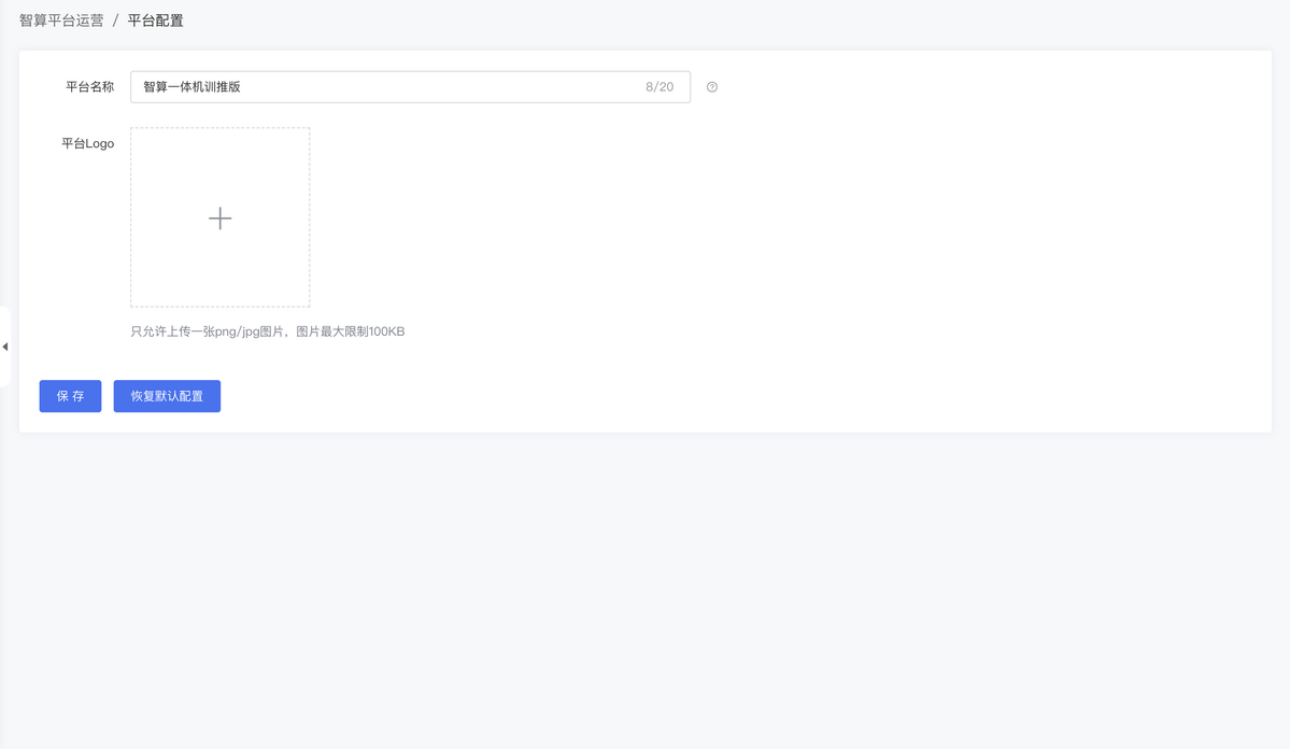
1.banner位管理

用户上传、修改模型广场banner。



2.平台配置

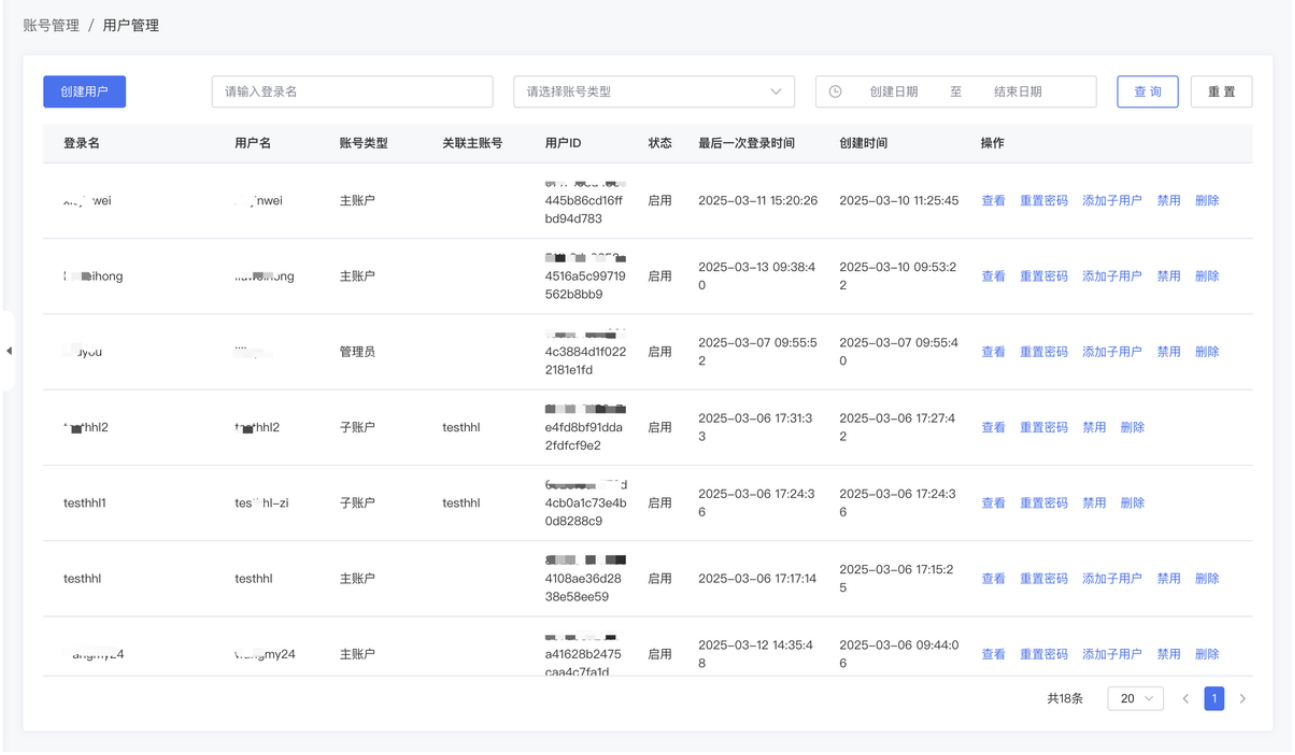
用户可以自定上传logo



2.7.2 账号管理

1. 账号信息纵览

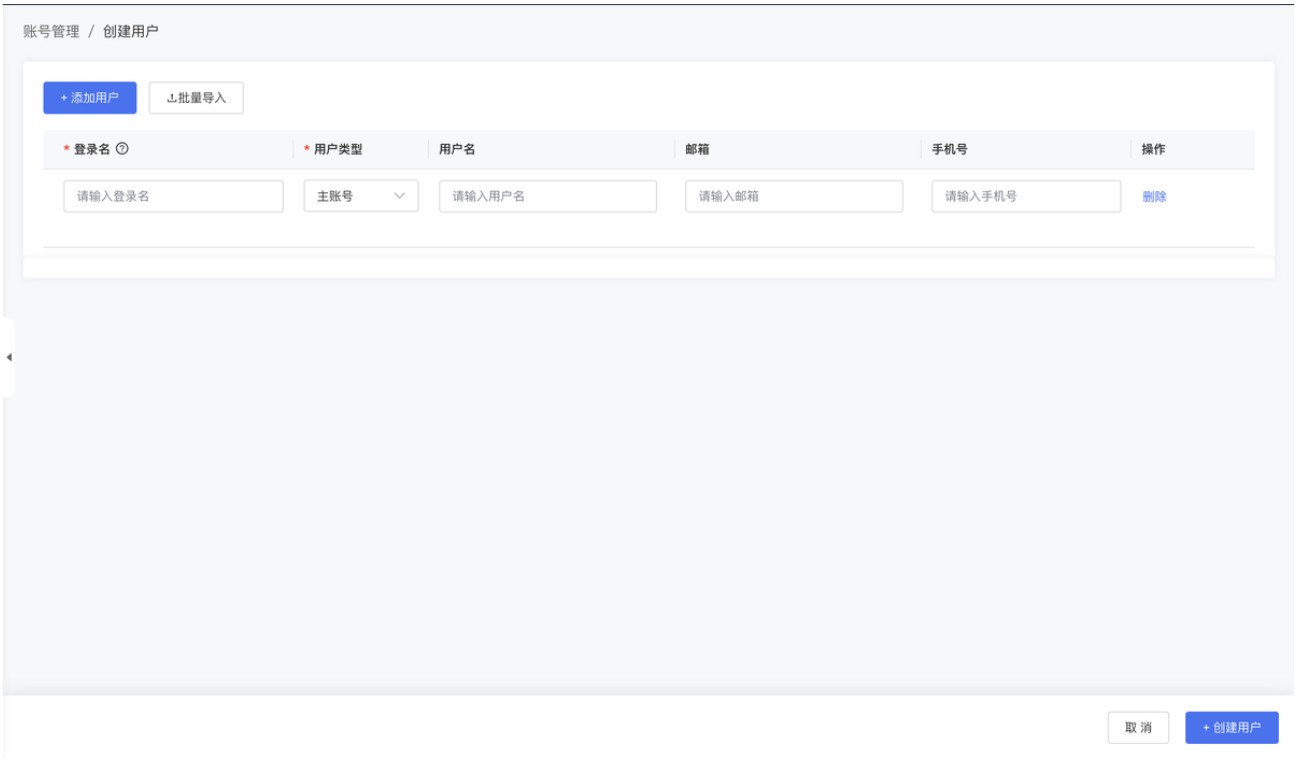
查看登录名、账号类型、注册时间、最近登录时间等信息以及对基本信息作出修改



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
请输入登录名
主账号
请输入用户名
请输入邮箱
请输入手机号
删除
取 消
+ 创建用户

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。