

EATP-标准版用户手册

一、产品概述

1.1 产品介绍

EATP (Enterprise AI Token Platform) 是企业专属 AI Token 运营平台，提供轻量化 Token 运营、模型服务及 GPU 算力部署方案。服务基于 “Token 全链路运营 + GPU 算力服务化” 两大核心模块，可在数小时内将现有异构 GPU 资源转化为具备计量计费、多租户管理与弹性扩缩能力的商业化 AI 推理平台 实现以最小化前置投入实现算力资产的高效变现与 AI 业务的敏捷上线。

1.2 产品核心能力

EATP 构建了以 AI 网关、推理网关、多引擎推理加速(vLLM/SGLang)及 Orchestrator 调度器为核心的统一服务平面。通过抽象底层异构算力环境，实现对 GPU 裸金属、GPU 云主机等 IaaS 资源的统一纳管与精细化调度 为上层应用提供标准化的推理接口与运营管理视图。

关键能力包含：

- **一键式推理部署**：支持标准化推理服务在 1 至 2 小时内完成搭建，服务接口自动兼容 OpenAI API 规范。
- **轻量化 Token 运营体系**：内置完整的 API Key 管理、配额控制、多租户隔离与用量可视化统计功能，原生支持按 Token 维度的精细化计量计费，为企业构建可商业变现的 AI 服务体系提供基础运营工具。
- **异构算力兼容**：广泛兼容多种主流 GPU 型号及计算实例形态。
- **自动化低门槛运维**：平台提供从模型加载、实例编排到弹性扩缩容的全流程自动化管控能力，有效屏蔽容器化部署与分布式调度的技术复杂性，普通运维人员即可完成日常管理与维护。

1.3 产品优势

- **快速交付**：EATP 套件以标准化产品形态，将传统自建方案 3 至 6 个月的交付周期压缩至 2 至 3 周，大幅缩短 AI 推理服务上线时间，帮助企业快速抢占业务窗口。
- **提升资源效能**：通过精细化 GPU 调度与推理加速引擎，EATP 将 GPU 集群平均利用率从传统模式的 30%~50%提升至 70%~90%，有效摊薄单位算力成本；引入自动化运维机制后，日常运维人力投入降低 60%以上，在 AI 业务规模化过程中实现显著降本增效。
- **安全合规**：依托天翼云安全可信的基础设施底座，EATP 套件提供兼顾合规遵从、性能极致与商业敏捷的 AI Token 全生命周期管理方案，确保企业 AI 服务安全、高效、可持续运营。

1.4 选型建议

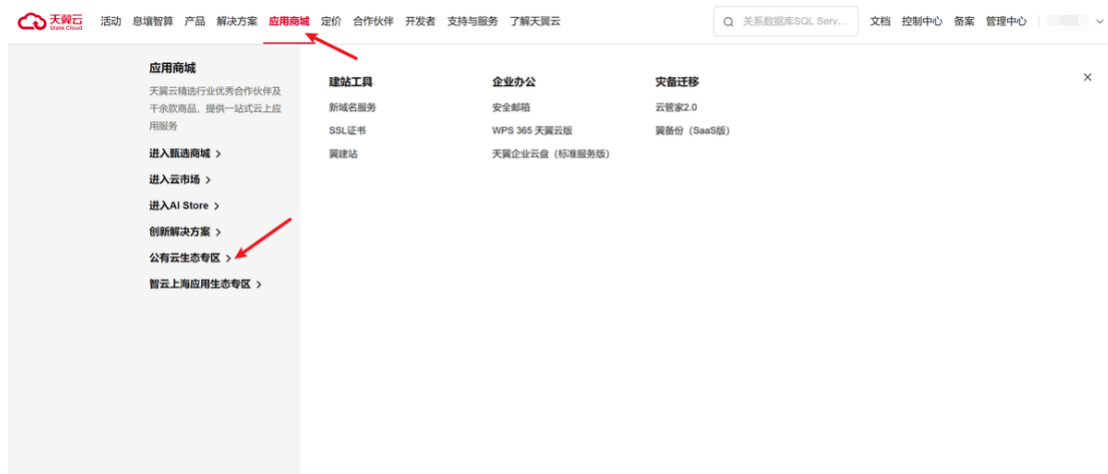
EATP 提供 **GPU 推理版** 和 **标准版**两个版本：

- **GPU 推理版**：自带 GPU 卡，开箱即用，支持单机部署模型推理，适合个人开发者、小型团队或快速验证（POC）场景，省去资源采购和运维成本；
- **标准版**：基于 CPU 云主机提供平台能力（含 Token 运营、模型部署、GPU 管理），但不含 GPU 推理服务，推理算力需用户自建或单独购买，适用于已有 GPU 资源池或需要生产级弹性扩展、高可用架构的中大型团队。

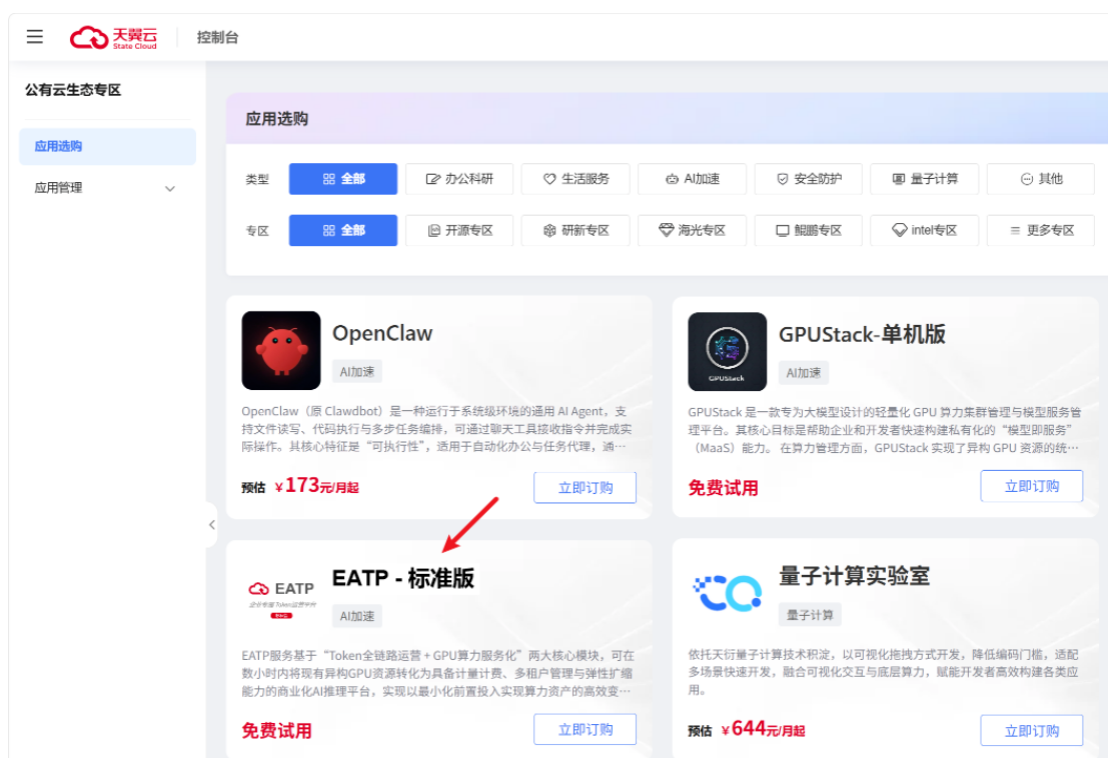
二、使用指南

2.1 购买云资源并部署 EATP 服务

1. 登录天翼云官网，选择【应用商城】-【[公有云生态专区](#)】，点击“立即选购”，进入应用专区页。



2. 选择“EATP-标准版”应用，点击立即订购。



3. 根据页面提示填写并确认“EATP-标准版本”订购信息，点击确认下单，支付后即可成功购买云资源及 EATP 应用服务。

三

天翼云
State Cloud

控制台

< 开通应用

加载应用

EATP - 标准版

EATP

Enterprise AI Token Platform

EATP (Enterprise AI Token Platform) 是企业专属 AI Token 运营平台，提供轻量化Token运营、模型服务及GPU算力部署方案

版本号: v1.0.0

详情

资源池

华北2

规格套餐

基础型

入门型

此套餐包含

EATP服务

规格名称

s5e 2xlarge.2

vCPU

8 核

内存

16 GB

系统盘

200 GB SSD-generic

带宽

10 Mbps

计费方式

包年包月

应用别名

EATP - 单机版-y2hw

13 / 30

实例名称

App-EATP-S-9970

15 / 53

实例密码

请输入实例密码

购买时长

1个月

云资源价格

¥

元

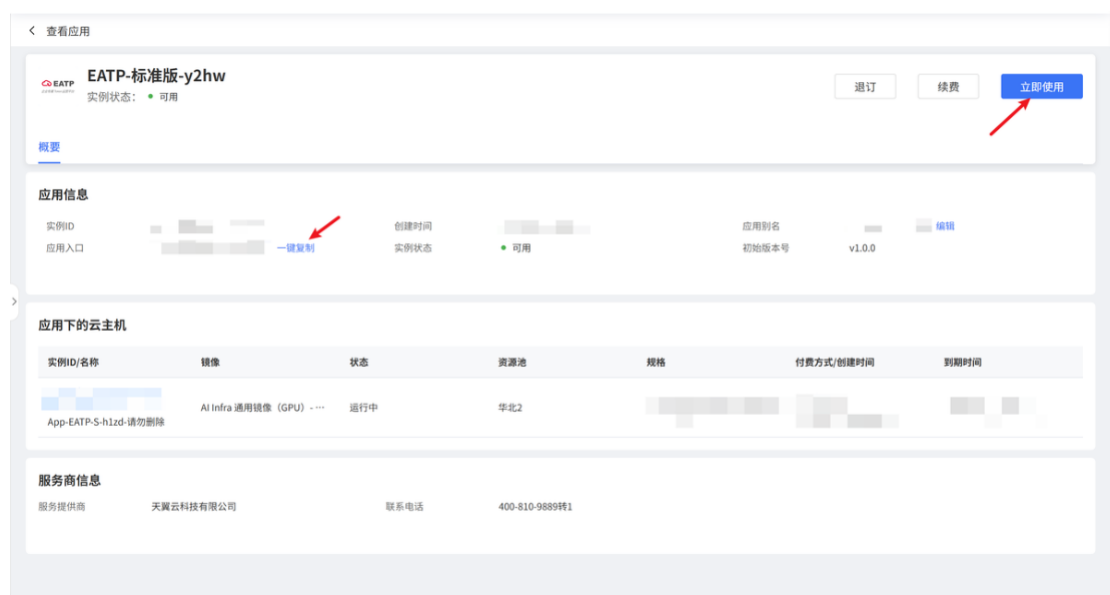
应用价格

免费

4. 返回【[我的应用](#)】页，查看应用状态，当应用状态为运行中时，代表服务部署完成。

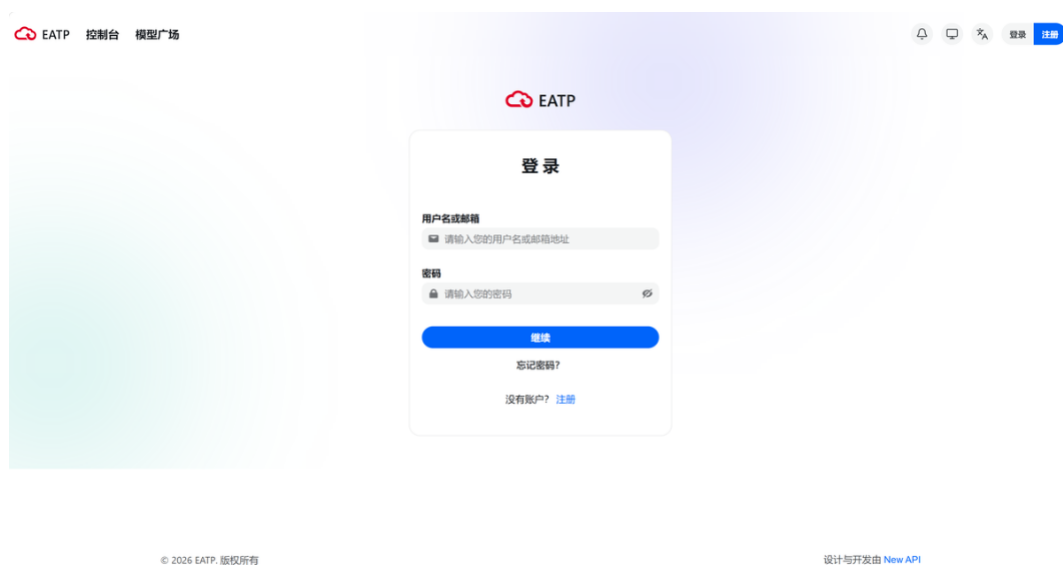


5. 点击【立即使用】或在浏览器中输入复制的应用入口，即可访问 EATP。



2.2 登录 EATP 服务

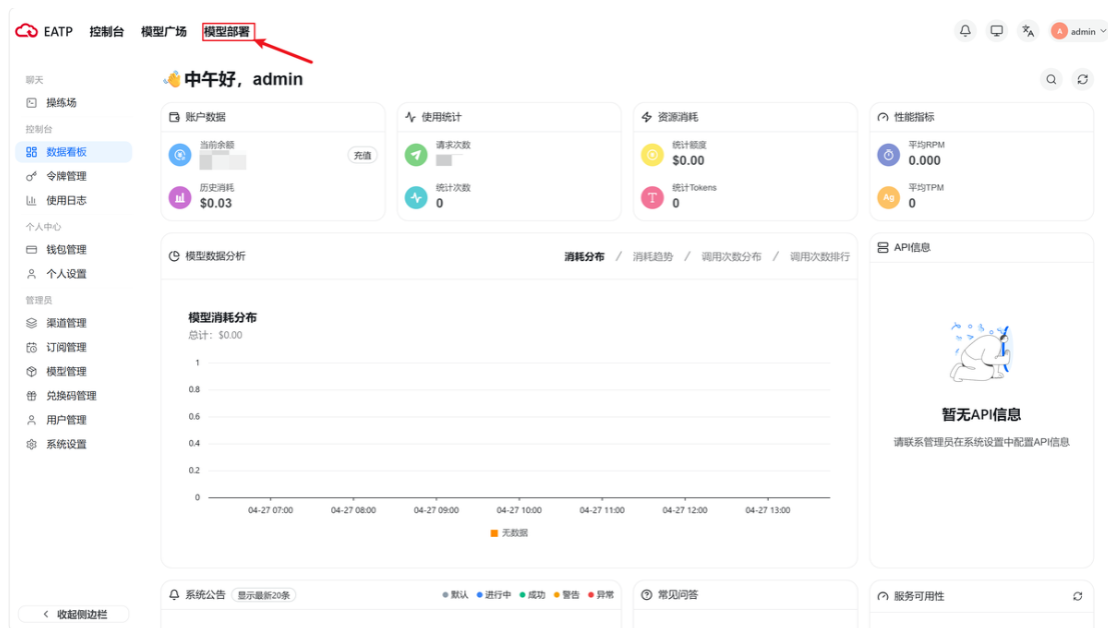
1. 登录 EATP。首次登录 EATP 默认用户名为 :admin ,密码为 :Eatp123@。(登录后 , 应用中可修改账号用户名及密码)



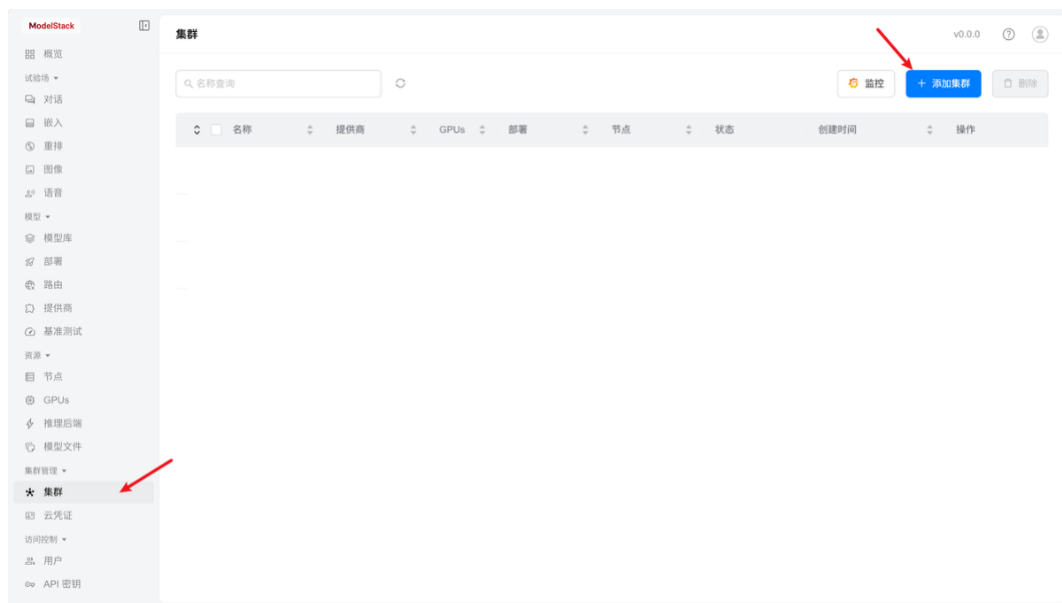
2.3 利用 EATP 添加 GPU 集群并部署模型

若您不需要模型部署、GPU 管理能力，可从 2.5 开始配置服务。

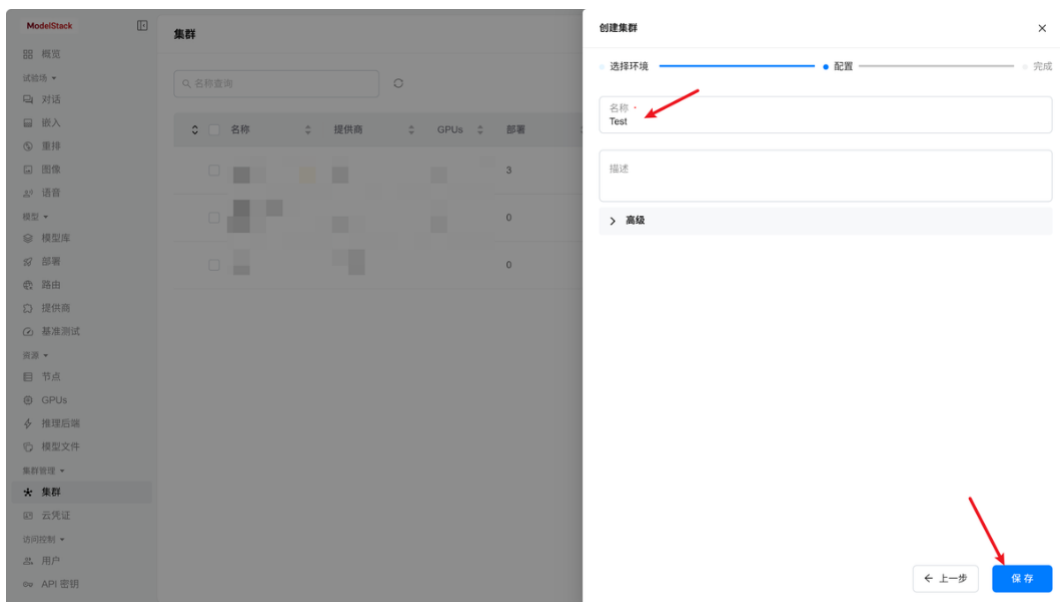
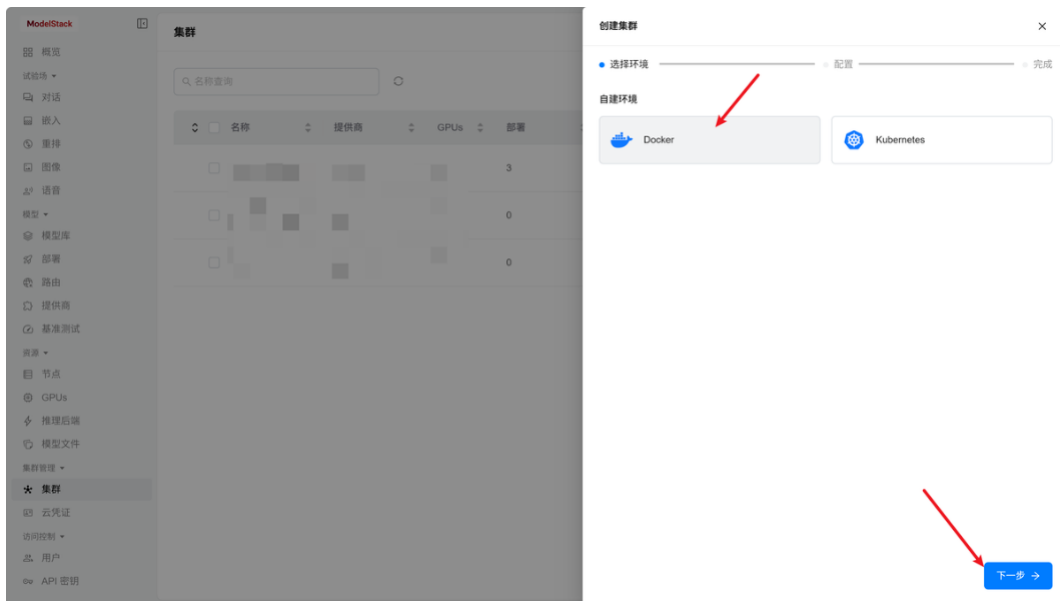
1. 登录后，在顶部导航中点击【模型部署】，从统一入口进入 ModelStack。



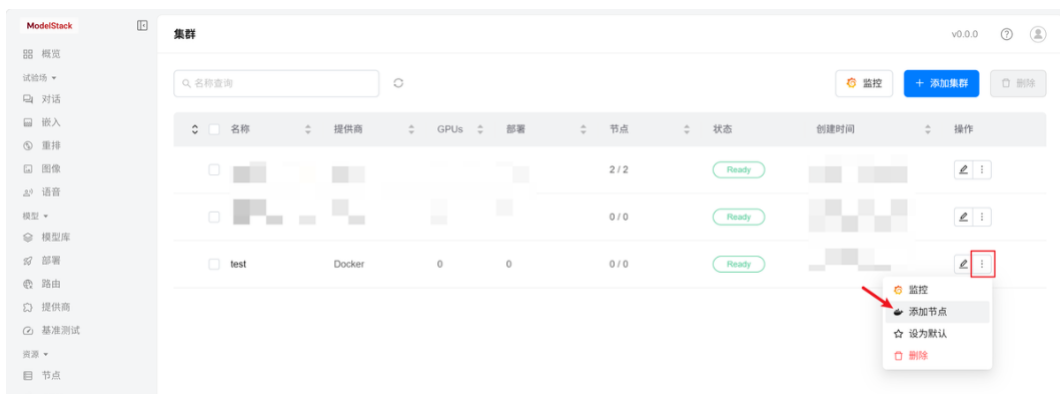
2. 进入ModelStack后，选择左侧导航【集群】，点击【添加集群】按钮，准备在ModelStack 中创建用于推理的集群。



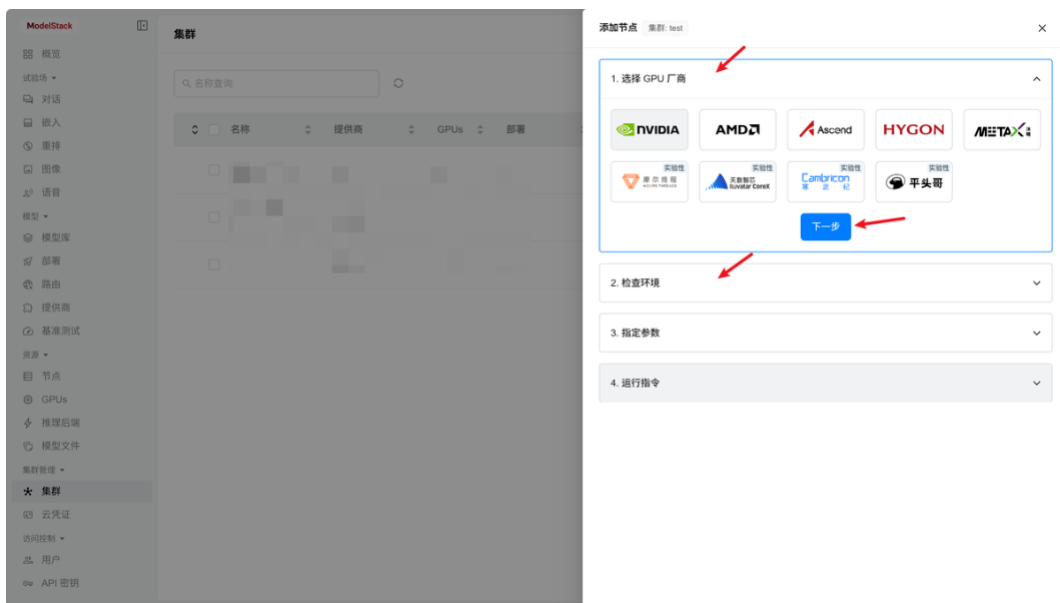
3. 创建集群时，选择“Docker”环境，并为集群命名，保存后即完成集群创建。



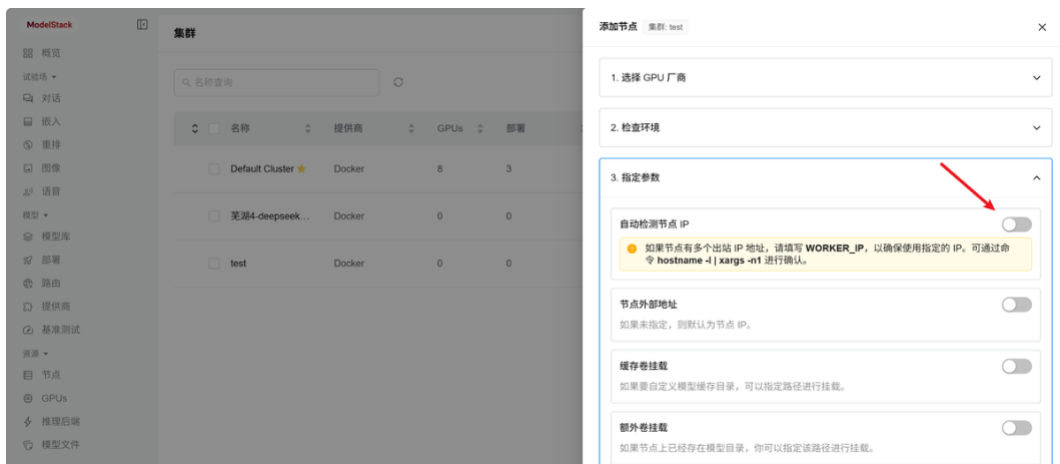
4. 为集群添加节点。您可以将需要用于推理的 GPU 节点信息添加至集群中，点击【添加节点】按钮。



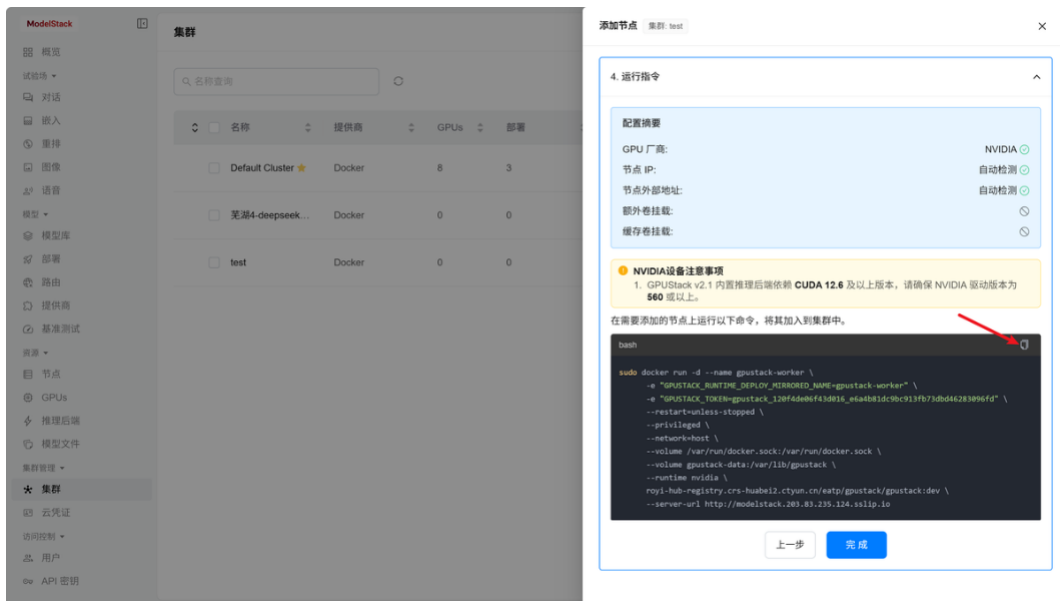
根据您的 GPU 设备型号选择对应的 GPU 厂商，并完成环境检查。



注意，指定参数配置中，请关闭手动配置 IP，选择自动检测节点 IP。



复制节点启动命令，并在登录 GPU 设备后执行。

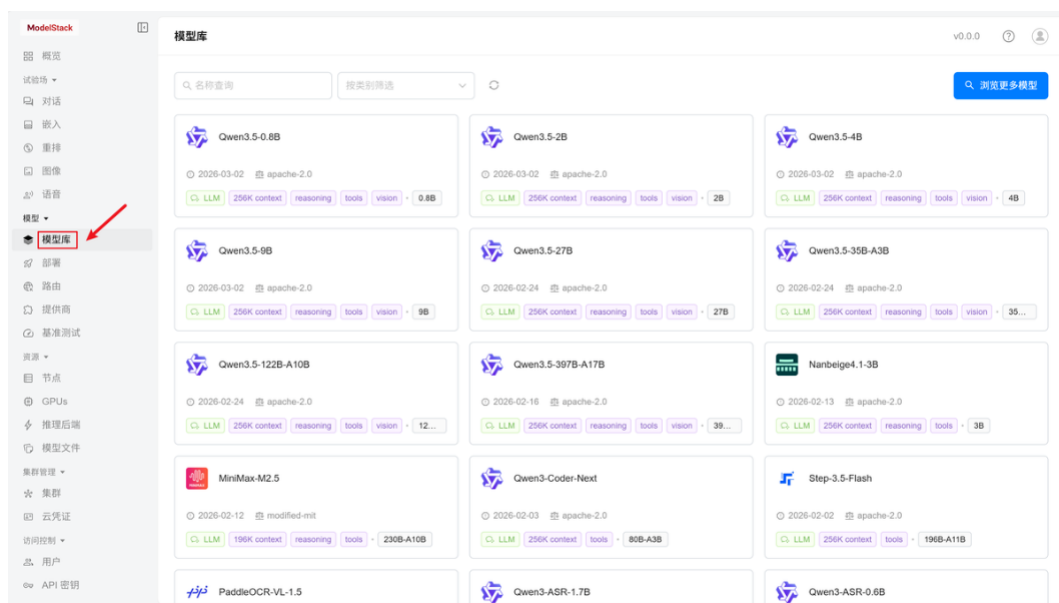


注意 :为保证 GPU 服务器与 EATP 正常通信 ,配置前 ,请确保已开放 GPU 服务器 10050、10051 端口。在规则 Tab 中 ,点击添加规则按钮 ,新增端口 10050 和 10051 允许策略。

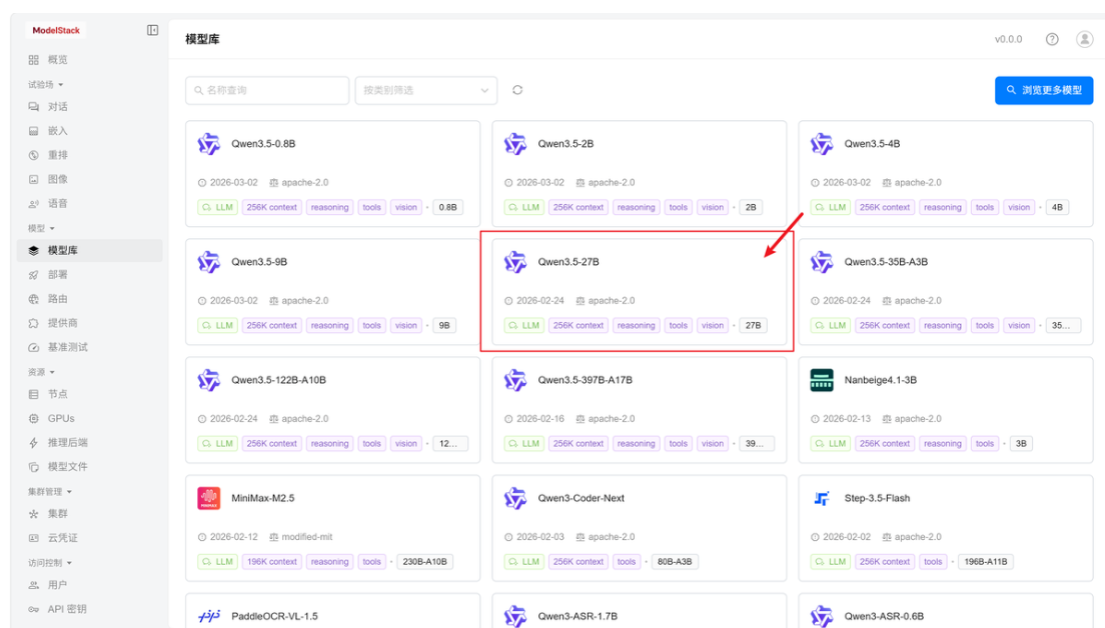
5. 确认平台里已经有可用的算力节点。添加成功后 ,可在集群列表查看节点添加结果 ,节点列表中可查看 GPU 服务器节点状态。



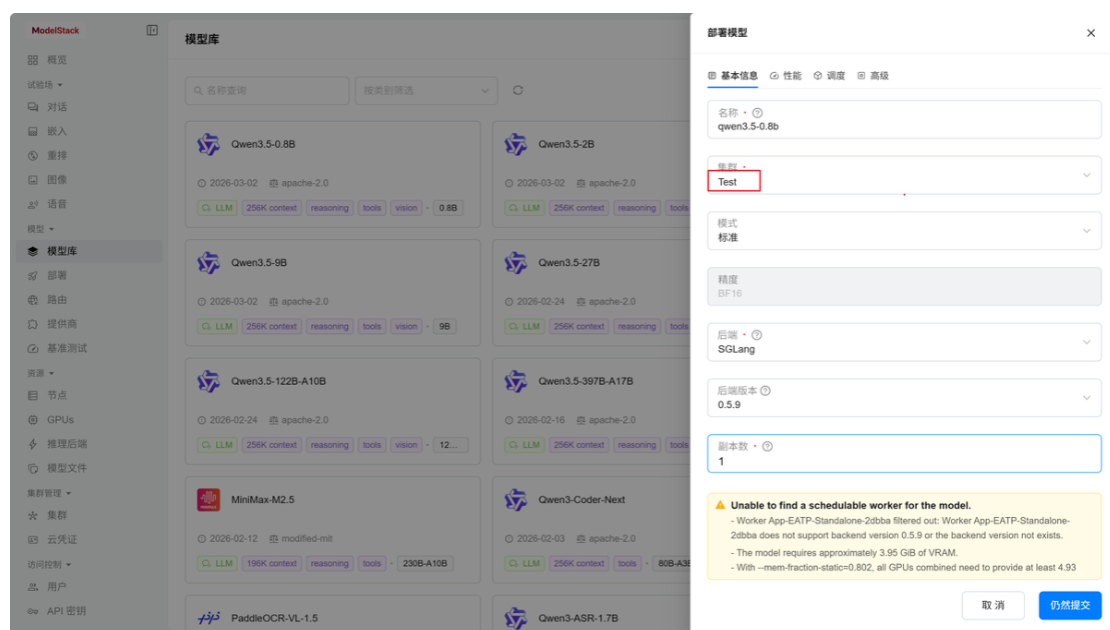
6. 部署模型。在左侧导航中 ,选择【模型库】 ,进入模型库页。



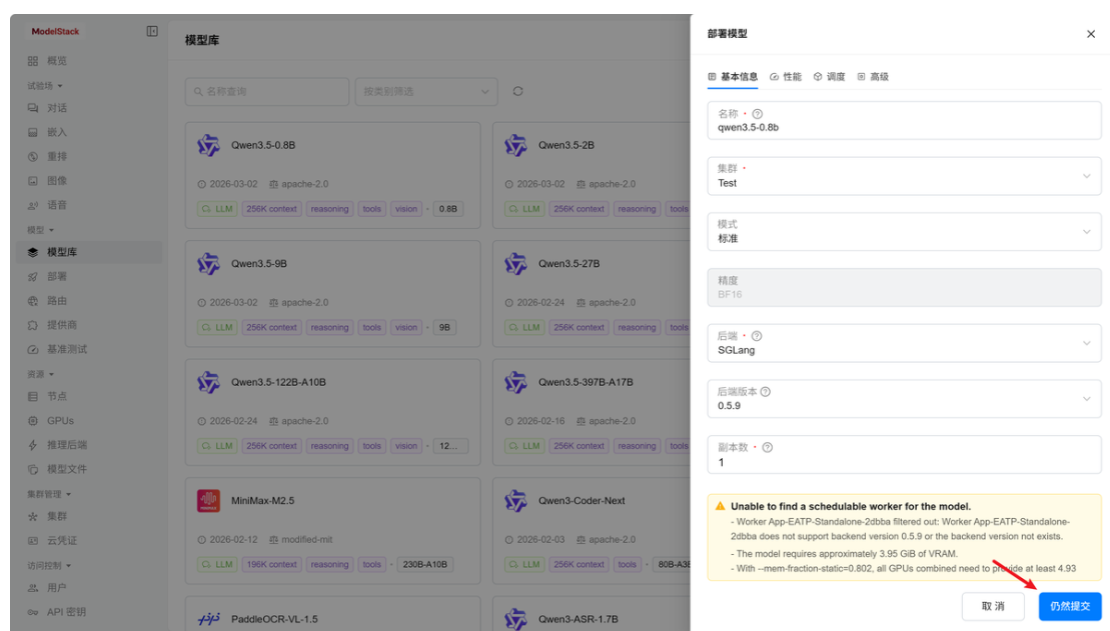
根据您的算力情况、使用场景，选择目标模型，如“Qwen3.5-0.8B” 点击即可部署。



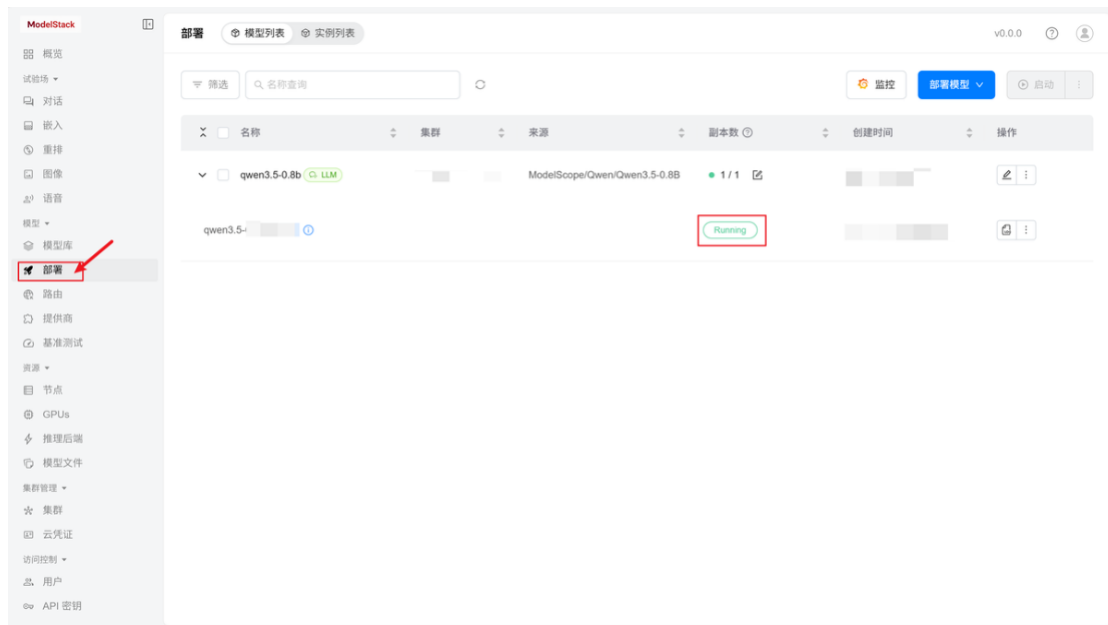
在弹出的表单中填写模型参数。标准版EATP服务组中，模型将默认部署在前置添加 GPU 服务器上，即“Test”。推理后端、版本如无特殊需求，使用默认推荐参数即可。



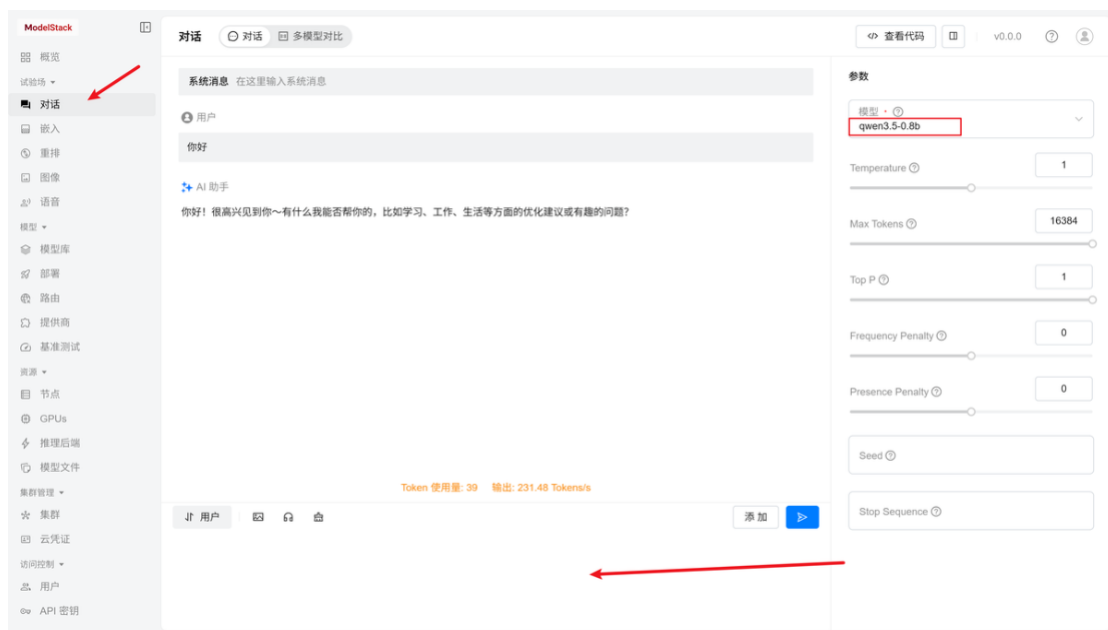
点击提交即开始模型部署。



模型部署时间根据模型大小有所不同，通常需要 15 分钟到 60 分钟不等。可在左侧导航中，点击【部署】，进入模型部署页面，查看模型部署进度。当节点状态为“Running”时代表模型部署完成。



7. 验证模型服务是否正常。左侧导航菜单中选择【对话】，在对话页面中，选择之前操作部署的模型，并输入对话。若能正常对话，即代表模型服务正常。



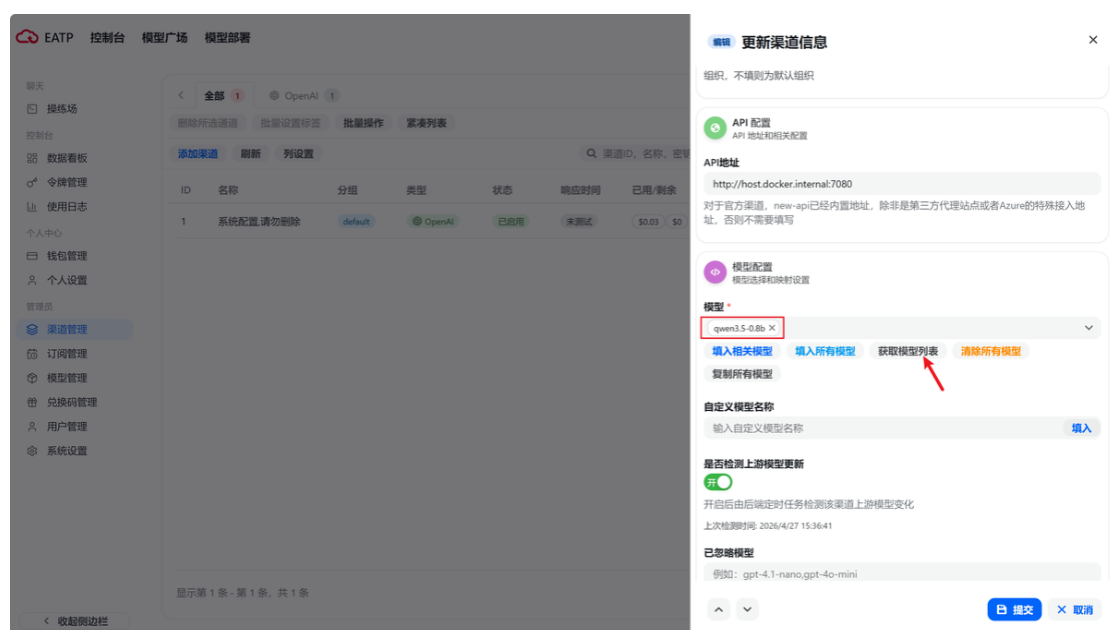
2.4 在 EATP 中同步模型配置

模型部署完后，需将模型部署信息同步至渠道后，才可对外提供模型服务。

1. 返回 EATP 主页面，左侧菜单中，点击【渠道管理】。EATP 中，已默认添加 2.3 步骤中在 ModelStack 中部署的模型渠道，可通过【编辑】弹窗中的信息确认模型是否同步完成。



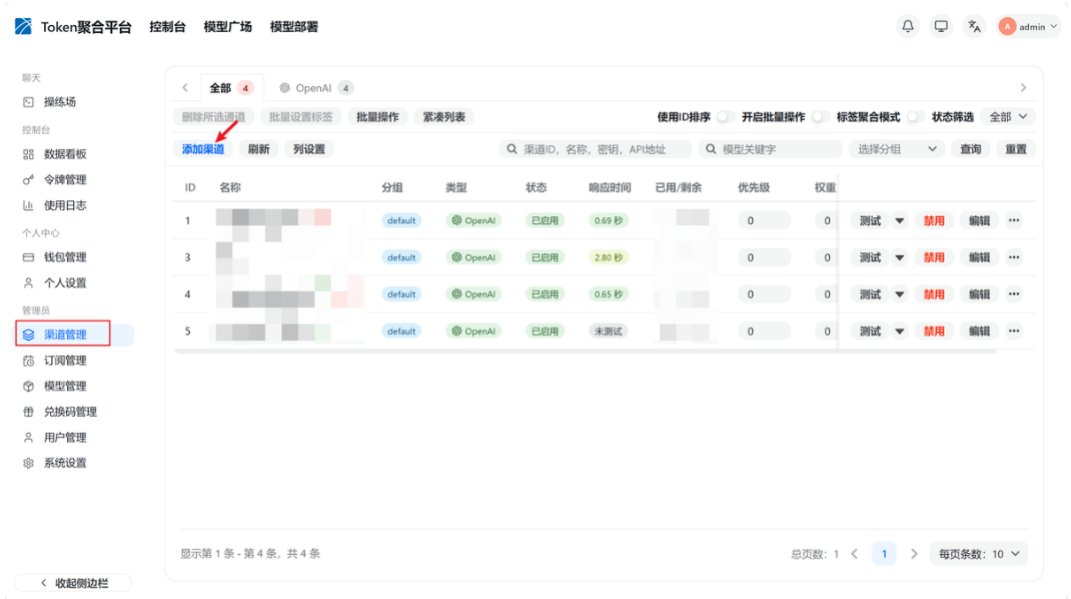
若“更新渠道信息”弹窗中正确显示2.3步骤中部署的模型信息,即代表模型部署且同步成功。若模型为空,可点击下方【获取模型列表】进行模型添加。



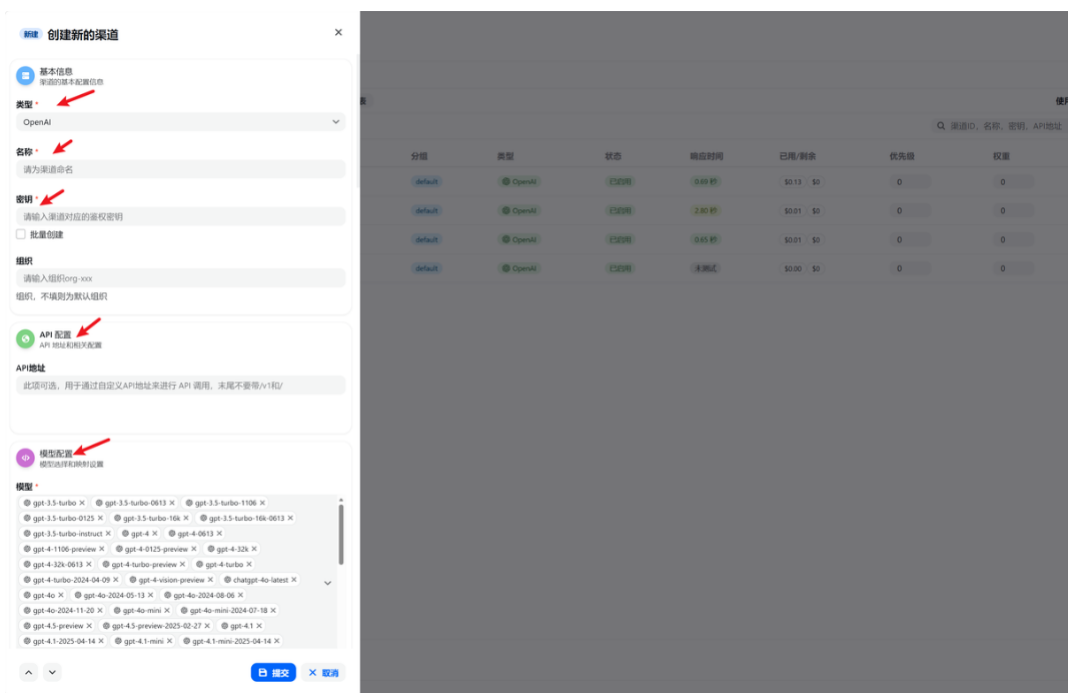
2.5 配置渠道

若需管理更多渠道来源的 API,需先将渠道 API 信息添加至 EATP 平台中。

1. 添加渠道。在左侧导航栏中,选择【渠道管理】,点击【添加渠道】按钮。



- 在添加渠道表单中，根据 API 供应商提供的信息，完成表单填写。点击【提交】，即可完成渠道配置。



2.6 验证模型是否正常对外提供服务

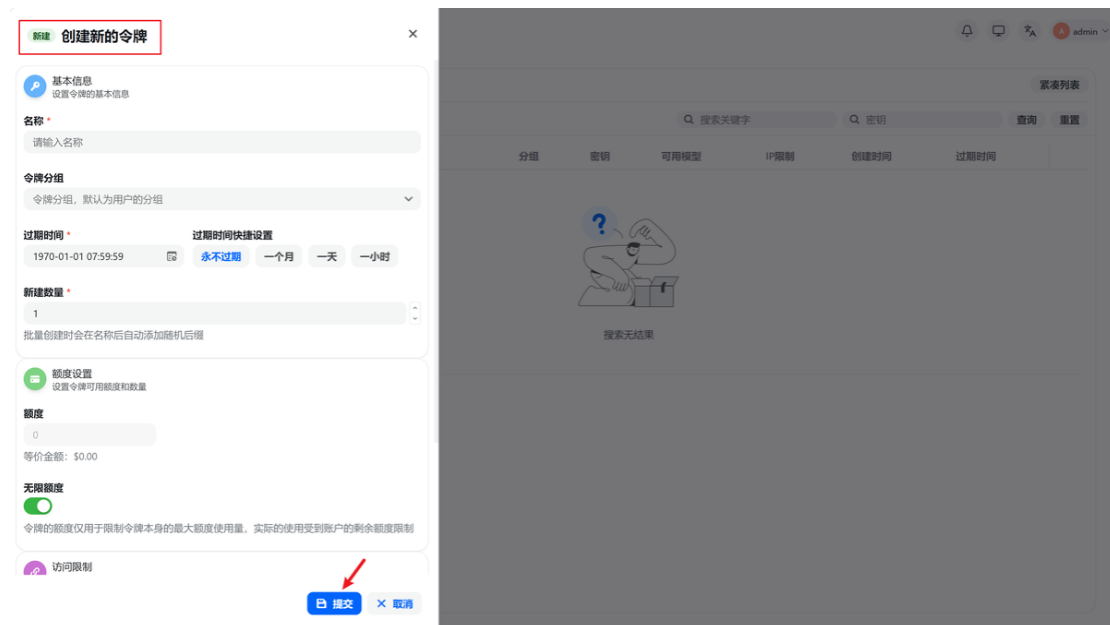
- 以管理员身份，在 EATP 平台中，左侧导航中选择【令牌管理】。



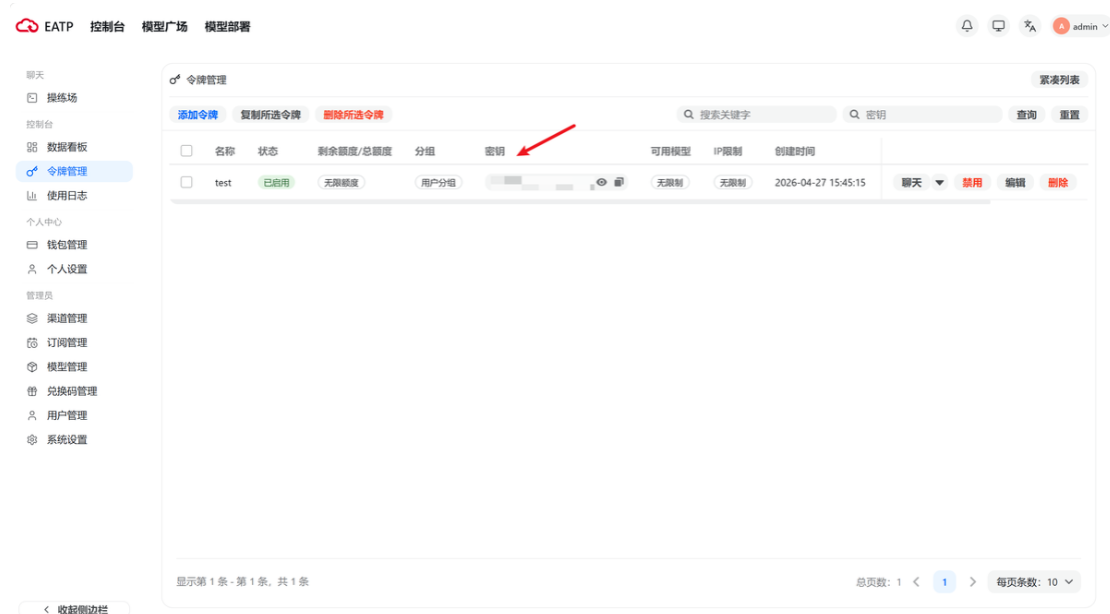
2. 点击添加令牌按钮，创建一个可用于外部调用模型服务的凭证。



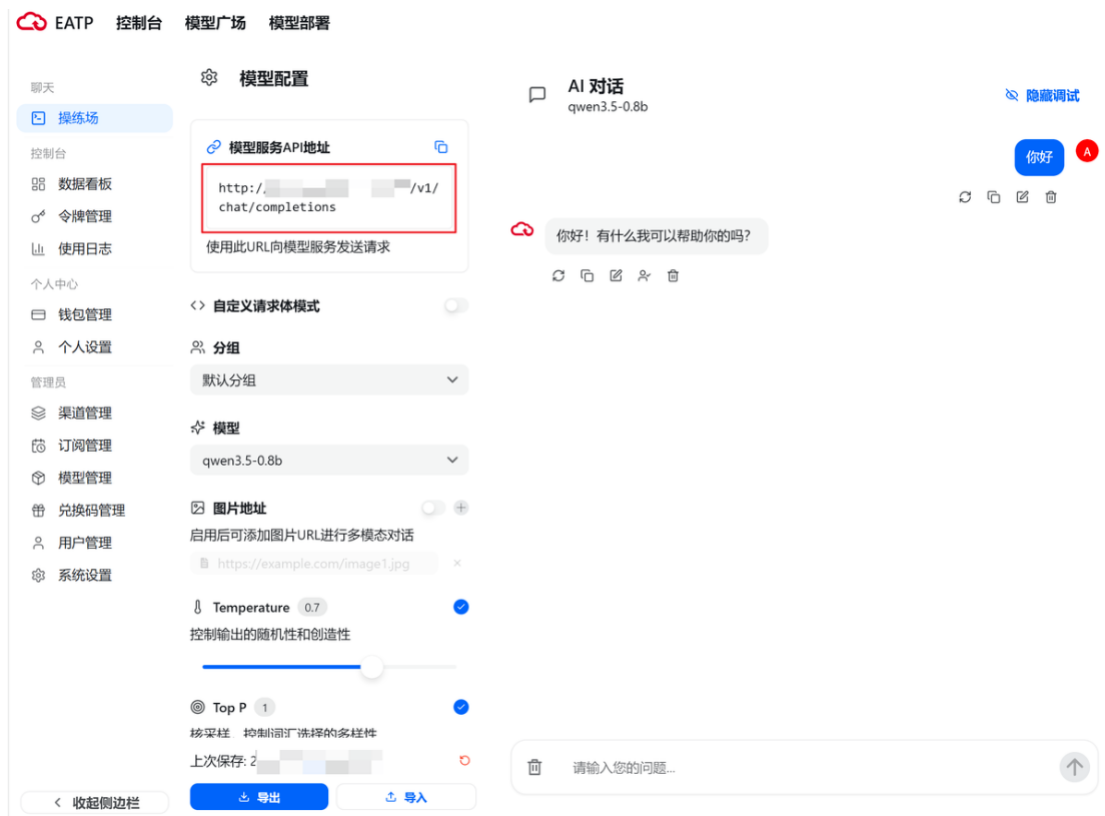
3. 在“创建新令牌”窗口中，根据需求填写令牌约束信息。



4. 创建完成后，可在列表中查看生成的令牌，密钥即模型服务的 API Key。



模型服务 API 的 URL，可在左侧菜单中【操场场】页面中获取。



您可通过在AI应用中配置模型服务API的URL即Key信息，已确认模型是否正常对外提供服务。同时，EATP平台中，【使用日志】页面也可查看调用相关的数据。



使用建议

使用 EATP 完成模型部署与配置后，您可直接创建令牌来运营模型服务；也可前往【用户管理】模块创建用户，由用户自行创建令牌来运营服务。