GPUStack-单机版 用户手册

一、产品概述

1.1 产品介绍

GPUStack 云服务是基于开源 GPUStack 构建的托管式 AI 模型部署平台，让您无需管理基础设施，即可在高性能 GPU 集群上轻松部署和运行各类 AI 模型。

1.2 产品核心能力

资源管理：提供自动化 GPU 资源调度与集群管理，支持异构 GPU 设备统一纳管，实现资源利用率最大化与成本最优化；

模型部署：支持主流开源大模型一键部署，兼容 Hugging Face、ModelScope 等模型源，集成 vLLM、SGLang 和 TensorRT-LLM 等高性能推理引擎，满足不同场景性能需求；

智能运维：内置自动扩缩容、故障转移与负载均衡机制，提供实时性能监控与告警，确保服务高可用性与稳定性；

安全管控：提供完善的认证授权体系与网络隔离策略，支持私有化部署与数据安全保障，满足企业级安全合规要求。

1.3 产品优势

自动处理底层 GPU 资源调度、模型优化和扩展，让您专注于应用开发而非运维。

零运维负担：无需管理 GPU 驱动、CUDA 版本或集群配置；

开箱即用： 集成 vLLM、SGLang 和 TensorRT-LLM 等高性能推理引擎，支持自定义推理框架；

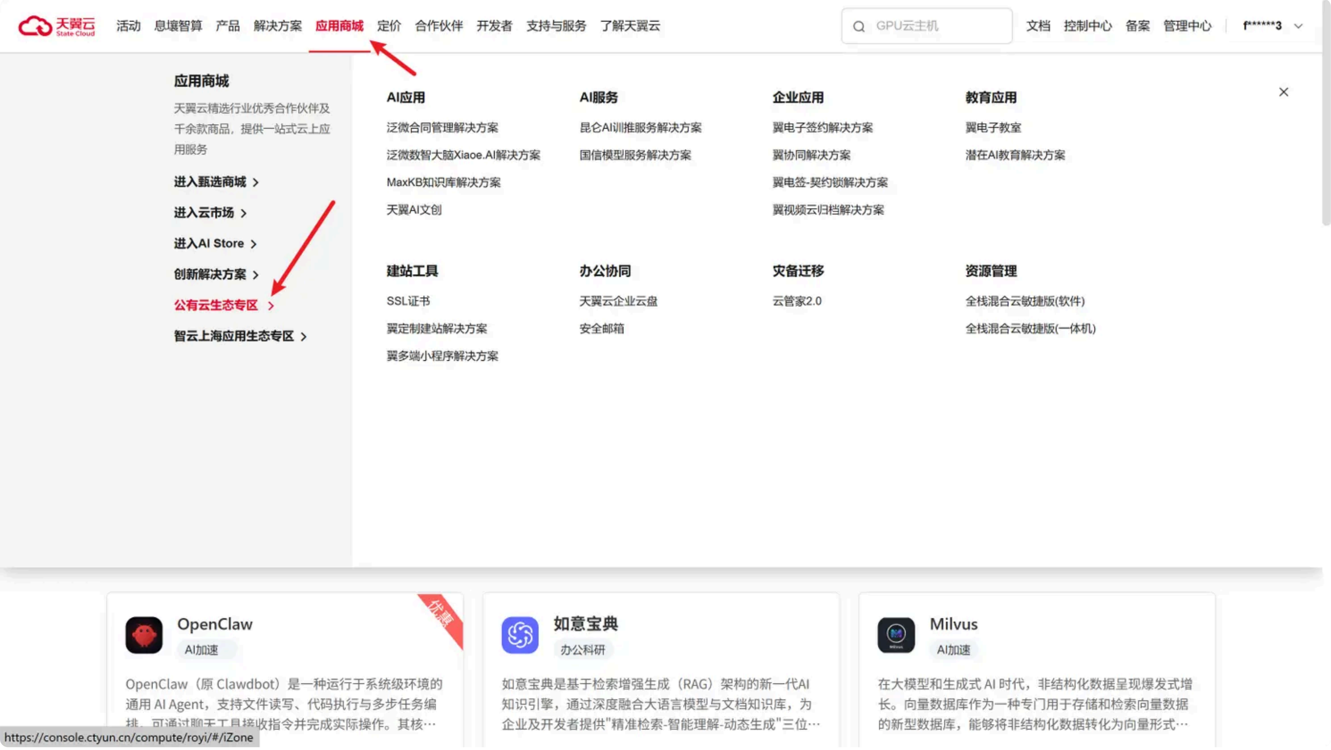
一键部署：支持从 Hugging Face、ModelScope、或本地直接部署，支持自动扩缩容、版本升降级；

性能优化配置：提供预调优模式，用于低延迟或高吞吐量；

运维能力： 支持自动故障恢复、负载均衡、监控、认证和访问控制。

二、使用指南

2.1 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



天翼云精选行业优秀合作伙伴及千余款商品，提供一站式云上应用服务

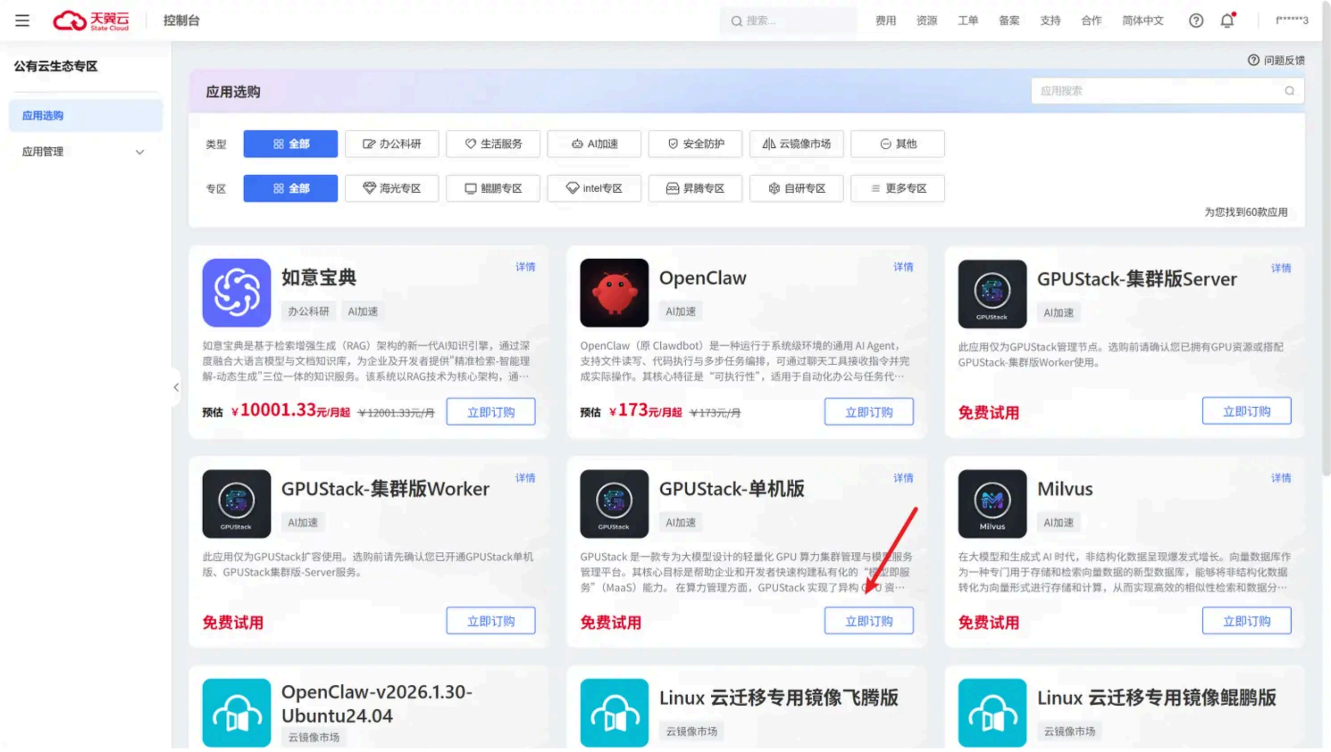
OpenClaw（原 Clawdbot）是一种运行于系统级环境的通用 AI Agent，支持文件读写、代码执行与多步任务编排，可通过聊天工具接收指令并完成实际操作。其核…

如意宝典是基于检索增强生成（RAG）架构的新一代AI知识引擎，通过深度融合大语言模型与文档知识库，为企业及开发者提供"精准检索-智能理解-动态生成"三位…

在大模型和生成式 AI 时代，非结构化数据呈现爆发式增长。向量数据库作为一种专门用于存储和检索向量数据的新型数据库，能够将非结构化数据转化为向量形式…

https://console.ctyun.cn/compute/royi/#/iZone

2.2 选择“GPUStack 单机版”，点击立即订购



如意宝典是基于检索增强生成（RAG）架构的新一代AI知识引擎，通过深度融合大语言模型与文档知识库，为企业及开发者提供"精准检索-智能理解-动态生成"三位一体的知识服务。该系统以RAG技术为核心架构，通…

预估 ￥10001.33元/月起 ~~￥12001.33元/月~~

OpenClaw（原 Clawdbot）是一种运行于系统级环境的通用 AI Agent，支持文件读写、代码执行与多步任务编排，可通过聊天工具接收指令并完成实际操作。其核心特征是“可执行性”，适用于自动化办公与任务代…

预估 ￥173元/月起 ~~￥173元/月~~

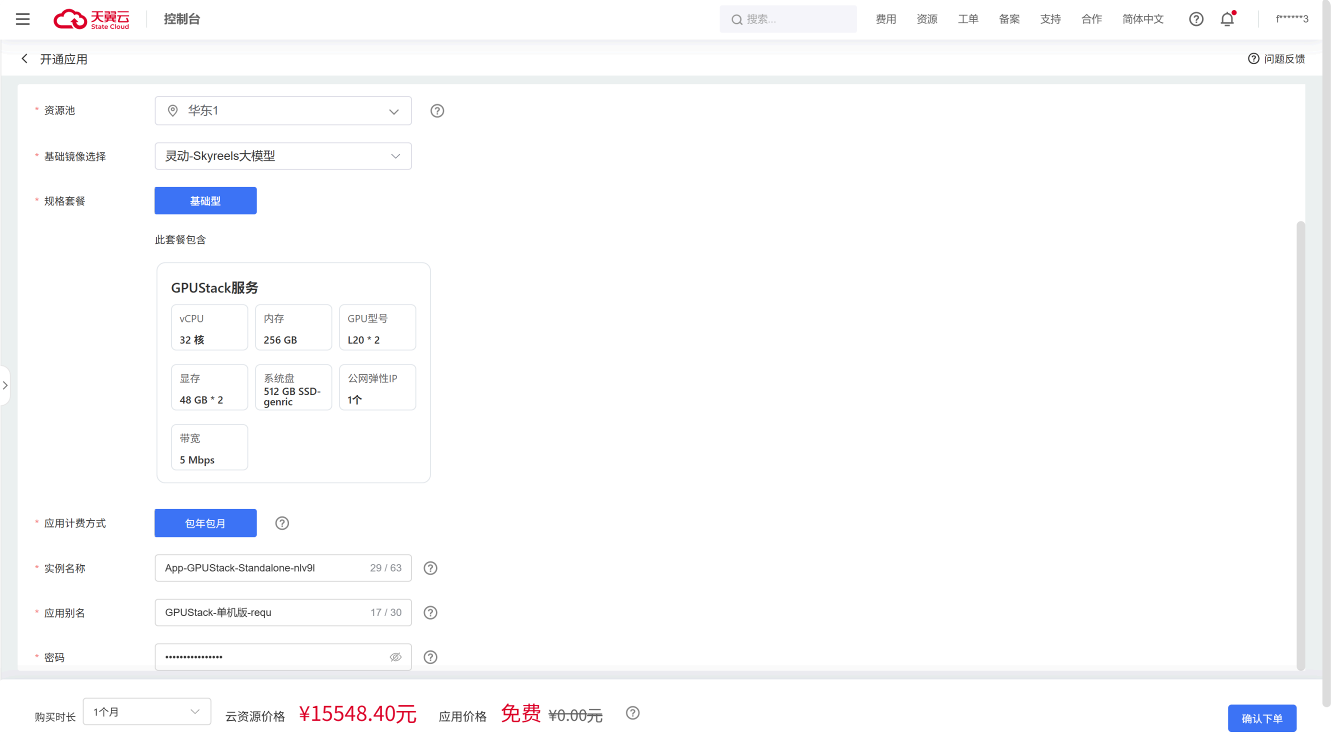
此应用仅为GPUStack管理节点。选购前请确认您已拥有GPU资源或搭配GPUStack-集群版Worker使用。

此应用仅为GPUStack扩容使用。选购前请先确认您已开通GPUStack单机版、GPUStack集群版-Server服务。

GPUStack 是一款专为大模型设计的轻量化 GPU 算力集群管理与模型服务管理平台。其核心目标是帮助企业和开发者快速构建私有化的 “模型即服务”（MaaS）能力。 在算力管理方面，GPUStack 实现了异构 GPU 资…

在大模型和生成式 AI 时代，非结构化数据呈现爆发式增长。向量数据库作为一种专门用于存储和检索向量数据的新型数据库，能够将非结构化数据转化为向量形式进行存储和计算，从而实现高效的相似性检索和数据分…

2.3 按提示订购“GPUstack 单机版”，点击确认下单即可成功购买服务。



* 资源池：华东1

* 基础镜像选择：灵动-Skyreels大模型

* 规格套餐：基础型

此套餐包含

GPUStack服务

vCPU 32 核	内存 256 GB	GPU型号 L20 * 2
显存 48 GB * 2	系统盘 512 GB SSD-genric	公网弹性IP 1个
带宽 5 Mbps		

* 应用计费方式：包年包月

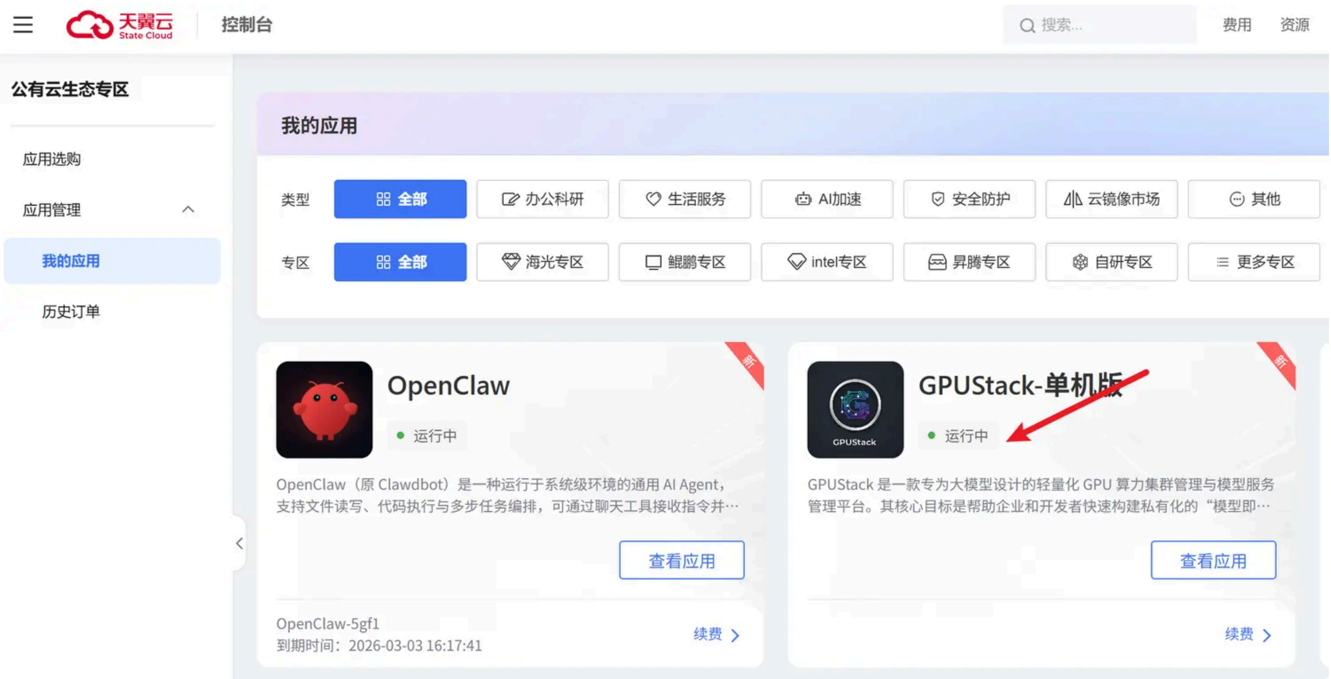
* 实例名称：App-GPUStack-Standalone-nlv9l 29 / 63

* 应用别名：GPUStack-单机版-requ 17 / 30

* 密码：••••••••••••••••

购买时长 1个月 云资源价格 ¥15548.40元 应用价格 免费 ~~¥0.00元~~

2.4 在【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。



OpenClaw

● 运行中

OpenClaw（原 Clawdbot）是一种运行于系统级环境的通用 AI Agent，支持文件读写、代码执行与多步任务编排，可通过聊天工具接收指令并…

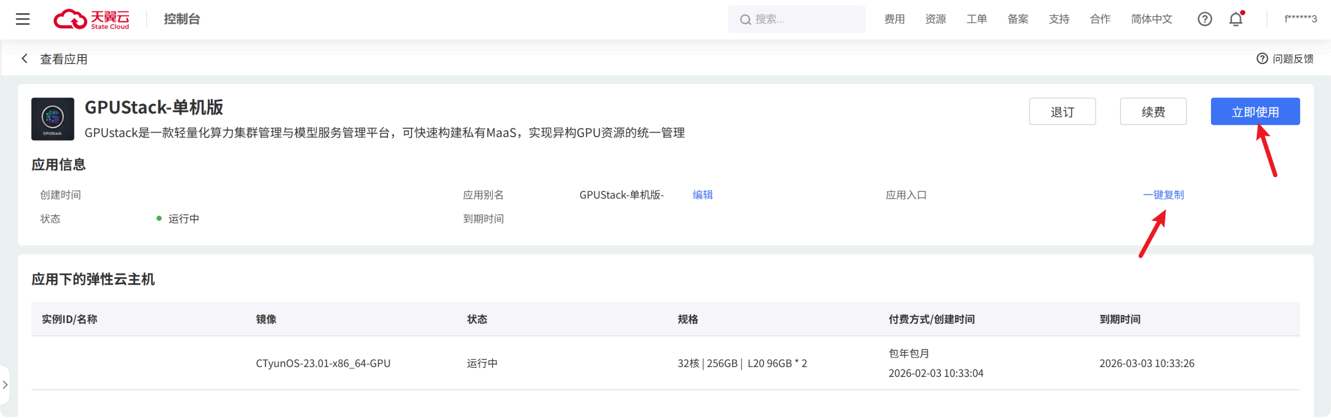
OpenClaw-5gf1
到期时间：2026-03-03 16:17:41

GPUStack-单机版

● 运行中

GPUStack 是一款专为大模型设计的轻量化 GPU 算力集群管理与模型服务管理平台。其核心目标是帮助企业和开发者快速构建私有化的“模型即…

2.5 点击查看应用按钮，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 GPUstack



GPUStack-单机版

GPUstack是一款轻量化算力集群管理与模型服务管理平台，可快速构建私有MaaS，实现异构GPU资源的统一管理

应用信息

创建时间

状态 ● 运行中

应用别名 GPUStack-单机版- 编辑

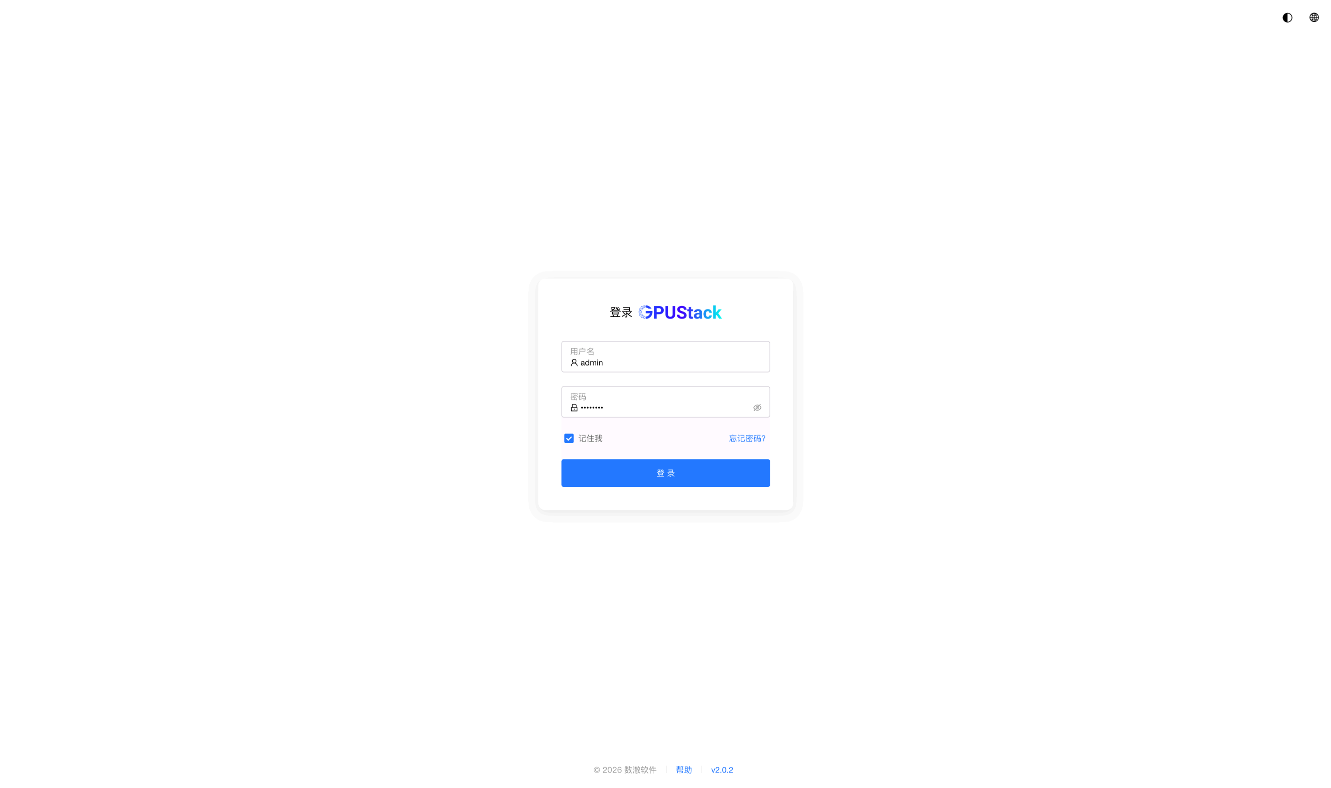
到期时间

应用入口 一键复制

应用下的弹性云主机

实例ID/名称	镜像	状态	规格	付费方式/创建时间	到期时间
	CTyunOS-23.01-x86_64-GPU	运行中	32核 | 256GB | L20 96GB * 2	包年包月 2026-02-03 10:33:04	2026-03-03 10:33:26

2.6 GPUstack 默认账号为 admin，密码为 gpustack。（可在应用中修改）

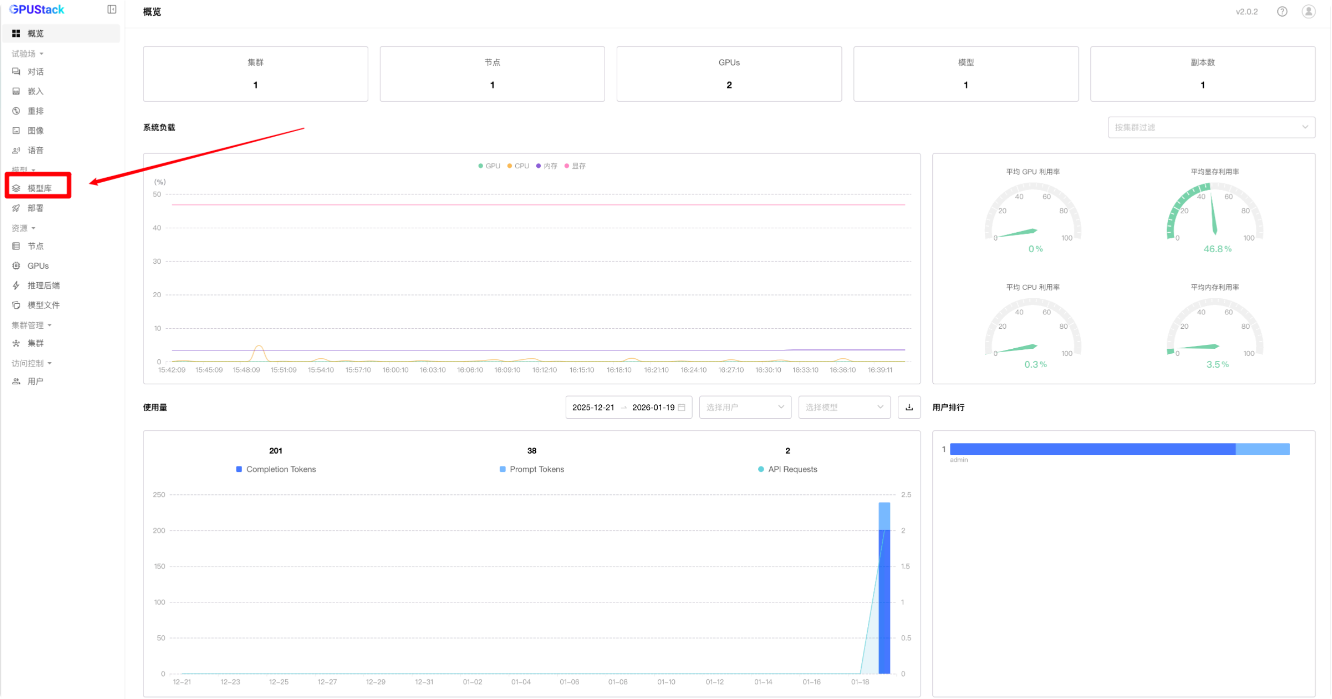


用户名 admin

密码 •••••••••

☑ 记住我

2.7 部署模型。进入模型库菜单，选择合适的模型开始部署；



概览

集群	节点	GPUs	模型	副本数
1	1	2	1	1

系统负载

按集群过滤

平均 GPU 利用率 0 %

平均显存利用率 46.8 %

平均 CPU 利用率 0.3 %

平均内存利用率 3.5 %

使用量

2025-12-21 → 2026-01-19

选择用户

选择模型

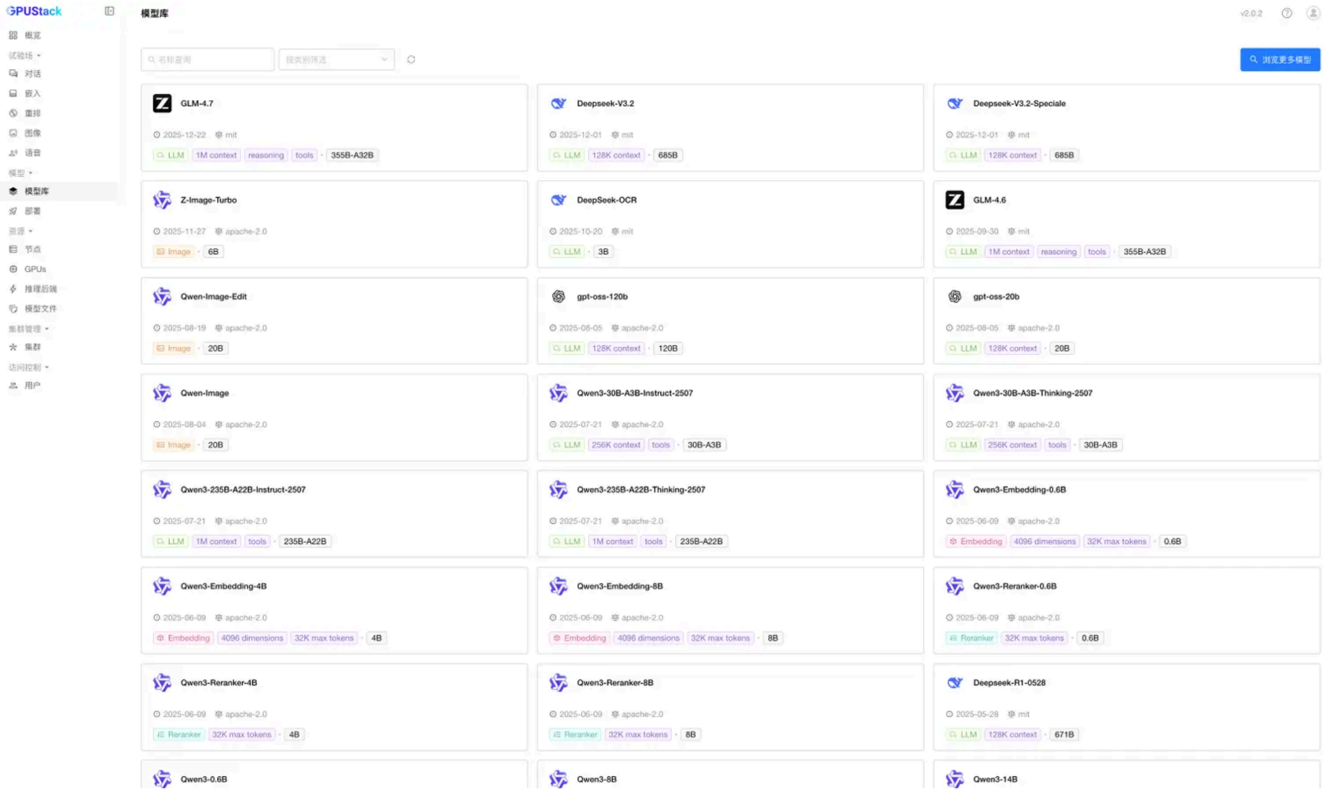
201 Completion Tokens

38 Prompt Tokens

2 API Requests

用户排行

1 admin



模型库

GLM-4.7
2025-12-22 mit
LLM 1M context reasoning tools · 355B-A32B

Deepseek-V3.2
2025-12-01 mit
LLM 128K context · 685B

Deepseek-V3.2-Speciale
2025-12-01 mit
LLM 128K context · 685B

Z-Image-Turbo
2025-11-27 apache-2.0
Image · 6B

DeepSeek-OCR
2025-10-20 mit
LLM · 3B

GLM-4.6
2025-09-30 mit
LLM 1M context reasoning tools · 355B-A32B

Qwen-Image-Edit
2025-08-19 apache-2.0
Image · 20B

gpt-oss-120b
2025-08-05 apache-2.0
LLM 128K context · 120B

gpt-oss-20b
2025-08-05 apache-2.0
LLM 128K context · 20B

Qwen-Image
2025-08-04 apache-2.0
Image · 20B

Qwen3-30B-A3B-Instruct-2507
2025-07-21 apache-2.0
LLM 256K context tools · 30B-A3B

Qwen3-30B-A3B-Thinking-2507
2025-07-21 apache-2.0
LLM 256K context tools · 30B-A3B

Qwen3-235B-A22B-Instruct-2507
2025-07-21 apache-2.0
LLM 1M context tools · 235B-A22B

Qwen3-235B-A22B-Thinking-2507
2025-07-21 apache-2.0
LLM 1M context tools · 235B-A22B

Qwen3-Embedding-0.6B
2025-06-09 apache-2.0
Embedding 4096 dimensions 32K max tokens · 0.6B

Qwen3-Embedding-4B
2025-06-09 apache-2.0
Embedding 4096 dimensions 32K max tokens · 4B

Qwen3-Embedding-8B
2025-06-09 apache-2.0
Embedding 4096 dimensions 32K max tokens · 8B

Qwen3-Reranker-0.6B
2025-06-09 apache-2.0
Reranker 32K max tokens · 0.6B

Qwen3-Reranker-4B
2025-06-09 apache-2.0
Reranker 32K max tokens · 4B

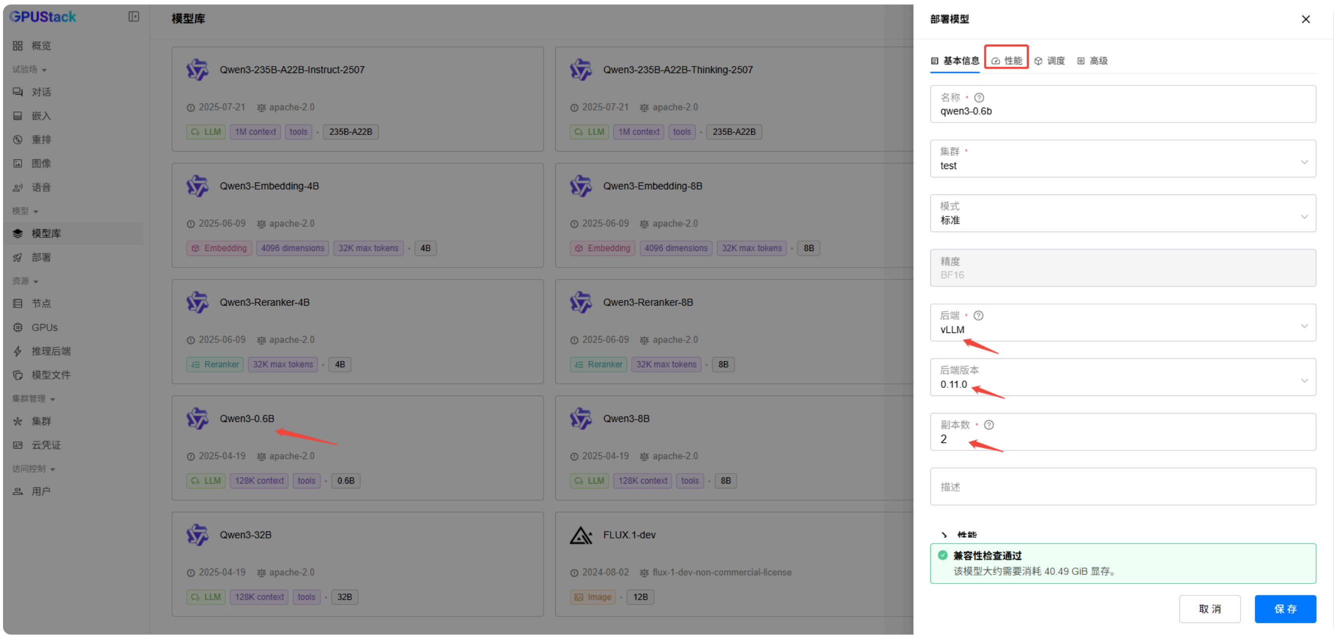
Qwen3-Reranker-8B
2025-06-09 apache-2.0
Reranker 32K max tokens · 8B

Deepseek-R1-0528
2025-05-28 mit
LLM 128K context · 671B

Qwen3-0.6B

Qwen3-8B

Qwen3-14B



Qwen3-235B-A22B-Instruct-2507 · 2025-07-21 · apache-2.0 · LLM · 1M context · tools · 235B-A22B

Qwen3-235B-A22B-Thinking-2507 · 2025-07-21 · apache-2.0 · LLM · 1M context · tools · 235B-A22B

Qwen3-Embedding-4B · 2025-06-09 · apache-2.0 · Embedding · 4096 dimensions · 32K max tokens · 4B

Qwen3-Embedding-8B · 2025-06-09 · apache-2.0 · Embedding · 4096 dimensions · 32K max tokens · 8B

Qwen3-Reranker-4B · 2025-06-09 · apache-2.0 · Reranker · 32K max tokens · 4B

Qwen3-Reranker-8B · 2025-06-09 · apache-2.0 · Reranker · 32K max tokens · 8B

Qwen3-0.6B · 2025-04-19 · apache-2.0 · LLM · 128K context · tools · 0.6B

Qwen3-8B · 2025-04-19 · apache-2.0 · LLM · 128K context · tools · 8B

Qwen3-32B · 2025-04-19 · apache-2.0 · LLM · 128K context · tools · 32B

FLUX.1-dev · 2024-08-02 · flux-1-dev-non-commercial-license · Image · 12B

部署模型

基本信息　性能　调度　高级

名称 *: qwen3-0.6b

集群 *: test

模式: 标准

精度: BF16

后端 *: vLLM

后端版本: 0.11.0

副本数 *: 2

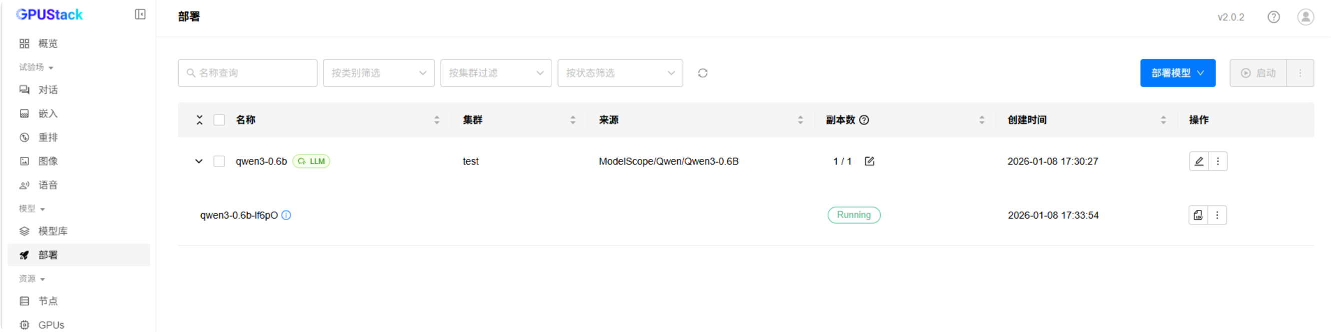
描述

性能

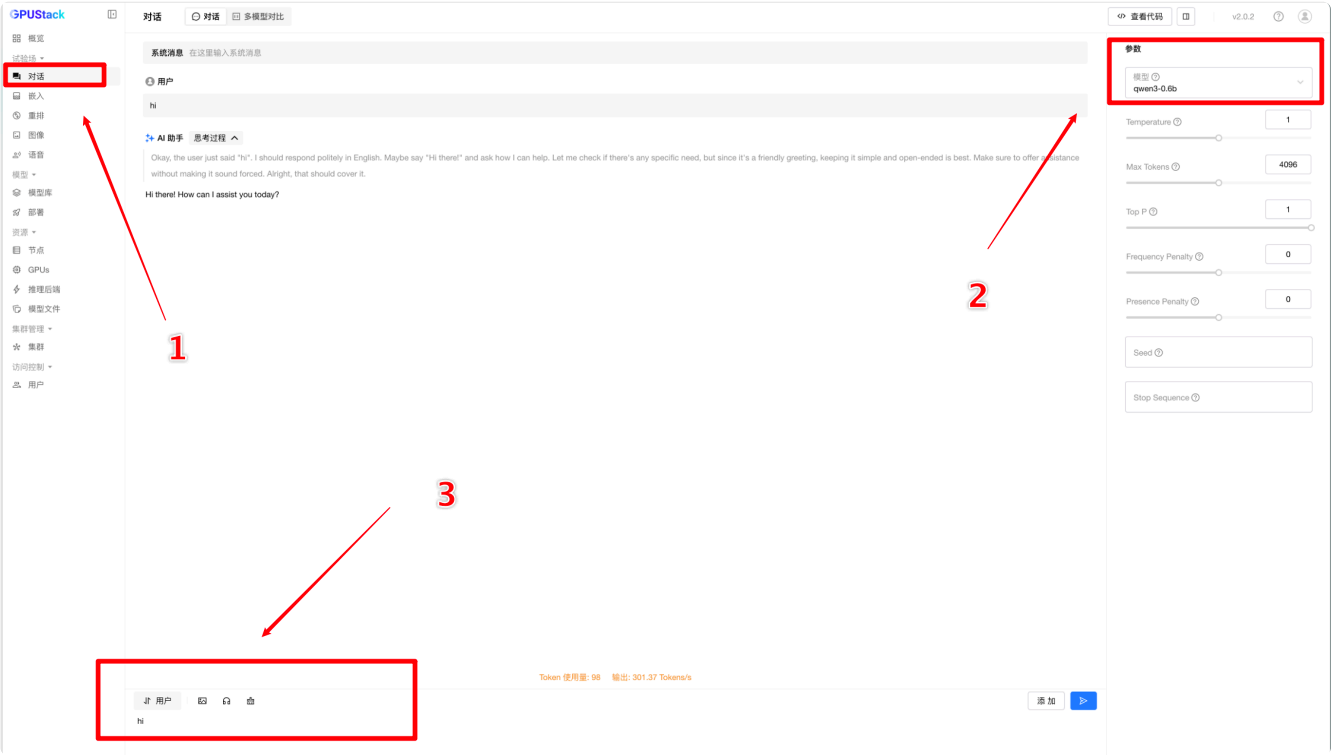
兼容性检查通过
该模型大约需要消耗 40.49 GiB 显存。

取 消　保 存

2.8 等待模型部署成功后，便可进入试验场，进行验证；



名称	集群	来源	副本数	创建时间	操作
qwen3-0.6b LLM	test	ModelScope/Qwen/Qwen3-0.6B	1 / 1	2026-01-08 17:30:27	
qwen3-0.6b-lf6pO			Running	2026-01-08 17:33:54	



系统消息　在这里输入系统消息

用户

hi

AI 助手　思考过程

Okay, the user just said "hi". I should respond politely in English. Maybe say "Hi there!" and ask how I can help. Let me check if there's any specific need, but since it's a friendly greeting, keeping it simple and open-ended is best. Make sure to offer assistance without making it sound forced. Alright, that should cover it.

Hi there! How can I assist you today?

Token 使用量: 98　输出: 301.37 Tokens/s

参数

模型: qwen3-0.6b

Temperature: 1

Max Tokens: 4096

Top P: 1

Frequency Penalty: 0

Presence Penalty: 0

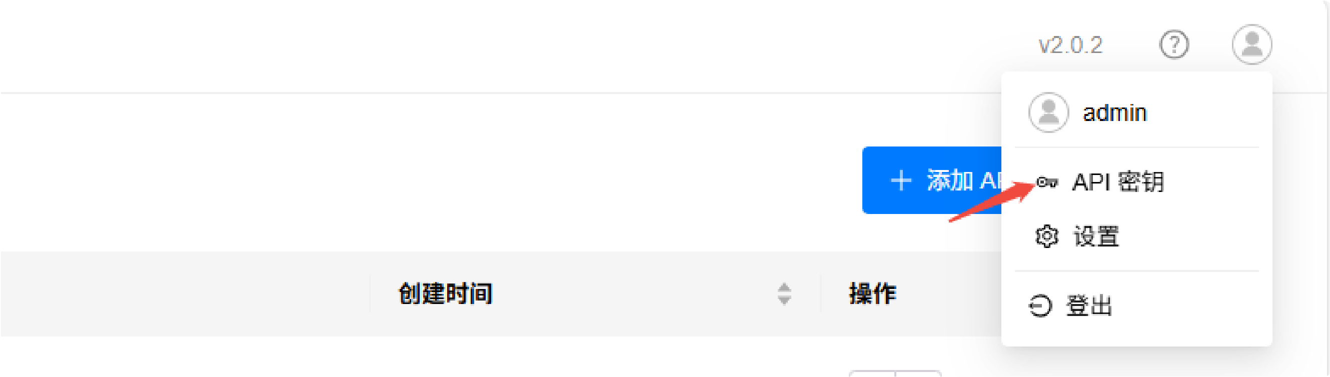
Seed

Stop Sequence

hi

2.9 API 密钥管理

1. 将鼠标悬停在用户头像上，选择"`API 密钥`"
2. 点击"`创建密钥`"生成新密钥
3. **安全提示**：该密钥仅在创建时可见一次，请妥善保存



三、常见问题

Q1: 如何升级或降级已部署的模型版本?

A: GPUStack 支持无缝模型版本切换

1. 在模型详情页点击"编辑"按钮
2. 在"模型版本"下拉菜单中选择目标版本
3. 手动删除原来的示例，即可自动创建新的副本
4. 支持新旧副本共存，可以在用户无感知情况下切换版本

Q2: 本地模型文件如何导入 GPUStack?

A：挂载共享存储，需要在集群创建之前完成挂载并存储，否则 worker 节点无法访问

1. 根据页面提示在创建集群时，进行缓存卷挂载，会自动生成对应的 worker 指令
2. 此时在部署模型时，可以选择本地模型进行部署

版权与声明

本文档内容仅供参考。