

大模型服务 用户指南

产品版本: v7.0.2

发布日期: 2026-10-09

目录

1 用户指南	1
1.1 模型广场	1
1.2 模型仓库	6
1.3 推理引擎	13
1.4 模型部署	20
1.5 AI 服务提供者	30
1.6 模型网关	37
1.7 模型体验	45
1.8 基准测试	52
1.9 套餐管理	58
1.10 API Key	65
1.11 Token 用量	74

1 用户指南

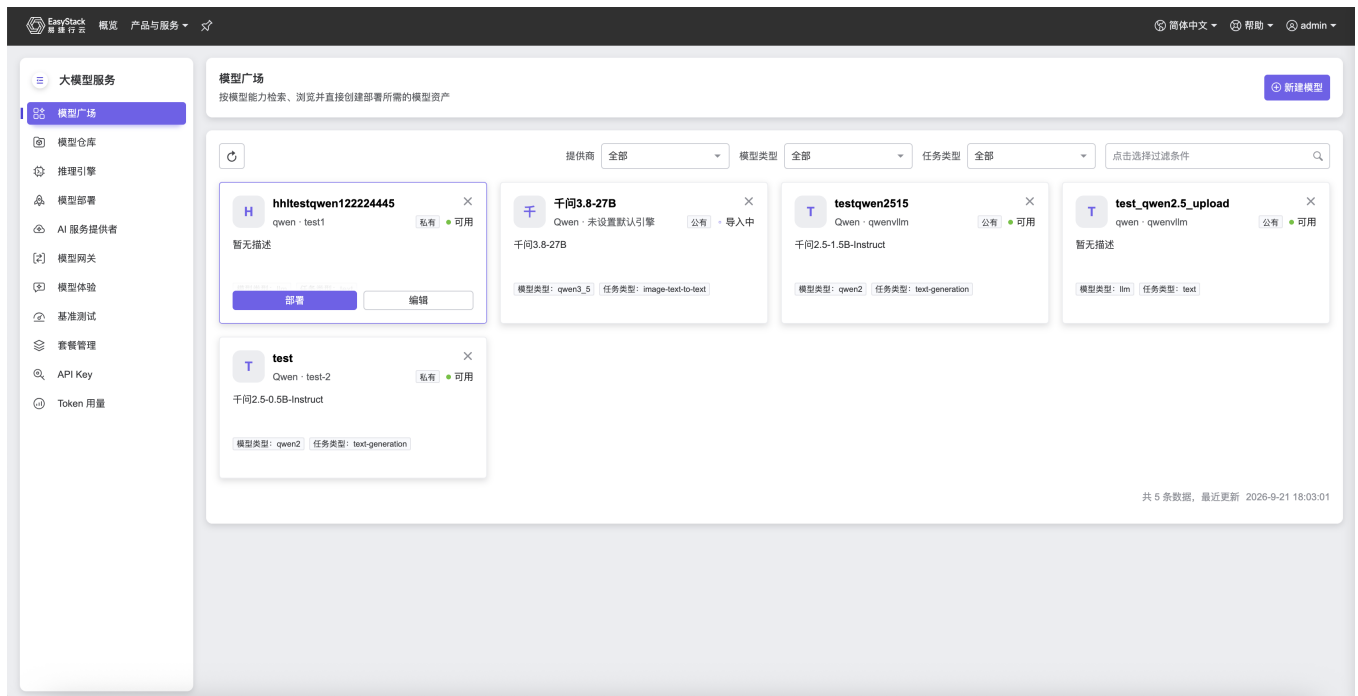
1.1 模型广场

本章节主要介绍在模型广场中浏览、筛选和创建模型，以及查看模型详情、推理引擎信息和发起模型部署等操作。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【模型广场】，即可进入“模型广场”页面。

查看模型

在“模型广场”页面中，可查看当前项目下已经创建的模型。模型卡片中显示模型名称、提供商、默认推理引擎、可见性、模型文件状态、来源名称、模型类型和任务类型等信息。

默认推理引擎未设置时，模型卡片中显示 未设置默认引擎。模型文件正在处理时，状态显示为 导入中；模型文件可以正常使用时，状态显示为 可用。



筛选模型

在“模型广场”页面中，可通过 提供商 、 模型类型 和 任务类型 等条件筛选模型。

设置筛选条件后，页面显示符合条件的模型。需要重新获取模型列表时，可单击页面中的“刷新”图标。

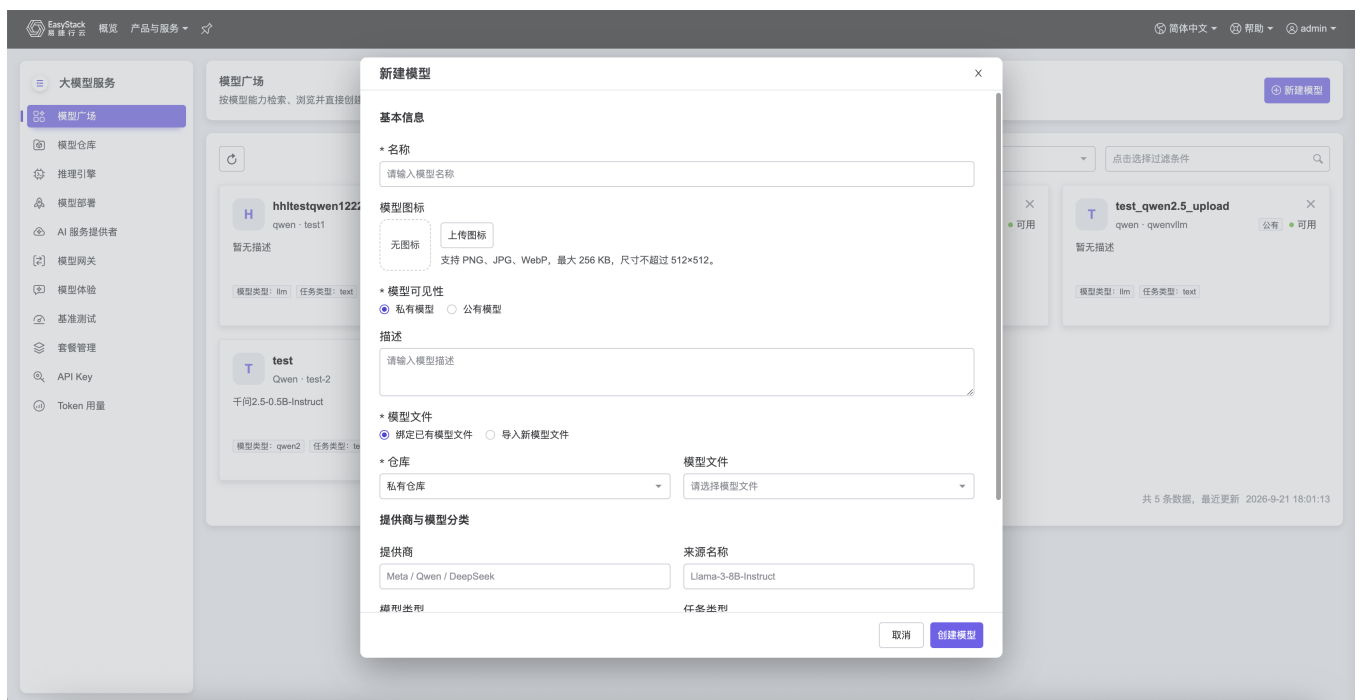
创建模型

前提条件

- 首次使用模型广场前，项目管理员已完成当前项目的Namespace初始化。
- 绑定已有模型文件时，目标模型文件已添加至模型仓库。
- 关联推理引擎时，目标推理引擎已经创建。

操作步骤

- 在“模型广场”页面中，单击 新建模型 ，弹出“新建模型”对话框。
- 配置参数后，单击 创建模型 ，完成操作。



参数	说明
名称	该模型在Beacon中的名称。

参数	说明
模型图标	该模型在模型广场中显示的图标。可选择 无图标 或单击 上传图标 上传图片。支持PNG、JPG和WebP格式，文件大小不能超过256 KB，图片尺寸不能超过512 × 512。
模型可见性	该模型的可见性。该参数值可选 私有模型 或 公有模型 。
描述	该模型的说明信息。
模型文件	该模型使用的模型文件。该参数值可选 绑定已有模型文件 或 导入新模型文件 。
仓库	选择 绑定已有模型文件 时，选择模型文件所在的模型仓库。
模型文件	选择 绑定已有模型文件 时，从所选仓库中选择需要绑定的模型文件。
提供商	该模型所属的提供商。
来源名称	该模型在原始来源中的名称。
模型类型	该模型所属的模型类型。
任务类型	该模型适用的任务类型。
推理引擎	该模型关联的推理引擎。

说明：

- 页面中带有“*”的参数为必填参数。
- 选择 **绑定已有模型文件** 后，需要继续设置模型文件所在的仓库和需要绑定的模型文件。
- 选择 **导入新模型文件** 后，根据页面显示的参数配置模型文件导入信息。

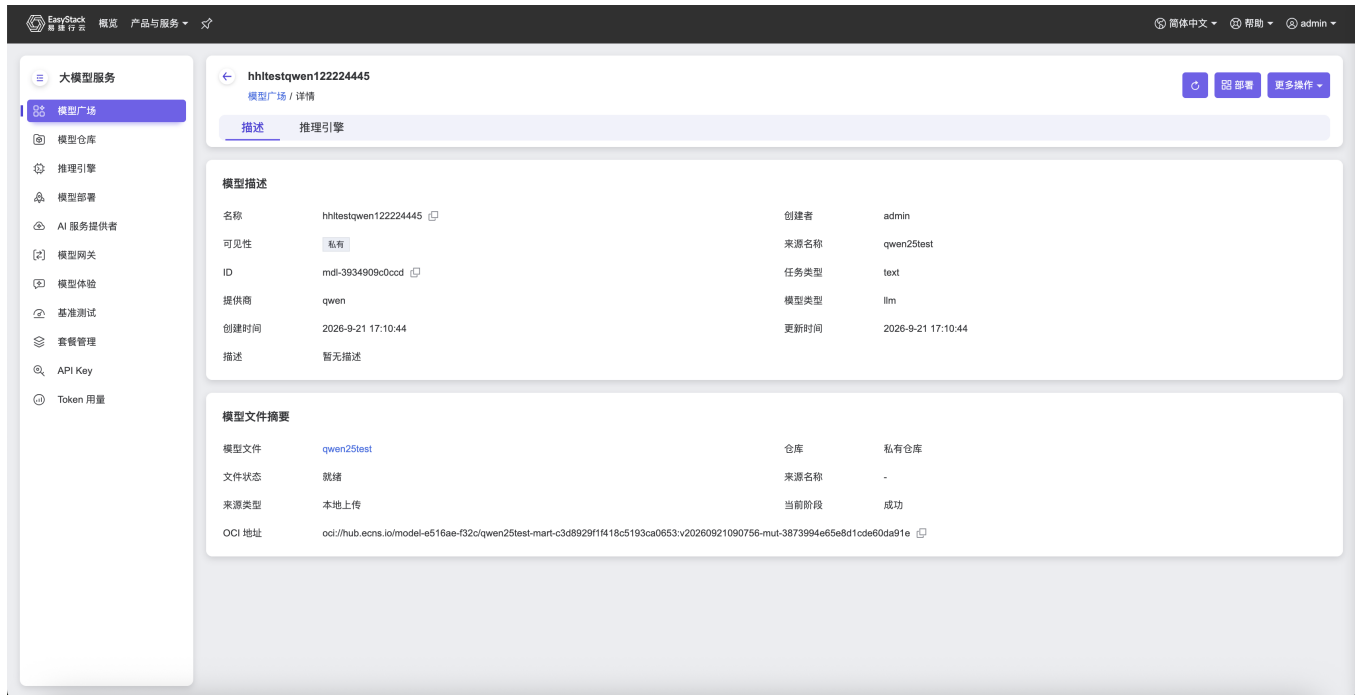
创建模型后，可返回“模型广场”页面查看模型。模型文件尚未处理完成时，模型卡片中的状态显示为 **导入中**；模型文件可以正常使用后，状态显示为 **可用**。

查看详情

在“模型广场”页面中，选择目标模型，可进入模型详情页面。模型详情页面包括 **描述** 和 **推理引擎** 两个页签。

查看模型描述

在模型详情页面中，选择 **描述** 页签，可查看模型基本信息和模型文件摘要。



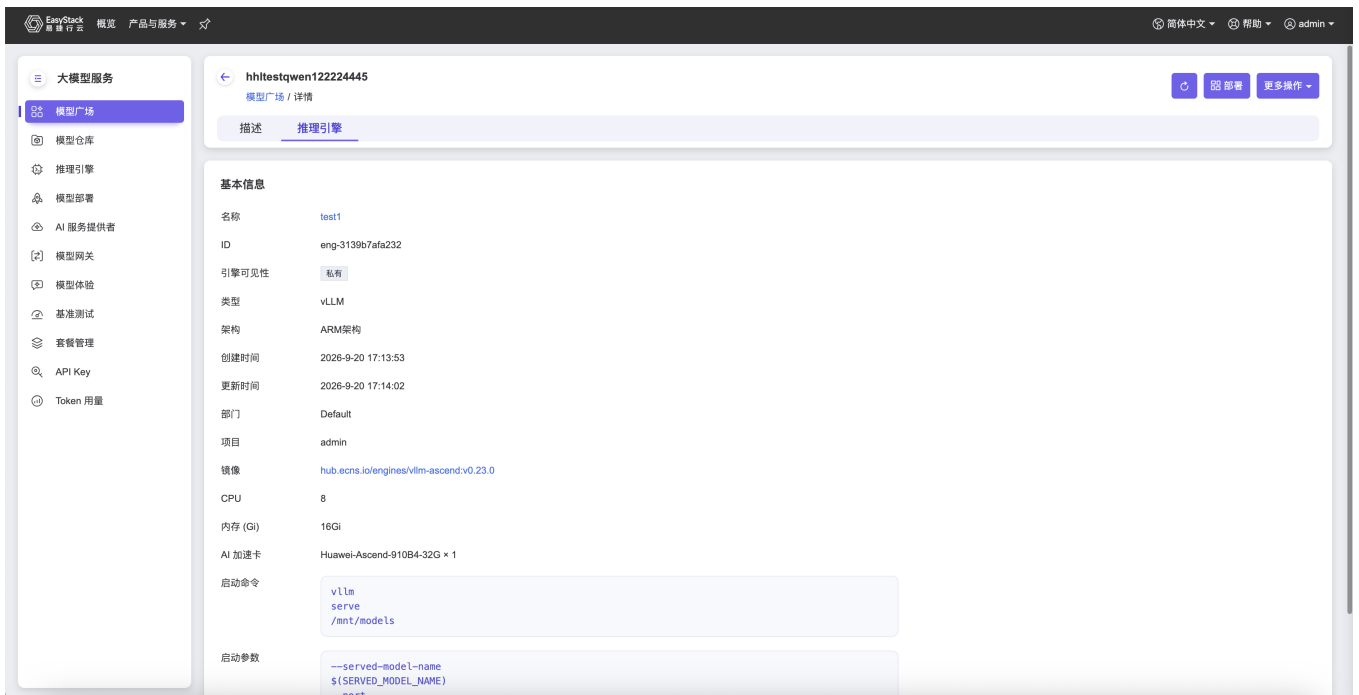
“模型描述”区域中显示名称、可见性、ID、提供商、创建时间、描述、创建者、来源名称、任务类型、模型类型和更新时间等信息。

“模型文件摘要”区域中显示模型文件、文件状态、来源类型、OCI地址、仓库、来源名称和当前阶段等信息。

单击模型文件名称，可进入对应的模型文件页面查看详细信息。

查看推理引擎

在模型详情页面中，选择 **推理引擎** 页签，可查看当前模型关联的推理引擎及其运行配置。



推理引擎信息包括名称、ID、引擎可见性、类型、架构、创建时间、更新时间、部门、项目、镜像、CPU、内存、AI加速卡、启动命令、启动参数和环境变量等。

部署模型

1. 在“模型广场”页面中，选择目标模型，进入模型详情页面。
2. 单击 **部署**，进入创建模型部署页面。
3. 根据页面提示配置模型部署信息并创建模型部署。

模型部署的具体配置和管理操作，请参考模型部署相关章节。

1.2 模型仓库

本章节主要介绍在模型仓库中查看模型文件、通过 ModelScope 或 Hugging Face 地址导入模型、上传本地模型、设置仓库容量，以及查看、编辑和删除模型文件等操作。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【模型仓库】，即可进入“模型仓库”页面。

查看模型文件

“模型仓库”页面左侧显示仓库列表，右侧显示所选仓库中的模型文件、任务状态和导入进度。

- 在“仓库列表”区域中，选择 **公共仓库** 或 **私有仓库**。
- 在右侧列表中查看所选仓库中的模型文件。

模型文件列表显示名称、来源、模型文件地址、状态、大小或进度，以及可执行的操作。模型文件处于处理过程中时，还会显示当前处理阶段和任务状态。

单击模型文件左侧的“展开”图标，可查看该模型文件包含的文件路径、各文件的传输进度和状态。

需要重新获取模型文件列表时，可单击列表上方的“刷新”图标。需要筛选模型文件时，可通过列表右上角的筛选框设置过滤条件。

The screenshot displays the 'Model Warehouse' (模型仓库) interface. On the left, a sidebar shows navigation options like 'Model Warehouse' (模型仓库), 'Model Marketplace' (模型广场), 'Inference Engine' (推理引擎), etc. The main area is divided into two sections: 'Warehouse List' (仓库列表) and 'Model Files' (模型文件).

Warehouse List: Shows two warehouses: 'Public Warehouse' (公共仓库) and 'Private Warehouse' (私有仓库). The 'Public Warehouse' is selected, showing its capacity (500 GB) and usage (2.98 GB).

Model Files: A table lists model files with columns for Name, Source, Model File Address, Status, Size/Progress, and Actions. The table shows three files: '千问3.8-27B', 'Qwen2.5-1.5B-Instruct', and 'qwentest2233'. The '千问3.8-27B' file is in the 'Importing' (导入中) state, while the others are 'Completed' (就绪).

File Details: Below the table, a detailed view of the 'qwentest2233' file shows its internal structure with columns for File Path, Progress, and Status. All listed files (config.json, configuration.json, generation_config.json, LICENSE, merges.txt, model.safetensors, README.md) show 100% progress and 'Success' (成功) status.

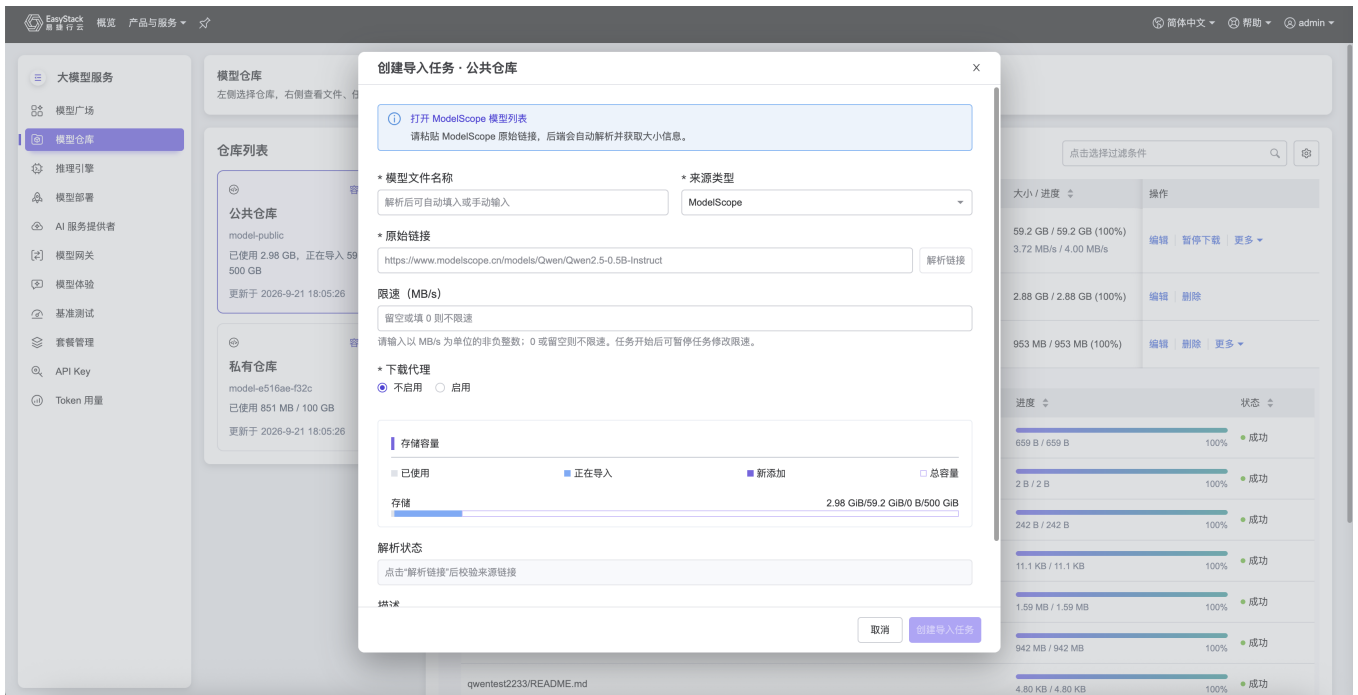
模型文件常见状态包括：

状态	说明
导入中	模型文件正在从模型来源下载并导入仓库。页面同时显示当前处理阶段、任务状态和导入进度。
就绪	模型文件已经处理完成，可以用于创建模型等后续操作。
成功	模型文件或对应任务已经执行成功。

创建导入任务

本操作用于通过 ModelScope 或 Hugging Face 模型链接，将模型文件导入模型仓库。

- 在“仓库列表”区域中，选择模型文件需要导入的仓库。
- 单击 **创建导入任务**，弹出“创建导入任务”对话框。对话框标题中同时显示当前选择的仓库。
- 配置参数后，单击 **创建导入任务**，完成操作。



参数	说明
模型文件名称	导入后模型文件在模型仓库中的名称。解析来源链接后，系统可自动填入该参数，也可以手动输入。
来源类型	模型文件的来源平台，可选ModelScope或Hugging Face。
原始链接	模型文件在所选来源平台中的链接。输入链接后，单击 解析链接 ，系统将解析模型信息并获取模型大小。
限速（MB/s）	模型文件的下载速度限制。请输入以MB/s为单位的非负整数，输入0或留空表示不限速。
下载代理	是否为当前导入任务启用下载代理。该参数值可选 不启用 或 启用 。
代理地址	启用下载代理后，输入用于来源链接解析和模型下载的代理地址。支持HTTP或HTTPS代理地址，地址中可以包含用户名和密码。
附加代理参数	启用下载代理后，根据需要设置HTTP_PROXY、HTTPS_PROXY、ALL_PROXY或NO_PROXY等代理参数。
描述	该模型文件的说明信息。

说明：

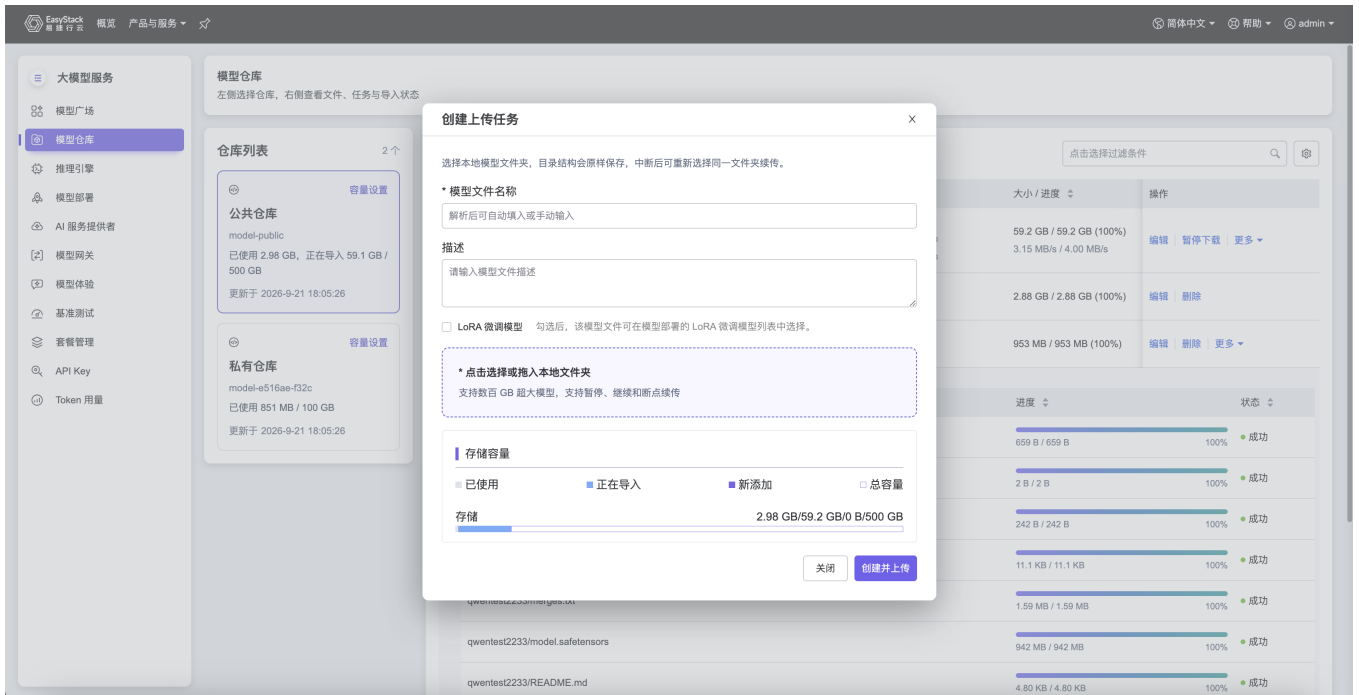
- 可单击对话框中的 **打开 ModelScope 模型列表**，进入ModelScope模型列表查找模型。
- 输入原始链接后，需要单击 **解析链接**。链接解析完成后，页面显示解析状态，并更新模型文件名称和存储容量信息。
- 任务开始后，可先暂停任务，再修改下载限速。
- “存储容量”区域显示仓库的已使用容量、正在导入容量、本次新增加容量和总容量。

创建任务后，可在模型文件列表中查看导入阶段、任务状态、下载速度和导入进度。模型文件正在下载时，可单击操作列中的 **暂停下载** 暂停任务。

创建上传任务

本操作用于将本地模型文件夹上传至模型仓库。上传时，本地文件夹的目录结构将原样保存。

1. 在“仓库列表”区域中，选择模型文件需要上传的仓库。
2. 单击 **创建上传任务**，弹出“创建上传任务”对话框。
3. 配置模型文件信息。
4. 单击文件选择区域，或者将本地模型文件夹拖入该区域。
5. 单击 **创建并上传**，完成操作。



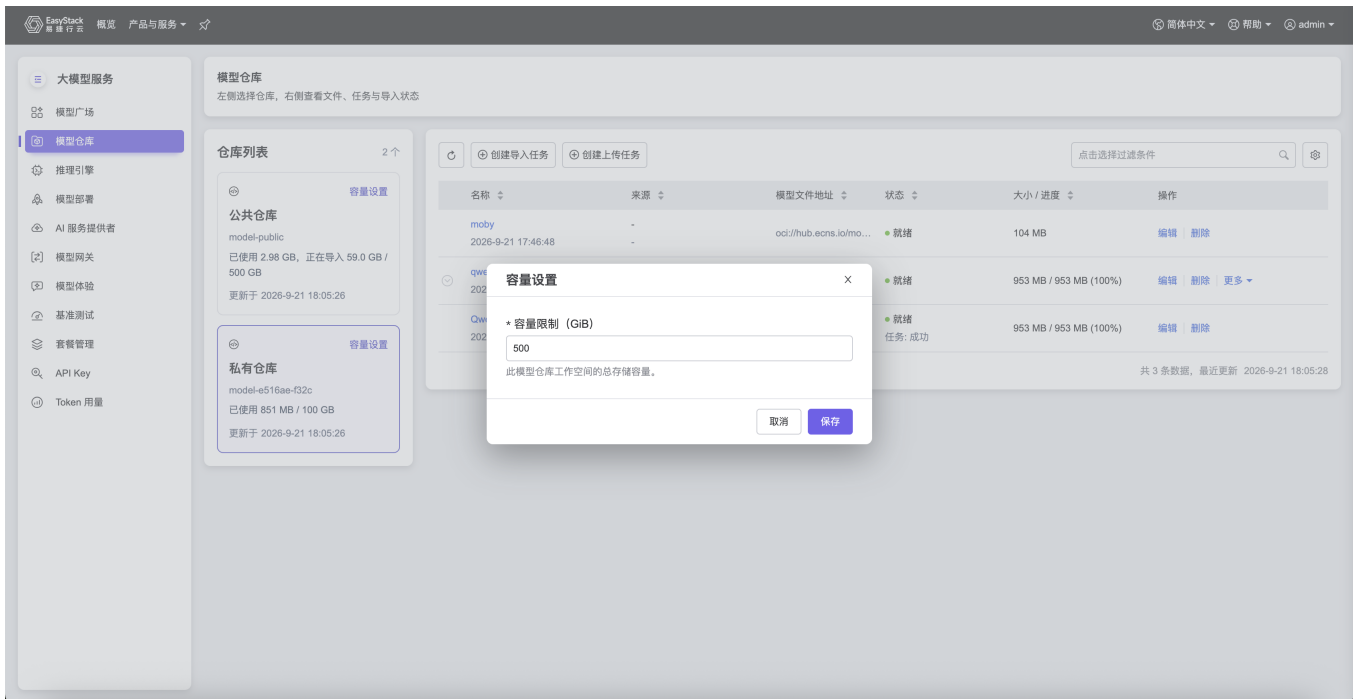
参数	说明
模型文件名称	上传后模型文件在模型仓库中的名称。系统解析文件夹后可自动填入该参数，也可以手动输入。
描述	该模型文件的说明信息。
LoRA微调模型	该模型文件是否为LoRA微调模型。勾选后，该模型文件可在模型部署的LoRA微调模型列表中选择。
本地文件夹	需要上传的本地模型文件夹。支持单击选择或直接拖入。

说明：

- 本地上传支持数百GB的大模型文件夹。
- 上传过程支持暂停、继续和断点续传。上传中断后，可重新选择同一本地文件夹继续上传。
- “存储容量”区域显示仓库的已使用容量、正在导入容量、本次新增加容量和总容量。

设置仓库容量

1. 在“仓库列表”区域中，找到待操作仓库。
2. 单击仓库卡片中的 **容量设置**，弹出“容量设置”对话框。
3. 配置参数后，单击 **保存**，完成操作。

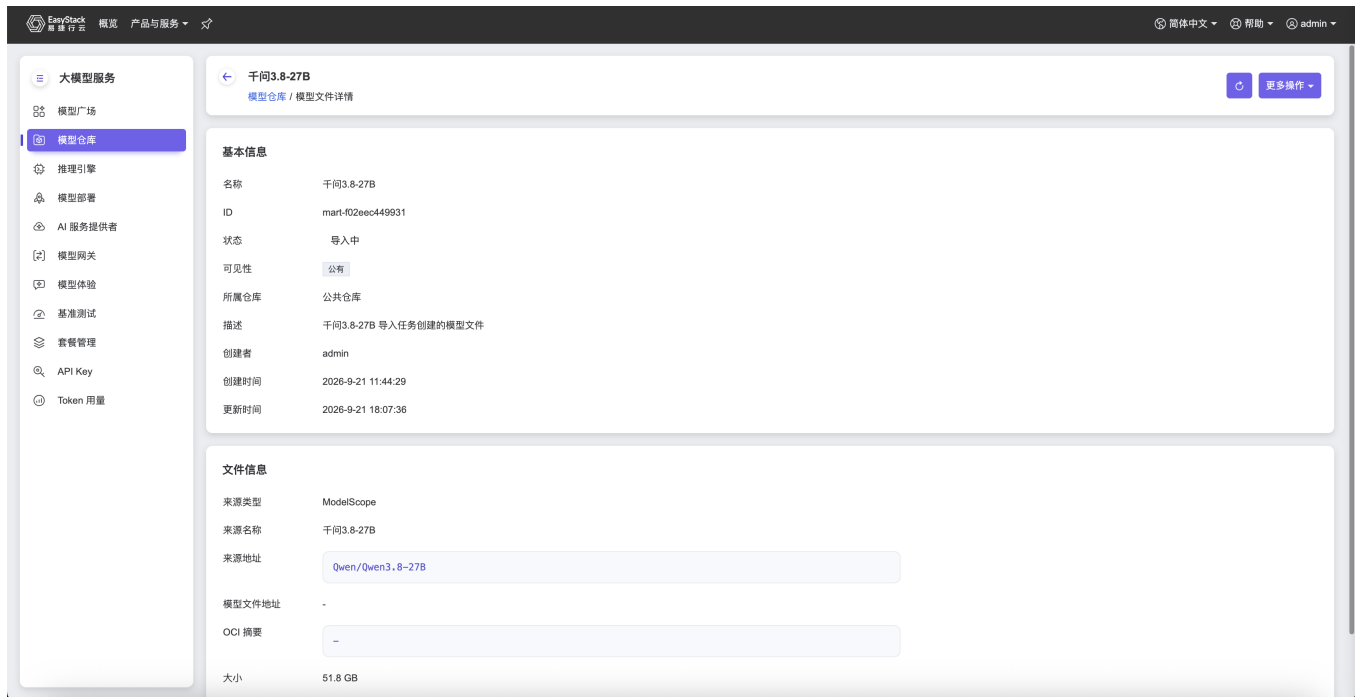


参数	说明
容量限制（GiB）	该模型仓库工作空间的总存储容量。

设置完成后，仓库卡片中显示已使用容量、正在导入容量和仓库总容量。

查看模型文件详情

在“模型仓库”页面中，单击模型文件名称，可进入“模型文件详情”页面。



“基本信息”区域中显示名称、ID、状态、可见性、所属仓库、描述、创建者、创建时间和更新时间等信息。

“文件信息”区域中显示来源类型、来源名称、来源地址、模型文件地址、OCI摘要、大小、处理阶段和关联模型等信息。

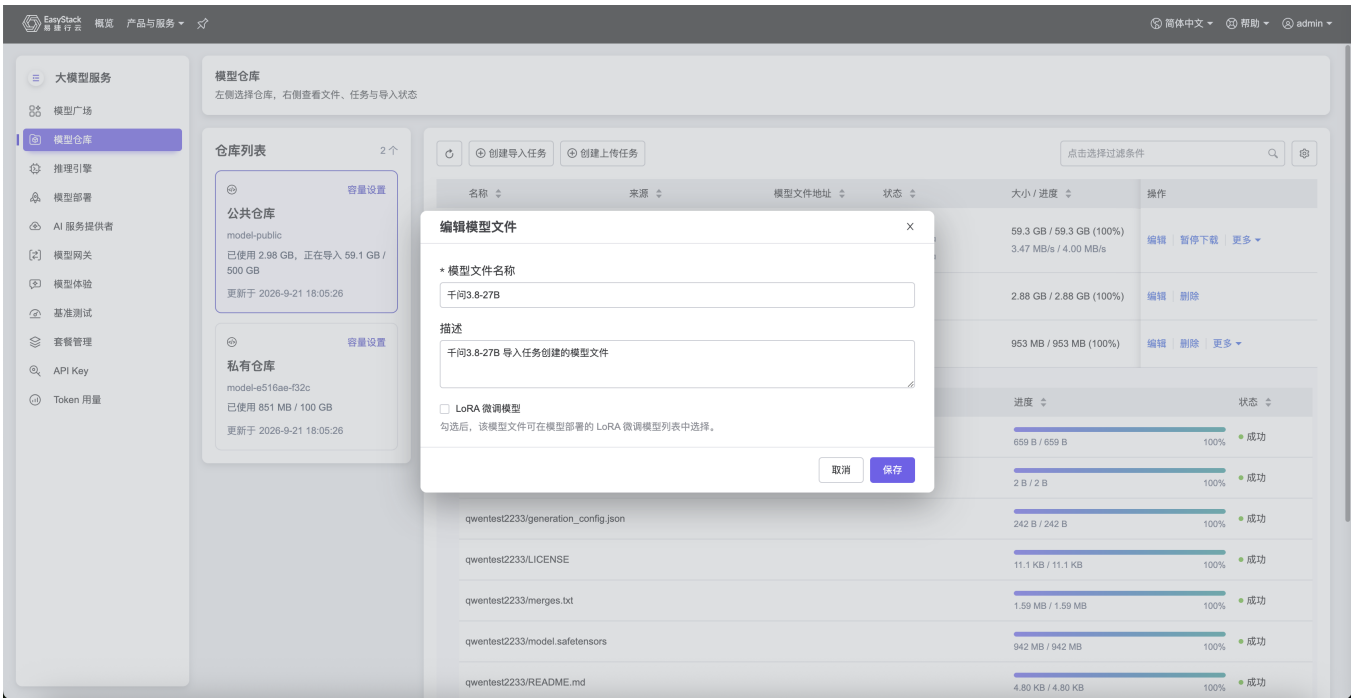
编辑模型文件

可以从模型文件列表或模型文件详情页面编辑模型文件。

- 在“模型仓库”页面中，找到待操作模型文件，单击操作列中的 **编辑**。
- 在“模型文件详情”页面中，依次单击 **更多操作** - **编辑**。

完成上述任一操作后，弹出“编辑模型文件”对话框。

- 修改模型文件信息。
- 单击 **保存**，完成操作。



参数	说明
模型文件名称	该模型文件在模型仓库中的名称。
描述	该模型文件的说明信息。
LoRA微调模型	该模型文件是否为LoRA微调模型。勾选后，该模型文件可在模型部署的LoRA微调模型列表中选择。

删除模型文件

- 在“模型仓库”页面中，找到待操作模型文件。
- 单击操作列中的 **删除** 。
- 根据页面提示完成删除操作。

1.3 推理引擎

本章节主要介绍在推理引擎页面中查看和创建推理引擎，以及查看详情、编辑、重命名、克隆和删除推理引擎等操作。推理引擎用于配置模型推理所需的镜像、启动命令、启动参数和算力资源，可在模型部署时复用。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【推理引擎】，即可进入“推理引擎”页面。

查看推理引擎

在“推理引擎”页面中，可查看当前已经创建的推理引擎。列表中显示名称、类型、镜像、CPU、内存、AI加速卡、部门/项目和创建时间等信息。

需要重新获取推理引擎列表时，可单击列表上方的“刷新”图标。需要查找推理引擎时，可通过列表右上角的筛选框设置过滤条件。

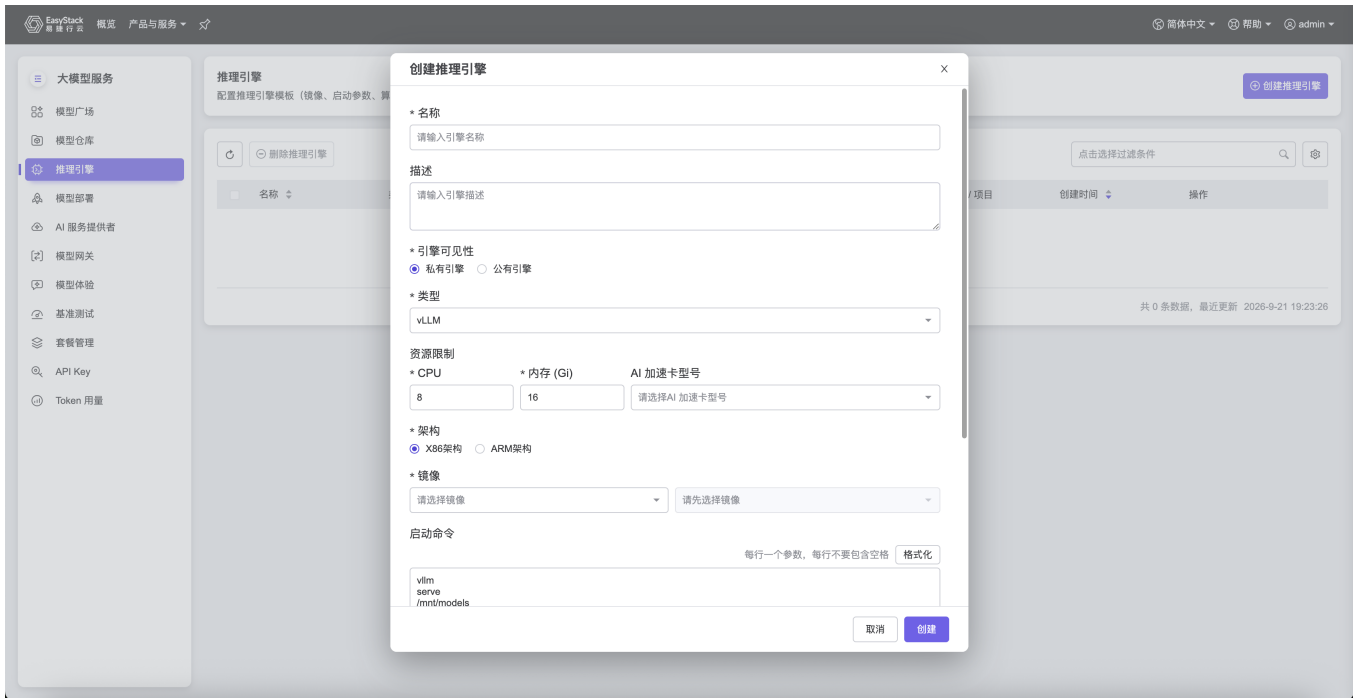
名称	类型	镜像	CPU	内存 (Gi)	AI 加速卡	部门 / 项目	创建时间	操作
推理引擎 2	公有	vLLM	8	16Gi	-	Default / admin	2026-9-21 19:24:16	编辑 重命名 更多
推理引擎 1	私有	vLLM	8	16Gi	-	Default / admin	2026-9-21 19:24:02	编辑 重命名 更多

共 2 条数据, 最近更新 2026-9-21 19:24:16

单击推理引擎名称，可进入对应的详情页面。

创建推理引擎

- 在“推理引擎”页面中，单击 **创建推理引擎**，弹出“创建推理引擎”对话框。
- 配置参数后，单击 **创建**，完成操作。



参数	说明
名称	该推理引擎的名称。
描述	该推理引擎的说明信息。
引擎可见性	该推理引擎的可见性。该参数值可选 私有引擎 或 公有引擎 。
类型	该推理引擎的类型，例如vLLM。
CPU	运行该推理引擎需要使用的CPU资源。
内存（Gi）	运行该推理引擎需要使用的内存资源。
AI加速卡型号	运行该推理引擎需要使用的AI加速卡型号。
数量（张）	运行该推理引擎需要使用的AI加速卡数量。
架构	推理引擎适用的处理器架构。该参数值可选 X86架构 或 ARM架构 。
镜像	推理引擎使用的容器镜像及镜像版本。
启动命令	容器启动时执行的命令。每行输入一个参数，每行不要包含空格。
启动参数	推理引擎启动时使用的参数。每行输入一个参数，每行不要包含空格。

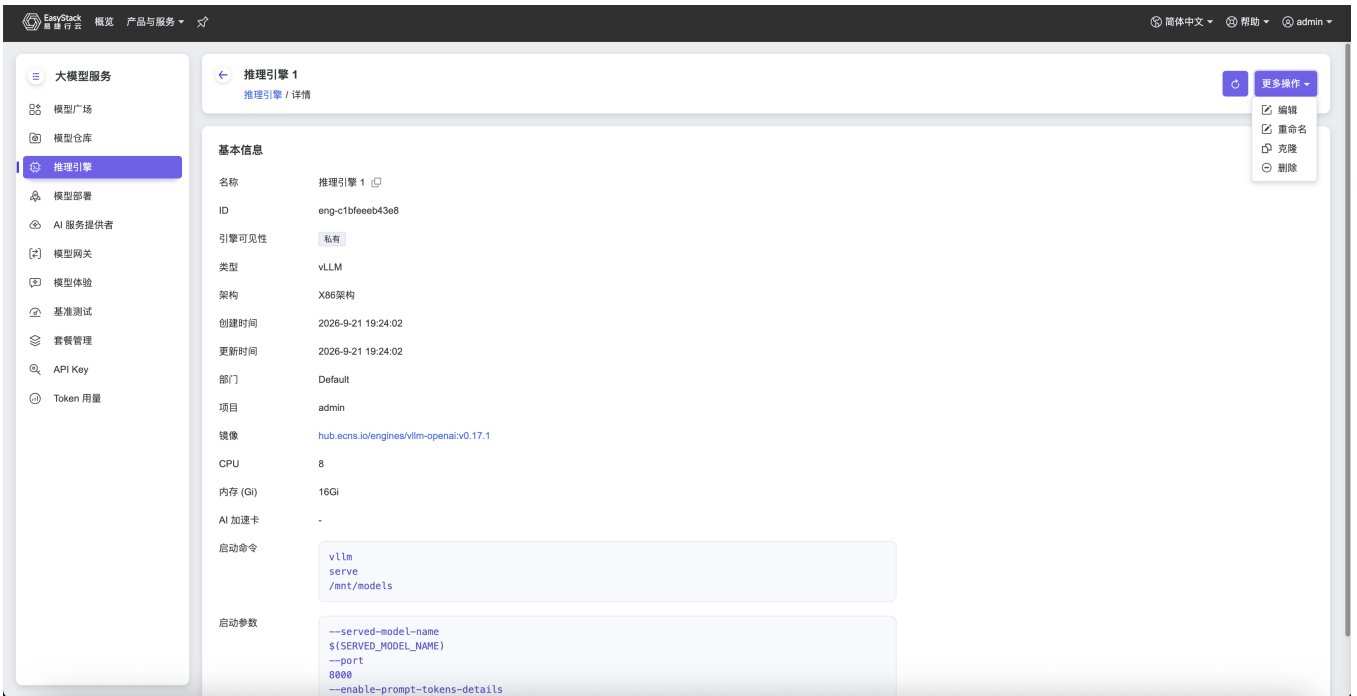
参数	说明
环境变量	推理引擎运行时使用的环境变量。分别输入变量名和值；需要配置多个环境变量时，可单击 添加 继续配置。

说明：

- 页面中带有“*”的参数为必填参数。
- 选择AI加速卡型号后，页面显示对应的资源名称及当前单节点可用数量。
- 配置多张AI加速卡时，请确认模型与推理后端兼容对应的张量并行数量。页面提示通常使用2张、4张或8张AI加速卡时兼容性更好。
- 默认模型根目录为 `/mnt/models`。
- 配置启动命令或启动参数后，可单击 [格式化](#) 整理输入内容。
- 平台仅支持使用8000端口。
- 启动参数支持使用 `${SERVED_MODEL_NAME}` 变量。在创建模型部署时，系统可将该变量动态替换为模型名称。

查看详情

在“推理引擎”页面中，单击推理引擎名称，可进入“推理引擎详情”页面。



详情页面显示以下信息：

信息	说明
名称	推理引擎的名称。
ID	推理引擎在系统中的唯一标识。
引擎可见性	推理引擎的可见性类型。
类型	推理引擎的类型。
架构	推理引擎适用的处理器架构。
创建时间	推理引擎的创建时间。
更新时间	推理引擎信息的最近更新时间。
部门	推理引擎所属的部门。
项目	推理引擎所属的项目。
镜像	推理引擎使用的容器镜像。
CPU	推理引擎配置的CPU资源。

信息	说明
内存 (Gi)	推理引擎配置的内存资源。
AI加速卡	推理引擎配置的AI加速卡型号和数量。
启动命令	推理引擎的容器启动命令。
启动参数	推理引擎启动时使用的参数。
环境变量	推理引擎运行时使用的环境变量。

需要重新获取详情信息时，可单击页面右上角的“刷新”图标。

编辑推理引擎

可以从推理引擎列表或推理引擎详情页面编辑推理引擎。

- 在“推理引擎”页面中，找到待操作推理引擎，单击操作列中的 **编辑**。
- 在“推理引擎详情”页面中，依次单击 **更多操作** - **编辑**。

修改推理引擎的配置后，根据页面提示完成保存。

重命名推理引擎

可以从推理引擎列表或推理引擎详情页面重命名推理引擎。

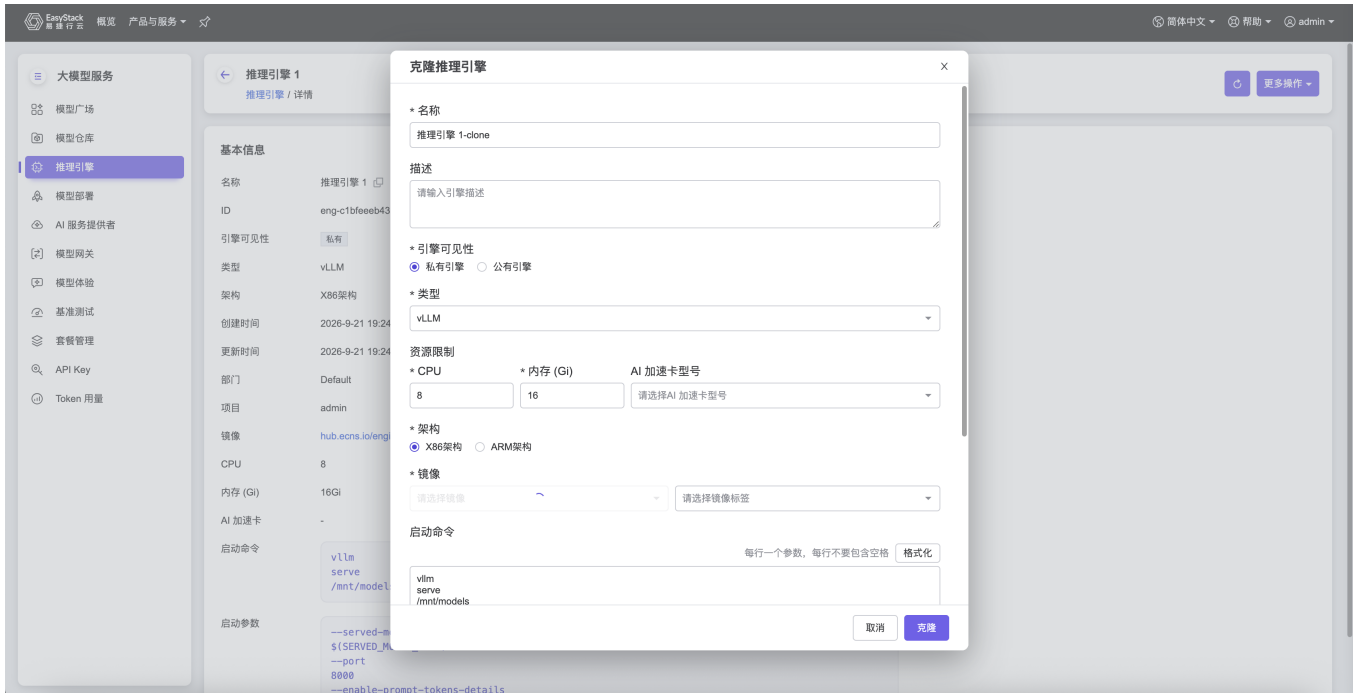
- 在“推理引擎”页面中，找到待操作推理引擎，单击操作列中的 **重命名**。
- 在“推理引擎详情”页面中，依次单击 **更多操作** - **重命名**。

输入新的推理引擎名称后，根据页面提示完成操作。

克隆推理引擎

克隆推理引擎时，系统将已有推理引擎的配置填入“克隆推理引擎”对话框。您可以在原配置的基础上修改名称、资源和运行参数，快速创建新的推理引擎。

1. 在“推理引擎”页面中，单击待操作推理引擎操作列中的 **更多** - **克隆**；也可以进入推理引擎详情页面，依次单击 **更多操作** - **克隆**。
2. 在“克隆推理引擎”对话框中修改名称、描述、引擎可见性、算力资源、架构、镜像、启动命令、启动参数或环境变量。
3. 单击 **克隆**，完成操作。



删除推理引擎

可以删除单个推理引擎，也可以在列表中批量删除推理引擎。

删除单个推理引擎

1. 在“推理引擎”页面中，找到待操作推理引擎，单击操作列中的 **更多** - **删除**；也可以进入推理引擎详情页面，依次单击 **更多操作** - **删除**。
2. 根据页面提示完成删除操作。

批量删除推理引擎

1. 在“推理引擎”页面中，勾选一个或多个待操作推理引擎。
2. 单击列表上方的 **删除推理引擎**。

3. 根据页面提示完成删除操作。

1.4 模型部署

本章节主要介绍模型部署的创建和运行管理操作，包括查看模型部署、创建模型部署、查看基本信息和运行状态、查看监控指标与日志、使用终端，以及启动、停止、重命名和删除模型部署等。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【模型部署】，即可进入“模型部署”页面。

查看模型部署

在“模型部署”页面中，可查看当前已经创建的模型部署。列表中显示名称、模型文件、推理引擎、副本数、CPU/内存、AI加速卡、状态、部门/项目和创建时间等信息。

需要重新获取模型部署列表时，可单击列表上方的“刷新”图标。需要查找模型部署时，可通过列表右上角的筛选框设置过滤条件。

名称	模型文件	推理引擎	副本	CPU / 内存	AI 加速卡	状态	部门 / 项目	操作
hhtestqwen12224445-de...	qwen25test	test1	1	8 / 16Gi	Huawei-Ascend-910B4-32G × 1	未就绪	Default / admin	停止 启动 更多
fxj-rerank	bge-reranker-v2-m3	fxj-vllm	1	8 / 16Gi	Huawei-Ascend-910B4-32G × 1	未就绪	fxj / fxj	停止 启动 更多
testqwen2515-deploy	Qwen2.5-1.5B-Instruct	qwenvllm	1	8 / 16Gi	Huawei-Ascend-910B4-32G × 1	未就绪	Default / admin	停止 启动 更多
test-ordinary-module	Qwen2.5-0.5B-Instruct	test_ordinary	1	8 / 16Gi	Huawei-Ascend-910B4-32G × 1	未就绪	Default / admin	停止 启动 更多
test2	Qwen2.5-0.5B-Instruct	test-2	1	16 / 32Gi	Huawei-Ascend-910B4-32G × 2	未就绪	Default / admin	停止 启动 更多
ecms-npu-2	Qwen2.5-0.5B-Instruct	ecms-npu-2	1	16 / 32Gi	Huawei-Ascend-910B4-32G × 2	未就绪	Default / admin	停止 启动 更多

共 6 条数据，最近更新 2026-9-21 20:13:37

模型部署常见状态包括：

状态	说明
未就绪	模型部署尚未进入可以正常提供服务的状态。

状态	说明
就绪	模型部署已经准备完成，可以正常提供服务。

单击模型部署名称，可进入对应的详情页面。单击模型文件或推理引擎名称，可查看对应资源的信息。

创建模型部署

前提条件

- 模型文件已经添加至模型仓库，或者已经获取可用的模型文件OCI地址。
- 已经创建与目标模型和运行环境匹配的推理引擎。
- 当前项目具有模型部署所需的CPU、内存和AI加速卡资源。

操作步骤

1. 在“模型部署”页面中，单击 **新建模型部署**，弹出“新建模型部署”对话框。
2. 配置参数后，单击 **创建**，完成操作。

也可以在“模型广场”中进入目标模型的详情页面，单击 **部署**，进入模型部署创建页面。



参数	说明
部署名称	该模型部署的名称。
副本数	该模型部署需要创建的运行副本数量。
推理引擎	该模型部署使用的推理引擎。模型部署默认使用所选推理引擎的镜像、启动参数和算力资源等配置。
架构	所选推理引擎适用的处理器架构。
模型文件	模型文件的配置方式。该参数值可选 从模型仓库选择 或 手动输入 OCI 地址 。
模型仓库	选择 从模型仓库选择 时，选择模型文件所在的模型仓库。
模型文件	选择 从模型仓库选择 时，从所选仓库中选择用于部署的模型文件。
OCI地址	模型文件对应的OCI地址。选择 从模型仓库选择 后，页面显示所选模型文件的OCI地址；选择 手动输入 OCI 地址 时，需要输入模型文件的OCI地址。
SERVED_MODEL_NAME	模型部署对外提供服务时使用的模型名称。填写后，系统将以环境变量 SERVED_MODEL_NAME 的形式注入部署容器。
启动命令	当前模型部署使用的启动命令。每行输入一个参数，每行不要包含空格。

参数	说明
启动参数	当前模型部署使用的启动参数。每行输入一个参数，每行不要包含空格。
环境变量	当前模型部署使用的环境变量。分别输入变量名和值；需要配置多个环境变量时，可单击 添加 继续配置。

说明：

- 页面中带有“*”的参数为必填参数。
- 选择推理引擎后，模型部署默认使用该推理引擎的镜像、启动命令、启动参数和算力资源等配置。
- `SERVED_MODEL_NAME` 需要在推理引擎的启动参数中被引用后才会生效，例如 `--served-model-name ${SERVED_MODEL_NAME}`。
- 创建页面右侧的资源配额区域显示CPU和内存的已使用资源、本次新增资源及总配额。

高级配置

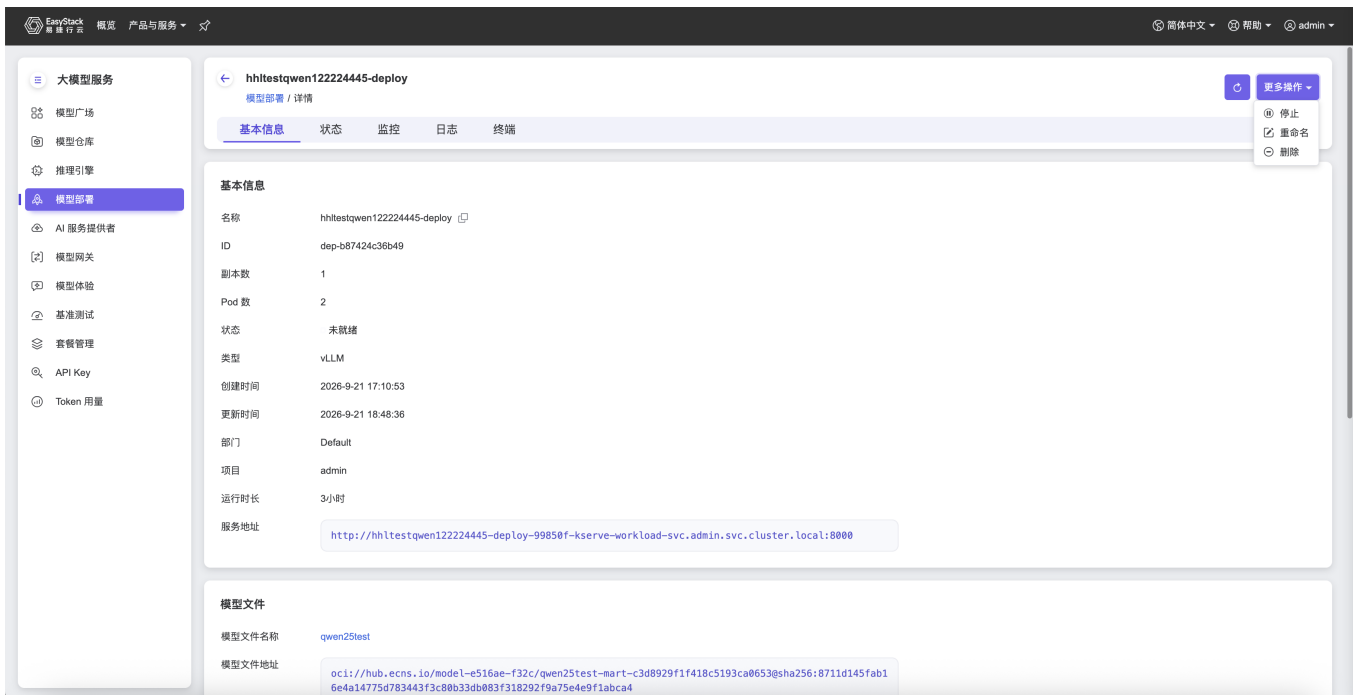
需要覆盖所选推理引擎的启动命令、启动参数或环境变量时，单击 [高级配置（覆盖启动命令/参数/环境变量）](#) 展开高级配置。

高级配置仅应用于当前模型部署，不会修改原推理引擎。配置项留空时，模型部署继续使用所选推理引擎中的默认值。

配置启动命令或启动参数后，可单击 [格式化](#) 整理输入内容。

查看基本信息

在“模型部署”页面中，单击模型部署名称进入详情页面，然后选择 [基本信息](#) 页签。



“基本信息”区域中显示名称、ID、副本数、Pod数、状态、类型、创建时间、更新时间、部门、项目、运行时长和服务地址等信息。

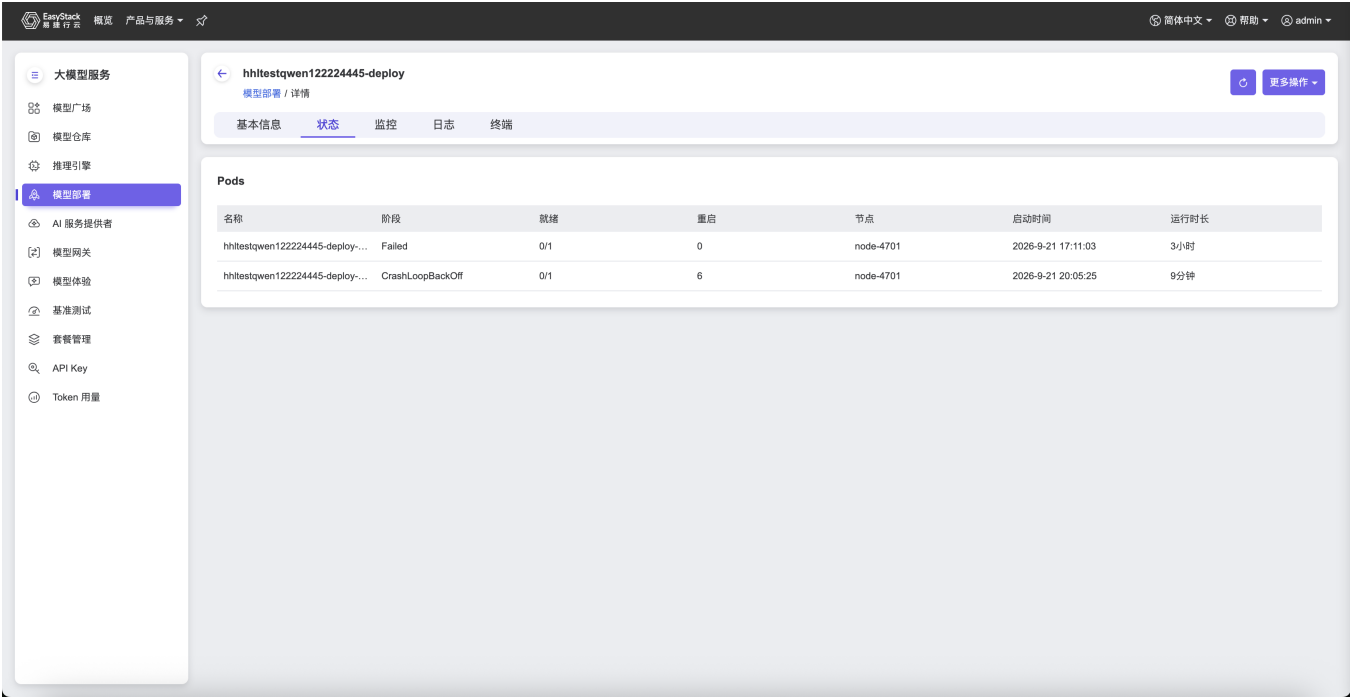
“模型文件”区域中显示模型文件名称、模型文件地址和模型名称等信息。

“推理引擎”区域中显示推理引擎名称、镜像、CPU、内存、AI加速卡、启动命令、启动参数和环境变量等信息。

需要重新获取模型部署信息时，可单击页面右上角的“刷新”图标。

查看运行状态

在模型部署详情页面中，选择 **状态** 页签，可查看该模型部署包含的Pod及其运行状态。

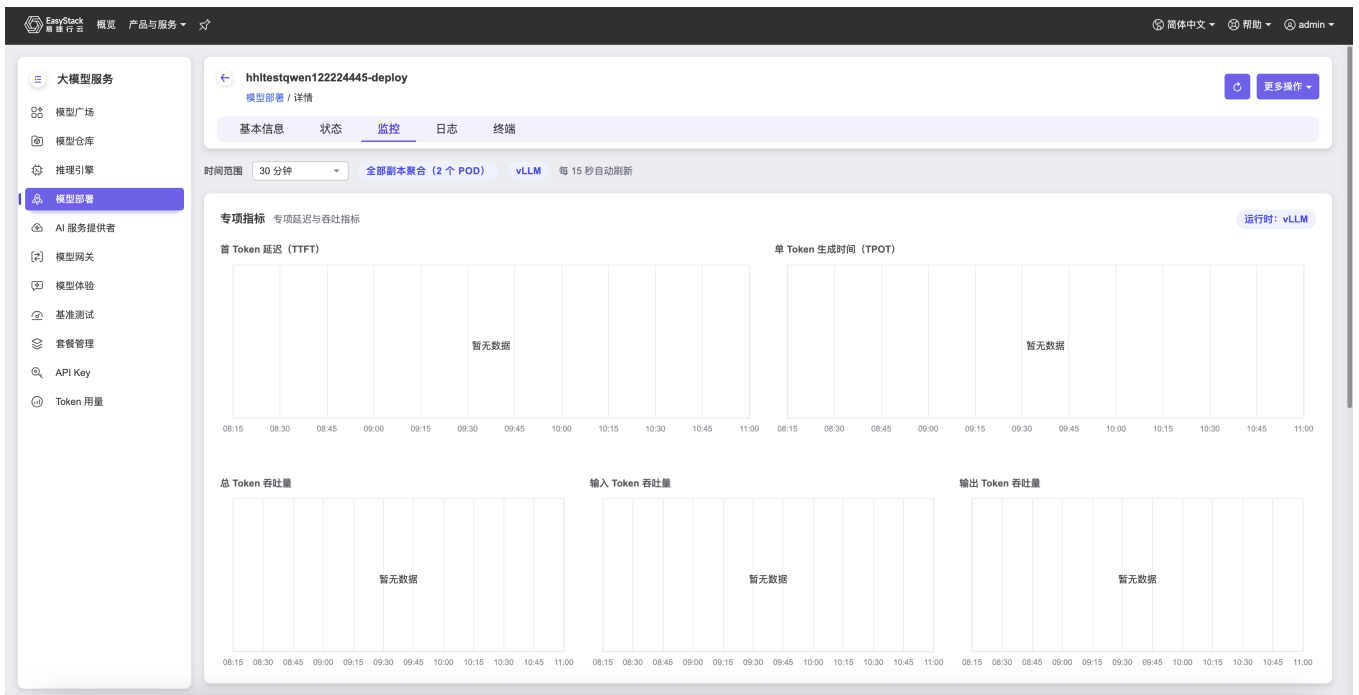


Pod列表显示以下信息：

信息	说明
名称	Pod的名称。
阶段	Pod当前所处的运行阶段。
就绪	Pod中已经就绪的容器数量和容器总数。
重启	Pod的重启次数。
节点	Pod所在的计算节点。
启动时间	Pod的启动时间。
运行时长	Pod从启动至当前的运行时长。

查看监控

在模型部署详情页面中，选择 监控 页签，可查看模型服务专项指标和运行时资源指标。



页面支持选择监控时间范围和目标副本，并按照固定间隔自动刷新监控数据。

专项指标包括：

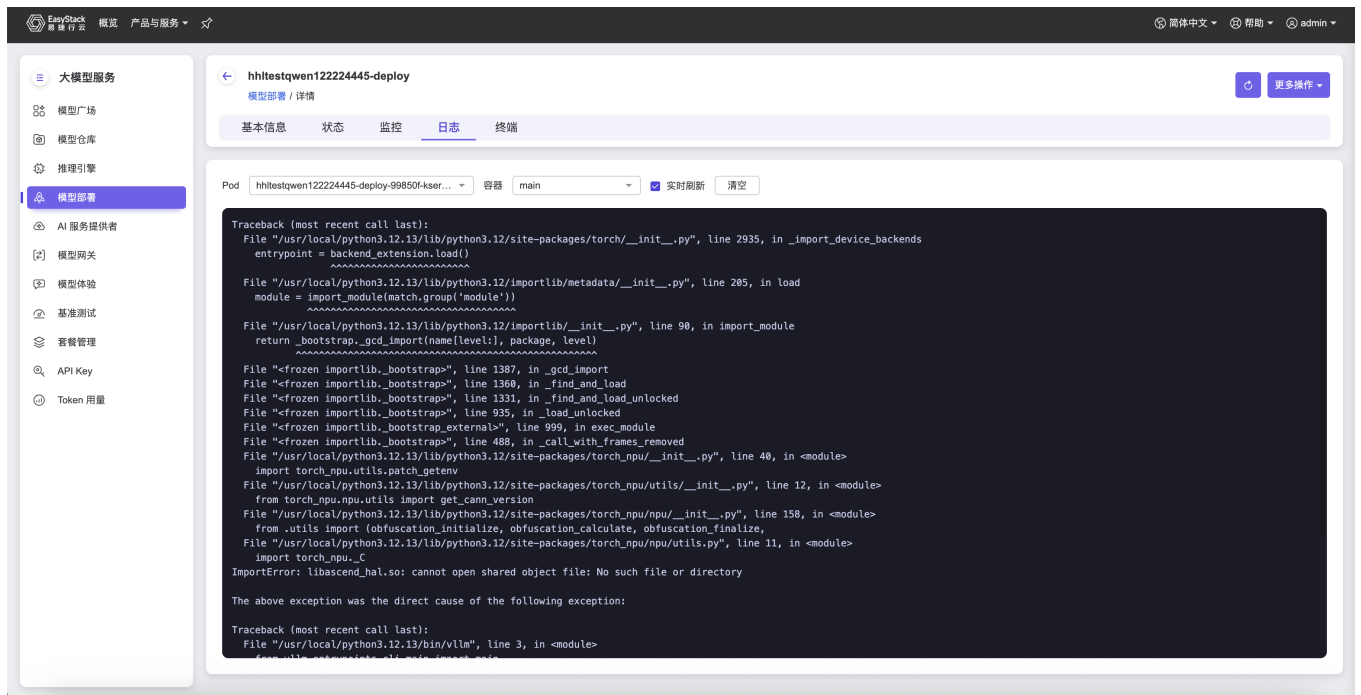
- 首Token延迟（TTFT）。
- 单Token生成时间（TPOT）。
- 总Token吞吐量。
- 输入Token吞吐量。
- 输出Token吞吐量。

运行时资源指标包括：

- CPU使用。
- 内存使用。
- 网络接收速率。
- 网络发送速率。

查看日志

在模型部署详情页面中，选择 **日志** 页签，可查看模型部署中指定Pod和容器的运行日志。

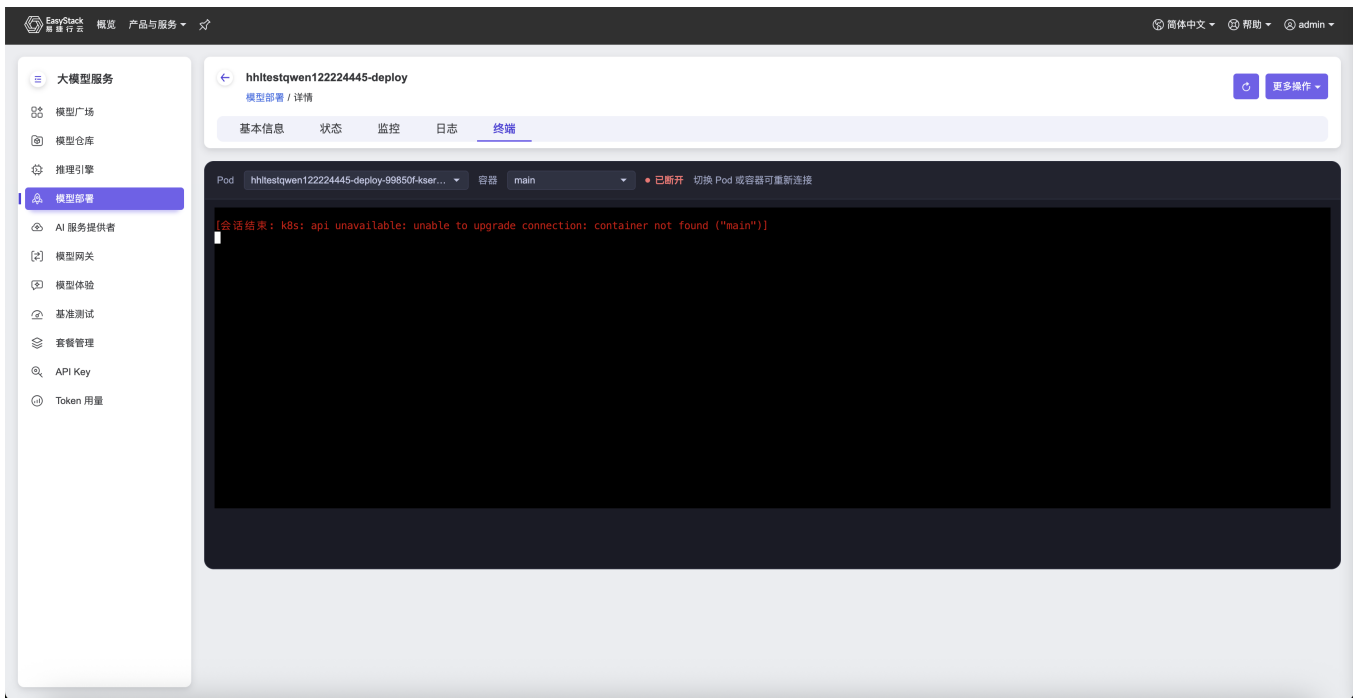


1. 在 **Pod** 下拉列表中选择需要查看的Pod。
2. 在 **容器** 下拉列表中选择需要查看的容器。
3. 查看所选容器输出的运行日志。

勾选 **实时刷新** 后，页面持续获取最新日志。需要清除页面中已经显示的日志时，可单击 **清空**。

使用终端

在模型部署详情页面中，选择 **终端** 页签，可进入模型部署中指定容器的终端。



1. 在 **Pod** 下拉列表中选择需要连接的Pod。
2. 在 **容器** 下拉列表中选择需要连接的容器。
3. 等待连接状态显示为 **已连接** 后，在终端中执行所需命令。

切换Pod或容器后，终端将重新建立连接。

启动模型部署

模型部署处于停止状态时，可执行启动操作。

1. 在“模型部署”页面中，找到待操作模型部署。
2. 单击操作列中的 **启动**。
3. 根据页面提示完成操作。

停止模型部署

可以从模型部署列表或详情页面停止模型部署。

- 在“模型部署”页面中，找到待操作模型部署，单击操作列中的 **停止**。
- 在模型部署详情页面中，依次单击 **更多操作** - **停止**。

根据页面提示完成操作。

重命名模型部署

1. 在模型部署详情页面中，依次单击 **更多操作** - **重命名**。
2. 输入新的模型部署名称。
3. 根据页面提示完成操作。

删除模型部署

可以删除单个模型部署，也可以在列表中批量删除模型部署。

删除单个模型部署

1. 在“模型部署”页面中，找到待操作模型部署，单击操作列中的 **更多** - **删除**；也可以进入模型部署详情页面，依次单击 **更多操作** - **删除**。
2. 根据页面提示完成删除操作。

批量删除模型部署

1. 在“模型部署”页面中，勾选一个或多个待操作模型部署。
2. 单击列表上方的 **删除模型部署**。
3. 根据页面提示完成删除操作。

1.5 AI 服务提供者

本章节主要介绍AI服务提供者的创建和管理操作，包括查看、创建、编辑、查看详情和删除AI服务提供者等。AI服务提供者用于接入外部AI服务或云平台内部的模型部署，创建后可供模型网关选择和路由。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【AI 服务提供者】，即可进入“AI 服务提供者”页面。

查看AI服务提供者

在“AI 服务提供者”页面中，可查看当前已经创建的AI服务提供者。列表中显示名称、API规范、访问地址、关联部署、状态、部门/项目和创建时间等信息。

AI服务提供者名称后显示其类型：

- 外部** 表示接入OpenAI、DeepSeek等外部AI服务。
- 内部** 表示接入云平台内部的模型部署。

需要重新获取AI服务提供者列表时，可单击列表上方的“刷新”图标。需要查找AI服务提供者时，可通过列表右上角的筛选框设置过滤条件。

AI 服务提供者

接入外部（OpenAI / DeepSeek 等）或内部 AI 服务，供模型网关路由选用

创建 AI 服务提供者

删除提供者

点击选择过滤条件

名称	API 规范	访问地址	关联部署	状态	部门 / 项目	操作
testqwen2515-deploy 内部	OpenAI	http://testqwen2515-deploy-1357b0-kserve-worklo...	testqwen2515-deploy	就绪	Default / admin	编辑 删除
fxj-rerank 内部	Cohere	http://fxj-rerank-73498e-kserve-workload-svc.fxj.sv...	fxj-rerank	就绪	fxj / fxj	编辑 删除
test_ordinary-provider 内部	OpenAI	http://test-ordinary-module-65861a-kserve-workloa...	test-ordinary-module	就绪	Default / admin	编辑 删除
test-2-provider 内部	OpenAI	http://test2-0d2805-kserve-workload-svc.admin.svc...	test2	就绪	Default / admin	编辑 删除
ecms-npu 内部	OpenAI	http://ecms-npu-2-a89670-kserve-workload-svc.ad...	ecms-npu-2	就绪	Default / admin	编辑 删除
fxj-backup 外部	OpenAI	http://fxj-image-mock-fxj-agent-router-test.svc.clust...	-	就绪	fxj / fxj	编辑 删除
fxj-primary 外部	OpenAI	http://fxj-image-mock-primary-fxj-agent-router-test...	-	就绪	fxj / fxj	编辑 删除

共 7 条数据，最近更新 2026-9-21 20:17:01

单击AI服务提供者名称，可进入对应的详情页面。

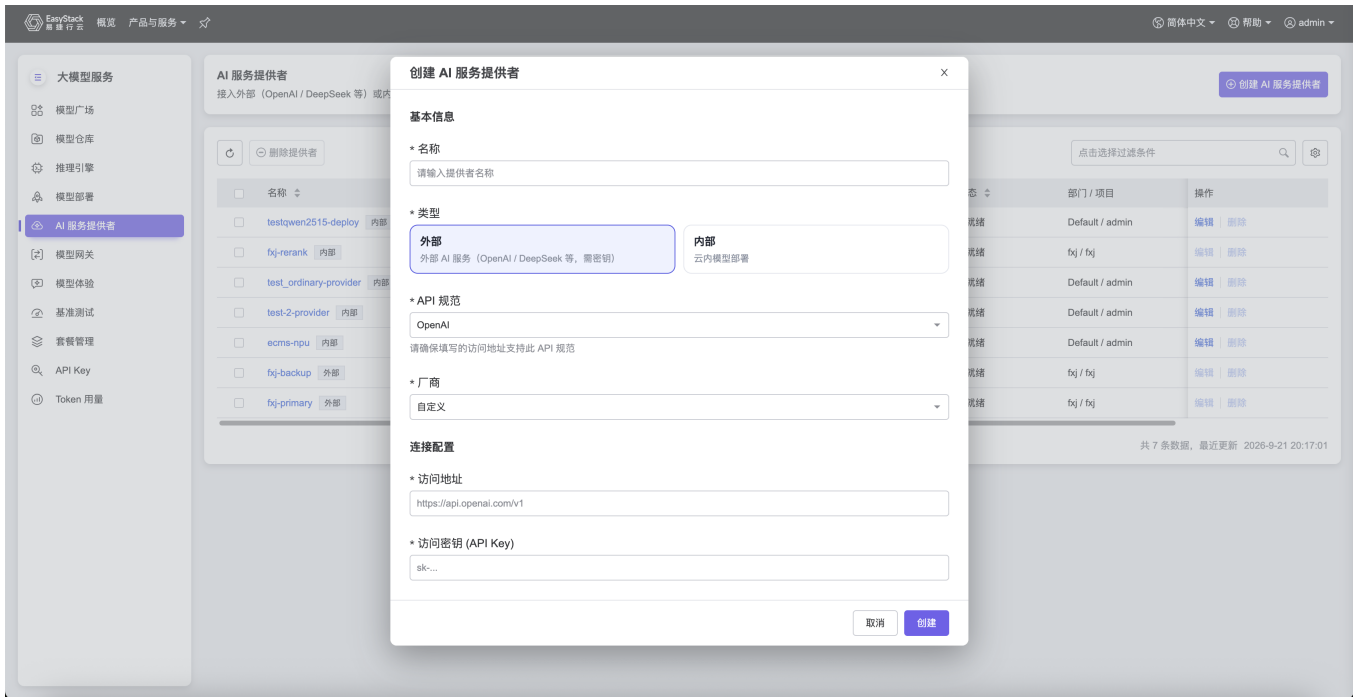
创建AI服务提供者

AI服务提供者分为 外部 和 内部 两种类型。不同类型需要配置的连接信息不同。

创建外部AI服务提供者

外部AI服务提供者用于接入OpenAI、DeepSeek等外部AI服务，需要配置服务访问地址和访问密钥。

- 在“AI 服务提供者”页面中，单击 创建 AI 服务提供者 ，弹出“创建 AI 服务提供者”对话框。
- 在 类型 中选择 外部 。
- 配置参数后，单击 创建 ，完成操作。



参数	说明
名称	该AI服务提供者的名称。
类型	AI服务提供者的类型。创建外部AI服务提供者时，选择 外部 。
API规范	外部AI服务使用的API规范，例如OpenAI。配置的访问地址需要支持所选API规范。

参数	说明
厂商	外部AI服务所属的厂商。选择厂商后，系统将自动填入兼容的API规范和访问地址，自动填入的内容仍可手动修改。
访问地址	外部AI服务的访问地址。
访问密钥（API Key）	调用外部AI服务时使用的访问密钥。

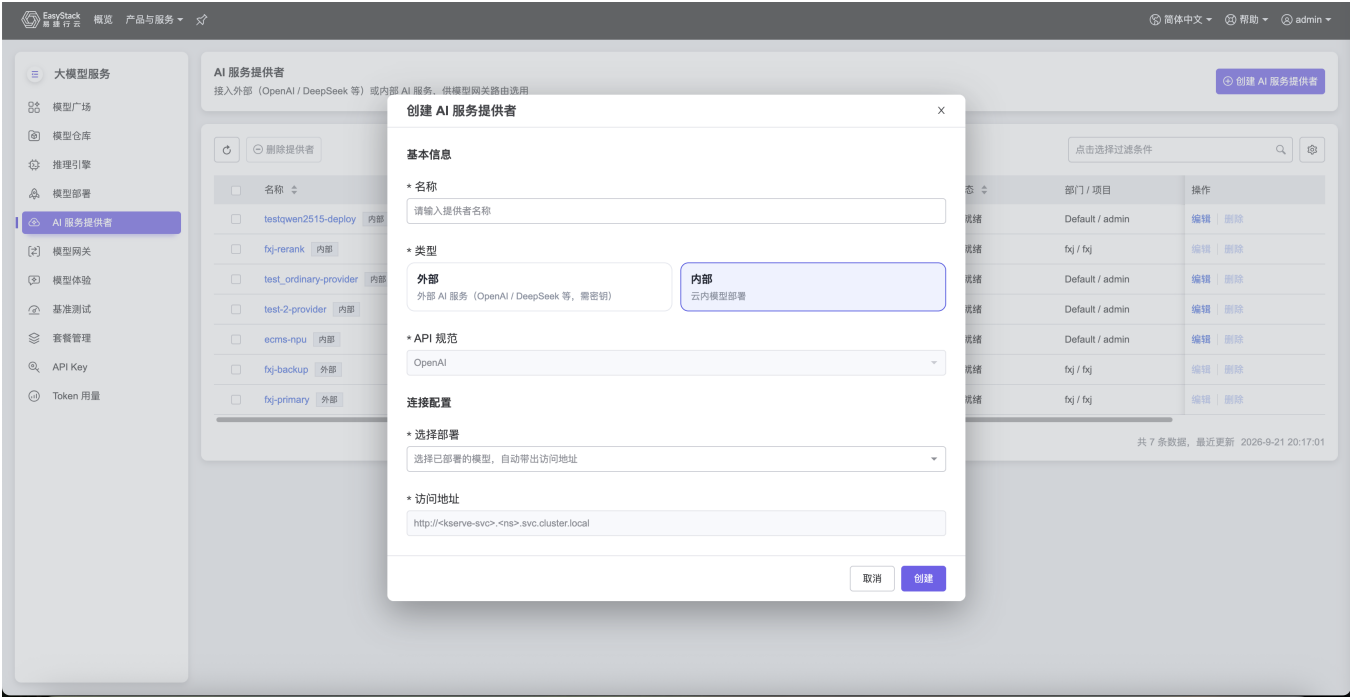
说明：

页面中带有“*”的参数为必填参数。

创建内部AI服务提供者

内部AI服务提供者用于接入云平台中已经创建的模型部署。选择模型部署后，系统将自动生成对应的访问地址。

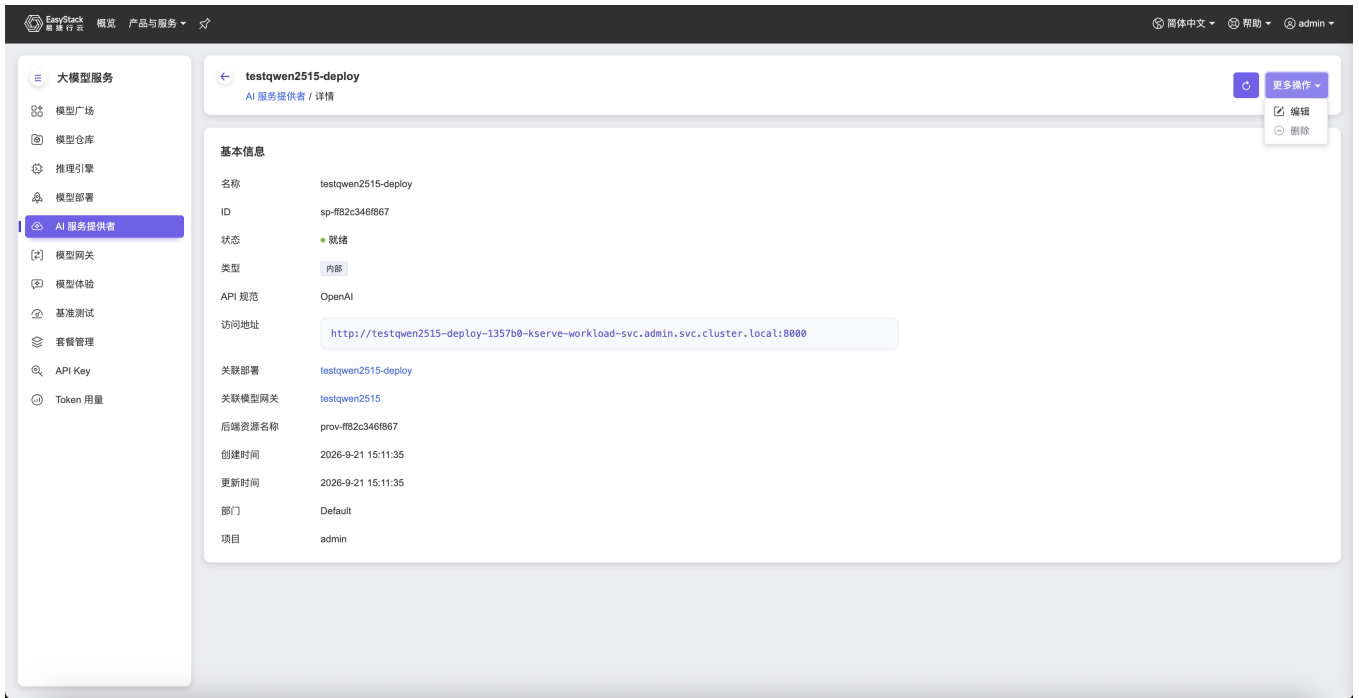
1. 在“AI 服务提供者”页面中，单击 **创建 AI 服务提供者**，弹出“创建 AI 服务提供者”对话框。
2. 在 **类型** 中选择 **内部**。
3. 配置参数后，单击 **创建**，完成操作。



参数	说明
名称	该AI服务提供者的名称。
类型	AI服务提供者的类型。创建内部AI服务提供者时，选择 内部 。
API规范	内部模型服务使用的API规范。
选择部署	选择需要接入的模型部署。选项中同时显示模型部署及其当前状态。
访问地址	所选模型部署对应的服务访问地址。选择模型部署后，系统自动填入该地址。

查看详情

在“AI 服务提供者”页面中，单击AI服务提供者名称，可进入“AI 服务提供者详情”页面。



详情页面显示以下信息：

信息	说明
名称	AI服务提供者的名称。
ID	AI服务提供者在系统中的唯一标识。
状态	AI服务提供者当前的状态。
类型	AI服务提供者的类型，包括内部和外部。
API规范	AI服务提供者使用的API规范。
访问地址	AI服务的访问地址。
关联部署	内部AI服务提供者关联的模型部署。
关联模型网关	当前AI服务提供者关联的模型网关。
后端资源名称	AI服务提供者对应的后端资源名称。
创建时间	AI服务提供者的创建时间。
更新时间	AI服务提供者信息的最近更新时间。

信息	说明
部门	AI服务提供者所属的部门。
项目	AI服务提供者所属的项目。

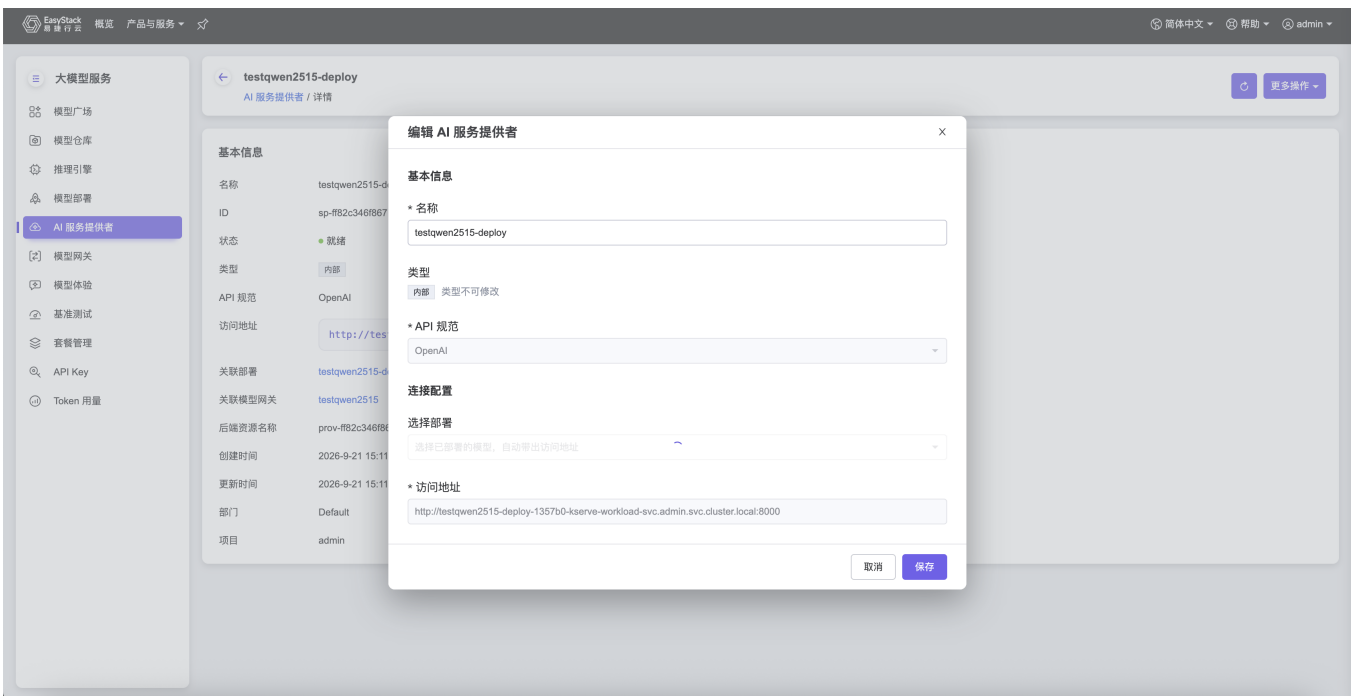
需要重新获取详情信息时，可单击页面右上角的“刷新”图标。

编辑AI服务提供者

可以从AI服务提供者列表或详情页面编辑AI服务提供者。

- 在“AI 服务提供者”页面中，找到待操作AI服务提供者，单击操作列中的 **编辑**。
- 在“AI 服务提供者详情”页面中，依次单击 **更多操作** - **编辑**。

修改配置后，单击 **保存**，完成操作。



编辑内部AI服务提供者时，可修改名称和关联的模型部署。重新选择模型部署后，系统将自动更新访问地址。

说明：

AI服务提供者创建后，不支持修改类型。

删除AI服务提供者

可以删除单个AI服务提供者，也可以在列表中批量删除AI服务提供者。

删除单个AI服务提供者

1. 在“AI 服务提供者”页面中，找到待操作AI服务提供者，单击操作列中的 **删除**；也可以进入AI服务提供者详情页面，依次单击 **更多操作** - **删除**。
2. 根据页面提示完成删除操作。

批量删除AI服务提供者

1. 在“AI 服务提供者”页面中，勾选一个或多个待操作AI服务提供者。
2. 单击列表上方的 **删除提供者**。
3. 根据页面提示完成删除操作。

1.6 模型网关

本章节主要介绍模型网关中的路由管理操作，包括修改根域名、创建路由、查看路由使用方法和详情、编辑路由以及删除路由等。模型网关用于对外提供统一的模型访问入口，根据请求中的模型名称将请求分发到对应的AI服务提供者。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【模型网关】，即可进入“模型网关”页面。

查看路由

在“模型网关”页面中，可查看当前已经创建的路由。列表中显示名称、域名、模型、AI服务提供者、状态、部门/项目和创建时间等信息。

需要重新获取路由列表时，可单击列表上方的“刷新”图标。需要查找路由时，可通过列表右上角的筛选框设置过滤条件。

The screenshot displays the 'Model Gateway' (模型网关) management interface. On the left is a sidebar with navigation options like 'Large Model Services' (大模型服务), 'Model Marketplace' (模型广场), 'Model Warehouse' (模型仓库), 'Inference Engine' (推理引擎), 'Model Deployment' (模型部署), 'AI Service Providers' (AI 服务提供者), 'Model Gateway' (模型网关), 'Model Experience' (模型体验), 'Benchmark Testing' (基准测试), 'Package Management' (套餐管理), 'API Key', and 'Token Usage' (Token 用量). The 'Model Gateway' option is selected. The main content area shows a table of routes. At the top right of this area, there's a filter for 'Root Domain: beacon.ai' (根域名: beacon.ai) and a 'Create Route' (创建路由) button. Below the table header, there are buttons for 'Refresh' (刷新) and 'Delete Route' (删除路由), and a search filter 'Click to select filter conditions' (点击选择过滤条件). The table lists several routes with columns for Name, Domain, Model, AI Service Provider, Status, Department/Project, and Actions. The routes listed are: testqwen2515, fxj-rerank, test-ordinary, test-2-router, ecms-npu-ai, and fxj-route. Each row includes links for 'Usage' (使用方法), 'Edit' (编辑), and 'More' (更多). At the bottom right, it indicates 'Total 6 records, last updated 2026-9-21 20:19:38'.

名称	域名	模型	AI 服务提供者	状态	部门 / 项目	操作
testqwen2515	testqwen2515.beacon.ai	Qwen2.5-1.5B-Instruct	testqwen2515-deploy	就绪	Default / admin	使用方法 编辑 更多
fxj-rerank	fxj-rerank.beacon.ai	bge-reranker-v2-m3	fxj-rerank	就绪	fxj / fxj	使用方法 编辑 更多
test-ordinary	my-ordinary.beacon.ai	Qwen2.5-0.5B-Instruct	test_ordinary-provider	就绪	Default / admin	使用方法 编辑 更多
test-2-router	my-modles.beacon.ai	Qwen2.5-0.5B-Instruct	test-2-provider	就绪	Default / admin	使用方法 编辑 更多
ecms-npu-ai	ecms-npu-ai.beacon.ai	Qwen2.5-0.5B-Instruct	ecms-npu	就绪	Default / admin	使用方法 编辑 更多
fxj-route	fxj-test.beacon.ai	fxj-test-model	fxj-primary, fxj-backup	就绪	fxj / fxj	使用方法 编辑 更多

单击路由名称，可进入对应的详情页面。单击AI服务提供者名称，可进入对应的AI服务提供者详情页面。

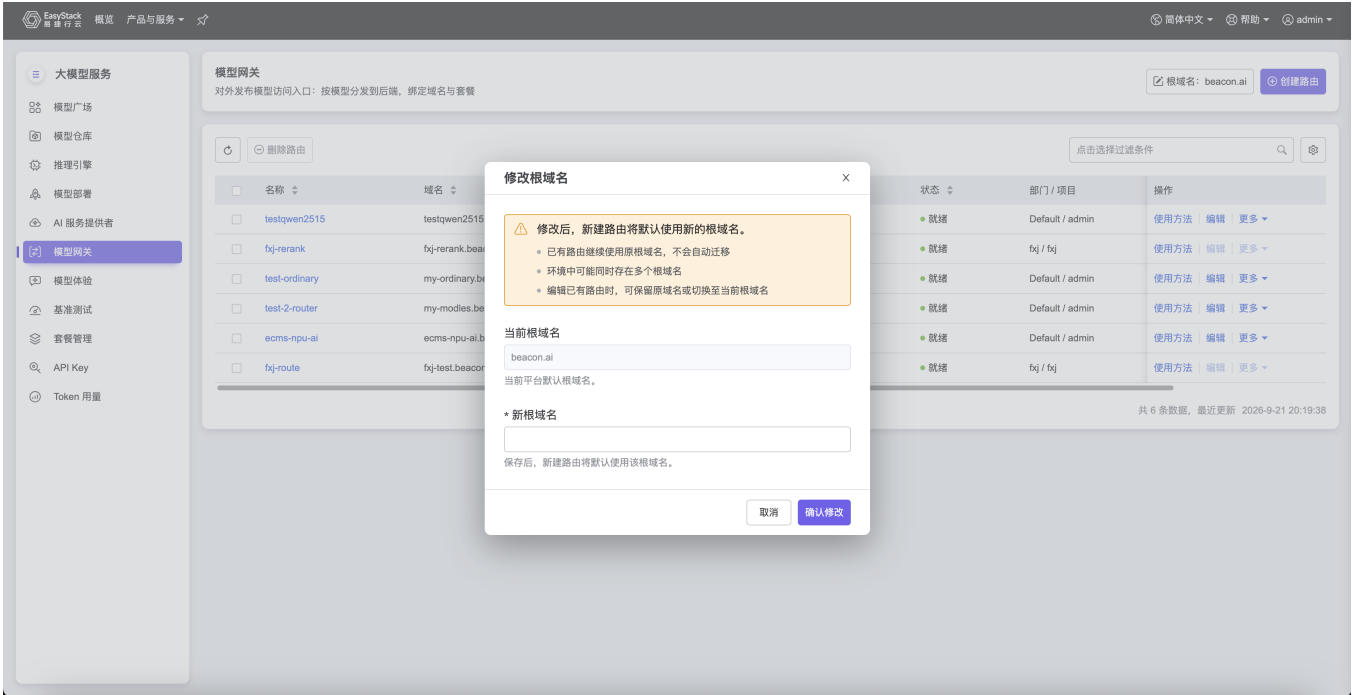
修改根域名

根域名用于生成新建路由的访问域名。路由域名由二级域名和根域名组成。

1. 在“模型网关”页面中，单击页面右上角显示的 **根域名** 。

2. 在“修改根域名”对话框中输入新的根域名。

3. 单击 **确认修改** ，完成操作。



参数	说明
当前根域名	云平台当前使用的默认根域名。
新根域名	修改后新建路由默认使用的根域名。

说明：

- 修改根域名后，新建路由将默认使用新的根域名。
- 已经创建的路由继续使用原根域名，不会自动迁移。
- 云平台中可能同时存在多个根域名。
- 编辑已有路由时，可以保留原域名，也可以切换至当前根域名。

创建路由

前提条件

- 已经创建可供路由使用的AI服务提供者。
- 如果需要限制路由的访问来源，已经准备相应的IP访问控制配置。

操作步骤

- 在“模型网关”页面中，单击 **创建路由**，弹出“创建路由”对话框。
- 配置路由及模型后端。
- 单击 **创建**，完成操作。



基本配置

参数	说明
名称	该路由的名称。
二级域名	该路由使用的二级域名。云平台将二级域名与当前根域名拼接为完整路由域名。二级域名需要在整个云平台中保持唯一。

参数	说明
IP访问控制（黑白名单）	该路由使用的IP访问控制配置。

模型和后端配置

一条路由可以配置多个模型。模型网关根据请求体中的 `model` 字段，将请求转发至对应模型配置的后端。

每个模型可以配置多个后端。后端根据优先级和权重参与请求分发：

- 优先级数值越小，后端越优先被使用，可用于配置主用和备用后端。
- 多个后端的优先级相同时，系统按照权重进行流量分配。

参数	说明
对外模型名称	调用路由时，请求体 <code>model</code> 字段使用的模型名称。该参数留空时，选择后端模型后，系统自动填入相同的对外模型名称。
后端	为当前模型提供服务的AI服务提供者。
后端模型名称	AI服务提供者中实际调用的模型名称。该模型需要从所选后端查询到的模型列表中选择。
优先级	当前后端的调用优先级。优先级数值越小，越优先被使用。
权重	多个后端优先级相同时，当前后端参与流量分配的权重。

需要为同一模型配置多个后端时，单击 [添加后端](#) 继续配置。需要在同一路由中发布多个模型时，单击 [添加模型](#) 继续配置。

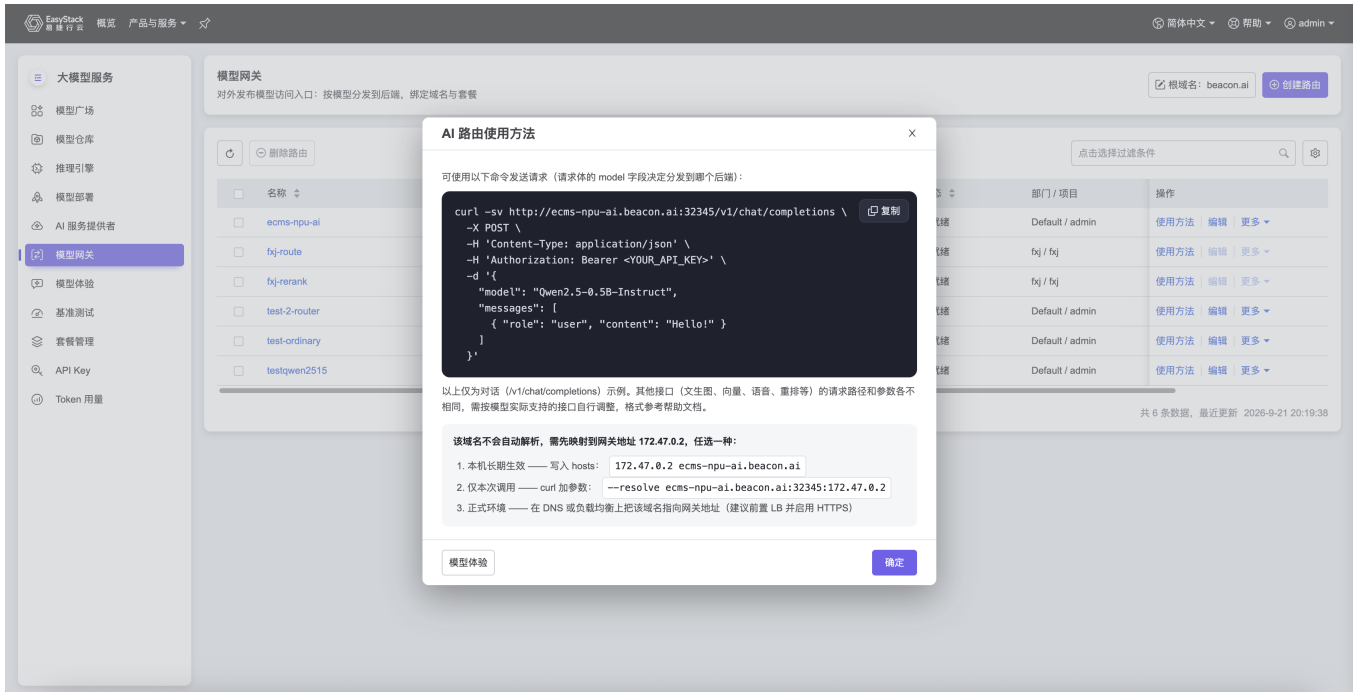
查看使用方法

可以从路由列表或路由详情页面查看路由的调用方法。

- 在“模型网关”页面中，找到目标路由，单击操作列中的 [使用方法](#)。

- 在路由详情页面中，依次单击 **更多操作** - **使用方法**。

完成上述任一操作后，弹出“AI 路由使用方法”对话框。



对话框中显示当前路由对应的请求命令。请求体中的 **model** 字段决定请求需要分发到的模型后端。

调用格式示例如下：

```
curl -sv http://<路由域名>:<端口>/v1/chat/completions \
-X POST \
-H 'Content-Type: application/json' \
-H 'Authorization: Bearer <YOUR_API_KEY>' \
-d '{
  "model": "<对外模型名称>",
  "messages": [
    {
      "role": "user",
      "content": "Hello!"
    }
  ]
}'
```

单击请求命令右上角的 **复制**，可复制完整命令。

说明：

- 对话框中的命令以对话接口为例。文本生成、向量嵌入、图像、语音和重排序等接口的请求路径与参数可能不同，需要根据目标模型实际支持的接口进行调整。
- 根域名不会自动解析，需要将路由域名映射至对话框中显示的网关地址。
- 本机长期使用时，可将网关地址与路由域名写入hosts文件。
- 仅当前请求使用时，可在curl命令中通过 `--resolve` 参数指定路由域名、端口和网关地址。
- 正式环境中，可通过DNS或负载均衡将路由域名指向网关地址。页面建议在网关前配置负载均衡并启用HTTPS。

需要使用当前路由验证模型调用效果时，可单击 [模型体验](#)。

查看详情

在“模型网关”页面中，单击路由名称，可进入“模型网关详情”页面。

The screenshot shows the 'Model Gateway Detail' page for 'ecms-npu-ai'. The left sidebar contains a navigation menu with options like 'Model Marketplace', 'Model Library', 'Inference Engine', 'Model Deployment', 'AI Service Providers', 'Model Gateway' (selected), 'Model Experience', 'Baseline Test', 'Quota Management', 'API Key', and 'Token Usage'. The main content area is divided into two sections: 'Basic Information' and 'Model -> Backend'. The 'Basic Information' section displays fields: Name (ecms-npu-ai), ID (route-59297afe0642), Status (就绪), Domain (ecms-npu-ai.beacon.ai), Access Control (无限制), Creation Time (2026-9-21 14:15:01), Update Time (2026-9-21 14:15:19), Department (Default), and Project (admin). The 'Model -> Backend' section shows a table with columns: AI Service Provider, Priority, Weight, and Backend Model Name. The table contains one entry: 'ecms-npu' with Priority 0, Weight 1, and Backend Model Name 'Qwen2.5-0.5B-Instruct'.

名称	ID	状态	域名	访问控制	创建时间	更新时间	部门	项目
ecms-npu-ai	route-59297afe0642	就绪	ecms-npu-ai.beacon.ai	无限制	2026-9-21 14:15:01	2026-9-21 14:15:19	Default	admin

模型名称	Qwen2.5-0.5B-Instruct		
AI 服务提供者	优先级	权重	后端模型名称
ecms-npu	0	1	Qwen2.5-0.5B-Instruct

“基本信息”区域中显示名称、ID、状态、域名、访问控制、创建时间、更新时间、部门和项目等信息。

“模型 → 后端”区域中显示路由包含的对外模型，以及每个模型对应的AI服务提供者、优先级、权重和后端模型名称。

需要重新获取路由信息时，可单击页面右上角的“刷新”图标。

编辑路由

可以从路由列表或路由详情页面编辑路由。

- 在“模型网关”页面中，找到待操作路由，单击操作列中的 **编辑**。
- 在路由详情页面中，依次单击 **更多操作** - **编辑**。

修改路由名称、二级域名、模型后端或IP访问控制配置后，单击 **保存**，完成操作。



编辑已有路由时，可以继续使用原根域名，也可以将路由切换至当前根域名。

删除路由

可以删除单个路由，也可以在列表中批量删除路由。

删除单个路由

- 在“模型网关”页面中，找到待操作路由，单击操作列中的 **更多** - **删除**；也可以进入路由详情页面，依次单击 **更多操作** - **删除**。

2. 根据页面提示完成删除操作。

批量删除路由

1. 在“模型网关”页面中，勾选一个或多个待操作路由。
2. 单击列表上方的 **删除路由** 。
3. 根据页面提示完成删除操作。

1.7 模型体验

本章节主要介绍如何通过模型体验页面测试模型能力，包括对话、图片生成、视觉/OCR和语音等。模型体验通过模型网关调用目标模型，需要选择网关路由和模型并填写API Key。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【模型体验】，即可进入“模型体验”页面。

准备工作

使用模型体验前，请确认已经完成以下操作：

- 已经创建模型网关路由，并为路由配置需要体验的模型和后端。
- 已经创建可用于调用模型网关的API Key。
- 目标模型支持需要体验的能力，例如对话、图片生成、视觉/OCR或语音。

选择体验模型

进入“模型体验”页面后，根据需要选择 **对话** 、 **图片生成** 、 **视觉/OCR** 或 **语音** 页签。

不同体验类型均需要完成以下配置：

1. 在 **API Key** 中粘贴用于调用模型网关的API Key。
2. 在模型体验区域中选择目标路由。
3. 选择该路由中需要体验的模型。
4. 根据当前体验类型输入请求内容并发送。

说明：

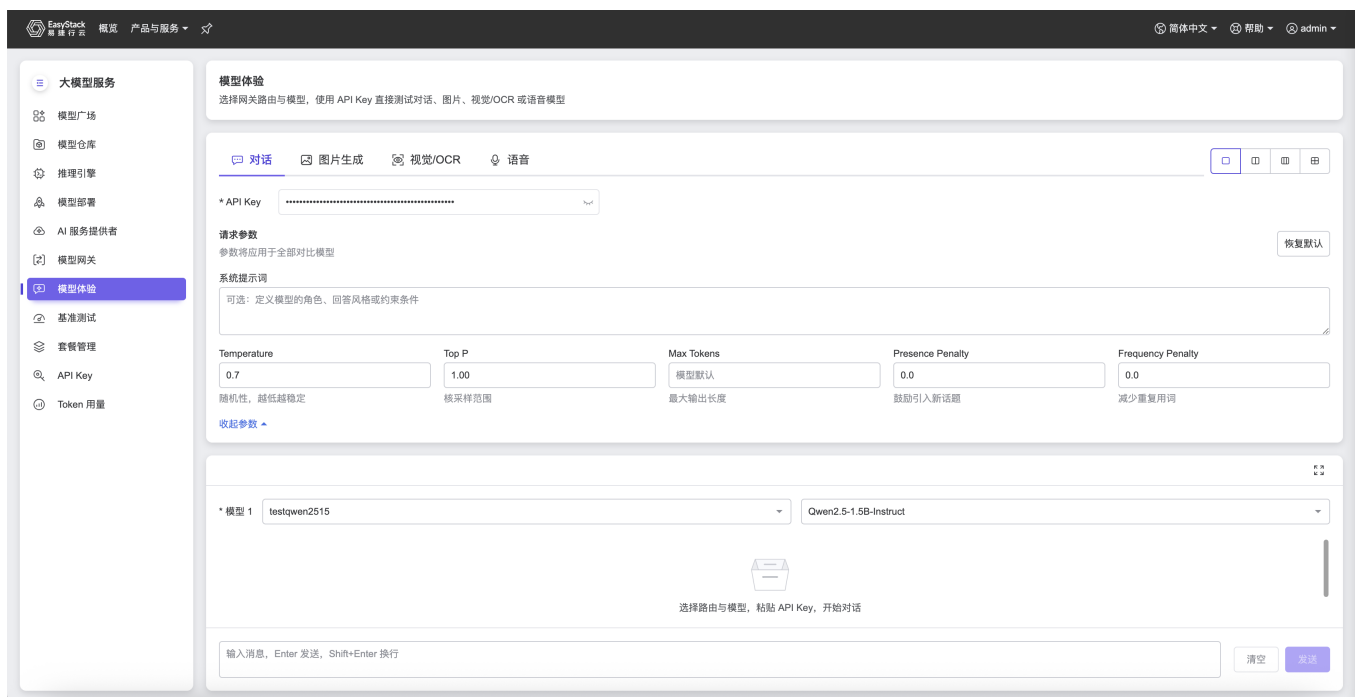
- API Key仅在当前会话中保留。
- 页面右上角提供不同的体验区布局，可用于切换同时显示的模型体验区域数量。
- 单击模型体验区域右上角的“最大化”图标，可展开当前体验区域。

对话体验

对话体验用于向对话模型发送消息并查看模型回复。

1. 在“模型体验”页面中，选择 **对话** 页签。
2. 填写API Key。
3. 根据需要配置请求参数。
4. 选择目标路由和对话模型。
5. 在页面底部输入消息。
6. 单击 **发送** 或按 **Enter** 键，向模型发送消息。

输入多行消息时，按 **Shift+Enter** 换行。



配置请求参数

对话请求参数将应用于当前页面中参与对比的全部模型。

参数	说明
系统提示词	用于定义模型角色、回答风格或其他约束条件。
Temperature	控制回复的随机性。参数值越低，生成结果越稳定。
Top P	控制模型生成内容时的核采样范围。

参数	说明
Max Tokens	设置模型回复的最大输出长度。未设置时使用模型默认值。
Presence Penalty	控制模型是否倾向于引入新的话题。
Frequency Penalty	控制模型是否减少重复用词或重复内容。

单击 **收起参数** 可隐藏请求参数区域。需要恢复默认配置时，可单击 **恢复默认**。

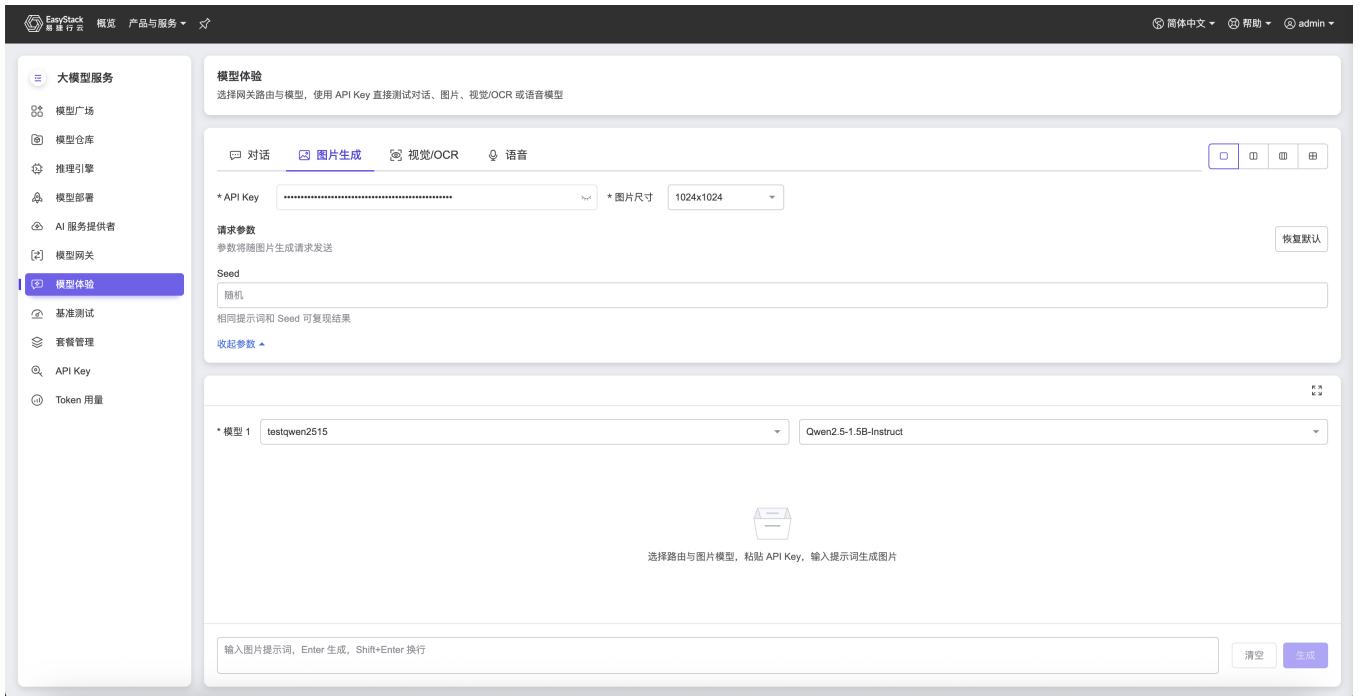
需要清除当前对话内容时，可单击 **清空**。

图片生成体验

图片生成体验用于根据文本提示词生成图片。

1. 在“模型体验”页面中，选择 **图片生成** 页签。
2. 填写API Key。
3. 配置图片生成参数。
4. 选择目标路由和图片生成模型。
5. 在页面底部输入图片提示词。
6. 单击 **生成** 或按 **Enter** 键，生成图片。

输入多行提示词时，按 **Shift+Enter** 换行。



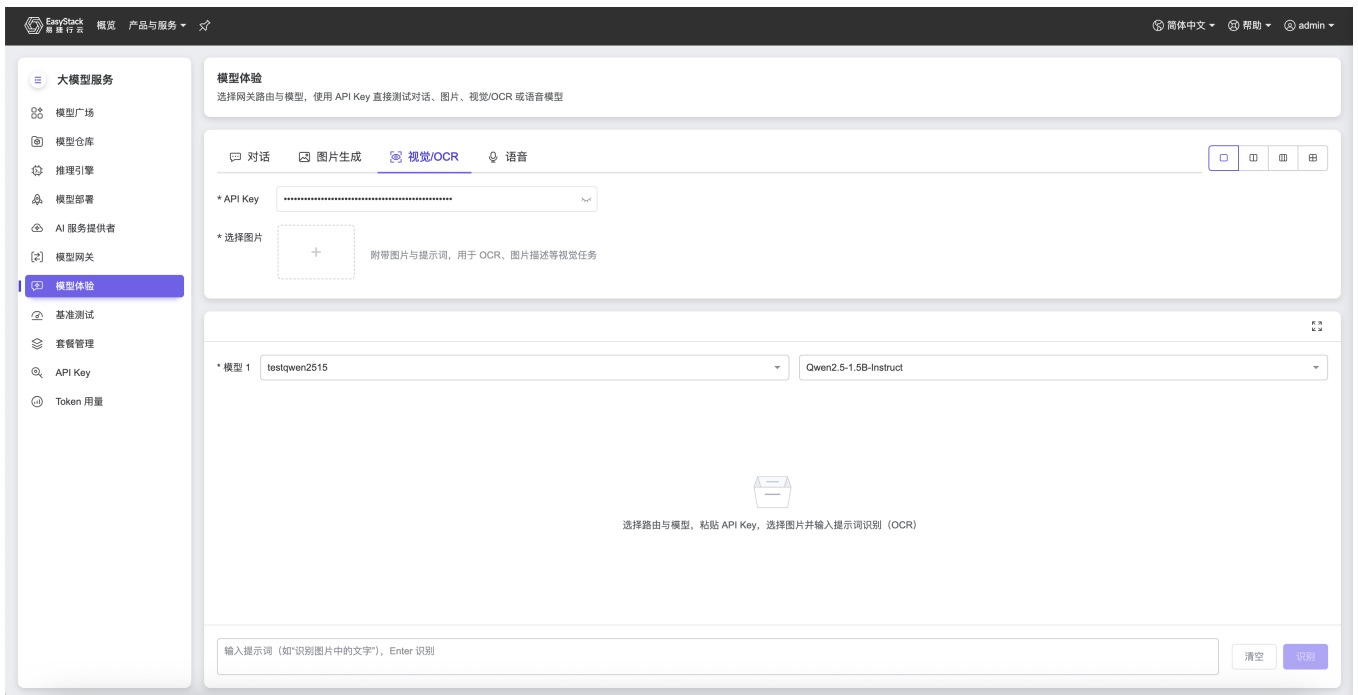
参数	说明
图片尺寸	生成图片的尺寸，例如1024 × 1024。
Seed	图片生成使用的随机种子。使用相同的提示词和Seed，可以复现相同或相近的生成结果。

单击 **收起参数** 可隐藏请求参数区域。需要清除当前生成内容时，可单击 **清空**。

视觉/OCR体验

视觉/OCR体验用于上传图片并执行文字识别、图片描述等视觉任务。

- 在“模型体验”页面中，选择 **视觉/OCR** 页签。
- 填写API Key。
- 单击图片选择区域，上传需要处理的图片。
- 选择目标路由和视觉模型。
- 在页面底部输入提示词，例如“识别图片中的文字”。
- 单击 **识别** 或按 **Enter** 键，发送视觉识别请求。



上传的图片将与提示词一同发送至目标模型。请根据具体任务，在提示词中说明需要模型执行的操作。

需要清除当前图片和识别内容时，可单击 **清空**。

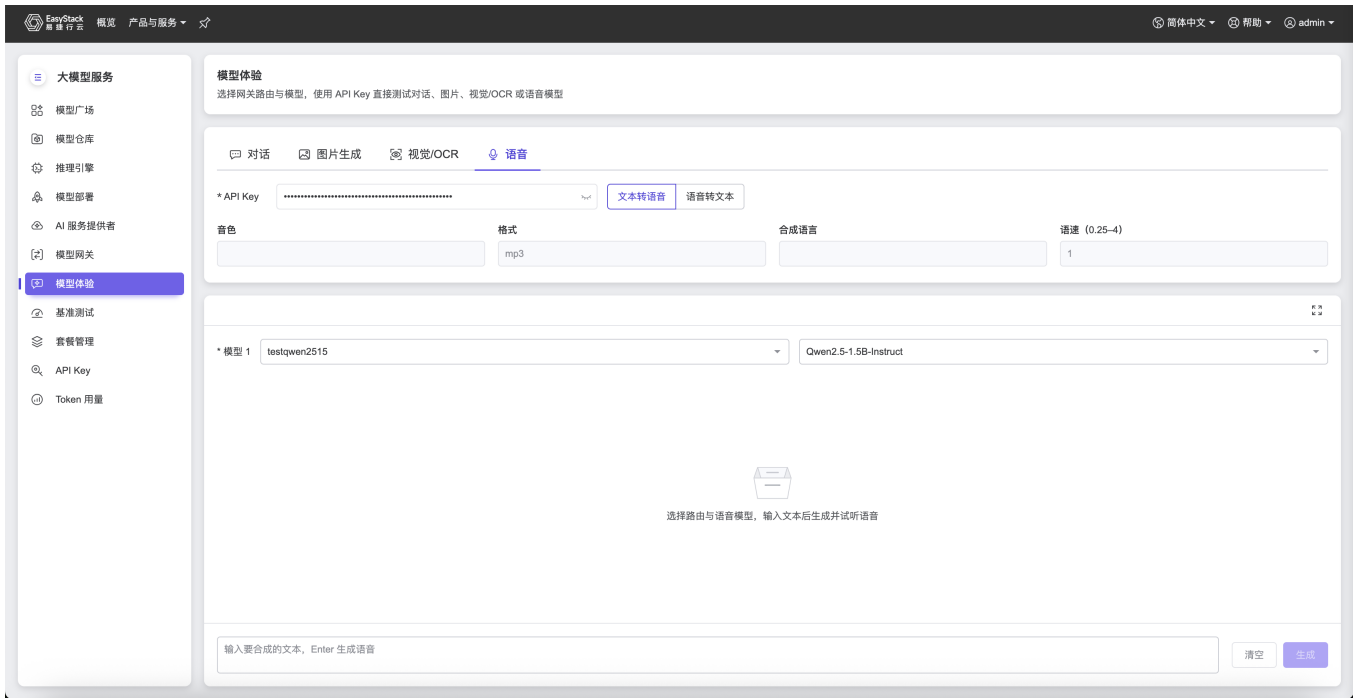
语音体验

语音体验支持 **文本转语音** 和 **语音转文本** 两种模式。

文本转语音

文本转语音用于将输入文本转换为语音。

1. 在“模型体验”页面中，选择 **语音** 页签。
2. 选择 **文本转语音**。
3. 填写API Key。
4. 配置语音生成参数。
5. 选择目标路由和语音模型。
6. 在页面底部输入需要合成的文本。
7. 单击 **生成** 或按 **Enter** 键，生成并试听语音。



参数	说明
音色	生成语音使用的音色。
格式	生成语音的文件格式，例如MP3。
合成语言	生成语音使用的语言。
语速	生成语音的播放速度，可配置范围为0.25 ~ 4。

如果当前选择的模型不支持文本转语音，页面将显示相应提示。此时需要重新选择支持该能力的路由和模型。

需要清除当前文本和生成结果时，可单击 清空。

语音转文本

语音转文本用于将语音内容转换为文本。

- 在“模型体验”页面中，选择 语音 页签。
- 选择 语音转文本。
- 填写API Key。

-
4. 选择目标路由和支持语音转文本的模型。
 5. 根据页面提示添加需要识别的语音并执行转换。

1.8 基准测试

本章节主要介绍基准测试的创建、结果查看、停止和删除操作。基准测试使用固定测试方法直连单个vLLM副本，用于测量模型部署的响应延迟和吞吐能力。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【基准测试】，即可进入“基准测试”页面。

查看基准测试

在“基准测试”页面中，可查看已经创建的测试任务。列表中显示名称、模型部署、测试类型、输出吞吐、TTFT P95、状态、部门/项目和创建时间等信息。

需要重新获取基准测试列表时，可单击列表上方的“刷新”图标。需要查找测试任务时，可通过列表右上角的筛选框设置过滤条件。

基准测试

使用固定方法直连单副本 vLLM，测量延迟与吞吐。

创建测试

刷新 删除基准测试

点击选择过滤条件

☐	名称	模型部署	测试类型	输出吞吐	TTFT P95	状态	部门 / 项目	创建时间	操作
☐	test	ecms-gpu-2	吞吐	2,155.7	36,578.1 ms	● 已完成	Default / admin	2026-9-21 14:32:50	停止 删除
☐	1	-	延迟	80.4	52.2 ms	● 已完成	Default / admin	2026-9-21 11:15:30	停止 删除

共 2 条数据，最近更新 2026-9-21 20:25:24

单击基准测试名称，可进入对应的详情页面。单击模型部署名称，可进入对应的模型部署详情页面。

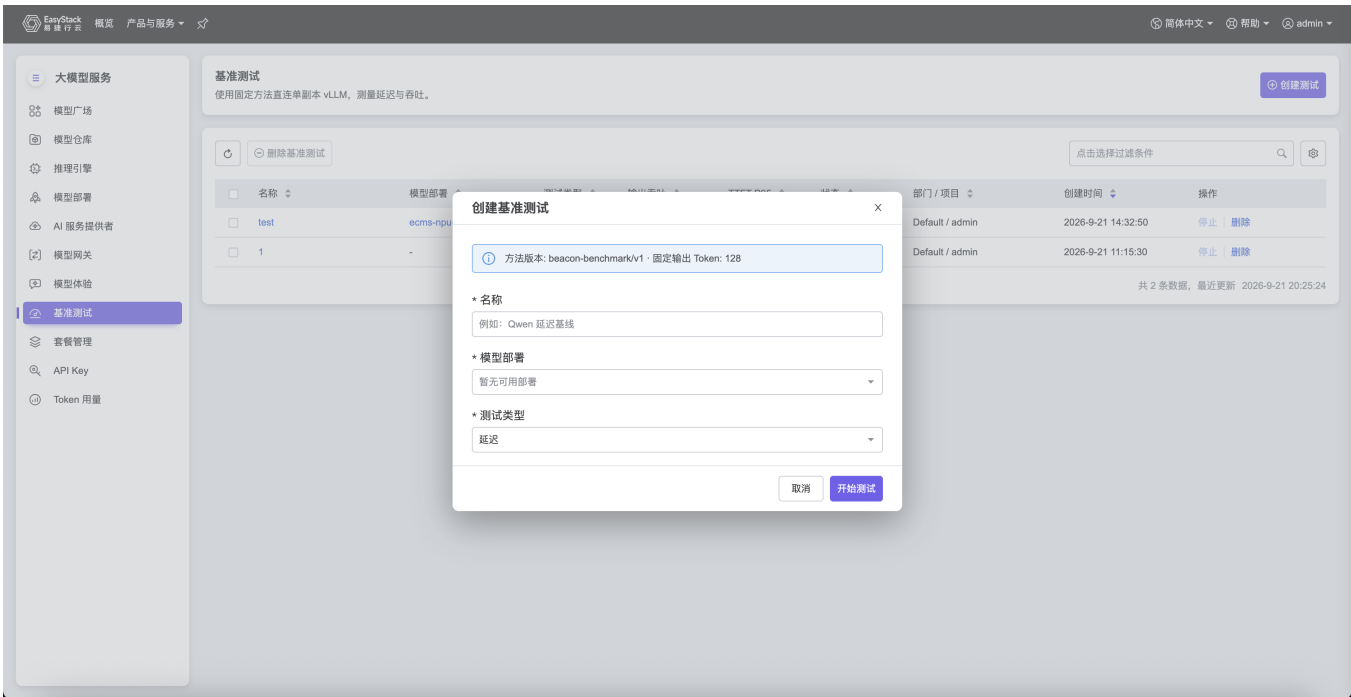
创建基准测试

前提条件

- 已经创建使用vLLM推理引擎的模型部署。
- 目标模型部署已经进入 就绪 状态。

操作步骤

- 在“基准测试”页面中，单击 创建测试 ，弹出“创建基准测试”对话框。
- 配置参数后，单击 开始测试 。



参数	说明
名称	该基准测试任务的名称。
模型部署	需要执行基准测试的模型部署。选项中同时显示模型部署的当前状态。
测试类型	基准测试使用的测试类型，例如延迟测试。

说明：

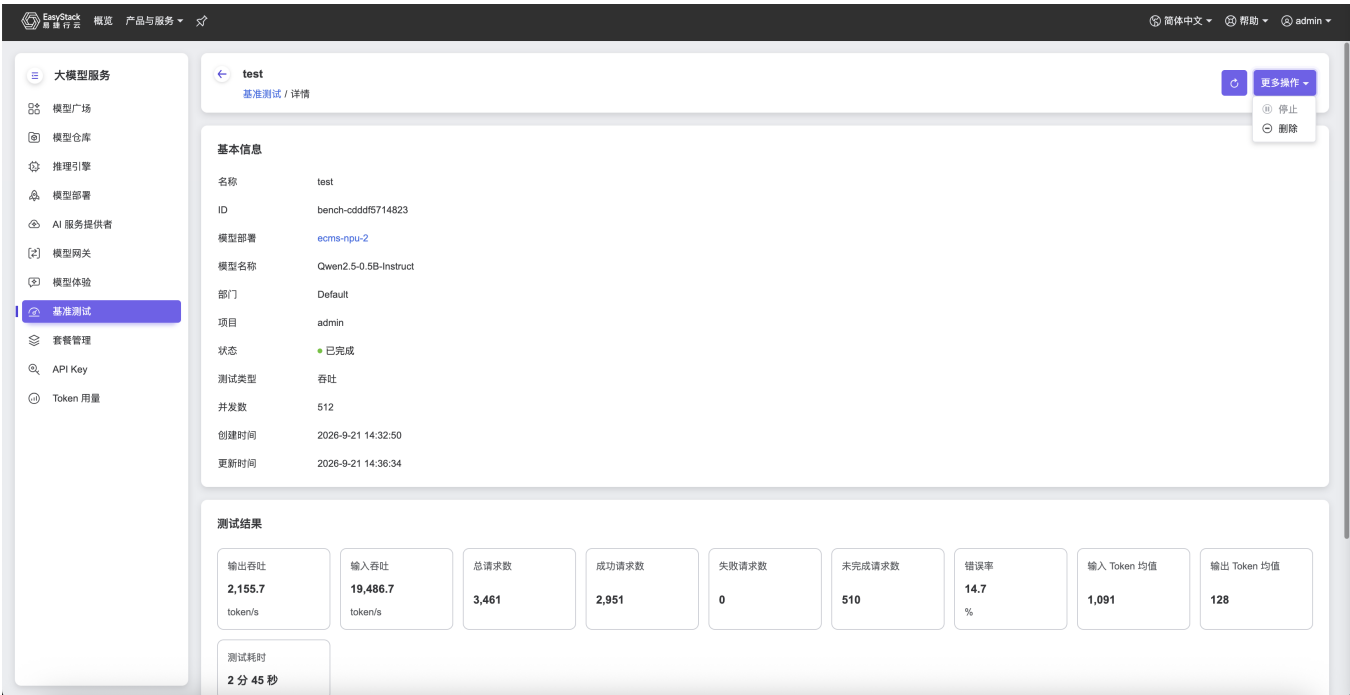
- 页面中带有“*”的参数为必填参数。
- 当前页面显示的方法版本为 beacon-benchmark/v1 。

- 测试方法的固定输出Token数为128。

开始测试后，可返回“基准测试”页面查看测试状态和结果摘要。测试完成后，状态显示为 **已完成**。

查看测试结果

在“基准测试”页面中，单击基准测试名称，可进入“基准测试详情”页面。



查看基本信息

“基本信息”区域显示以下内容：

信息	说明
名称	基准测试任务的名称。
ID	基准测试任务在系统中的唯一标识。
模型部署	当前测试使用的模型部署。
模型名称	当前测试使用的模型名称。

信息	说明
部门	基准测试任务所属的部门。
项目	基准测试任务所属的项目。
状态	基准测试任务当前的运行状态。
测试类型	当前测试使用的测试类型。
并发数	测试请求的并发数量。
创建时间	基准测试任务的创建时间。
更新时间	基准测试任务的最近更新时间。

查看结果指标

“测试结果”区域显示测试任务的吞吐、请求和延迟指标。

吞吐及请求指标包括：

指标	说明
输出吞吐	模型每秒输出的Token数量。
输入吞吐	模型每秒处理的输入Token数量。
总请求数	测试期间发送的请求总数。
成功请求数	测试期间成功完成的请求数量。
失败请求数	测试期间执行失败的请求数量。
未完成请求数	测试结束时尚未完成的请求数量。
错误率	失败请求在总请求中的占比。
输入Token均值	每个请求包含的平均输入Token数量。
输出Token均值	每个请求生成的平均输出Token数量。

指标	说明
测试耗时	当前基准测试任务的总运行时长。

延迟指标包括：

指标	说明
TTFT	Time to First Token，模型从收到请求到输出首个Token所需的时间。
TPOT	Time Per Output Token，模型生成每个输出Token所需的平均时间。
请求延迟	请求从发送到完成所需的时间。

TTFT、TPOT和请求延迟分别显示均值、P50、P95和P99数据。

测试结果底部同时显示当前模型部署使用的AI加速卡和推理引擎镜像。

需要重新获取测试结果时，可单击页面右上角的“刷新”图标。

停止基准测试

基准测试正在运行时，可以停止测试。

- 在“基准测试”页面中，找到待操作测试任务，单击操作列中的 **停止**。
- 在“基准测试详情”页面中，依次单击 **更多操作** - **停止**。

根据页面提示完成操作。

删除基准测试

可以删除单个基准测试，也可以在列表中批量删除测试任务。

删除单个基准测试

1. 在“基准测试”页面中，找到待操作测试任务，单击操作列中的 **删除**；也可以进入基准测试详情页面，依次单击 **更多操作** - **删除**。
2. 根据页面提示完成删除操作。

批量删除基准测试

1. 在“基准测试”页面中，勾选一个或多个待操作测试任务。
2. 单击列表上方的 **删除基准测试**。
3. 根据页面提示完成删除操作。

1.9 套餐管理

本章节主要介绍套餐的创建和管理操作，包括配置请求与Token限速、授权访问路由、共享套餐，以及查看已加入套餐的API Key等。套餐可以为API Key配置可访问的模型网关路由和调用限速。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【套餐管理】，即可进入“套餐管理”页面。

查看套餐

在“套餐管理”页面中，可查看当前已经创建的套餐。列表中显示套餐名称、请求限速、Token限速、授权路由、已加入API Key、共享项目、状态、部门/项目和更新时间等信息。

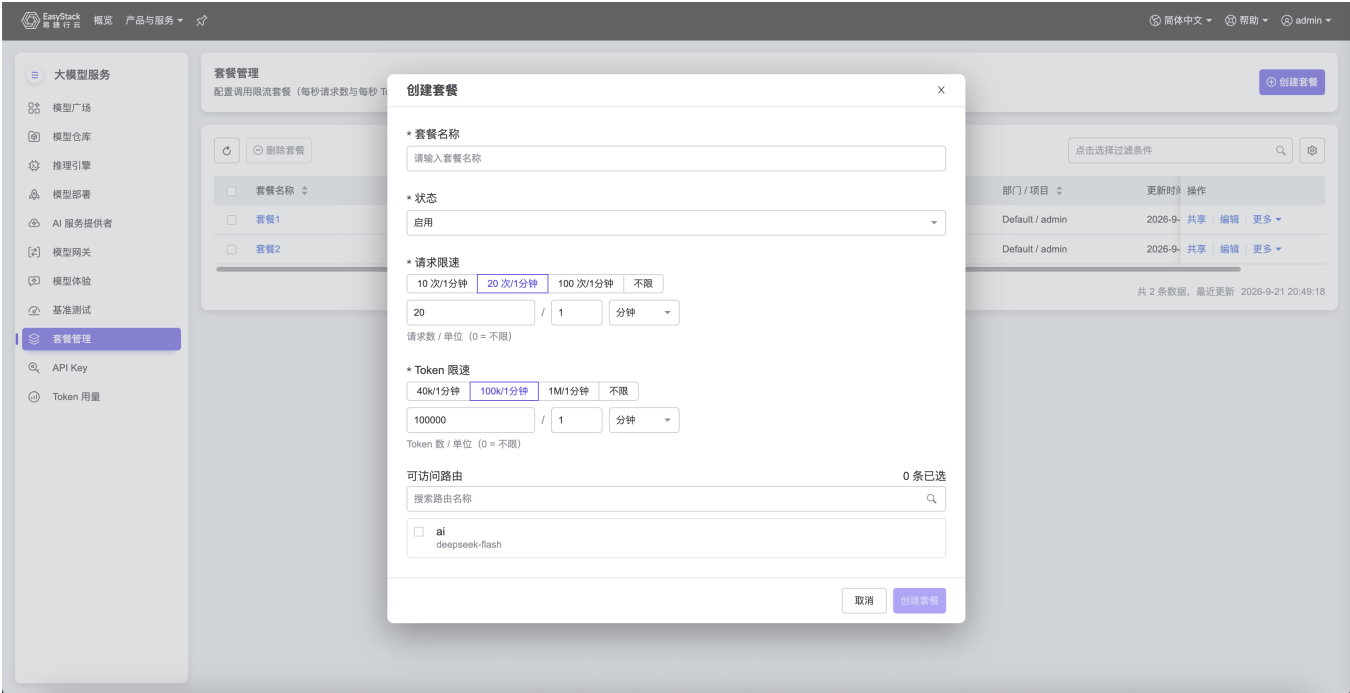
需要重新获取套餐列表时，可单击列表上方的“刷新”图标。需要查找套餐时，可通过列表右上角的筛选框设置过滤条件。

套餐名称	请求限速	Token 限速	授权路由	已加入 API Key	共享项目	状态	部门 / 项目	更新时间	操作
套餐1	20 次 / 1分钟	100,000 token / 1分钟	1	0	0	启用	Default / admin	2026-9-	共享 编辑 更多
套餐2	20 次 / 1分钟	100,000 token / 1分钟	1	0	0	启用	Default / admin	2026-9-	共享 编辑 更多

单击套餐名称，可进入对应的详情页面。

创建套餐

- 在“套餐管理”页面中，单击 **创建套餐**，弹出“创建套餐”对话框。
- 配置参数后，单击 **创建套餐**，完成操作。



参数	说明
套餐名称	该套餐的名称。
状态	该套餐的状态。
请求限速	API Key在指定周期内允许发送的请求数量。可选择预设值或自行设置请求数量和限速周期。设置为0时表示不限制。
Token限速	API Key在指定周期内允许使用的Token数量。可选择预设值或自行设置Token数量和限速周期。设置为0时表示不限制。
可访问路由	使用该套餐的API Key可以访问的模型网关路由。支持通过路由名称搜索并选择一个或多个路由。

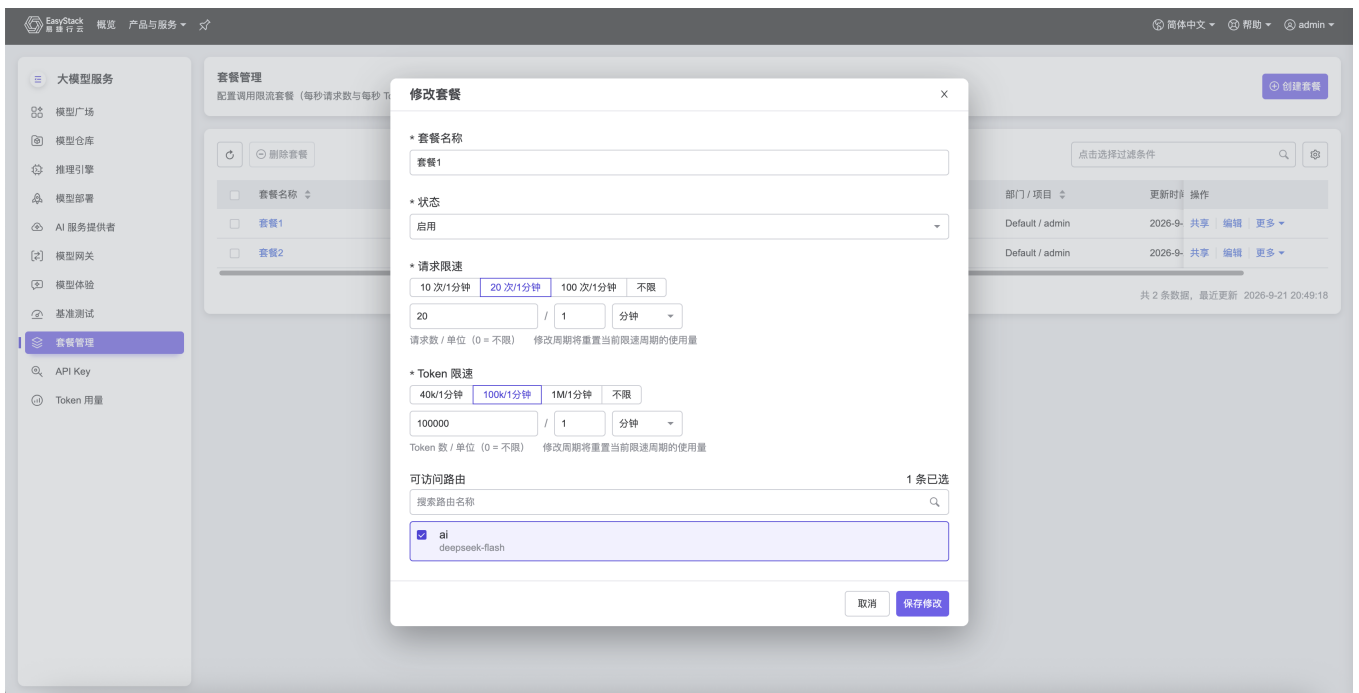
请求限速提供 10 次/1分钟 、 20 次/1分钟 、 100 次/1分钟 和 不限 等预设值。

Token限速提供 40k/1分钟 、 100k/1分钟 、 1M/1分钟 和 不限 等预设值。

说明：
页面中带有“*”的参数为必填参数。

编辑套餐

- 在“套餐管理”页面中，找到待操作套餐，单击操作列中的 **编辑**；也可以进入套餐详情页面，依次单击 **更多操作** - **编辑**。
- 在“修改套餐”对话框中修改套餐名称、状态、请求限速、Token限速或可访问路由。
- 单击 **保存修改**，完成操作。



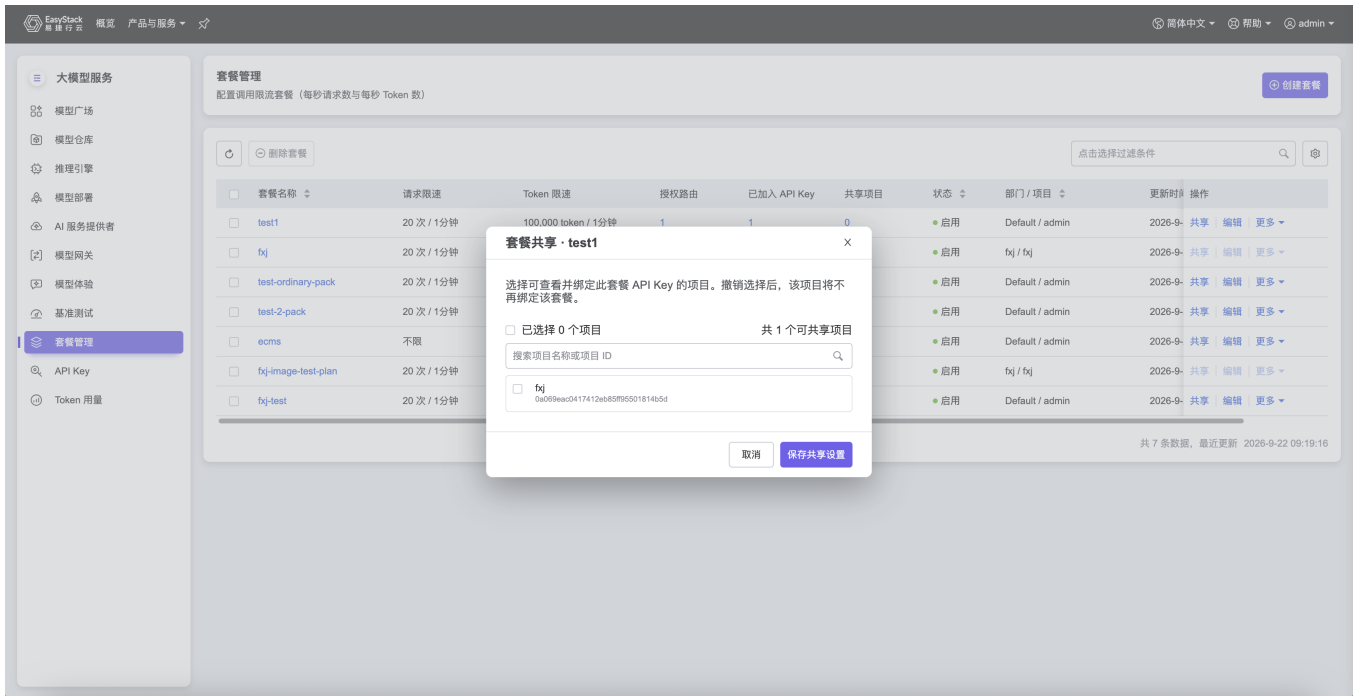
说明：

修改请求限速或Token限速的周期后，将重置当前限速周期内的使用量。

共享套餐

套餐可以共享给其他项目。共享后，目标项目可以查看该套餐，并将项目中的API Key加入该套餐。

- 在“套餐管理”页面中，找到待操作套餐，单击操作列中的 **共享**；也可以进入套餐详情页面，依次单击 **更多操作** - **共享**。
- 在“套餐共享”对话框中选择需要共享的项目。
- 单击 **保存共享设置**，完成操作。



支持通过项目名称或项目ID搜索目标项目。对话框中同时显示已选择项目数量和当前可共享项目数量。

说明：

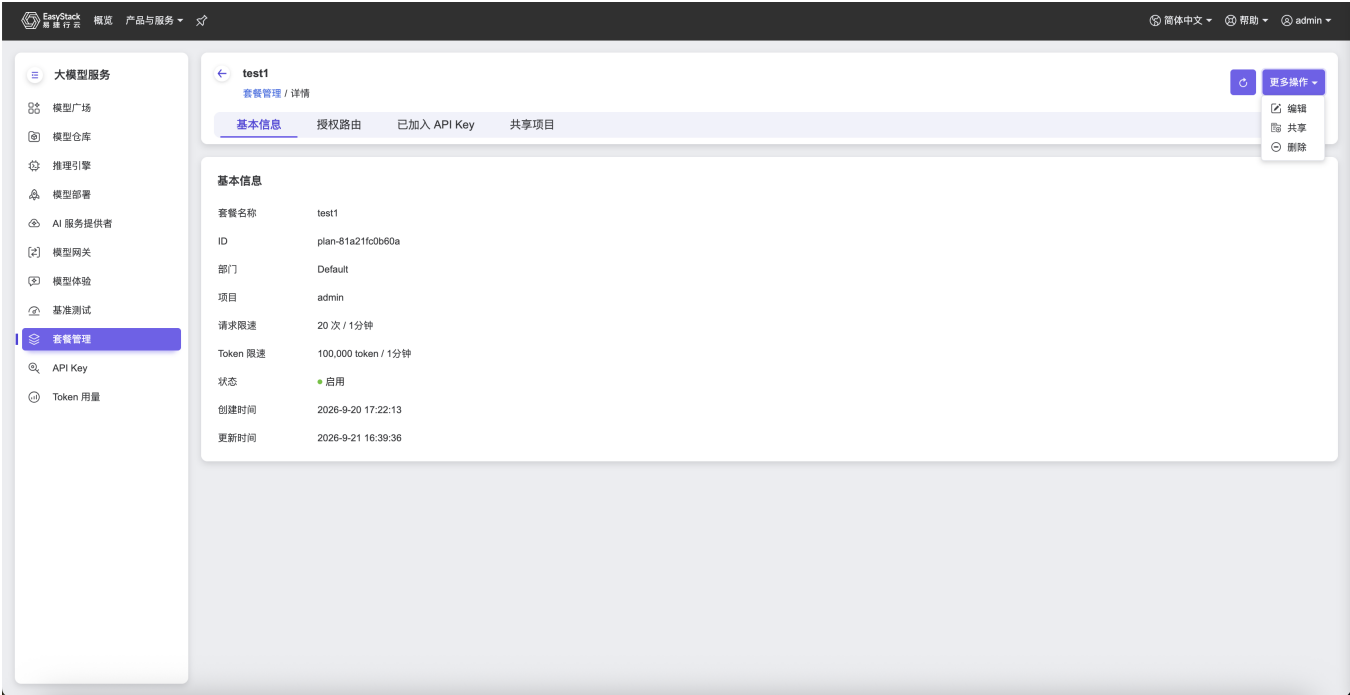
取消选择已经共享的项目后，该项目将不能继续将API Key绑定至该套餐。

查看详情

在“套餐管理”页面中，单击套餐名称，可进入“套餐管理详情”页面。详情页面包括 **基本信息**、**授权路由**、**已加入 API Key** 和 **共享项目** 页签。

查看基本信息

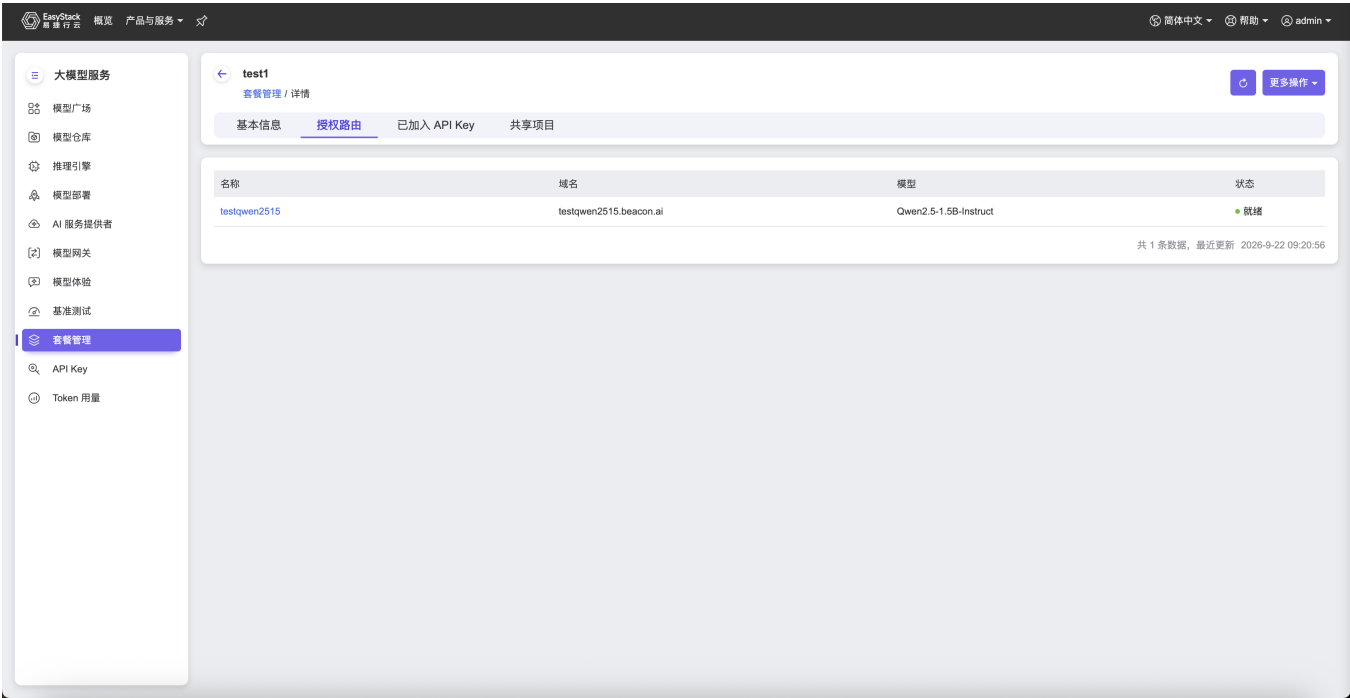
在套餐详情页面中，选择 **基本信息** 页签，可查看套餐的限速配置和当前状态。



基本信息包括套餐名称、ID、部门、项目、请求限速、Token限速、状态、创建时间和更新时间等。

查看授权路由

在套餐详情页面中，选择 授权路由 页签，可查看该套餐允许访问的模型网关路由。



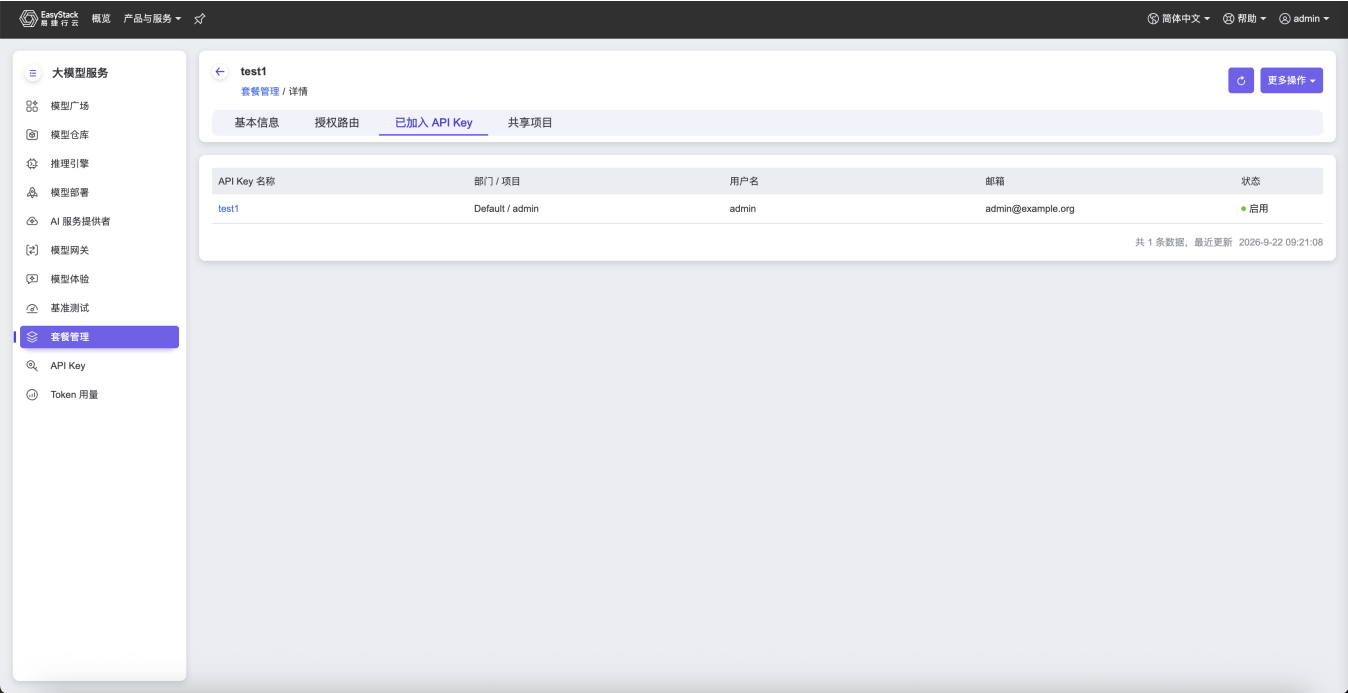
授权路由列表显示以下信息：

信息	说明
名称	模型网关路由的名称。
域名	模型网关路由使用的访问域名。
模型	路由对外提供的模型。
状态	模型网关路由当前的状态。

授权路由需要通过编辑套餐进行添加或删除。

查看已加入的API Key

在套餐详情页面中，选择 **已加入 API Key** 页签，可查看当前已经加入该套餐的API Key。

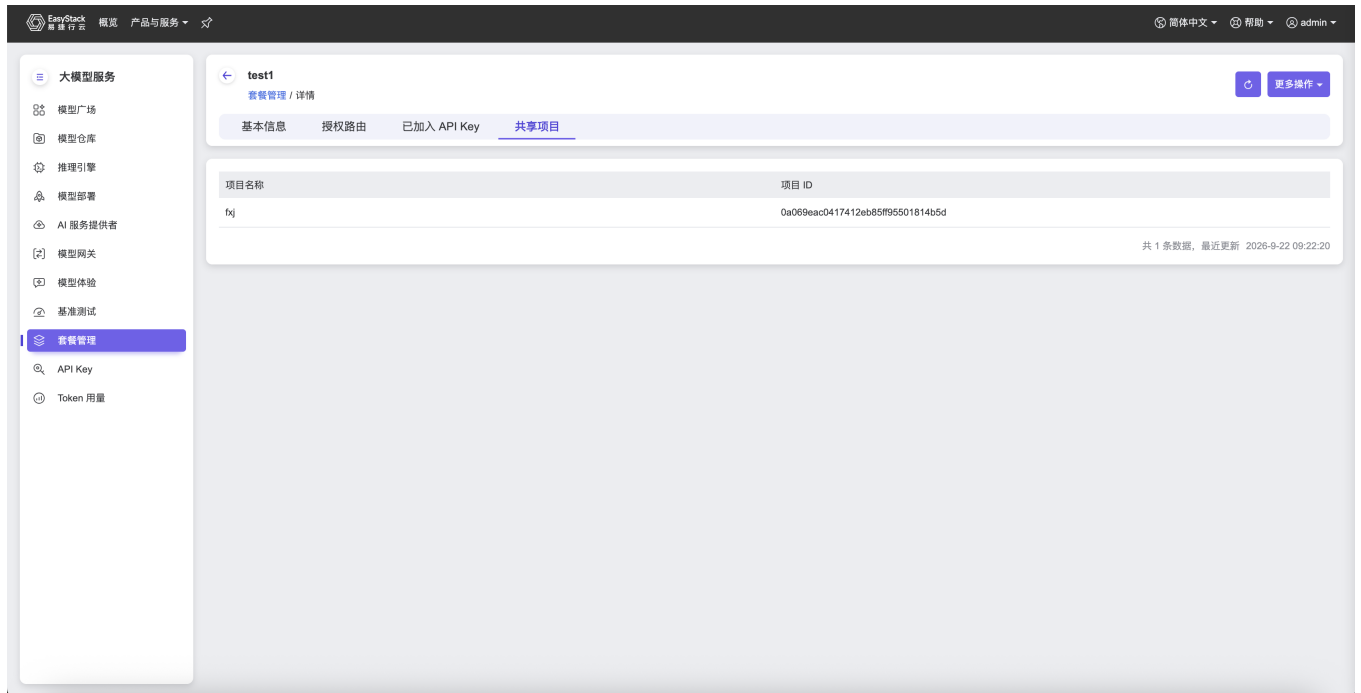


列表显示API Key名称、部门/项目、用户名、邮箱和状态等信息。

API Key与套餐的绑定操作在API Key管理页面中完成。

查看共享项目

在套餐详情页面中，选择 **共享项目** 页签，可查看该套餐当前已经共享至的项目。



列表显示项目名称和项目ID。需要调整共享范围时，可执行共享套餐操作。

删除套餐

可以删除单个套餐，也可以在列表中批量删除套餐。

删除单个套餐

- 在“套餐管理”页面中，找到待操作套餐，单击操作列中的 **更多** - **删除**；也可以进入套餐详情页面，依次单击 **更多操作** - **删除**。
- 根据页面提示完成删除操作。

批量删除套餐

- 在“套餐管理”页面中，勾选一个或多个待操作套餐。
- 单击列表上方的 **删除套餐**。
- 根据页面提示完成删除操作。

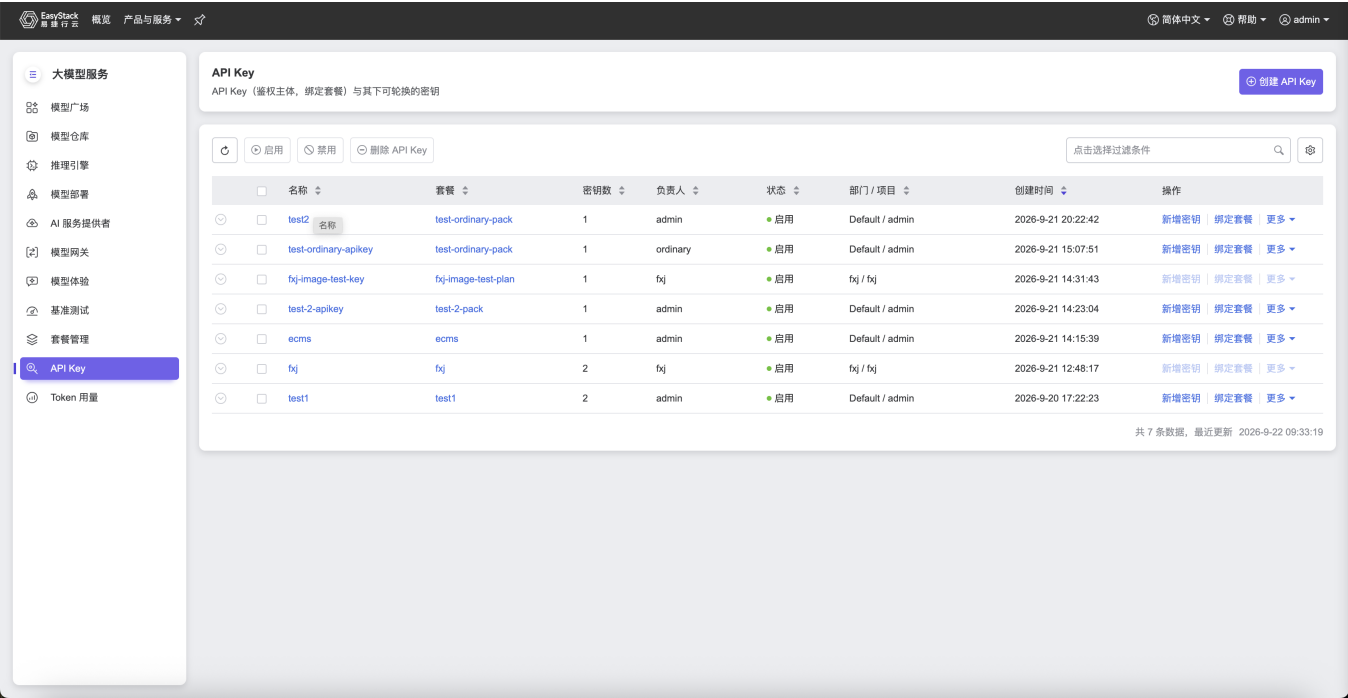
1.10 API Key

本章节主要介绍 API Key 的创建、查看详情、绑定套餐、新增密钥、启用、禁用、重命名和删除等管理操作。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【API Key】，即可进入“API Key”页面。

API Key 是访问模型网关的鉴权主体，可绑定套餐并持有多把密钥。其中，密钥是业务系统实际使用的访问凭证；同一 API Key 下的不同密钥可独立禁用或删除，便于进行密钥轮换。

查看API Key

在“API Key”页面中，可查看已创建的 API Key，并根据名称、状态等条件筛选目标 API Key。



列表中主要显示以下信息。

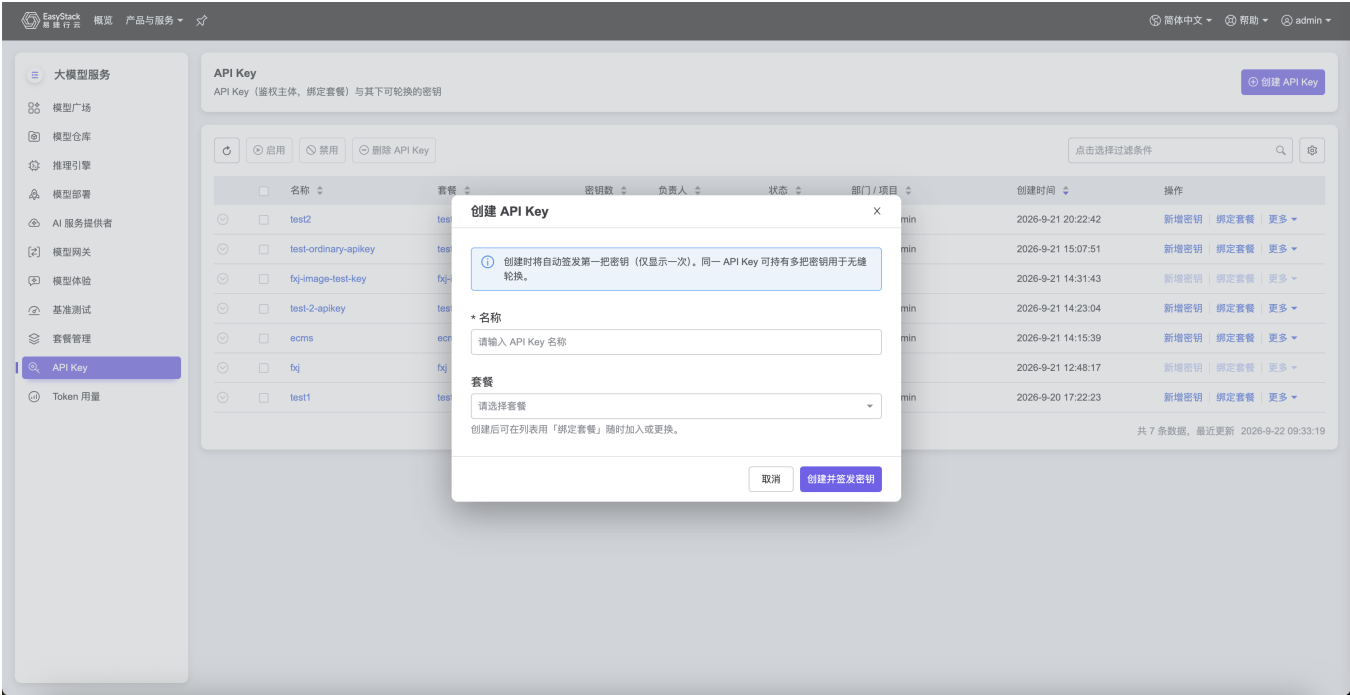
参数	说明
名称	API Key 的名称。单击名称可进入对应的详情页面。

参数	说明
套餐	API Key 当前绑定的套餐。套餐用于控制可访问的模型路由以及请求数、Token 数等限速额度。
密钥数	API Key 当前持有的密钥数量。
负责人	API Key 的负责人。
状态	API Key 当前的状态，包括 启用 和 禁用 。
部门/项目	API Key 所属的部门或项目。
创建时间	API Key 的创建时间。
操作	支持执行 新增密钥 、 绑定套餐 及其他管理操作。

单击 API Key 所在行前的展开图标，可查看该 API Key 已签发的密钥。密钥列表显示密钥值、状态、创建时间及可执行的操作。出于安全考虑，已创建的密钥值将以脱敏形式显示，例如 `sk-beacon-...1544`。

创建API Key

1. 在“API Key”页面中，单击 **创建 API Key**，弹出“创建 API Key”对话框。
2. 配置参数后，单击 **创建并签发密钥**。
3. 创建完成后，复制并妥善保存页面显示的密钥。

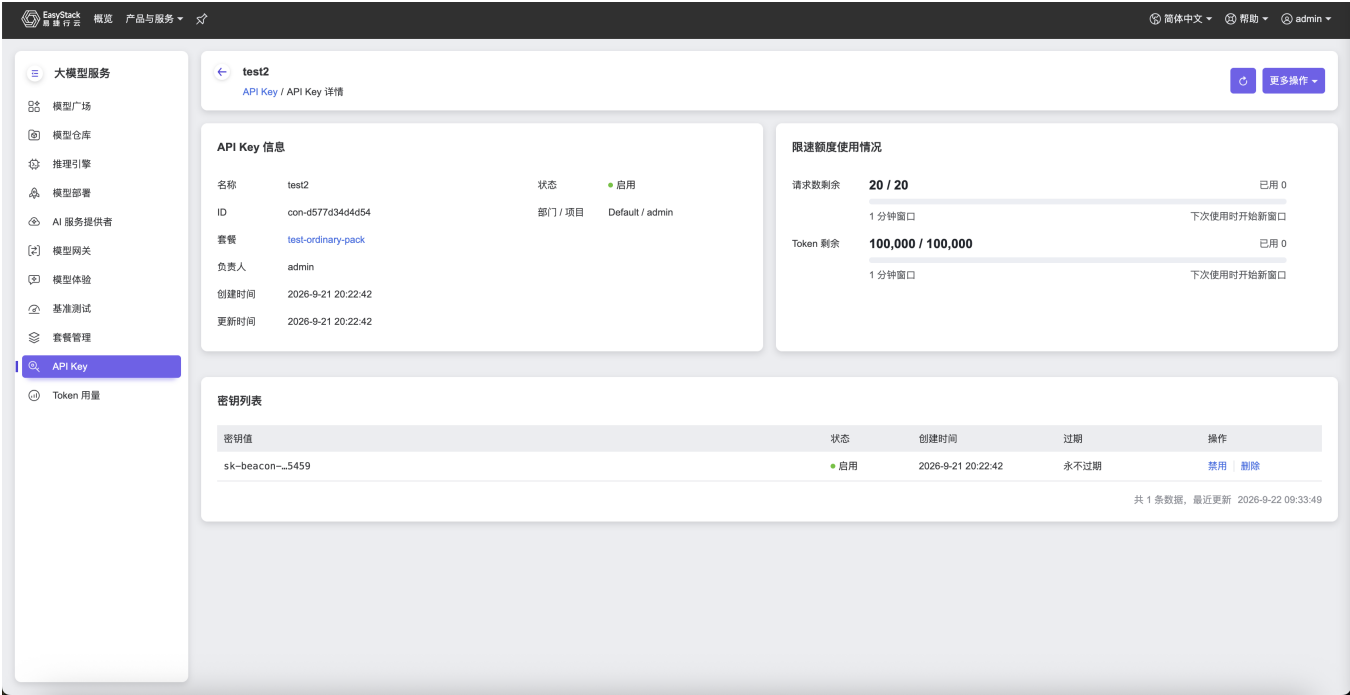


参数	说明
名称	API Key 的名称，用于标识使用该 API Key 的业务、应用或调用方。
套餐	API Key 需要绑定的套餐。绑定套餐后，API Key 将按照套餐配置获得相应的模型路由访问权限和限速额度。创建时也可暂不绑定，后续通过 绑定套餐 完成配置。

- 说明：
- 创建 API Key 时，系统将自动签发第一把密钥。
 - 新签发的密钥仅在创建完成后完整显示一次，请立即复制并存放在安全位置。关闭页面后，无法再次查看完整密钥。
 - 请勿将密钥直接写入公开代码仓库、前端代码或其他不安全的位置。
 - 同一个 API Key 可持有多把密钥，以便进行无缝轮换。

查看详情

在“API Key”页面中，单击 API Key 名称，可进入“API Key 详情”页面。



详情页面主要包含以下区域。

API Key信息

“API Key 信息”区域显示以下内容：

参数	说明
名称	API Key 的名称。
ID	API Key 的唯一标识。
套餐	API Key 当前绑定的套餐。
负责人	API Key 的负责人。
创建时间	API Key 的创建时间。
更新时间	API Key 最近一次更新的时间。
状态	API Key 当前的状态。
部门/项目	API Key 所属的部门或项目。

限速额度使用情况

“限速额度使用情况”区域显示当前 API Key 的额度使用情况，包括：

- 请求数剩余量、已使用量及限速窗口。
- Token 剩余量、已使用量及限速窗口。
- 当前限速窗口的状态。

请求数和 Token 限额由 API Key 绑定的套餐决定。新的限速窗口将在符合套餐规则的情况下重新开始计算。

密钥列表

“密钥列表”区域显示当前 API Key 持有的全部密钥，包括密钥值、状态、创建时间和过期时间等信息，并支持对指定密钥执行 **禁用** 和 **删除** 操作。

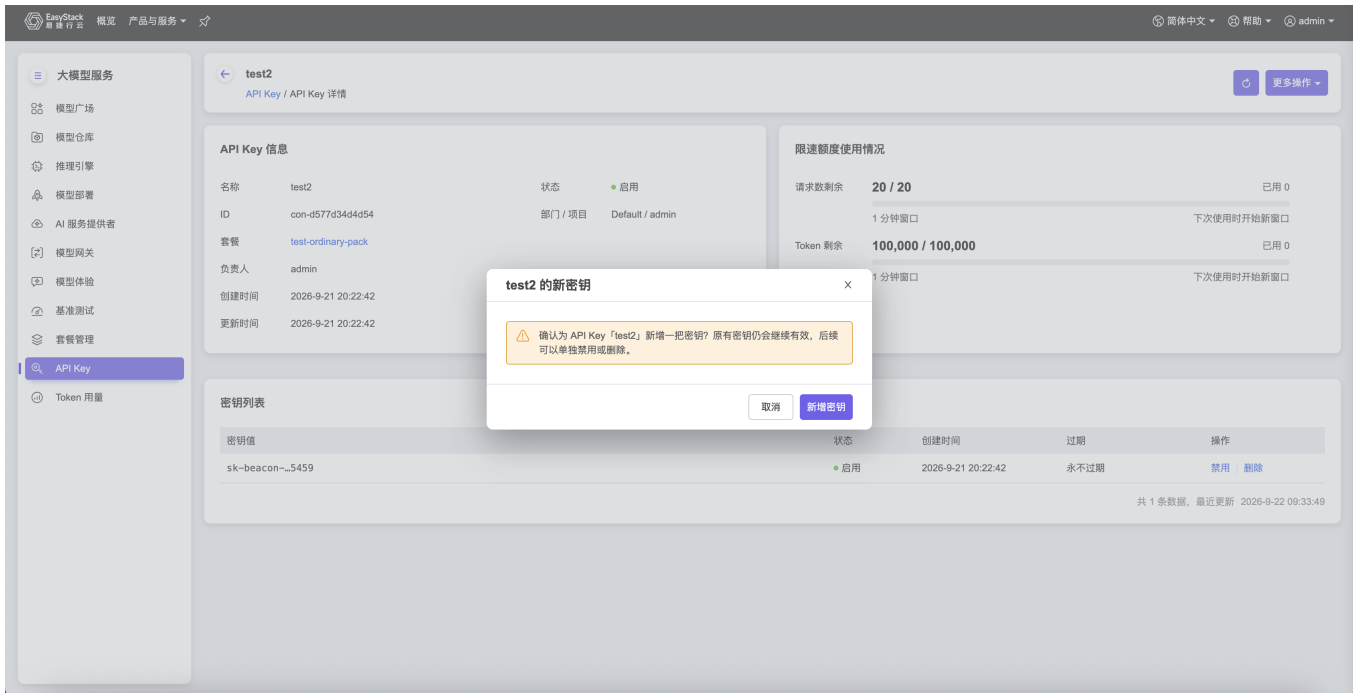
新增密钥

本操作用于为已有 API Key 签发新的密钥。新增密钥后，原有密钥仍然有效，可用于业务系统的无缝密钥轮换。

可通过以下任一方式进入新增密钥操作：

- 在“API Key”页面中，选择目标 API Key，单击 **新增密钥**。
- 在“API Key 详情”页面中，单击 **更多操作** - **新增密钥**。

1. 在新增密钥提示框中，确认待操作的 API Key。
2. 单击 **新增密钥**，完成操作。
3. 按照页面提示复制并妥善保存新签发的密钥。



说明：

- 新增密钥不会自动禁用或删除原有密钥。
- 建议先将新密钥配置到业务系统并确认调用正常，再禁用或删除旧密钥，避免业务中断。
- 不同密钥归属于同一个 API Key，共享该 API Key 绑定的套餐、状态和限速额度。

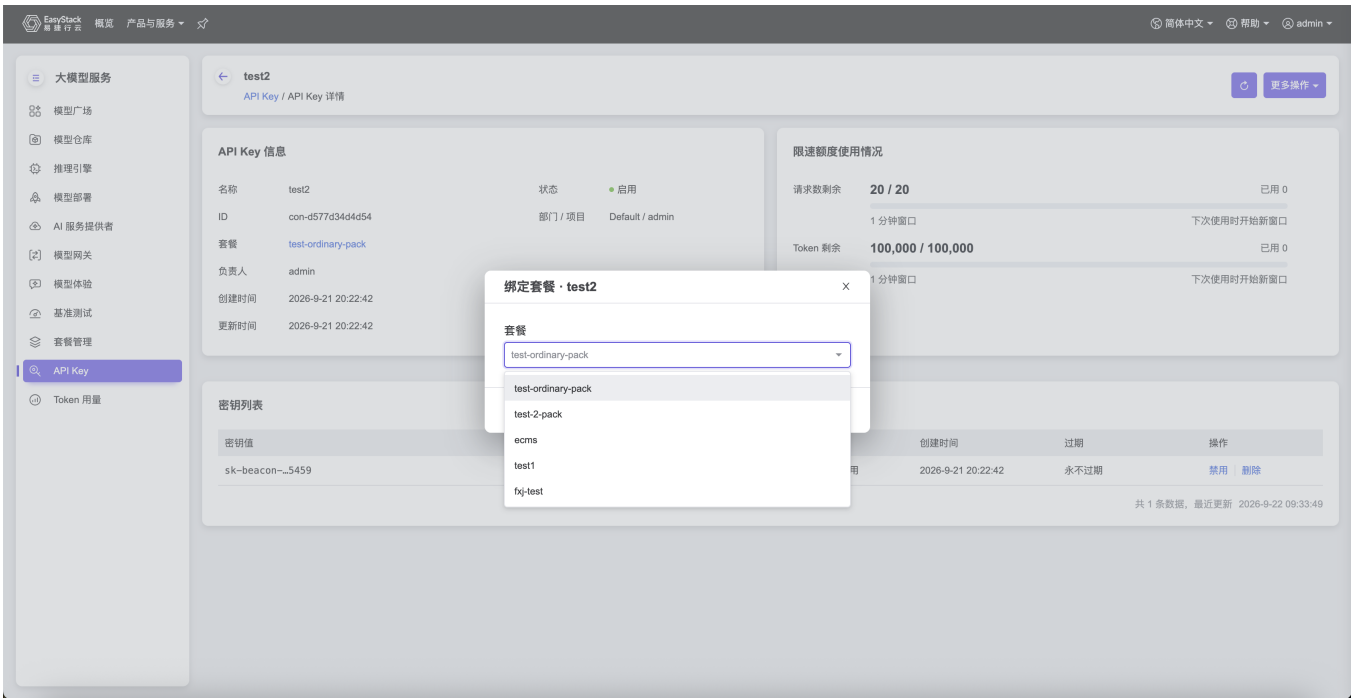
绑定套餐

本操作用于为 API Key 绑定或更换套餐。

可通过以下任一方式进入绑定套餐操作：

- 在“API Key”页面中，选择目标 API Key，单击 **绑定套餐**。
- 在“API Key 详情”页面中，单击 **更多操作** - **绑定套餐**。

1. 在“绑定套餐”对话框中，选择需要绑定的套餐。
2. 单击 **保存**，完成操作。



参数	说明
套餐	API Key 需要绑定的套餐。套餐决定该 API Key 可访问的模型路由以及请求数、Token 数等限速额度。

说明：

更换套餐后，API Key 将按照新套餐的配置获得相应的模型访问权限和限速额度。请在操作前确认新套餐能够满足现有业务的调用需求。

重命名API Key

- 在“API Key”页面中，单击目标 API Key 的名称，进入“API Key 详情”页面。
- 单击 **更多操作** - **重命名** 。
- 输入新的名称后，根据页面提示完成操作。

启用API Key

1. 在“API Key”页面中，勾选状态为 **禁用** 的 API Key。
2. 单击列表上方的 **启用**。
3. 根据页面提示完成操作。

API Key 启用后，其有效密钥可继续用于模型网关鉴权。

禁用API Key

可通过以下任一方式禁用 API Key：

- 在“API Key”页面中，勾选目标 API Key，单击列表上方的 **禁用**。
- 在“API Key 详情”页面中，单击 **更多操作** - **禁用**。

根据页面提示确认后，完成操作。

说明：

禁用 API Key 可能影响使用其密钥的业务系统。操作前，请确认相关业务已停止调用或已切换至其他可用的 API Key。

禁用密钥

本操作仅禁用 API Key 下的指定密钥，不影响同一 API Key 下的其他密钥。

可通过以下任一方式禁用密钥：

- 在“API Key”页面中，展开目标 API Key，在密钥列表中单击目标密钥后的 **禁用**。
- 在“API Key 详情”页面的“密钥列表”区域中，单击目标密钥后的 **禁用**。

根据页面提示确认后，完成操作。

说明：

禁用密钥前，请确认业务系统已停止使用该密钥，或已完成新旧密钥切换。

删除密钥

1. 在“API Key”页面中展开目标 API Key，或进入对应的“API Key 详情”页面。
2. 在密钥列表中，单击目标密钥后的 **删除**。
3. 根据页面提示确认后，完成操作。

说明：

删除密钥后，使用该密钥的业务系统将无法继续通过鉴权。请确认密钥已停止使用后再执行删除操作。

删除API Key

可通过以下任一方式删除 API Key：

- 在“API Key”页面中，勾选目标 API Key，单击列表上方的 **删除 API Key**。
- 在“API Key 详情”页面中，单击 **更多操作** - **删除**。

根据页面提示确认后，完成操作。

说明：

删除 API Key 前，请确认其已不再被任何业务系统使用。对于仍在使用的 API Key，建议先完成业务迁移并验证调用正常，再执行删除操作。

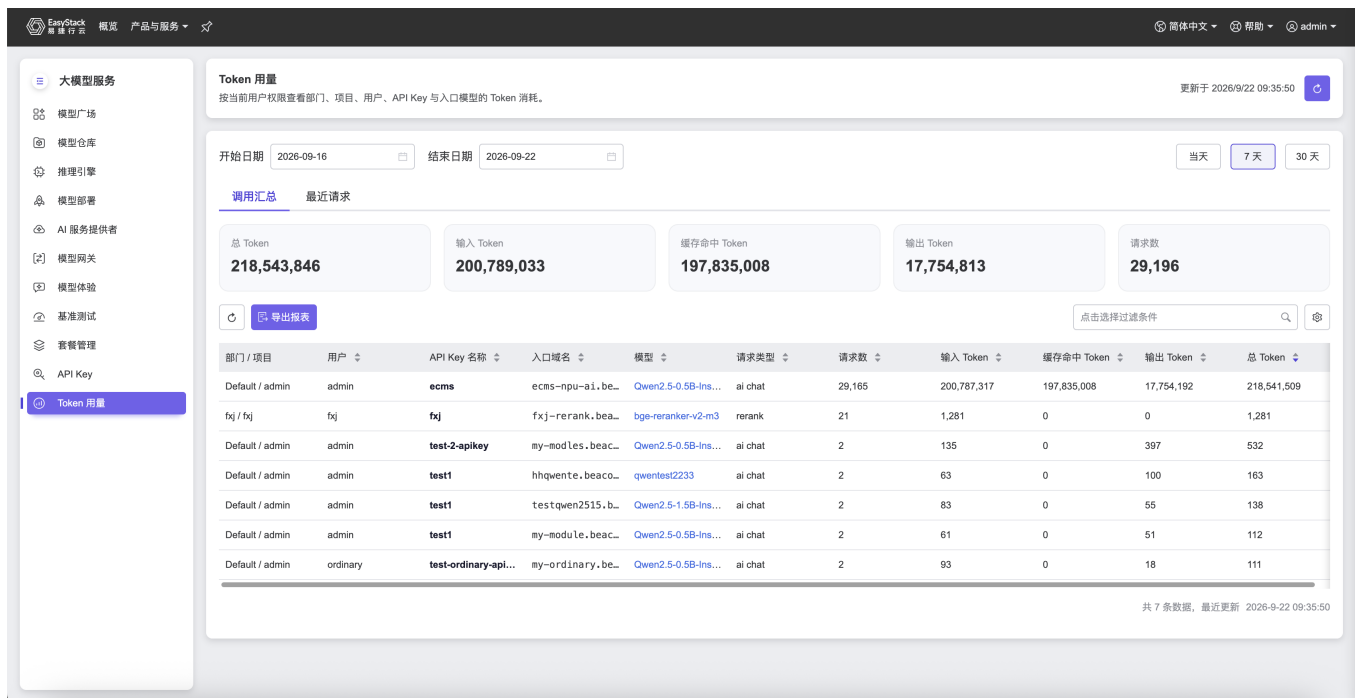
1.11 Token 用量

本章节主要介绍在 Token 用量页面中，查看调用汇总、最近请求和模型用量详情，以及筛选和导出报表等操作。在云平台顶部导航栏中，依次选择【产品与服务】-【大模型服务】-【Token 用量】，即可进入“Token 用量”页面。

“Token 用量”页面按照当前用户的权限，展示部门、项目、用户、API Key 与入口模型的 Token 消耗情况。

查看调用汇总

- 在“Token 用量”页面中，设置 开始日期 和 结束日期 ，或选择 当天 、 7 天 、 30 天 ，指定统计时间范围。
- 选择 调用汇总 页签，查看所选时间范围内的 Token 用量汇总数据和调用明细。



页面上方显示以下汇总数据。

参数	说明
总 Token	所选时间范围内消耗的 Token 总数。

参数	说明
输入 Token	请求中输入内容消耗的 Token 数量。
缓存命中 Token	请求中命中缓存的 Token 数量。
输出 Token	模型生成内容消耗的 Token 数量。
请求数	所选时间范围内产生的请求总数。

调用明细列表显示以下信息。

参数	说明
部门/项目	请求所属的部门或项目。
用户	发起请求的用户。
API Key 名称	请求使用的 API Key 名称。
入口域名	请求访问模型服务时使用的入口域名。
模型	请求调用的模型。单击模型名称可进入该模型的用量详情页面。
请求类型	请求对应的调用类型。
请求数	当前汇总记录包含的请求数量。
输入 Token	请求中输入内容消耗的 Token 数量。
缓存命中 Token	请求中命中缓存的 Token 数量。
输出 Token	模型生成内容消耗的 Token 数量。
总 Token	当前汇总记录消耗的 Token 总数。

查看最近请求

1. 在“Token 用量”页面中，设置 开始日期 和 结束日期 ，或选择 当天 、 7 天 、 30 天 ，指定查询时间范围。
2. 选择 最近请求 页签，查看所选时间范围内的请求记录。

大模型服务

模型广场

模型仓库

推理引擎

模型部署

AI 服务提供者

模型网关

模型体验

基准测试

套餐管理

API Key

Token 用量

Token 用量

按当前用户权限查看部门、项目、用户、API Key 与入口模型的 Token 消耗。

更新于 2026/9/22 09:35:50

开始日期 2026-09-16 结束日期 2026-09-22

调用汇总 最近请求

导出报表

点击选择过滤条件

时间	部门 / 项目	用户	API Key 名称	入口域名	模型	请求类型	输入 Token	缓存命中 Token	输出 Token	总 Token	状态
2026-9-21 16:43... d52c02fe-7c64...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	841	7,726	200
2026-9-21 16:43... 1676b554-53ee...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	769	7,654	200
2026-9-21 16:43... c0491bc4-a16d...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	910	7,795	200
2026-9-21 16:43... aca313ec-0077...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	708	7,593	200
2026-9-21 16:43... 191a7c4d-717f...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	894	7,779	200
2026-9-21 16:43... f911adb7-e662...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	699	7,584	200
2026-9-21 16:43... 6f64a3db-6287...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	611	7,496	200
2026-9-21 16:43... 8ff3275a-a009...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	646	7,531	200
2026-9-21 16:43... 0ca8ad0b-c133...	Default / admin	admin	ecms	ecms-npu-ai.b...	Qwen2.5-0.5B-In...	ai chat	6,885	6,784	554	7,439	200

最近请求列表显示以下信息。

参数	说明
时间	请求发生的时间及对应的请求标识。
部门/项目	请求所属的部门或项目。
用户	发起请求的用户。
API Key 名称	请求使用的 API Key 名称。
入口域名	请求访问模型服务时使用的入口域名。
模型	请求调用的模型。
请求类型	请求对应的调用类型。
输入 Token	本次请求中输入内容消耗的 Token 数量。

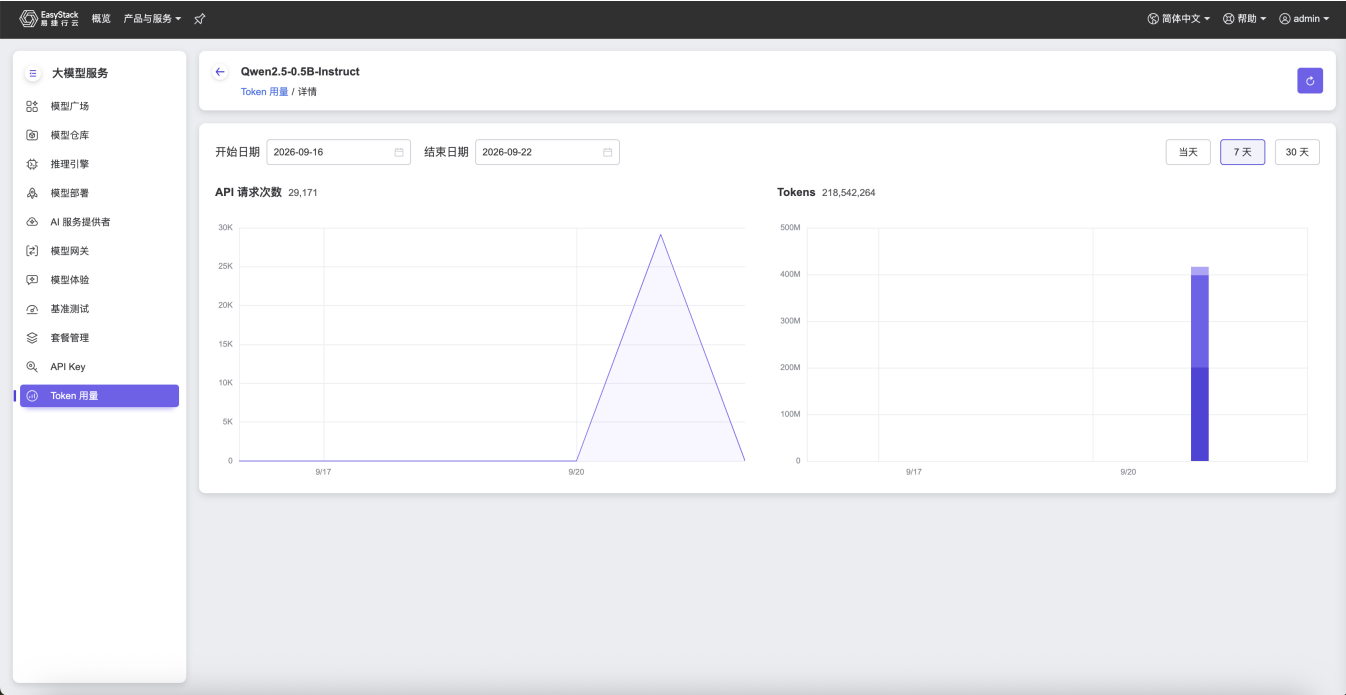
参数	说明
缓存命中 Token	本次请求中命中缓存的 Token 数量。未产生缓存命中时显示 - 。
输出 Token	本次请求中模型生成内容消耗的 Token 数量。
总 Token	本次请求消耗的 Token 总数。
状态	本次请求的响应状态，例如 200 。

查看模型用量详情

1. 在“Token 用量”页面中，选择 调用汇总 页签。

2. 在调用明细列表中，单击目标模型的名称，进入该模型的 Token 用量详情页面。

3. 设置 开始日期 和 结束日期 ，或选择 当天 、 7 天 、 30 天 ，查看指定时间范围内的模型调用数据。



模型用量详情页面通过以下图表展示数据：

图表	说明
----	----

图表	说明
API 请求次数	展示指定时间范围内该模型的 API 请求次数及其变化趋势。
Tokens	展示指定时间范围内该模型的 Token 消耗量及其变化趋势。

筛选数据

1. 在“Token 用量”页面的列表上方，单击 [点击选择过滤条件](#)。
2. 根据需要选择过滤条件并配置对应的条件值。
3. 完成配置后，查看符合条件的调用汇总或最近请求数据。

说明：

页面仅展示当前用户有权限查看的部门、项目、用户、API Key 和模型数据。

导出报表

1. 在“Token 用量”页面中，设置统计时间范围，并根据需要配置过滤条件。
2. 选择 [调用汇总](#) 或 [最近请求](#) 页签。
3. 单击 [导出报表](#)，导出当前页面的统计结果。

刷新数据

在“Token 用量”页面中，单击“刷新”图标，可获取最新的 Token 用量数据。页面右上角同时显示数据的最近更新时间。

咨询热线：400-100-3070

北京易捷思达科技发展有限公司：

北京市海淀区西北旺东路10号院东区23号楼华胜天成科研大楼一层东侧120-123

南京分公司：

江苏省南京市雨花台区软件大道168号润和创智中心B栋一楼西101

上海office：

上海黄浦区西藏中路336号华旭大厦22楼2204

成都分公司：

成都市高新区天府五街168号德必天府五街WE602

邮箱：

contact@easystack.cn (业务咨询)

partners@easystack.cn(合作伙伴咨询)

marketing@easystack.cn (市场合作)

training@easystack.cn (培训咨询)

hr@easystack.cn (招聘咨询)