

大模型服务 快速入门

产品版本: v7.0.2
发布日期: 2026-10-08

目录

1 快速入门	1
1.1 操作指引	1
1.2 前置条件准备	5
1.3 模型创建/导入	7
1.4 推理实例部署	18
1.5 模型服务及访问配置	22
1.6 API Key创建及调用	31

1 快速入门

1.1 操作指引

产品逻辑总览

在动手操作之前，先理解平台的分层设计与角色分工，可以避免在“模型、推理引擎、部署、服务提供者、网关、套餐、密钥”之间反复跳转。本平台与 OpenAI、火山方舟、阿里云百炼等主流 MaaS 产品的设计思路一致：把“把模型跑起来”和“让外部安全地调用模型”拆成两条链路，用网关做统一收口。

平台分层架构

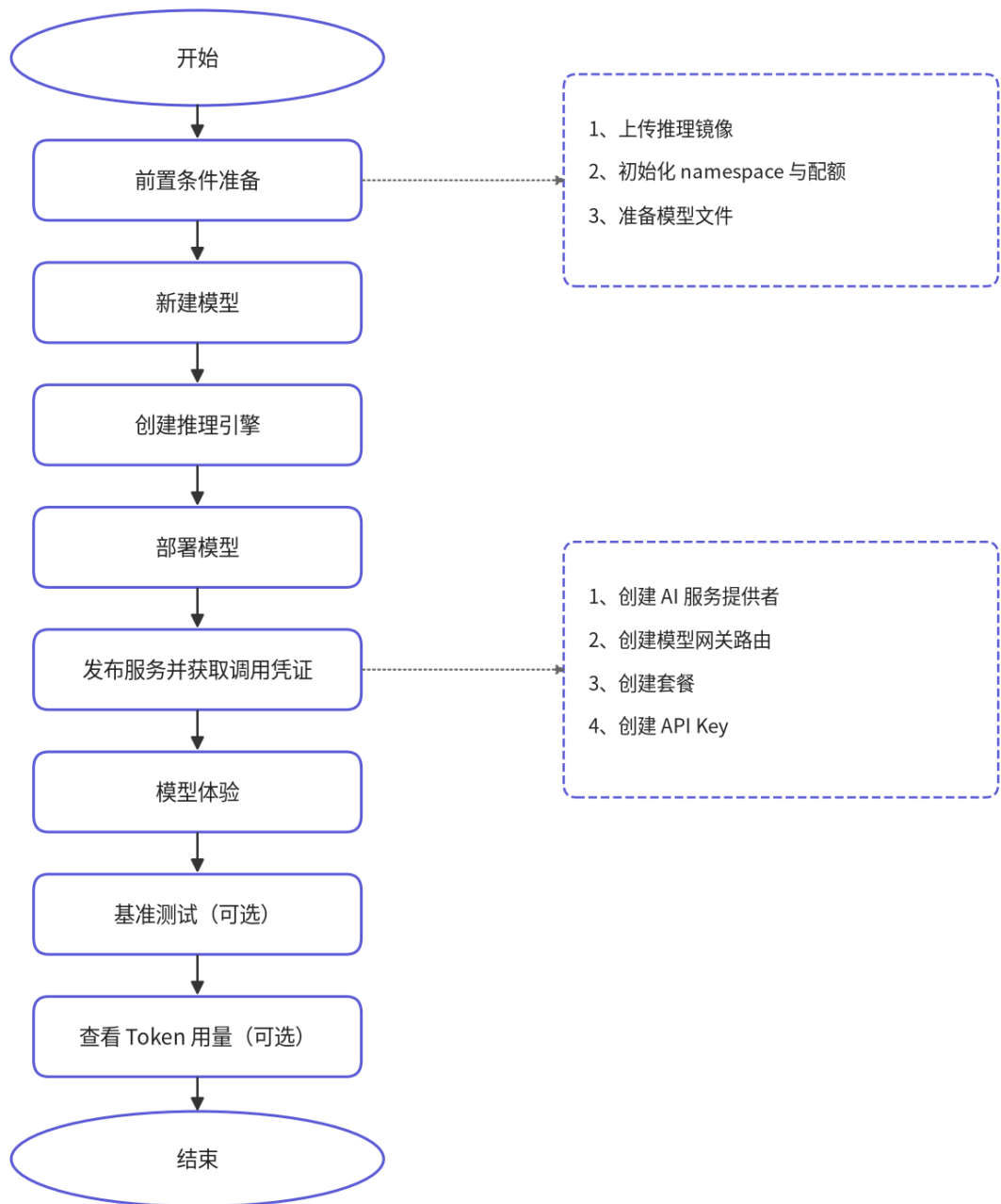
一次模型调用从下到上分为六层，每层只做一件事：模型文件层：模型权重本身，存放在《模型仓库》，可在线导入或手动上传。推理引擎层：定义模型用什么运行时（镜像、CPU/内存/加速卡、启动命令与参数），存放在《推理引擎》。模型部署层：把模型文件与推理引擎绑定成一个可运行实例，设置副本数与服务名，存放在《模型部署》。AI 服务提供者层：把运行中的部署封装成可被网关发现的“后端服务”，并区分内部部署与外部公有云 API（OpenAI、DeepSeek 等），存放在《AI 服务提供者》。模型网关层：对外暴露统一域名，按模型名把请求路由到一个或多个后端，支持优先级、权重负载均衡与故障切换，存放在《模型网关》。套餐与密钥层：套餐规定所管理的路由的请求数与 Token 数上限；API Key 绑定套餐后即成为调用凭证，存放在《套餐管理》与《API Key》。

一次调用的数据流

以一次对话请求为例：应用向 `https://<路由域名>/v1/chat/completions` 发起 HTTPS 请求，带上 `Authorization: Bearer sk-...` 请求头；模型网关校验密钥身份，确认该 Key 所属套餐允许访问此路由且未超限流；随后按请求体中的 `model` 字段，在路由配置中找到对应的后端 AI 服务提供者；请求被转发到模型部署实例上的 vLLM 服务，推理结果沿原路返回。理解这条链路后，后面每一步操作的目的都会很清晰：前面的菜单都在为这条链路“接线”。

大模型服务使用流程

大模型服务云产品的主线使用流程及具体说明如下：



快速上线操作流程

快速上线操作流程如下所示：

操作流程		描述
前置条件准备	上传推理镜像	通过 ctr 或 docker 将推理镜像上传到容器镜像服务引擎仓库，需打标签的镜像在详情页设置。
	初始化 namespace 与配额	项目管理员在《模型广场》初始化项目 namespace，并在《配额设置》按安全容器服务配置项目所需的加速卡与资源配额。
	准备模型文件	通过 ModelScope 在线导入，或下载模型包解压后在《模型仓库》创建上传任务，为新建模型准备可绑定的模型文件。
新建模型	新建模型	在《模型广场》点击“新建模型”，选择“绑定已有模型文件”，填写模型名称并选择上一步准备好的模型文件。
发布服务并获取调用凭证	创建推理引擎	在《推理引擎》选择 AI 加速卡型号，选择自动带出的镜像，使用平台默认启动命令与启动参数，按需调整资源规格。
	部署模型	在《模型部署》新建部署，选择模型与推理引擎，设置副本数后点击创建，等待模型部署状态变为就绪。
发布服务并获取调用凭证	创建 AI 服务提供者	在《AI 服务提供者》类型选择“内部”，选择刚部署的模型，集群内部访问地址自动带出。
	创建模型网关路由	在《模型网关》创建路由，填写二级域名，将模型映射到刚创建的 AI 服务提供者，按需配置优先级、权重与访问控制。
	创建套餐	在《套餐管理》创建套餐，设置每秒请求数与每秒 Token 数上限，可访问路由选择刚刚创建的路由。
	创建 API Key	在《API Key》创建 API Key，选择刚刚创建的套餐并“创建并签发密钥”，成功后立即复制密钥（密钥仅显示一次）。
模型体验	模型体验	在《模型体验》粘贴 API Key，选择路由与模型，按能力使用对话、图片生成、视觉 OCR 或语音页签验证服务可用。

操作流程		描述
基准测试（可选）	基准测试（可选）	在《基准测试》创建测试，选择已部署模型与测试类型，测量延迟与吞吐，无压测需求可跳过。
查看 Token 用量（可选）	查看 Token 用量（可选）	在《Token 用量》查看调用汇总与最近请求明细，确认计费与调用情况，上线验证后可按需查看。

1.2 前置条件准备

在使用大模型服务前，请先完成以下准备工作。

推理镜像准备

通过 `ctr` 或 `docker` 命令将推理镜像上传到容器镜像服务，建议在工作空间创建 `engines` 共享空间统一存放。
`omni` 为多模态推理引擎镜像，模型输出为语音、图片、视频的均属多模态；模型输出为文字时，无论输入是文字、语音还是图片，均不属多模态，应使用非 `omni` 镜像。镜像清单如下表，环境需要时拉取、重新打 tag 后 `push` 到目标环境；需打标签的镜像在容器镜像服务镜像详情页面设置。

推理镜像清单

镜像地址	需打标签	适用环境
vllm/vllm-openai:v0.29.0 hub.easystack.cn/engines/vllm-openai:v0.29.0	vllm、nvidia	英伟达 GPU
vllm/vllm-omni:v0.28.0 hub.easystack.cn/engines/vllm-omni:v0.28.0	vllm、nvidia	英伟达 GPU (多模态)
harbor.sourcefind.cn:5443/dcu/admin/base/vllm:0.18.1-ubuntu22.04-dtk26.04-py3.10 hub.easystack.cn/engines/vllm:0.18.1-ubuntu22.04-dtk26.04-py3.10	vllm、hygon	海光 HCU
quay.io/ascend/vllm-ascend:v0.23.0 hub.easystack.cn/engines/vllm-ascend:v0.23.0	vllm、huawei-ascend	昇腾 NPU
quay.io/ascend/vllm-omni:v0.28.0 hub.easystack.cn/engines/ascend-vllm-omni:v0.28.0	vllm、huawei-ascend	昇腾 NPU (多模态)
lmsysorg/sglang:v0.5.19-runtime hub.easystack.cn/engines/sglang:v0.5.19-runtime	sglang、nvidia	英伟达 GPU (SGLang)

说明：

访问hub.easystack.cn需要联系EasyStack相关人员下载

项目命名空间与配额

项目管理员登录后进入《模型广场》，初始化项目所需 namespace；随后在《配额设置》中为项目设置所需配额，使用安全容器服务的配置。

模型文件准备

模型文件有两种获取方式：一是在线导入，例如使用 <https://www.modelscope.cn/models/Qwen/Qwen2.5-0.5B-Instruct> 直接导入；二是手动上传，下载模型包（如 <http://fs.easystack.cn/models/Qwen2.5-0.5B.tar.gz>）后解压，在《模型仓库》创建上传任务上传到环境。新建模型时选择“绑定已有模型文件”。

1.3 模型创建/导入

推理引擎配置

1. 功能概述

推理引擎是大模型服务平台中用于定义模型推理运行环境的基础模板，通过配置镜像、启动参数、算力资源等信息，为后续模型部署提供标准化的推理运行环境。您可以通过创建推理引擎，自定义适配不同框架、不同算力需求的推理运行方案。

2. 操作入口

- 登录大模型服务平台，在左侧主导航栏点击 **推理引擎**，进入推理引擎管理列表页。
- 点击页面右上角的 **创建推理引擎** 按钮，即可打开创建配置弹窗。

3. 配置项详解

创建推理引擎的配置项分为基础信息、资源限制、运行架构、镜像选择、启动命令、启动参数、环境变量七大类，带 * 标识为必填项。

3.1 基础信息

表格

配置项	必填性	说明
名称	必填	输入推理引擎的名称，建议使用便于识别的命名，描述引擎用途或框架类型。
描述	选填	补充输入推理引擎的详细说明、适用场景等描述信息。
引擎可见性	必填	单选：- 私有引擎 ：仅创建者及被授权的用户可见、可使用- 公有引擎 ：平台内所有用户均可见、可使用
类型	必填	下拉选择推理引擎的技术框架类型，例如 vLLM 等，平台将基于该类型过滤匹配的镜像。

3.2 资源限制

用于指定推理引擎实例运行时分配的算力资源规格。

表格

配置项	必填性	说明
CPU	必填	设置实例分配的 CPU 核数，单位：核，示例值：8。
内存 (Gi)	必填	设置实例分配的内存容量，单位：Gi，示例值：10。
AI 加速卡型号	选填	下拉选择适配的 AI 加速卡（如 GPU 型号），用于提升推理性能；无加速卡需求可留空。

3.3 架构

- 必填，单选：**X86 架构 / ARM 架构**
- 注意：所选架构必须与后续选择的镜像架构保持一致，否则无法匹配到可用镜像。

3.4 镜像

镜像为推理引擎运行的基础操作系统与环境包，平台会根据您选择的引擎类型、架构自动过滤匹配的镜像。

- 若提示「暂无匹配的引擎镜像」，请检查架构选择是否正确，或确认平台是否已上传带有对应引擎标签（如 vllm）的镜像。
- 需先完成引擎类型、架构的选择后，再选择对应的引擎镜像。

3.5 启动命令

用于定义推理引擎服务的启动入口指令：

- 格式要求：**每行填写一个参数，每行内不要包含空格。**
- 示例默认命令：

```
vllm
serve
/mnt/models
```

- 说明：平台默认模型权重文件存放路径为 `/mnt/models`。
- 点击右侧 **格式化** 按钮，可自动规范命令格式。

3.6 启动参数

用于配置推理服务的运行参数：

- 格式要求：**每行填写一个参数，每行内不要包含空格。**
- 示例默认参数：

```
--served-model-name  
${SERVED_MODEL_NAME}  
--port  
8000  
--enable-prompt-tokens-details
```

- 关键说明：
 1. 平台固定支持 8000 端口，请勿修改 `--port` 后的端口值，否则服务无法正常暴露。
 2. `${SERVED_MODEL_NAME}` 为平台内置变量，在模型部署时会自动替换为实际的模型名称，无需手动填写具体值。
- 点击右侧 **格式化** 按钮，可自动规范参数格式。

3.7 环境变量

- 点击 **添加** 按钮，可新增环境变量的键值对，为推理引擎运行注入自定义环境配置（如路径、密钥、功能开关等）。

4. 操作步骤

1. 进入【推理引擎】列表页，点击右上角 **创建推理引擎**。
2. 按照上述说明，依次完成所有必填项（带*）的配置，按需配置选填项。
3. 确认所有配置无误后，点击弹窗右下角的 **创建** 按钮，完成推理引擎创建。
4. 创建成功后，返回推理引擎列表页，即可查看新建的引擎，后续可在模型部署时选择该引擎作为运行环境。
5. 若需放弃创建，点击 **取消** 按钮即可关闭弹窗。

5. 列表管理说明

在推理引擎列表页，您还可以进行以下操作：

- 查看所有引擎的名称、类型、所属部门 / 项目、创建时间等信息。
- 通过筛选条件快速查找目标引擎。
- 勾选单个或多个引擎后，点击 **删除推理引擎** 进行批量删除。
- 在操作列对单个引擎进行管理操作。

6. 注意事项

1. 镜像架构必须与所选 CPU 架构（X86/ARM）一致，且镜像需带有对应引擎类型的标签（如 vllm），否则无法匹配到可用镜像。
2. 启动命令与启动参数均需严格遵循「每行一个参数、参数内无空格」的格式要求，建议使用格式化功能校验格式。
3. 推理服务端口固定为 8000，禁止私自修改端口参数，避免服务无法正常访问。
4. 请根据模型体积、并发推理需求合理配置 CPU、内存、加速卡资源，资源不足会导致服务启动失败或推理性能低下。
5. 公有引擎会对全平台用户开放，请勿在引擎中配置敏感信息。

模型仓库配置

1. 功能概述

模型仓库是大模型服务平台的模型文件存储与管理中心，提供**公共仓库**与**私有仓库**两类存储空间，支持从 ModelScope、Hugging Face 远程导入模型，以及本地文件夹上传模型两种入库方式，同时支持仓库容量配置、下载代理设置、断点续传等能力，为后续模型部署提供稳定的文件存储支撑。

2. 操作入口

1. 登录大模型服务平台，在左侧主导航栏点击 **模型仓库**，进入模型仓库管理页面。
2. 页面左侧为「仓库列表」，展示所有可访问的仓库；右侧为选中仓库的模型文件与任务列表，支持查看、筛选与管理操作。

3. 仓库列表说明

平台默认提供两类仓库，不同仓库的访问范围与初始容量不同：

表格

仓库类型	仓库标识	默认容量	访问范围
公共仓库	model-public	500 GiB	平台内所有用户可见、可使用，共享模型存储空间
私有仓库	model-xxxxxx（唯一 ID）	100 GiB	仅创建者及被授权用户可见、可使用，专属私有存储空间

每个仓库卡片展示信息：仓库名称、仓库 ID、已使用容量 / 总容量、最近更新时间，卡片右上角提供 **容量设置** 入口。

4. 仓库容量设置

您可根据业务需求调整每个仓库的总存储容量：

1. 在左侧仓库列表中，点击目标仓库卡片右上角的 **容量设置**，弹出容量设置弹窗。
2. 在「容量限制（GiB）」输入框中填写目标总容量，单位为 GiB。
3. 点击 **保存** 生效，点击 **取消** 放弃修改。

说明：该容量为仓库工作空间的总存储上限，已使用容量不可超过设置的总容量；调小容量前请确保仓库新容量不低于仓库内文件总大小。

5. 创建导入任务（远程拉取模型）

通过导入任务，可从 ModelScope、Hugging Face 等开源模型平台远程拉取模型文件到当前仓库。

5.1 操作入口

选中目标仓库后，点击页面上方的 **创建导入任务** 按钮，打开创建配置弹窗。

5.2 配置项详解

带 * 标识为必填项。

基础信息

表格

配置项	必填性	说明
来源类型	必填	下拉选择模型来源：- ModelScope ：魔搭社区（国内平台，推荐）- Hugging Face ：Hugging Face 模型平台
模型文件名称	必填	模型在当前仓库中的显示名称；粘贴原始链接后可自动解析填充，也可手动修改。
原始链接	必填	输入对应平台的模型详情页原始链接，例如 <code>https://www.modelscope.cn/models/Qwen/Qwen2.5-0.5B-Instruct</code> 。点击右侧 解析链接 ，系统自动校验链接有效性并解析模型大小，结果显示在「解析状态」区域。
限速（MB/s）	选填	设置下载最大速度，单位为 MB/s。留空或填写 <code>0</code> 表示不限速；任务运行中暂停后可修改限速。

下载代理

用于网络受限环境下的模型下载，单选「不启用」或「启用」：

- **不启用**：直接通过公网下载模型，默认选项。
- **启用**：需配置代理参数：
 - 代理地址：输入代理服务器地址，支持 `http://` / `https://` 协议，可携带用户名和密码，示例：`http://user:pass@proxy.example.com:8080`。
 - 附加代理参数：支持配置 `HTTP_PROXY`、`HTTPS_PROXY`、`ALL_PROXY`、`NO_PROXY` 环境变量，点击 **新增参数** 可添加多条，格式为「参数名 = 参数值」。

存储容量预览

可视化展示当前仓库的**已使用容量**、**本次新增容量**与**总容量**，便于提前判断存储空间是否充足，避免因容量不足导致任务失败。

其他信息

- 解析状态：显示链接解析结果与校验状态，链接解析成功后会展示模型文件大小。
- 描述：选填，补充模型的用途、版本、说明等信息。

5.3 任务提交

所有配置确认无误后，点击弹窗右下角 **创建导入任务**，系统后台自动开始下载模型；点击 **取消** 关闭弹窗。

6. 创建上传任务（本地模型上传）

通过上传任务，可将本地的模型文件夹上传至平台仓库，支持大文件断点续传。

6.1 操作入口

选中目标仓库后，点击页面上方的 **创建上传任务** 按钮，打开上传配置弹窗。

6.2 配置项详解

表格

配置项	必填性	说明
模型文件名称	必填	模型上传后在仓库中的显示名称。
描述	选填	模型文件的补充说明信息。
LoRA 微调模型	选填	勾选后，该模型将被标记为 LoRA 微调权重，可在模型部署的 LoRA 微调模型列表中被选择。
本地文件夹选择	必填	点击选择区域，选取本地的模型文件夹。特性：1. 支持数百 GB 超大模型上传2. 支持暂停、继续与断点续传3. 上传后目录结构原样保存4. 中断后重新选择同一文件夹即可续传
存储容量预览	-	展示当前仓库存储使用情况，确认容量充足后再提交。

6.3 任务提交

配置完成后，点击 **创建并上传** 开始上传；点击 **关闭** 放弃上传并关闭弹窗。

7. 模型列表与任务管理

在页面右侧的模型列表中，可查看当前仓库内所有模型文件与任务：

- 列表字段：名称、来源、模型文件地址、状态、大小 / 进度、操作。
- 支持通过右上角筛选框快速搜索目标模型。

- 点击页面左上角刷新按钮，可同步最新的任务进度与状态。

8. 注意事项

1. **容量限制**：请确保仓库剩余容量充足，模型文件大小超出剩余容量会导致导入 / 上传失败。
2. **网络要求**：远程导入模型时，请确保原始链接可正常访问；启用代理请提前验证代理地址的连通性。
3. **断点续传**：本地上传中断后，请勿修改本地文件夹的路径与内容，重新选择同一文件夹即可继续上传。
4. **公共仓库规范**：公共仓库为全平台共享空间，请勿上传敏感、未授权或违规的模型文件。
5. **LoRA 标记**：「LoRA 微调模型」仅为分类标记，仅勾选后可在部署微调场景中被检索，普通基础模型无需勾选。

模型广场配置

1. 功能概述

模型广场是大模型服务平台的模型资产统一管理与分发入口，支持按模型能力检索、浏览全平台可用的模型资产，并可快速创建可直接用于部署的模型资产。平台提供两种模型创建方式：绑定模型仓库中已有的模型文件，或直接从 ModelScope 等远程源导入新模型文件并同步生成模型资产。

2. 操作入口与列表说明

1. 登录大模型服务平台，在左侧主导航栏点击 **模型广场**，进入模型广场列表页面。
2. 页面顶部提供筛选与搜索能力：
 - 筛选维度：提供商、模型类型、任务类型，支持下拉选择对应条件过滤模型。
 - 搜索框：输入关键词快速检索目标模型。
 - 刷新按钮：点击同步最新的模型资产数据。
3. 页面右上角点击 **新建模型**，即可打开模型创建弹窗。
4. 列表区域展示所有可见的模型资产，包含模型名称、提供商、类型、任务类型等信息，支持按条件筛选与查看。

3. 新建模型

创建模型分为「绑定已有模型文件」和「导入新模型文件」两种方式，均需先完成基础信息配置。

3.1 基础信息配置（通用）

两类创建方式均需配置以下基础信息，带 * 为必填项。

表格

配置项	必填性	说明
名称	必填	输入模型资产的显示名称，建议命名清晰可识别。
模型图标	选填	点击 上传图标 上传自定义模型图标。格式要求：支持 PNG、JPG、WebP 格式；文件大小最大 256 KB；像素尺寸不超过 512×512。
模型可见性	必填	单选：- 私有模型 ：仅创建者及被授权用户可见、可使用- 公有模型 ：平台内所有用户均可见、可使用
描述	选填	输入模型的功能说明、适用场景、版本信息等描述内容。

3.2 方式一：绑定已有模型文件

适用于模型文件已上传 / 导入至模型仓库，需将其封装为可部署的模型资产的场景。

配置项说明：

表格

配置项	必填性	说明
仓库	必填	下拉选择模型文件所在的仓库：公共仓库 / 私有仓库。
模型文件	必填	下拉选择对应仓库中已存在的模型文件。
提供商与模型分类	-	用于模型分类检索与能力标识：- 提供商 ：模型所属的厂商或组织机构，如 Meta、Qwen、DeepSeek 等- 来源名称 ：模型的原始官方名称，如 Llama-3-8B-Instruct- 模型类型 ：模型能力类型，如 llm（大语言模型）、embedding（向量嵌入）、reranker（重排序）等- 任务类型 ：支持的任务场景，如 text-generation（文本生成）、multimodal（多模态）等

配置项	必填性	说明
推理引擎	选填	为模型绑定默认的推理引擎模板，部署时可直接复用该引擎配置。

配置完成后，点击弹窗右下角 **创建模型**，即可生成模型资产并展示在模型广场列表中。

3.3 方式二：导入新模型文件

适用于需要从远程平台拉取模型文件，同时完成模型资产创建的场景。系统会自动将模型文件导入目标仓库，并同步生成模型资产。

配置项说明：

表格

配置项	必填性	说明
导入目标仓库	必填	选择模型文件导入的目标仓库：公共仓库 / 私有仓库。
来源类型	必填	选择模型来源平台，支持 ModelScope 等。
原始链接	必填	粘贴对应平台的模型详情页原始链接，示例： <code>https://www.modelscope.cn/models/Qwen/Qwen2.5-0.5B-Instruct</code> 点击右侧 解析链接 ，系统自动校验链接有效性并解析模型信息。可点击「打开 ModelScope 模型列表」跳转至模型库选择目标模型。
下载代理	必填	单选：- 不启用 ：直接通过公网下载，默认选项- 启用 ：网络受限环境下使用，需配置：① 代理地址：支持 http:// 或 https:// 协议，可包含用户名和密码 ② 附加代理参数：支持配置 HTTP_PROXY、HTTPS_PROXY、ALL_PROXY、NO_PROXY 环境变量，点击「新增参数」可添加多条
限速（MB/s）	选填	设置下载最大速度，单位 MB/s；留空或填写 0 表示不限速，任务运行中可暂停修改。
存储容量	-	预览目标仓库的已用容量与总容量，确保存储空间充足。

配置完成后，点击 **创建模型**，系统将自动启动模型文件导入任务，并在导入完成后生成对应的模型资产。

4. 模型列表管理

在模型广场列表页，您可以进行以下操作：

- 通过「提供商」「模型类型」「任务类型」下拉筛选，快速定位对应能力的模型。
- 在搜索框输入关键词，模糊搜索匹配的模型资产。
- 点击列表中的模型，可查看模型详情、进行部署等后续操作。

5. 注意事项

1. **可见性规范**：设置为「公有模型」的资产将对全平台用户开放，请勿在模型信息与文件中包含敏感、未授权内容。
2. **文件完整性**：绑定已有模型文件前，请确认模型仓库中的文件完整可用，避免部署失败。
3. **导入容量**：选择导入新模型文件时，请确保目标仓库剩余容量充足，超出容量将导致导入失败。
4. **网络连通性**：远程导入模型时，请确保原始链接可正常访问；启用代理请提前验证代理地址的连通性。
5. **图标规范**：上传的模型图标请遵守格式与大小限制，不符合要求的图片将无法上传成功。

1.4 推理实例部署

模型部署配置

1. 功能概述

模型部署是大模型服务平台的推理服务上线模块，用于将模型文件与推理引擎模板结合，部署为可对外提供服务的在线推理实例。平台支持两种模型来源：从平台模型仓库中选择已有模型文件，或直接通过 OCI 地址拉取外部模型；同时支持自定义副本数、覆盖启动配置，实现推理服务的快速创建、运行管理与全生命周期运维。

2. 操作入口与列表说明

1. 登录大模型服务平台，在左侧主导航栏点击 **模型部署**，进入模型部署管理列表页。
2. 页面右上角点击 **新建模型部署**，即可打开部署配置弹窗。
3. 列表字段说明：
 - 名称、模型文件、推理引擎：展示部署服务的基础标识与依赖资源。
 - 副本、CPU / 内存、AI 加速卡：展示部署实例的算力资源配置。
 - 状态：展示服务当前运行状态（创建中、运行中、异常、已停止等）。
 - 部门 / 项目、创建时间：展示归属信息与创建时间。
 - 操作列：支持对单个部署进行查看、启停、删除等操作。
4. 列表顶部功能：
 - 刷新按钮：同步最新的部署状态与数据。
 - 删除模型部署：勾选多个部署后，可批量删除。
 - 筛选 / 搜索框：按条件快速检索目标部署服务。

3. 新建模型部署

点击「新建模型部署」后，按以下步骤完成配置，带  标识为必填项。

3.1 基本配置

表格

配置项	必填性	说明
部署名称	必填	输入推理部署服务的名称，用于标识与管理该服务。
副本数	必填	设置推理服务的实例副本数量，默认为 1；增加副本可提升服务并发处理能力。
推理引擎	必填	下拉选择已创建的推理引擎模板。部署默认复用所选引擎的镜像、启动参数、资源规格等全部配置；架构信息会根据所选引擎自动带出。

提示：如需自定义本次部署的启动命令、参数、环境变量，可展开「高级配置」进行覆盖，不会修改原推理引擎模板。

3.2 模型文件配置

支持两种模型来源方式，二选一配置。

方式一：从模型仓库选择

适用于模型文件已导入平台模型仓库的场景。

表格

配置项	必填性	说明
模型仓库	必填	下拉选择模型所在的仓库：公共仓库 / 私有仓库。
模型文件	必填	下拉选择对应仓库内的目标模型文件。

方式二：手动输入 OCI 地址

适用于模型存放在外部镜像仓库，无需提前导入平台仓库的场景。

表格

配置项	必填性	说明
模型文件地址 (OCI URI)	必填	输入模型的 OCI 标准地址，示例： <code>oci://hub.ecns.io/models/qwen2.5-0.5b-instruct:latest</code>

3.3 SERVED_MODEL_NAME

- 选填项，填写后会作为环境变量 `SERVED_MODEL_NAME` 注入部署容器。
- 生效条件：必须在推理引擎的「启动参数」中引用该变量，例如配置 `--served-model-name ${SERVED_MODEL_NAME}`，才会动态替换为实际值。
- 作用：实现部署时动态指定服务对外暴露的模型名称，无需修改引擎模板。

3.4 高级配置（覆盖启动命令 / 参数 / 环境变量）

点击展开「高级配置」，可针对本次部署自定义运行配置，配置仅写入本次部署，不会修改原推理引擎模板；留空则沿用所选推理引擎的默认值。

表格

配置项	说明
启动命令	自定义服务启动命令，格式要求： 每行一个参数，每行内不要包含空格 。平台默认模型权重文件存放路径为 <code>/mnt/models</code> 。点击右侧「格式化」可自动规范命令格式。
启动参数	自定义服务启动参数，格式要求： 每行一个参数，每行内不要包含空格 。注意：平台固定使用 8000 端口，请勿修改端口参数，否则服务无法正常暴露。 <code>\${SERVED_MODEL_NAME}</code> 变量在此处引用生效。
环境变量	点击「添加」可新增自定义环境变量，以「变量名 = 值」的形式配置，为部署实例注入运行所需的环境配置。

3.5 提交部署

所有配置确认无误后，点击弹窗右下角 **创建**，系统开始创建部署实例；点击 **取消** 放弃创建并关闭弹窗。

4. 部署服务管理

在模型部署列表页，可对已创建的服务进行管理：

1. **状态查看**：实时查看服务的运行状态，部署完成后状态变为「运行中」即可对外提供推理服务。
2. **服务启停**：在操作列可对单个服务执行启动、停止操作，调整服务运行状态。
3. **批量删除**：勾选多个部署服务，点击顶部「删除模型部署」批量释放资源；也可在操作列删除单个服务。
4. **详情查看**：点击部署名称可进入详情页，查看服务端点、配置信息、运行日志等。

5. 注意事项

1. **架构匹配**：所选推理引擎的架构必须与模型文件的运行架构一致，否则会导致部署失败。
2. **配置作用范围**：高级配置仅对本次部署生效，不会修改原推理引擎模板；若需通用配置，请直接修改对应推理引擎。
3. **端口限制**：推理服务固定使用 8000 端口，禁止在启动参数中修改端口值，避免服务无法正常访问。
4. **资源配额**：副本数与资源规格请根据平台资源配额与业务并发需求合理设置，资源不足会导致部署启动失败。
5. **OCI 地址连通性**：使用手动输入 OCI 地址方式时，请确保平台网络可正常访问对应的镜像仓库地址。
6. **变量生效条件**：`SERVED_MODEL_NAME` 必须在启动参数中通过 `${SERVED_MODEL_NAME}` 引用才会生效，仅填写变量值不引用无法生效。

1.5 模型服务及访问配置

在使用模型服务前，需要先进行访问配置，故先做模型网关配置再做模型体验配置。

模型网关配置

1. 功能概述

模型网关是大模型服务平台的对外服务统一接入层，负责将内部部署的模型服务封装为标准化的对外访问入口，支持按模型名称分发请求到对应后端服务，提供多后端优先级调度、权重分流、域名管理、IP 访问控制等能力。通过网关可实现模型服务的统一对外发布、流量路由与安全管控，同时可与套餐体系联动实现调用限流与配额管理。

2. 操作入口与列表说明

1. 登录大模型服务平台，在左侧主导航栏点击 **模型网关**，进入路由管理列表页。
2. 页面右上角功能：
 - **根域名**：点击可查看与修改平台网关的根域名，所有路由的二级域名均基于根域名拼接生成。
 - **创建路由**：点击打开路由创建弹窗，新建对外访问路由。
3. 列表字段说明：
 - **名称**：路由的标识名称。
 - **域名**：路由对外提供服务的完整访问域名。
 - **模型**：该路由下挂载的对外模型数量与标识。
 - **AI 服务提供者**：路由后端对接的 AI 服务来源。
 - **状态**：路由当前运行状态。
 - **部门 / 项目、创建时间**：归属信息与创建时间。
 - **操作列**：支持对单个路由进行编辑、删除、详情查看等操作。
4. 列表顶部功能：
 - **刷新按钮**：同步最新的路由配置与状态数据。
 - **删除路由**：勾选多个路由后，可执行批量删除。
 - **筛选与搜索**：通过右上角筛选条件与搜索框快速定位目标路由。

3. 根域名管理

根域名是所有网关路由的基础域名后缀，平台支持自定义根域名，满足不同业务的域名规范要求。

3.1 修改根域名

1. 点击页面右上角的 **根域名** 按钮，弹出修改根域名弹窗。
2. 配置项说明：
 - 当前根域名：展示当前平台默认使用的根域名，如 `beacon.ai`。
 - 新根域名：输入要修改的目标根域名。
3. 修改影响说明：
 - 修改后，**新建路由将默认使用新的根域名**。
 - 已有路由继续使用原根域名，不会自动迁移。
 - 环境中可同时存在多个根域名；编辑已有路由时，可选择保留原域名或切换至当前根域名。
4. 确认无误后，点击 **确认修改** 生效；点击 **取消** 放弃修改。

4. 创建路由

点击「创建路由」后，按以下步骤完成配置，带 ***** 标识为必填项。

4.1 基础配置

表格

配置项	必填性	说明
名称	必填	输入路由的名称，用于标识与管理该路由。
二级域名	必填	输入路由的二级域名前缀，平台将自动拼接为 <code>二级域名.根域名</code> 的完整访问地址。要求：全平台唯一，不可与已有路由重复。

4.2 模型 → 后端映射配置

路由核心配置，定义请求分发规则：一条路由可配置多个模型，按请求体中的 `model` 参数将请求路由到对应 AI 服务；每个模型下可挂载多个后端实例，实现容灾与流量分流。

分发规则

- **模型路由**：根据请求中的模型名称，匹配到对应模型分组，转发至该分组下的后端服务。
- **后端优先级**：优先级数值越小越优先，作为主用后端；高优先级后端故障时，自动切换到低优先级后端（兜底）。
- **同优先级分流**：同一优先级的多个后端，按配置的权重比例分配流量。
- 后端模型必须从所选后端服务可查询的模型列表中选择。

模型配置项

每个模型分组包含以下配置：

表格

配置项	必填性	说明
对外模型名称	选填	调用方使用的模型标识名称。若留空，选择后端模型名称后将自动填入相同的名称作为对外模型名。
后端	必填	下拉选择后端的 AI 服务提供者 / 部署服务。
后端模型名称	必填	选择对应后端服务中实际部署的模型名称。
优先级	必填	填写整数数值，数值越小优先级越高；默认值为 0。
权重	必填	同优先级下的流量分配权重，数值越大分配的流量越多；默认值为 1。

操作说明

- 点击 **添加后端**：为当前模型新增一个后端实例，用于配置多后端容灾或流量分流。
- 点击 **添加模型**：为当前路由新增一个模型分组，支持多模型共用同一个网关域名。
- 点击模型分组右上角的 **删除**：移除整个模型分组。
- 点击后端左侧的红色删除按钮：移除单个后端实例。

4.3 IP 访问控制（黑白名单）

用于对路由的访问来源进行安全管控，下拉选择控制模式：

- **白名单（只允许列出的 IP）**：仅白名单内的 IP 地址可以访问该路由，其他 IP 均被拦截。
- **黑名单（屏蔽列出的 IP）**：黑名单内的 IP 地址被禁止访问，其他 IP 均可正常访问。

- 不选择则默认不开启 IP 访问控制，所有来源均可访问。

4.4 提交创建

所有配置确认无误后，点击弹窗右下角 **创建**，完成路由创建；点击 **取消** 放弃创建并关闭弹窗。

5. 路由管理

5.1 编辑路由

在操作列点击「编辑」，可修改路由的域名、模型后端映射、IP 访问控制等配置；修改保存后实时生效。

5.2 删除路由

- 单个删除：在操作列执行删除操作。
- 批量删除：勾选多个路由后，点击顶部「删除路由」批量删除。
- 注意：删除后路由对外访问地址立即失效，无法恢复，请确认业务无影响后再操作。

5.3 状态管理

可对路由执行启用 / 禁用操作，禁用后路由对外停止服务，适用于紧急下线场景。

6. 注意事项

1. **域名唯一性**：二级域名全平台唯一，创建前请确认无重名；根域名修改后，新建路由请留意域名后缀变化。
2. **后端匹配**：后端模型名称必须与所选后端服务中实际部署的模型一致，否则会导致请求转发失败。
3. **优先级与权重**：合理配置后端优先级实现主备容灾，同优先级按权重分流；建议至少配置 2 个不同优先级的后端，保障服务可用性。
4. **IP 名单规范**：黑白名单请准确填写 IP 地址，配置错误可能导致合法用户无法访问或异常访问未被拦截。
5. **根域名修改**：修改根域名不会自动迁移已有路由，如需统一域名需逐个编辑已有路由进行切换。
6. **服务可用性**：路由创建后，请验证接口连通性与模型转发正确性，再正式对外提供服务。

模型体验配置

1. 功能概述

模型体验是大模型服务平台的在线测试与效果对比工具，支持通过网关路由与 API Key 直接调用已部署的模型服务，覆盖**对话**、**图片生成**、**视觉 / OCR**、**语音**四大类模型场景。平台支持 1~4 列多模型同屏对比，相同

输入可同步返回多个模型的输出结果，帮助快速验证模型效果、完成选型对比与功能调试。

2. 通用操作说明

2.1 操作入口

登录大模型服务平台，在左侧主导航栏点击 **模型体验**，即可进入体验页面。

2.2 前置配置：API Key

- 所有体验场景均需先配置 **API Key**，为必填项。
- 在页面顶部的 API Key 输入框中，粘贴 `sk-` 开头的有效密钥；密钥仅在当前会话保留，关闭页面后不会持久化存储。
- 注意：请确保 API Key 已绑定有效套餐，且拥有对应路由的访问权限，否则调用会失败。

2.3 对比布局切换

页面右上角提供布局切换按钮，支持四种对比模式：

- 1 列布局：单模型体验，适合单个模型深度测试。
- 2 列布局：双模型横向对比，适合 A/B 效果验证。
- 3 列布局：三模型横向对比。
- 4 列布局：四模型田字格对比，适合多模型集中选型。

切换布局后，所有对比模型使用相同的输入参数与请求内容，同步返回结果。

2.4 模型选择通用规则

每个对比模型栏均需按以下步骤选择目标模型：

1. 先选择**路由**：下拉选择已在模型网关中创建的访问路由。
2. 再选择**模型**：下拉选择该路由下挂载的对应模型。
3. 所有对比模型配置完成后，输入统一的请求内容，即可同步调用并展示结果。

3. 对话体验

对话体验用于测试大语言模型的对话、问答、内容生成等文本交互能力，支持自定义推理参数，多模型同步对比回复效果。

3.1 请求参数配置

参数将应用于所有对比模型，点击「收起参数」可折叠参数面板；点击「恢复默认」可一键还原为初始值。

表格

参数	默认值	说明
系统提示词	空	可选，定义模型的角色定位、回答风格、约束条件等，会作用于整个对话上下文。
Temperature	0.7	控制输出的随机性，取值范围 0~2；数值越低输出越稳定、越确定，数值越高创意性越强、越随机。
Top P	1.00	核采样范围，控制候选词的概率累积阈值；取值越小输出越集中，取值 1.0 表示不限制。
Max Tokens	模型默认	最大输出 Token 长度，限制模型单次回复的最大生成长度。
Presence Penalty	0.0	话题新颖度惩罚，取值范围 -2.0~2.0；正值会鼓励模型引入新话题，降低内容重复度。
Frequency Penalty	0.0	频率惩罚，取值范围 -2.0~2.0；正值会减少模型的重复用词，降低内容重复率。

3.2 操作步骤

1. 选择「对话」标签页，配置有效的 API Key。
2. 根据需求调整对话请求参数，或保持默认值。
3. 切换到目标对比布局，为每个模型栏选择对应的路由与模型。
4. 在底部输入框中输入对话消息，按 `Enter` 发送，或点击「发送」按钮。
5. 各模型栏会同步返回回复结果，可横向对比回答质量、风格与响应速度。
6. 点击「清空」可清空当前对话内容，重新开始测试。

4. 图片生成体验

图片生成体验用于测试文生图模型的生成效果，支持自定义尺寸与随机种子，多模型同步出图对比。

4.1 请求参数配置

表格

参数	说明
图片尺寸	下拉选择生成图片的分辨率，支持：512×512、768×768、1024×1024、1024×1792、1792×1024。
Seed	随机种子，默认随机；填写固定数值后，相同提示词与 Seed 可复现一致的生成结果，便于效果复现与调试。

4.2 操作步骤

1. 选择「图片生成」标签页，配置有效的 API Key。
2. 设置图片尺寸、Seed 等参数。
3. 切换对比布局，为每个模型栏选择对应的文生图路由与模型。
4. 在底部输入框中输入图片生成提示词，按 `Enter` 或点击「生成」按钮。
5. 各模型栏同步返回生成的图片，可对比画面效果、风格还原度与生成速度。
6. 点击「清空」可清空历史生成记录。

5. 视觉 / OCR 体验

视觉 / OCR 体验用于测试多模态模型的图片理解、文字识别、图像描述等能力，支持「上传图片 + 提示词」的方式调用，多模型同步对比识别效果。

5.1 前置配置

- **选择图片**：必填，点击「+」上传本地图片，所有对比模型共用同一张测试图片。
- 支持常见图片格式，可用于 OCR 文字识别、图片内容描述、视觉问答等场景。

5.2 操作步骤

1. 选择「视觉 / OCR」标签页，配置有效的 API Key。
2. 点击上传区域，选择本地测试图片。
3. 切换对比布局，为每个模型栏选择对应的视觉模型路由与模型。
4. 在底部输入框中输入指令提示词，例如“识别图片中的所有文字”“描述这张图片的内容”。

5. 点击「识别」按钮，各模型同步返回识别 / 理解结果，对比效果差异。
6. 点击「清空」可重置图片与输入内容。

6. 语音体验

语音体验支持**文本转语音（TTS）与语音转文本（ASR）**两种模式，用于测试语音类模型的合成与识别效果。

6.1 文本转语音

将输入的文本合成为语音音频，支持多参数调整与效果试听。

参数配置

表格

参数	说明
音色	选择合成语音的音色风格，可选音色由所选模型决定。
格式	输出音频格式，默认 mp3。
合成语言	选择某一国家语言，如英语。
语速	调整合成语速，默认 1 为正常语速。

操作步骤

1. 选择「语音」标签页，点击「文本转语音」子模式。
2. 配置有效的 API Key，设置音色、格式、语言、语速等参数。
3. 选择对应的语音合成路由与模型。
4. 在底部输入框中输入需要合成的文本内容。
5. 点击「生成」按钮，生成完成后可在线试听语音效果。

6.2 语音转文本

将音频文件或实时录音识别为文字，支持多模型对比识别准确率。

音频输入方式

- **上传文件**：上传本地音频文件进行识别。
- **麦克风录音**：通过设备麦克风实时录制音频进行识别。

操作步骤

1. 选择「语音」标签页，点击「语音转文本」子模式。
2. 配置有效的 API Key，选择音频输入方式（上传文件 / 麦克风录音）。
3. 切换对比布局，为每个模型栏选择对应的语音识别路由与模型。
4. 上传音频文件或完成录音后，点击「识别」按钮。
5. 各模型同步返回文字识别结果，对比识别准确率与效果。

7. 注意事项

1. **权限校验**：请确保使用的 API Key 拥有对应路由的访问权限，且套餐配额充足，否则会出现调用失败、限流等问题。
2. **参数一致性**：同一场景下的请求参数会应用于所有对比模型，确保对比条件一致，结果具备参考性。
3. **响应时长**：多模型同时调用时，响应速度取决于最慢的模型；大参数量模型、大图、长音频场景响应时间会相应延长。
4. **数据安全**：API Key 仅在当前会话保存，请勿在公共设备使用；上传的测试图片、音频请确保合规，不包含敏感信息。
5. **模型匹配**：请选择与场景匹配的模型，例如对话场景选语言模型、图片生成选文生图模型，选错模型会导致调用失败。
6. **结果对比**：建议控制变量测试，固定参数、输入内容后对比模型输出，更易评估模型效果差异。

1.6 API Key创建及调用

在API Key创建及调用之前，需要先配置好套餐管理模块内容，以便在API Key创建时选择调用已配置好的套餐。

套餐管理配置

1. 功能概述

套餐管理是大模型服务平台的流量管控与权限配置模块，用于定义 API 调用的限流规则（单位时间请求数与 Token 处理量）与可访问路由范围。套餐通过与 API Key 绑定，实现对接口调用的流量控制、访问权限管控与配额管理，保障平台服务稳定运行，同时实现资源的精细化分配与用量管控。

2. 操作入口与列表说明

1. 登录大模型服务平台，在左侧主导航栏点击 **套餐管理**，进入套餐管理列表页。

2. 页面右上角点击 **创建套餐**，即可打开套餐配置弹窗。

3. 列表字段说明：

- 套餐名称：套餐的标识名称。
- 请求限速：单位时间内的请求次数上限。
- Token 限速：单位时间内的 Token 处理总量上限。
- 授权路由：该套餐允许访问的模型服务路由范围。
- 已加入 API Key：当前绑定该套餐的 API Key 数量。
- 共享项目：套餐共享的项目范围。
- 状态：展示套餐当前启用 / 禁用状态。
- 部门 / 项目、更新时间：展示归属信息与最近更新时间。
- 操作列：支持对单个套餐进行编辑、删除等管理操作。

4. 列表顶部功能：

- 刷新按钮：同步最新的套餐配置与状态数据。
- 删除套餐：勾选多个套餐后，可执行批量删除操作。
- 筛选与搜索：通过右上角筛选条件与搜索框，快速定位目标套餐。

3. 创建套餐

点击「创建套餐」后，按以下说明完成配置，带 * 标识为必填项。

3.1 基础信息

表格

配置项	必填性	说明
套餐名称	必填	输入套餐的名称，建议按限流规格、业务场景命名，便于区分与管理。
状态	必填	下拉选择套餐状态：- 启用 ：套餐规则生效，绑定的 API Key 可按规则调用服务- 禁用 ：套餐规则失效，所有绑定该套餐的 API Key 均无法正常调用服务

3.2 请求限速

用于控制单位时间内的接口请求次数，超出限制后调用请求将被拦截。

- **预设档位**：提供「10 次 / 1 分钟」「20 次 / 1 分钟」「100 次 / 1 分钟」「不限」四个快捷选项，点击即可自动填充数值。
- **自定义配置**：
 - 左侧输入框：填写请求数量数值。
 - 右侧下拉框：选择时间单位（如分钟）。
 - 填写 0 代表该维度不限流。

3.3 Token 限速

用于控制单位时间内的 Token 处理总量（包含输入 Token 与输出 Token），超出限制后调用将被限流。

- **预设档位**：提供「40k/1 分钟」「100k/1 分钟」「1M/1 分钟」「不限」四个快捷选项，点击即可自动填充数值。
- **自定义配置**：
 - 左侧输入框：填写 Token 数量数值。
 - 右侧下拉框：选择时间单位（如分钟）。
 - 填写 0 代表该维度不限流。

3.4 可访问路由

用于配置该套餐允许访问的模型服务路由范围：

- 在搜索框输入路由名称，可快速筛选目标路由。
- 勾选对应路由，即可为该套餐授权访问权限；未勾选的路由，绑定该套餐的 API Key 无法调用。
- 支持批量选择多条路由，实现精细化的权限管控。

3.5 提交创建

所有配置确认无误后，点击弹窗右下角 **创建套餐**，完成套餐创建；点击 **取消** 放弃创建并关闭弹窗。

4. 套餐管理

4.1 编辑套餐

在操作列点击「编辑」，可修改套餐的限流规则、授权路由与状态；修改完成后保存，配置将实时对所有绑定该套餐的 API Key 生效。

4.2 启用 / 禁用

- 启用套餐：套餐规则正常生效，绑定的 API Key 可按规则调用服务。
- 禁用套餐：立即终止该套餐的所有限流与授权规则，绑定的 API Key 将无法调用服务，适用于紧急管控场景。

4.3 删除套餐

- 单个删除：在操作列执行删除操作。
- 批量删除：勾选多个套餐后，点击顶部「删除套餐」批量删除。
- 注意：删除前请确认无 API Key 绑定该套餐，删除后规则永久失效且无法恢复。

4.4 绑定 API Key

套餐创建完成后，可在「API Key」管理页面，为指定 API Key 绑定对应套餐，绑定后限流与路由规则立即生效。

5. 注意事项

1. **限流规则**：请求限速与 Token 限速为并行限制，任意一项达到阈值，调用都会被限流拦截。

2. **路由授权**：仅授权范围内的路由可被调用，新增模型服务路由后需手动添加至对应套餐，否则无法通过该套餐访问。
3. **生效范围**：套餐配置修改后实时生效，建议在业务低峰期调整配置，避免影响在线业务。
4. **禁用影响**：禁用套餐会导致所有绑定的 API Key 失去调用权限，请确认业务影响后再操作。
5. **配额独立**：同一套餐绑定多个 API Key 时，各 API Key 的限流配额独立计算，互不占用。
6. **数值规范**：限流数值填写 `0` 代表对应维度不限流；请勿填写负数，避免配置失效。

API Key 配置

1. 功能概述

API Key 是大模型服务平台的接口调用鉴权主体，承担身份认证、服务调用权限控制与调用配额计量的作用。每个 API Key 可绑定对应配额套餐，并支持多把密钥无缝轮换，在保障调用安全的同时实现业务无感知的密钥更新，是对外调用模型推理服务的必备凭证。

2. 操作入口与列表说明

1. 登录大模型服务平台，在左侧主导航栏点击 **API Key**，进入 API Key 管理列表页。
2. 页面右上角点击 **创建 API Key**，即可打开创建配置弹窗。
3. 列表字段说明：
 - 名称：API Key 的标识名称。
 - 套餐：当前绑定的配额套餐。
 - 密钥数：该 API Key 下生效的密钥数量。
 - 负责人：API Key 的所属负责人。
 - 状态：显示当前运行状态（启用 / 禁用）。
 - 部门 / 项目、创建时间：展示归属信息与创建时间。
 - 操作列：支持对单个 API Key 进行密钥管理、套餐绑定、删除等操作。
4. 列表顶部功能：
 - 刷新按钮：同步最新的密钥与状态数据。
 - 批量操作：勾选多个 API Key 后，可批量执行**启用**、**禁用**、**删除 API Key**操作。
 - 筛选与搜索：通过右上角筛选条件与搜索框，快速定位目标 API Key。

3. 创建 API Key

点击「创建 API Key」后，按以下说明完成配置与创建。

3.1 前置提示

创建时系统将自动签发第一把密钥，**密钥明文仅显示一次**；同一 API Key 可持有多把密钥，用于业务无缝轮换。

3.2 配置项说明

带  标识为必填项。

表格

配置项	必填性	说明
名称	必填	输入 API Key 的名称，建议按业务场景、使用方命名，便于区分管理。
套餐	选填	下拉选择绑定的配额套餐；也可创建完成后，在列表中随时绑定或更换套餐。

3.3 提交与密钥保存

配置完成后，点击 **创建并签发密钥**，系统自动生成密钥并弹出展示。

重要提醒：密钥明文仅在创建时展示一次，平台不留存明文，请立即复制并妥善保存。若密钥丢失，无法找回，需重新签发新密钥。

4. API Key 与密钥管理

4.1 状态管理

- **启用**：激活 API Key，其下所有密钥可正常调用服务。
- **禁用**：停用 API Key，其下所有密钥立即失效，无法调用服务；适用于异常场景快速止损。
- **删除**：删除后 API Key 及所有密钥永久失效，已绑定的配额将释放，请谨慎操作。

4.2 密钥轮换管理

每个 API Key 支持多密钥并存，可实现无缝轮换，保障业务连续性：

1. 在操作列进入密钥管理页面，可新增密钥、删除旧密钥。
2. 标准轮换流程：新增一把新密钥 → 业务端逐步替换为新密钥 → 确认旧密钥无流量后，删除旧密钥。
3. 适用场景：密钥泄露、定期安全更换、密钥过期等场景，替换过程中业务不中断。

4.3 套餐绑定

- 每个 API Key 需绑定对应套餐后，才可正常调用服务并计量 Token 消耗。
- 支持在列表操作中随时绑定、更换套餐，灵活调整配额规格。

5. 注意事项

1. **密钥安全**：密钥明文仅创建时可见，请妥善保管，避免泄露；密钥泄露可能导致配额被盗用、服务被非法调用。
2. **丢失处理**：若密钥丢失或确认泄露，请立即新增密钥替换业务调用，再删除泄露的旧密钥；紧急情况可直接禁用对应 API Key。
3. **配额要求**：未绑定套餐的 API Key 无法正常调用服务，请确保绑定有效套餐且配额充足。
4. **批量操作风险**：批量禁用 / 删除会影响所有关联业务调用，请确认范围后再操作。
5. **安全建议**：建议定期轮换密钥，降低安全风险；轮换时采用双密钥过渡方案，避免业务中断。

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)