

大模型服务

API参考

产品版本: v7.0.2

发布日期: 2026-10-09

目录

1 API参考	1
1.1 API简介	1
1.2 调用方式	3
1.3 模型广场	9
1.4 模型导入	26
1.5 模型仓库	74
1.6 模型上传	84
1.7 容器镜像	111
1.8 推理引擎	116
1.9 模型部署	132
1.10 服务提供者	172
1.11 模型网关	191
1.12 模型体验	216
1.13 套餐管理	228
1.14 基准测试	250
1.15 租户和配额	269
1.16 平台设置	278
1.17 消费者与API Key	282
1.18 AI 加速卡资源	306

1.19 Token用量	309
1.20 发布记录	327

1 API参考

1.1 API简介

欢迎使用 Beacon 大模型服务平台 API。您可以通过 API 管理模型与模型制品、部署推理服务、接入外部或集群内部 AI 服务，并配置服务路由和访问权限，将大模型能力集成到自己的应用程序中。

本文档提供了 Beacon 各类 API 的功能描述、调用方式、参数说明及请求与响应示例，涵盖模型管理、推理部署、模型体验和基准测试等功能。在调用 API 之前，请先了解以下相关术语。

术语	说明
模型	用于描述 AI 模型的名称、类型、来源和能力等信息，可关联默认推理引擎与模型制品，作为部署和服务管理的基础。
模型仓库	用于存放和管理模型制品的存储资源，提供仓库查询与容量管理能力。
模型制品	可供部署使用的模型文件及其发布信息，记录文件来源、版本、存储位置和处理状态。
模型导入任务	将远程来源的模型文件导入模型仓库的异步任务，可查询执行进度、结果和失败原因。
模型上传	将客户端的模型文件上传到平台的过程，支持上传会话及进度管理，完成后可用于生成模型制品。
推理引擎	运行 AI 模型所需的推理运行配置，包括容器镜像、启动参数和计算资源要求。
加速器	为模型推理提供加速计算能力的设备资源，例如 GPU；可查询库存及相关资源信息。
容器镜像	封装推理服务运行环境及依赖的镜像，可用于推理引擎和模型部署。
模型部署	将模型制品与推理引擎组合为可运行的推理服务，管理其副本、资源配置和运行状态。
AI 服务提供者	供模型网关调用的上游服务配置，可以接入平台外部 AI 服务，也可以接入集群内部模型部署；可配置 API 规范、访问地址和认证信息。

术语	说明
网关路由	将对外模型名称、访问域名等入口信息与一个或多个 AI 服务提供者后端关联的转发配置。
套餐	定义 API Key 可访问的模型网关路由，以及按周期生效的请求数和 Token 数限流。
消费者	通过套餐获得模型服务访问权限的调用主体，可关联用于访问网关的 API Key。
API Key	消费者访问模型网关时使用的密钥，用于识别调用方及应用套餐的路由授权和限流配置，不用于替代管理接口的登录令牌。
模型体验	通过交互式请求体验模型能力的功能，涵盖对话、向量、图像及音频等接口，具体能力取决于目标服务。
基准测试	使用平台预定义的测试方法对模型部署发起测试请求，统计响应时延、吞吐量等指标，用于评估和对比模型服务性能。
配额	项目可使用资源的限制，用于约束资源分配；用量用于反映已经使用的资源或服务调用情况。
命名空间	承载项目模型部署等运行资源的逻辑隔离空间。
Token	身份认证令牌。管理接口通过登录令牌识别用户及其项目、部门与角色；模型用量中的 Token 指模型处理的文本计量单位。
部门	部门是云平台中用户权限的一个划分层级，用户不能横跨多个部门。
项目	项目是定义资源所有权的基本单元，可理解为租户。所有资源（如云主机等）都要隶属于某个项目中。项目必须隶属于一个部门。项目名称在单个部门中的管理范围内是唯一的，但在整个云平台中可以不唯一。
用户	用户可以被云管理员、部门管理员创建。用户通过登录后，可以操作云平台提供的各项资源，如云主机 / 云硬盘等。

1.2 调用方式

请求结构

Beacon API 通过平台提供的访问地址接收 HTTP 请求。请求使用 GET、POST、PUT、PATCH 或 DELETE 方法，具体方法以接口说明为准。查询参数应放在 URL 中。以下示例使用当前环境的 Beacon 接入路径。

结构示例

以下为请求 URL 示例：

```
https://{IP地址}/beacon/api/v1/models?page=1&page_size=20
```

- `https` 为当前环境使用的通信协议。
- `{IP地址}` 为平台提供的访问地址；如需指定端口，应写在 IP 地址之后。
- `/beacon` 为当前环境对外提供的 Beacon 访问前缀。
- `/api/v1/models` 为模型列表接口路径。
- `page=1&page_size=20` 为查询参数，多个参数使用 `&` 连接；参数值中的特殊字符应进行 URL 编码。

通信协议

当前环境通过 HTTPS 对外提供 Beacon API。其他部署环境的协议以平台管理员提供的访问地址为准；传输登录令牌、API Key 等敏感信息时应使用 HTTPS。

服务网址

当前环境的 Beacon 对外访问基地址为 `https://{IP地址}/beacon`。请向平台管理员确认实际 IP 地址和端口。

产品接口路径统一以 `/api/v1` 开头。将对外访问基地址与接口详情中的路径拼接即可，不要重复添加 `/beacon` 或 `/api/v1`。例如，`GET /api/v1/models` 对应 `GET https://{IP地址}/beacon/api/v1/models`；`GET /api/v1/health` 对应 `GET https://{IP地址}/beacon/api/v1/health`。

其他接口文档中的 `https://{endpoint}/api/v1/...` 为同一接入方式的简写：在当前环境，`{endpoint}` 应替换为 `{IP地址}/beacon`，不能只替换为 IP 地址。模型上传涉及的内部回调由平台服务调用，开发者无需直接请求。

请求方法

HTTP请求方法（也称为操作或动词），它告诉服务你正在请求什么类型的操作。

方法	说明
GET	查询指定资源；返回内容以具体接口说明为准。
PUT	向指定资源提交配置，更新或替换的范围以具体接口说明为准。
POST	创建资源、提交任务或执行指定操作。
PATCH	对指定资源进行部分更新，通常只修改请求中提供的字段。
DELETE	请求服务器删除指定资源。

字符编码

请求及返回结果都使用UTF-8字符集编码。文件、音频等二进制内容按具体接口约定传输。

公共参数

公共参数用于描述请求内容、身份认证和请求追踪。各参数的适用条件如下；接口特有参数在对应接口中说明。

参数必选性

参数表中的“是”表示完成该业务操作应提供的字段；“条件必选”表示满足描述中的条件时必须提供；“否”表示可省略。系统或客户端自动填充的必选字段仍应出现在最终请求中，默认选项不等于允许提交空值。

必选性按完整业务配置约定说明，不代表服务端对所有缺失字段均已实施强制校验。嵌套字段的必选性以使用其所属对象为前提。PATCH 接口允许省略未修改字段，但更新后的资源仍须满足业务约束；不能将完整编辑表单的必填项直接解释为每次 PATCH 都必须提交。

公共请求参数

名称	类型	是否必选	描述
Host	String	是	请求的服务器信息，从服务API的URI中获取，值为hostname[:port]，通常由HTTP客户端自动设置。
x-keystone-token	String	使用 Token 认证时必选	有效的 Keystone Token，直接填写 Token 值，不添加 Bearer 前缀。
Content-Type	String	有请求体时按接口要求	JSON请求使用application/json；音频转写使用带boundary的multipart/form-data。
Content-Length	String	否	请求body长度，单位为Byte，通常由HTTP客户端自动计算。
X-Request-Id	String	否	请求追踪标识。
X-Experience-Key	String	模型体验接口按 要求提供	用于访问模型网关的API Key。
X-Experience-Host	String	模型体验接口按 要求提供	用于选择目标路由域名。

项目与部门上下文来自登录身份，不能通过添加项目查询参数或请求头任意切换资源归属。内部上传回调的共享密钥要求见对应接口。

公共返回参数

参数名称	参数类型	描述
X-Request-Id	String	响应头中的请求追踪标识。
request_id	String	普通资源接口错误响应中的请求追踪标识；是否出现以具体响应为准。

参数名称	参数类型	描述
Content-Type	String	响应内容类型，例如application/json、text/event-stream或音频类型。

身份认证

Beacon 管理接口使用 Keystone Token识别调用者；无需认证的接口以各接口说明为准。Token由平台身份认证流程提供，Beacon 的当前用户接口只查询登录身份，不签发令牌。调用需要认证的管理接口时，在请求头中添加 `x-keystone-token: <Keystone Token>`；通过平台网关访问Beacon管理接口时，使用有效的Keystone Token，在请求头中添加`x-keystone-token:`，不添加Bearer前缀。Token由平台身份认证服务签发，应具有目标项目的作用域。Beacon当前用户接口仅查询登录身份，不签发令牌。也支持使用平台登录后的有效会话Cookie访问。

```
GET https://{IP地址}/beacon/api/v1/me
x-keystone-token: <Keystone Token>
X-Request-Id: beacon-doc-example
```

认证通过后，平台根据用户角色、项目、资源所有权及共享关系检查操作权限。读取与修改操作需要相应授权，具体访问要求见各接口说明。网关API Key用于模型服务调用，不能替代管理接口的Token。

返回结果

请求发送以后，您会收到响应，包含状态码、响应消息头和消息体。

状态码是一组从1xx到5xx的数字代码，状态码表示了请求响应的状态。为了便于查看，API文档返回示例均有换行和缩进等处理。示例中的地址、ID、令牌及业务数据仅用于说明，请按实际调用替换。

正确返回结果

普通接口调用成功时，HTTP状态码为2xx，响应体包含具体接口约定的数据。204表示成功且无响应体；WebSocket升级和缓存校验按各接口约定返回。

以获取当前登录用户为例，调用成功可能返回：

```
{
  "id": "user-example",
  "name": "示例用户",
  "email": "user@example.com",
  "roles": ["member"],
  "domainId": "domain-example",
  "domainName": "示例部门",
  "projectId": "project-example",
  "projectName": "示例项目"
}
```

异步任务已接受或创建成功，不表示模型服务已经就绪，应继续查询任务或部署状态。分页参数及默认值以各列表接口为准；部分更新时，省略字段表示不修改该项，空字符串、空数组与null的含义以具体字段说明为准。

错误返回结果

普通资源接口错误通常包含error、message和request_id，部分错误还包含details。例如：

```
{
  "error": "invalid_request",
  "message": "name is required",
  "request_id": "beacon-doc-example"
}
```

认证失败可返回error和login_url；模型体验及内部回调的错误格式见各接口说明。模型体验的上游错误可能以HTTP 200返回，需要结合响应体中的error和响应头X-Upstream-Status判断结果。

公共错误码

状态码	含义
400	请求参数或格式错误。
401	未认证或登录令牌无效。
403	无权执行该操作。

状态码	含义
404	请求的资源或接口不存在。
409	资源引用、状态或容量冲突。
413	请求体过大。
422	请求未通过业务校验。
500	服务内部错误。
502	下游服务异常。
503	服务或其依赖暂不可用。

具体错误码和触发条件以各接口说明为准。

1.3 模型广场

模型广场

模型、默认引擎和默认制品的可见性必须一致。公共模型的修改需要相应公共资源写权限。

创建时 modelArtifactId 与 importArtifact 互斥；importArtifact 必须给出 repositoryId、sourceType、sourceUri。

图标使用 iconData（Data URL）写入，通过 /models/{id}/icon 读取；图标响应支持 ETag 与 If-None-Match，命中时返回 304。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/models	分页查询模型
POST	/models	创建模型
GET	/models/{id}	获取模型详情
GET	/models/{id}/icon	获取模型图标
PATCH	/models/{id}	更新模型
DELETE	/models/{id}	删除模型

分页查询模型

功能介绍

分页查询模型。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/models
```

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始
page_size	integer	否	每页数量；默认 20，最多 200
q	string	否	关键词搜索
provider	string	否	模型提供者
modelType	string	否	模型类型
task	string	否	模型任务类型或导入任务对象
visibility	string	否	可见性筛选
status	string	否	资源状态；具体取值见业务规则
label	string	否	标签筛选
label_key	string	否	标签键筛选
label_value	string	否	标签值筛选
sort	string	否	排序表达式

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<model.Response>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始

参数	参数类型	描述
page_size	integer	每页数量
facets	model.FacetOptions	模型筛选项集合

请求示例

```
GET https://{endpoint}/api/v1/models?page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "name": "示例资源",
      "description": "",
      "status": "ready",
      "projectId": "projectId-example",
      "ownerUserId": "ownerUserId-example",
      "properties": {},
      "summary": {},
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:00:00Z"
    }
  ],
  "total": 1,
  "page": 1,
  "page_size": 1,
  "facets": {
    "providers": [
      ""
    ]
  },
}
```

```
"modelTypes": [  
  ""  
],  
"tasks": [  
  ""  
]  
}  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to list models

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建模型

功能介绍

创建模型。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/models
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
----	------	------	----

参数	参数类型	是否必选	描述
name	string	是	显示名称
description	string	否	描述
provider	string	否	模型提供者，可留空，也可由模型元数据填充。
sourceName	string	否	模型来源名称，可留空，也可由模型元数据填充。
modelType	string	否	模型类型，可留空，也可由模型元数据填充。
task	string	否	模型任务类型，可留空，也可由模型元数据填充。
engineId	string	否	默认推理引擎 ID，允许暂不绑定。
modelArtifactId	string	否	已有模型制品 ID，允许暂不绑定；同步导入新制品时省略。
labels	map<string, string>	是	标签键值对；创建时需包含 visibility，取值 private 或 public。其他标签可选。
iconData	string	否	图标 Data URL；支持格式及大小受校验约束；空值时可能省略
importArtifact	model.importArtifactRequest / null	条件必选	同步导入新制品时必选；不导入时省略。对象内必选字段见下表。

同步导入时的嵌套参数：

参数	参数类型	是否必选	描述
importArtifact.repositoryId	string	是	目标模型仓库 ID。

参数	参数类型	是否必选	描述
importArtifact.sourceType	string	是	modelscope 或 huggingface。
importArtifact.sourceUri	string	是	解析后的远程模型 URI，不能为空。
importArtifact.name	string	否	制品名称；省略时使用模型名称。
importArtifact.labels	object	是	包含 visibility，取值与目标仓库的可见性一致。
importArtifact.proxyUrl	string	否	启用代理时可提供的代理地址。
importArtifact.proxyEnv	object	否	启用代理时可提供的代理环境变量。
importArtifact.speedLimit	integer	否	下载限速，字节/秒；提供时必须大于 0。

嵌套表的必选性以提供 importArtifact 为前提。选择已有制品时，仓库选择用于筛选制品，不作为顶层 repositoryId 发送；不要将表单选择项当作 API 参数。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
ownerUserId	string	创建者用户 ID

参数	参数类型	描述
ownerName	string	创建者名称；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
properties	model.Properties	模型属性
summary	model.Summary	摘要信息
defaultEngine	model.LinkedResource / null	默认推理引擎；空值时可能省略
defaultArtifact	model.LinkedResource / null	默认模型制品；空值时可能省略
readme	string	模型说明 Markdown；空值时可能省略
iconUrl	string	图标读取 URL；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
POST https://{endpoint}/api/v1/models
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "示例模型",
  "provider": "huggingface",
  "modelType": "llm",
  "task": "text-generation",
  "labels": {
    "visibility": "private"
  }
}
```

正常响应示例

HTTP/1.1 201
Content-Type: application/json

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "properties": {},
  "summary": {},
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
403	forbidden	write permission required
400	invalid_request	malformed JSON body
409	conflict	model already exists
404	not_found	default engine or model artifact not found
409	base_image_missing	base image is missing and automatic initialization failed; configure Harbor base image before importing model files
502	backend_unavailable	base image check failed; verify Harbor base image configuration
409	capacity_exceeded	model repository capacity would be exceeded
500	internal_error	failed to create model

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取模型详情

功能介绍

获取模型详情。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/models/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
ownerUserId	string	创建者用户 ID

参数	参数类型	描述
ownerName	string	创建者名称；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
properties	model.Properties	模型属性
summary	model.Summary	摘要信息
defaultEngine	model.LinkedResource / null	默认推理引擎；空值时可能省略
defaultArtifact	model.LinkedResource / null	默认模型制品；空值时可能省略
readme	string	模型说明 Markdown；空值时可能省略
iconUrl	string	图标读取 URL；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/models/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "properties": {},
```

```
"summary": {},  
"createdAt": "2026-09-21T00:00:00Z",  
"updatedAt": "2026-09-21T00:00:00Z"  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to resolve default engine
404	not_found	model not found

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取模型图标

功能介绍

获取模型图标。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/models/{id}/icon
```

参数	类型	必选	描述
id	string	是	目标资源 ID

接口请求头

请求头	说明
If-None-Match	可选，图标 ETag 缓存校验。

请求消息

无请求体。

响应消息

200 返回图标二进制，Content-Type 为图标 MIME 类型，含 ETag / Cache-Control；304 不返回响应体。

请求示例

```
GET https://{endpoint}/api/v1/models/resource-id/icon
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: image/png
ETag: ""

<图标字节>
```

状态码与错误码

成功状态码：200, 304。

HTTP	error	说明
404	not_found	model not found
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

更新模型

功能介绍

更新模型。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
PATCH /api/v1/models/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string / null	否	仅修改名称时提供，不能为空。
description	string / null	否	描述
provider	string / null	否	模型提供者
sourceName	string / null	否	模型来源名称
modelType	string / null	否	模型类型
task	string / null	否	模型任务类型或导入任务对象
engineId	string / null	否	推理引擎 ID
modelArtifactId	string / null	否	模型制品 ID

参数	参数类型	是否必选	描述
labels	map<string, string> / null	否	省略时不修改；修改可见性时须在 labels.visibility 中提供 private 或 public。
iconData	string / null	否	图标 Data URL；支持格式及大小受校验约束

更新接口允许只提交待修改字段。完整编辑表单要求名称和可见性有值，不表示每次 PATCH 必须重复发送所有字段。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
properties	model.Properties	模型属性
summary	model.Summary	摘要信息
defaultEngine	model.LinkedResource / null	默认推理引擎；空值时可能省略
defaultArtifact	model.LinkedResource / null	默认模型制品；空值时可能省略
readme	string	模型说明 Markdown；空值时可能省略

参数	参数类型	描述
iconUrl	string	图标读取 URL；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
PATCH https://{endpoint}/api/v1/models/resource-id
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "更新后的名称"
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "properties": {},
  "summary": {},
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
403	forbidden	write permission required
400	invalid_request	malformed JSON body
404	not_found	default engine or model artifact not found
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除模型

功能介绍

删除模型。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
DELETE /api/v1/models/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/models/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
403	forbidden	write permission required
404	not_found	model not found
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.4 模型导入

模型导入

模型文件是保存在模型仓库中、可供模型注册和部署使用的 OCI 制品。远程导入是一项异步操作：服务端从 ModelScope 或 Hugging Face 下载文件，打包并发布到 Harbor，任务完成后模型文件状态变为 `ready`。

本文档同时说明远程来源解析、模型文件管理和导入任务控制。接口字段因兼容性原因继续使用 `artifact`、`modelArtifactId` 等名称，中文说明统一使用“模型文件”。本地目录上传使用另一套分片和断点续传流程，见[本地模型上传](#)。

模型文件的读取范围为当前项目和公共模型文件，不存在共享模型文件。普通项目只能修改本项目的私有模型文件；公共仓库和公共模型文件仅允许具有公共资源写权限的管理员修改。导入任务始终按当前项目隔离。

推荐的远程导入调用顺序如下：

1. 调用 `/model-source-metadata` 解析来源，取得规范化模型信息、文件列表和总大小。
2. 调用 `/model-artifacts/import`，原子创建模型文件与导入任务。
3. 使用返回的任务 ID 查询进度，必要时暂停、修改限速、恢复或重试。
4. 模型文件 `status` 变为 `ready` 后，再用于模型注册或部署。

方法	路径（产品接口统一前缀 <code>/api/v1</code> ）	功能
GET	<code>/model-source-metadata</code>	通过查询参数解析远程模型信息
POST	<code>/model-source-metadata</code>	通过 JSON 解析远程模型信息，可指定代理
POST	<code>/model-artifacts/import</code>	原子创建模型文件与导入任务（推荐）
GET	<code>/model-artifacts</code>	查询模型文件
POST	<code>/model-artifacts</code>	创建模型文件记录（高级用法）
GET	<code>/model-artifacts/{id}</code>	获取模型文件详情
PATCH	<code>/model-artifacts/{id}</code>	更新模型文件元数据

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/model-artifacts/{id}/linked-models	查询模型文件关联的模型
DELETE	/model-artifacts/{id}	删除模型文件
POST	/model-artifacts/tasks	为已有模型文件创建导入任务（高级用法）
GET	/model-artifacts/tasks	查询指定模型文件的导入任务摘要
GET	/model-artifacts/tasks/{taskId}	获取导入任务详情
POST	/model-artifacts/tasks/{taskId}/pause	暂停导入任务
PATCH	/model-artifacts/tasks/{taskId}	修改已暂停任务的限速
POST	/model-artifacts/tasks/{taskId}/resume	恢复导入任务
POST	/model-artifacts/tasks/{taskId}/retry	重试失败的导入任务
DELETE	/model-artifacts/tasks/{taskId}	取消或删除导入任务

公共数据结构

模型文件对象

模型文件接口返回以下字段：

参数	参数类型	描述
id	string	模型文件 ID
name	string	模型文件显示名称
description	string	模型文件说明
status	string	模型文件状态，取值见状态说明

参数	参数类型	描述
projectId	string	所属项目 ID
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时省略
labels	object	标签；空值时省略。 <code>visibility</code> 为 <code>private</code> 或 <code>public</code>
backend	object	后端制品引用，包含 <code>catalog</code> 、 <code>type</code> 和 <code>id</code>
summary	object	模型文件来源、导入进度和 OCI 信息
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

`summary` 的字段如下：

参数	参数类型	描述
repositoryId	string	所属模型仓库 ID
repositoryName	string	所属模型仓库名称
sourceType	string	来源类型： <code>modelscope</code> 、 <code>huggingface</code> 或 <code>local_upload</code>
taskType	string	关联任务类型；空值时省略
uploadTaskId	string	本地上传任务 ID；不适用于远程导入时省略
rootName	string	本地上传根目录名称；不适用于远程导入时省略
sourceName	string	来源模型名称；空值时省略
sourceUri	string	来源模型 URI

参数	参数类型	描述
revision	string	来源版本或提交标识；空值时省略
ociRepositoryName	string	服务端生成的 OCI 仓库名称
ociUri	string	发布后的 OCI URI；尚未发布时省略
ociDigest	string	OCI 摘要；尚未发布时省略
sizeBytes	integer	模型文件总大小，单位为字节；未知时省略
downloads	integer	远程来源下载次数；未知时省略
sourceUpdatedAt	integer	远程来源最后更新时间，Unix 秒时间戳；未知时省略
stage	string	当前阶段，例如 <code>pending</code> 、 <code>downloading</code> 、 <code>packing</code> 、 <code>publishing</code> 、 <code>succeeded</code> 或 <code>failed</code>
progress	integer	总体进度百分比，范围 0 ~ 100
totalFiles	integer	文件总数；未知时省略
completedFiles	integer	已完成文件数；未知时省略
totalBytes	integer	本次处理的总字节数；未知时省略
completedBytes	integer	已完成字节数；未知时省略
readmeReady	boolean	README 是否已经可用；值为 <code>false</code> 时可能省略
errorMessage	string	失败原因；无错误时省略
sourceFiles	array	来源文件清单；未知时省略
sourceFiles[].name	string	文件相对路径
sourceFiles[].size	integer	文件大小，单位为字节；未知时省略

参数	参数类型	描述
sourceFiles[].sha256	string	文件 SHA-256；未知时省略

导入任务对象

参数	参数类型	描述
id	string	导入任务 ID
modelArtifactId	string	关联的模型文件 ID
taskType	string	固定为 <code>import_and_publish</code>
sourceType	string	来源类型： <code>modelscope</code> 或 <code>huggingface</code>
sourceUri	string	远程模型 URI
status	string	任务状态，取值见状态说明
workerId	string	当前执行节点 ID；未分配时省略
totalBytes	integer	总字节数；未知时省略
completedBytes	integer	已完成字节数；未知时省略
progressPercent	integer	进度百分比；未知时省略
speedLimit	integer	下载限速，单位为 bytes/s；未限速时省略
downloadSpeed	integer	实时下载速度，单位为 bytes/s；没有实时数据时省略
activeTasks	integer	系统当前活跃任务数；仅待调度任务可能返回，值为 0 时省略
waitingTasksAhead	integer	当前任务之前的排队任务数；值为 0 时省略
errorMessage	string	失败原因；无错误时省略

参数	参数类型	描述
startedAt	string(date-time)	开始执行时间；尚未开始时省略
finishedAt	string(date-time)	完成或失败时间；尚未结束时省略

状态说明

模型文件常见状态：

状态	描述
pending	已创建，等待导入任务执行
importing	正在下载远程模型文件
publishing	正在打包或发布到 Harbor
paused	关联导入任务已暂停
ready	已发布，可以用于模型注册或部署
failed	导入失败，具体原因见 <code>summary.errorMessage</code>
deleting	正在取消任务并清理模型文件

导入任务常见状态：

状态	描述
pending	等待执行
running	正在下载、打包或发布
paused	已收到暂停请求； <code>workerId</code> 清空后才能修改限速或恢复

状态	描述
retry_waiting	等待自动重试，也可以调用重试接口重新排队
succeeded	导入成功，对应模型文件通常为 <code>ready</code>
failed	导入失败，可以调用重试接口
deleting	正在取消任务并清理相关数据

典型状态转换如下：

```
pending → running → succeeded
    ↳ paused → pending
    ↳ failed / retry_waiting → pending
    ↳ deleting
```

通过查询参数解析远程模型信息

功能介绍

实时读取 ModelScope 或 Hugging Face 上的模型信息，用于在创建导入任务前校验来源 URI、展示模型信息、取得文件列表和计算总大小。该接口不在 Beacon 中创建或保存任何资源。

访问要求：有效 Token；需要模型读取权限。

URI

```
GET /api/v1/model-source-metadata
```

查询参数

参数	类型	必选	描述
sourceType	string	是	来源类型： <code>modelscope</code> 或 <code>huggingface</code>

参数	类型	必选	描述
sourceUri	string	是	模型 ID 或模型页面 URL；Hugging Face 模型 ID 例如 <code>Qwen/Qwen2.5-0.5B-Instruct</code>

请求消息

无请求体。

响应消息

参数	参数类型	描述
sourceType	string	来源类型
sourceUri	string	请求中的来源 URI
modelId	string	解析后的模型 ID
displayName	string	模型显示名称；空值时省略
storageSize	integer	所有已识别文件的总大小，单位为字节；未知时省略
downloads	integer	来源站点记录的下载次数；未知时省略
revision	string	来源版本或提交标识；未知时省略
lastUpdatedTime	integer	最后更新时间，Unix 秒时间戳；未知时省略
license	string	许可证；未知时省略
modelTypes	array	模型类型；未知时省略
libraries	array	依赖库；未知时省略
languages	array	支持语言；未知时省略
tags	array	远程模型标签；未知时省略
tasks	array	模型任务类型；未知时省略

参数	参数类型	描述
publisher	string	发布者；未知时省略
files	array	文件列表；未知时省略
files[].name	string	文件相对路径
files[].size	integer	文件大小，单位为字节；未知时省略
files[].sha256	string	文件 SHA-256；来源未提供时省略

请求示例

```
GET https://{endpoint}/api/v1/model-source-metadata?
sourceType=huggingface&sourceUri=Qwen%2FQwen2.5-0.5B-Instruct
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "modelId": "Qwen/Qwen2.5-0.5B-Instruct",
  "displayName": "Qwen/Qwen2.5-0.5B-Instruct",
  "storageSize": 1073741824,
  "downloads": 1250000,
  "revision": "8de5b4c7",
  "lastUpdatedTime": 1789948800,
  "license": "apache-2.0",
  "libraries": ["transformers"],
  "tags": ["text-generation"],
  "tasks": ["text-generation"],
  "publisher": "Qwen",
  "files": [
    {
```

```
{
  "name": "config.json",
  "size": 735,
  "sha256":
    "0123456789abcdef0123456789abcdef0123456789abcdef0123456789abcdef"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	<code>sourceType</code> 或 <code>sourceUri</code> 缺失、不受支持或格式无效
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型读取权限
502	upstream_error	远程站点不可访问、返回错误或响应无法解析；具体原因见 <code>message</code>

错误返回格式见[调用方式](#)。

通过 JSON 解析远程模型信息

功能介绍

与 GET 接口功能相同，但使用 JSON 请求体，并允许为本次远程访问指定 HTTP/HTTPS 代理。即使不使用代理，也可以使用此接口提交 JSON。

当前权限中间件将该 POST 请求按写操作鉴权，因此需要模型写入权限；接口本身不会创建或修改资源。

URI

```
POST /api/v1/model-source-metadata
```

查询参数

参数	类型	必选	描述
sourceType	string	条件必选	请求体未提供非空值时，从查询参数读取
sourceUri	string	条件必选	请求体未提供非空值时，从查询参数读取

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
sourceType	string	条件必选	来源类型： <code>modelscope</code> 或 <code>huggingface</code> ；优先于同名查询参数
sourceUri	string	条件必选	模型 ID 或模型页面 URL；优先于同名查询参数
proxyUrl	string	否	完整的 HTTP 或 HTTPS 代理 URL，例如 <code>http://proxy.example.com:8080</code>

响应消息

响应字段与“通过查询参数解析远程模型信息”相同。

请求示例

```
POST https://{endpoint}/api/v1/model-source-metadata
x-keystone-token:
Content-Type: application/json
```

```
{
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
```

```
"proxyUrl": "http://proxy.example.com:8080"
}
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "modelId": "Qwen/Qwen2.5-0.5B-Instruct",
  "displayName": "Qwen/Qwen2.5-0.5B-Instruct",
  "storageSize": 1073741824,
  "revision": "8de5b4c7",
  "publisher": "Qwen",
  "files": [
    {
      "name": "config.json",
      "size": 735
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	JSON 格式错误、字段缺失、来源 URI 无效或代理 URL 无效
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型写入权限
502	upstream_error	远程站点不可访问、返回错误或响应无法解析；具体原因见 message

错误返回格式见[调用方式](#)。

原子创建模型文件与导入任务

功能介绍

在一个事务中创建模型文件及其首个远程导入任务，是第三方开发者开始远程导入时的推荐接口。201 表示两项资源已创建并进入队列，不表示模型文件已经下载完成。

`artifact.sourceType`、`artifact.sourceUri` 必须分别与 `task.sourceType`、`task.sourceUri` 完全相同。任务的 `modelArtifactId` 由服务端关联，不需要提交。

访问要求：有效 Token；需要模型文件写入权限。普通项目只能写入本项目的私有仓库，公共仓库需要公共资源写入权限。

URI

```
POST /api/v1/model-artifacts/import
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
artifact	object	是	要创建的模型文件
artifact.repositoryId	string	是	目标模型仓库 ID
artifact.name	string	是	模型文件名称
artifact.description	string	否	模型文件说明
artifact.sourceType	string	是	<code>modelscope</code> 或 <code>huggingface</code>

参数	参数类型	是否必选	描述
artifact.sourceName	string	否	来源模型显示名称
artifact.sourceUri	string	是	经过解析确认的远程模型 URI
artifact.sourceSizeBytes	integer	否	远程文件总大小，单位为字节；为保证容量校验准确，建议传入解析接口返回的正数 <code>storageSize</code>
artifact.labels	object	否	自定义标签； <code>visibility</code> 由目标仓库决定，服务端会补充或覆盖
task	object	是	导入任务配置
task.sourceType	string	是	必须与 <code>artifact.sourceType</code> 相同
task.sourceUri	string	是	必须与 <code>artifact.sourceUri</code> 相同
task.proxyUrl	string	否	HTTP/HTTPS 下载代理 URL
task.proxyEnv	object	否	代理环境变量，仅支持 <code>HTTP_PROXY</code> 、 <code>HTTPS_PROXY</code> 、 <code>ALL_PROXY</code> 、 <code>NO_PROXY</code> ；值不能为空或包含换行符
task.speedLimit	integer / null	否	下载限速，单位为 bytes/s；非空时必须大于 0

响应消息

参数	参数类型	描述
artifact	object	新建的模型文件对象，初始状态为 <code>pending</code>
task	object	新建的导入任务对象，初始状态为 <code>pending</code>

字段定义见“公共数据结构”。

请求示例

```
POST https://{endpoint}/api/v1/model-artifacts/import
x-keystone-token:
Content-Type: application/json
```

```
{
  "artifact": {
    "repositoryId": "repository-private",
    "name": "Qwen2.5-0.5B-Instruct",
    "description": "从 Hugging Face 导入",
    "sourceType": "huggingface",
    "sourceName": "Qwen/Qwen2.5-0.5B-Instruct",
    "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
    "sourceSizeBytes": 1073741824
  },
  "task": {
    "sourceType": "huggingface",
    "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
    "speedLimit": 10485760
  }
}
```

正常响应示例

```
HTTP/1.1 201 Created
Content-Type: application/json
```

```
{
  "artifact": {
    "id": "artifact-7f3a",
    "name": "Qwen2.5-0.5B-Instruct",
    "description": "从 Hugging Face 导入",
    "status": "pending",
    "projectId": "project-a",
```

```
"ownerUserId": "user-a",
"labels": {
  "visibility": "private"
},
"backend": {
  "catalog": "harbor",
  "type": "image",
  "id": "pending-import:artifact-7f3a"
},
"summary": {
  "repositoryId": "repository-private",
  "repositoryName": "私有仓库",
  "sourceType": "huggingface",
  "sourceName": "Qwen/Qwen2.5-0.5B-Instruct",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "ociRepositoryName": "qwen2.5-0.5b-instruct-artifact-7f3a",
  "sizeBytes": 1073741824,
  "stage": "pending"
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
},
"task": {
  "id": "task-91bc",
  "modelArtifactId": "artifact-7f3a",
  "taskType": "import_and_publish",
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "status": "pending",
  "totalBytes": 1073741824,
  "speedLimit": 10485760
}
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
------	-------	----

HTTP	error	说明
400	invalid_request	JSON 格式错误、必填字段缺失、来源类型不受支持、文件与任务来源不一致、代理或限速配置无效
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限
409	capacity_exceeded	目标仓库容量不足；详细容量信息见 <code>details</code>
500	internal_error	模型仓库、基础镜像或服务端配置异常，导致导入任务创建失败

错误返回格式见[调用方式](#)。

查询模型文件

功能介绍

查询当前项目和公共模型文件。传入 `repositoryId` 时，服务端会先尝试同步该仓库中的 Harbor 制品；Harbor 暂时不可用或同步超时时返回已有缓存数据，其他同步错误返回 500。

访问要求：有效 Token；需要模型文件读取权限。

URI

```
GET /api/v1/model-artifacts
```

查询参数

参数	类型	必选	描述
repositoryId	string	否	模型仓库 ID；提供时触发该仓库的 Harbor 同步

参数	类型	必选	描述
page	integer	否	页码，从 1 开始；省略或小于 1 时按 1 返回
page_size	integer	否	每页数量，最大 200；省略或传 0 时返回全部匹配项
status	string	否	按模型文件状态精确过滤
q	string	否	按名称、说明或来源名称进行不区分大小写的模糊搜索

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array	模型文件列表，字段见“模型文件对象”
total	integer	匹配总数
page	integer	当前页码
page_size	integer	请求的每页数量；未限制时为 0

请求示例

```
GET https://{endpoint}/api/v1/model-artifacts?repositoryId=repository-private&page=1&page_size=20&q=Qwen
x-keystone-token:
```

正常响应示例

HTTP/1.1 200 OK
Content-Type: application/json

```
{
  "items": [
    {
      "id": "artifact-7f3a",
      "name": "Qwen2.5-0.5B-Instruct",
      "description": "从 Hugging Face 导入",
      "status": "ready",
      "projectId": "project-a",
      "ownerUserId": "user-a",
      "labels": {
        "visibility": "private"
      },
      "backend": {
        "catalog": "harbor",
        "type": "image",
        "id": "registry.example.com/project-a/qwen2.5-0.5b-instruct-artifact-7f3a:v20260921000500-task-91bc"
      },
      "summary": {
        "repositoryId": "repository-private",
        "repositoryName": "私有仓库",
        "sourceType": "huggingface",
        "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
        "ociUri": "registry.example.com/project-a/qwen2.5-0.5b-instruct-artifact-7f3a:v20260921000500-task-91bc",
        "ociDigest":
"sha256:0123456789abcdef0123456789abcdef0123456789abcdef0123456789abcdef",
        "sizeBytes": 1073741824,
        "stage": "succeeded",
        "progress": 100,
        "readmeReady": true
      },
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:10:00Z"
    }
  ],
  "total": 1,
}
```

```
"page": 1,  
"page_size": 20  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件读取权限
500	internal_error	查询模型文件或同步 Harbor 制品失败

错误返回格式见[调用方式](#)。

创建模型文件记录

功能介绍

只创建状态为 `pending` 的模型文件记录，不创建下载任务，也不会自动变为 `ready`。通常应改用原子导入接口；仅在调用方需要自行控制两步创建流程时使用本接口，然后调用 `/model-artifacts/tasks` 创建任务。

访问要求：有效 Token；需要模型文件写入权限。普通项目只能写入本项目的私有仓库，公共仓库需要公共资源写入权限。

URI

```
POST /api/v1/model-artifacts
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
repositoryId	string	是	目标模型仓库 ID
name	string	是	模型文件名称
description	string	否	模型文件说明
sourceType	string	是	modelscope 或 huggingface
sourceName	string	否	来源模型显示名称
sourceUri	string	是	远程模型 URI
sourceSizeBytes	integer	否	来源文件总大小，单位为字节；不能为负数，建议传入正数以便容量校验
labels	object	否	自定义标签；visibility 由目标仓库决定，服务端会补充或覆盖

响应消息

返回新建的模型文件对象，status 和 summary.stage 均为 pending。字段见“模型文件对象”。

请求示例

```
POST https://{endpoint}/api/v1/model-artifacts
x-keystone-token:
Content-Type: application/json
```

```
{
  "repositoryId": "repository-private",
  "name": "Qwen2.5-0.5B-Instruct",
  "sourceType": "huggingface",
  "sourceName": "Qwen/Qwen2.5-0.5B-Instruct",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "sourceSizeBytes": 1073741824
}
```

正常响应示例

```
HTTP/1.1 201 Created
Content-Type: application/json
```

```
{
  "id": "artifact-7f3a",
  "name": "Qwen2.5-0.5B-Instruct",
  "description": "",
  "status": "pending",
  "projectId": "project-a",
  "ownerUserId": "user-a",
  "labels": {
    "visibility": "private"
  },
  "backend": {
    "catalog": "harbor",
    "type": "image",
    "id": "pending-import:artifact-7f3a"
  },
  "summary": {
    "repositoryId": "repository-private",
    "repositoryName": "私有仓库",
    "sourceType": "huggingface",
    "sourceName": "Qwen/Qwen2.5-0.5B-Instruct",
    "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
    "ociRepositoryName": "qwen2.5-0.5b-instruct-artifact-7f3a",
    "sizeBytes": 1073741824,
    "stage": "pending"
  },
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
------	-------	----

HTTP	error	说明
400	invalid_request	JSON 格式错误、必填字段缺失、来源类型不受支持或 <code>sourceSizeBytes</code> 为负数
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限，或当前调用方不能写入公共仓库
404	not_found	模型仓库不存在或不可见
409	capacity_exceeded	模型仓库容量不足
409	conflict	模型文件与已有资源冲突
500	internal_error	创建模型文件失败

错误返回格式见[调用方式](#)。

获取模型文件详情

功能介绍

获取当前项目或公共模型文件的完整信息。

访问要求：有效 Token；需要模型文件读取权限。

URI

```
GET /api/v1/model-artifacts/{id}
```

参数	类型	必选	描述
id	string	是	模型文件 ID

请求消息

无请求体。

响应消息

返回模型文件对象，字段见“公共数据结构”。

请求示例

```
GET https://{endpoint}/api/v1/model-artifacts/artifact-7f3a
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "artifact-7f3a",
  "name": "Qwen2.5-0.5B-Instruct",
  "description": "从 Hugging Face 导入",
  "status": "ready",
  "projectId": "project-a",
  "ownerUserId": "user-a",
  "labels": {
    "visibility": "private"
  },
  "backend": {
    "catalog": "harbor",
    "type": "image",
    "id": "registry.example.com/project-a/qwen2.5-0.5b-instruct-artifact-7f3a:v20260921000500-task-91bc"
  },
  "summary": {
    "repositoryId": "repository-private",
    "repositoryName": "私有仓库",
    "sourceType": "huggingface",
    "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
    "ociUri": "registry.example.com/project-a/qwen2.5-0.5b-instruct-artifact-7f3a:v20260921000500-task-91bc",
    "ociDigest":
```

```
"sha256:0123456789abcdef0123456789abcdef0123456789abcdef",
  "sizeBytes": 1073741824,
  "stage": "succeeded",
  "progress": 100,
  "readmeReady": true
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:10:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件读取权限
404	not_found	模型文件不存在或对当前项目不可见
500	internal_error	获取模型文件失败

错误返回格式见[调用方式](#)。

更新模型文件元数据

功能介绍

按字段更新模型文件的名称、说明或自定义标签。标签采用合并语义：字符串值新增或覆盖标签，`null` 删除对应标签；未提交的标签保持不变。`visibility` 由所属仓库决定，不允许通过此接口修改。名称仅用于展示，修改后不会重命名已经生成的 OCI 仓库。

访问要求：有效 Token；需要模型文件写入权限。公共模型文件只允许具有公共资源写入权限的管理员更新。

URI

```
PATCH /api/v1/model-artifacts/{id}
```

参数	类型	必选	描述
id	string	是	模型文件 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	否	新名称；去除首尾空白后不能为空
description	string	否	新说明；空字符串表示清空
labels	object	否	要合并的标签；值为 <code>null</code> 时删除该标签，不允许包含 <code>visibility</code>

响应消息

返回更新后的模型文件对象，字段见“公共数据结构”。

请求示例

```
PATCH https://{endpoint}/api/v1/model-artifacts/artifact-7f3a
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "Qwen 对话模型",
  "labels": {
    "purpose": "chat",
    "legacy": null
  }
}
```

```
}  
}
```

正常响应示例

```
HTTP/1.1 200 OK  
Content-Type: application/json
```

```
{  
  "id": "artifact-7f3a",  
  "name": "Qwen 对话模型",  
  "description": "从 Hugging Face 导入",  
  "status": "ready",  
  "projectId": "project-a",  
  "ownerUserId": "user-a",  
  "labels": {  
    "visibility": "private",  
    "purpose": "chat"  
  },  
  "backend": {  
    "catalog": "harbor",  
    "type": "image",  
    "id": "registry.example.com/project-a/qwen2.5-0.5b-instruct-artifact-7f3a:v20260921000500-task-91bc"  
  },  
  "summary": {  
    "repositoryId": "repository-private",  
    "sourceType": "huggingface",  
    "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",  
    "ociUri": "registry.example.com/project-a/qwen2.5-0.5b-instruct-artifact-7f3a:v20260921000500-task-91bc",  
    "stage": "succeeded",  
    "progress": 100  
  },  
  "createdAt": "2026-09-21T00:00:00Z",  
  "updatedAt": "2026-09-21T01:00:00Z"  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	JSON 格式错误、名称为空、标签键为空或试图修改 <code>visibility</code>
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限，或公共模型文件对当前调用方只读
404	not_found	模型文件不存在或不可见
500	internal_error	更新模型文件失败

错误返回格式见[调用方式](#)。

查询模型文件关联的模型

功能介绍

查询直接引用指定模型文件的模型，用于在删除前确认影响范围。返回的模型与模型文件具有相同可见性。

访问要求：有效 Token；需要模型文件读取权限。

URI

```
GET /api/v1/model-artifacts/{id}/linked-models
```

参数	类型	必选	描述
id	string	是	模型文件 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array	关联模型列表
items[].id	string	模型 ID
items[].name	string	模型名称

请求示例

```
GET https://{endpoint}/api/v1/model-artifacts/artifact-7f3a/linked-models
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "model-qwen",
      "name": "Qwen2.5"
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效

HTTP	error	说明
403	forbidden	无模型文件读取权限
404	not_found	模型文件不存在或不可见
500	internal_error	查询关联模型失败

错误返回格式见[调用方式](#)。

删除模型文件

功能介绍

删除模型文件、相关导入任务以及 Harbor 中已发布的 OCI 制品。存在运行中或暂停中的任务时，接口提交取消请求并返回 204，模型文件和任务先进入 `deleting`，由工作节点异步停止下载并完成清理。

如果模型文件被模型引用，`deleteLinkedModels=false` 会保留模型、解除模型文件关联并将模型标记为不可用；`deleteLinkedModels=true` 会连带删除这些模型。

访问要求：有效 Token；需要模型文件写入权限。公共模型文件只允许具有公共资源写入权限的管理员删除。

URI

```
DELETE /api/v1/model-artifacts/{id}
```

参数	类型	必选	描述
id	string	是	模型文件 ID

查询参数

参数	类型	必选	描述
----	----	----	----

参数	类型	必选	描述
deleteLinkedModels	boolean	否	是否连带删除关联模型；仅 <code>true</code> 启用，默认 <code>false</code>

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。对于异步取消，204 仅表示删除请求已接受。

请求示例

```
DELETE https://{endpoint}/api/v1/model-artifacts/artifact-7f3a?
deleteLinkedModels=false
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限，或公共模型文件对当前调用方只读
404	not_found	模型文件不存在或不可见
409	conflict	并发任务状态变化或关联关系导致删除冲突

HTTP	error	说明
502	backend_unavailable	Harbor 不可用或拒绝删除远程 OCI 制品
500	internal_error	删除模型文件失败

错误返回格式见[调用方式](#)。

为已有模型文件创建导入任务

功能介绍

为已经通过 `POST /model-artifacts` 创建的模型文件记录创建导入任务。一个模型文件同一时间只能有一个 `pending`、`running`、`paused` 或 `retry_waiting` 任务。

这是两步创建流程的第二步。常规集成应优先使用原子导入接口，避免模型文件已创建但任务创建失败。

访问要求：有效 Token；需要当前项目的模型文件写入权限。

URI

```
POST /api/v1/model-artifacts/tasks
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
modelArtifactId	string	是	当前项目中的模型文件 ID
sourceType	string	是	<code>modelscope</code> 或 <code>huggingface</code>
sourceUri	string	是	远程模型 URI
proxyUrl	string	否	HTTP/HTTPS 下载代理 URL

参数	参数类型	是否必选	描述
proxyEnv	object	否	代理环境变量，仅支持 HTTP_PROXY 、 HTTPS_PROXY 、 ALL_PROXY 、 NO_PROXY ；值不能为空或包含换行符
speedLimit	integer / null	否	下载限速，单位为 bytes/s；非空时必须大于 0

响应消息

返回新建的导入任务对象，初始 `status` 为 `pending` ， `taskType` 为 `import_and_publish` 。

请求示例

```
POST https://{endpoint}/api/v1/model-artifacts/tasks
x-keystone-token:
Content-Type: application/json
```

```
{
  "modelArtifactId": "artifact-7f3a",
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "speedLimit": 10485760
}
```

正常响应示例

```
HTTP/1.1 201 Created
Content-Type: application/json
```

```
{
  "id": "task-91bc",
  "modelArtifactId": "artifact-7f3a",
  "taskType": "import_and_publish",
  "sourceType": "huggingface",
```

```
"sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
"status": "pending",
"totalBytes": 1073741824,
"speedLimit": 10485760
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
400	invalid_request	JSON 格式错误、必填字段缺失、来源类型不受支持、代理或限速配置无效
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限
404	not_found	当前项目中不存在指定模型文件
409	conflict	模型文件已有活动任务
409	capacity_exceeded	模型仓库容量不足
409	base_image_missing	导入所需基础镜像缺失且自动初始化失败
502	backend_unavailable	无法验证导入所需基础镜像
500	internal_error	创建导入任务失败

错误返回格式见[调用方式](#)。

查询指定模型文件的导入任务摘要

功能介绍

查询当前项目中一个或多个模型文件关联的导入任务摘要。每个 `modelArtifactId` 最多返回一项；此接口不是完整任务历史列表，也不保证返回全部历史任务。不传 `modelArtifactId` 时返回空列表。

访问要求：有效 Token；需要模型文件读取权限。

URI

```
GET /api/v1/model-artifacts/tasks
```

查询参数

参数	类型	必选	描述
<code>modelArtifactId</code>	array	是	模型文件 ID，可重复传入，例如 <code>modelArtifactId=id-a&modelArtifactId=id-b</code> ；空值和重复值会被忽略

请求消息

无请求体。

响应消息

参数	参数类型	描述
<code>items</code>	array	各指定模型文件返回的任务摘要，字段见“导入任务对象”
<code>total</code>	integer	返回的任务数量，不是历史任务总数

请求示例

```
GET https://{endpoint}/api/v1/model-artifacts/tasks?
modelArtifactId=artifact-7f3a&modelArtifactId=artifact-8b2c
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "task-91bc",
      "modelArtifactId": "artifact-7f3a",
      "taskType": "import_and_publish",
      "sourceType": "huggingface",
      "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
      "status": "running",
      "workerId": "worker-01",
      "totalBytes": 1073741824,
      "completedBytes": 536870912,
      "progressPercent": 40,
      "downloadSpeed": 5242880,
      "startedAt": "2026-09-21T00:01:00Z"
    }
  ],
  "total": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件读取权限
500	internal_error	查询导入任务失败

错误返回格式见[调用方式](#)。

获取导入任务详情

功能介绍

按任务 ID 获取当前项目中的导入任务。待调度任务还会尽可能返回活跃任务数和排队位置。

访问要求：有效 Token；需要模型文件读取权限。

URI

```
GET /api/v1/model-artifacts/tasks/{taskId}
```

参数	类型	必选	描述
taskId	string	是	导入任务 ID

请求消息

无请求体。

响应消息

返回导入任务对象，字段见“公共数据结构”。

请求示例

```
GET https://{endpoint}/api/v1/model-artifacts/tasks/task-91bc
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "task-91bc",
  "modelArtifactId": "artifact-7f3a",
```

```
"taskType": "import_and_publish",
"sourceType": "huggingface",
"sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
"status": "running",
"workerId": "worker-01",
"totalBytes": 1073741824,
"completedBytes": 536870912,
"progressPercent": 40,
"speedLimit": 10485760,
"downloadSpeed": 5242880,
"startedAt": "2026-09-21T00:01:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件读取权限
404	not_found	当前项目中不存在该任务或关联模型文件已删除
500	internal_error	获取导入任务失败

错误返回格式见[调用方式](#)。

暂停导入任务

功能介绍

暂停 `pending` 或 `running` 状态的任务，并保留已下载数据以便继续。接口成功后立即返回 `status: paused`；如果返回中仍有 `workerId`，表示工作节点还在退出，需等待 `workerId` 省略后才能修改限速或恢复。

访问要求：有效 Token；需要当前项目的模型文件写入权限。

URI

```
POST /api/v1/model-artifacts/tasks/{taskId}/pause
```

参数	类型	必选	描述
taskId	string	是	导入任务 ID

请求消息

无请求体。

响应消息

返回暂停后的导入任务对象。

请求示例

```
POST https://{endpoint}/api/v1/model-artifacts/tasks/task-91bc/pause
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "task-91bc",
  "modelArtifactId": "artifact-7f3a",
  "taskType": "import_and_publish",
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "status": "paused",
  "workerId": "worker-01",
  "totalBytes": 1073741824,
  "completedBytes": 536870912,
  "progressPercent": 40,
```

```
"startedAt": "2026-09-21T00:01:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限
404	not_found	当前项目中不存在该任务
409	conflict	任务状态不是 <code>pending</code> 或 <code>running</code> ，无法暂停
500	internal_error	暂停任务失败

错误返回格式见[调用方式](#)。

修改已暂停任务的限速

功能介绍

修改已完全暂停任务的下载限速。任务必须为 `paused`，并且 `workerId` 已清空。正整数设置 bytes/s 限速，`null` 取消限速。

访问要求：有效 Token；需要当前项目的模型文件写入权限。

URI

```
PATCH /api/v1/model-artifacts/tasks/{taskId}
```

参数	类型	必选	描述
taskId	string	是	导入任务 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
speedLimit	integer / null	否	正整数设置下载限速，单位为 bytes/s； 省略或传 <code>null</code> 均表示取消限速

响应消息

返回更新后的任务，`status` 仍为 `paused`。

请求示例

```
PATCH https://{endpoint}/api/v1/model-artifacts/tasks/task-91bc
x-keystone-token:
Content-Type: application/json
```

```
{
  "speedLimit": 20971520
}
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "task-91bc",
  "modelArtifactId": "artifact-7f3a",
  "taskType": "import_and_publish",
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "status": "paused",
```

```
"totalBytes": 1073741824,  
"completedBytes": 536870912,  
"progressPercent": 40,  
"speedLimit": 20971520,  
"startedAt": "2026-09-21T00:01:00Z"  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	JSON 格式错误，或非空 <code>speedLimit</code> 小于等于 0
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限
404	not_found	当前项目中不存在该任务
409	conflict	任务尚未完全暂停，或工作节点尚未释放任务
500	internal_error	修改任务限速失败

错误返回格式见[调用方式](#)。

恢复导入任务

功能介绍

将已完全暂停的任务重新加入队列。任务必须为 `paused` 且 `workerId` 已清空；成功后返回 `status: pending`。

访问要求：有效 Token；需要当前项目的模型文件写入权限。

URI

```
POST /api/v1/model-artifacts/tasks/{taskId}/resume
```

参数	类型	必选	描述
taskId	string	是	导入任务 ID

请求消息

无请求体。

响应消息

返回重新进入队列的任务对象，`status` 为 `pending`。

请求示例

```
POST https://{endpoint}/api/v1/model-artifacts/tasks/task-91bc/resume
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "task-91bc",
  "modelArtifactId": "artifact-7f3a",
  "taskType": "import_and_publish",
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "status": "pending",
  "totalBytes": 1073741824,
  "completedBytes": 536870912,
  "progressPercent": 40,
  "speedLimit": 20971520,
  "startedAt": "2026-09-21T00:01:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限
404	not_found	当前项目中不存在该任务
409	conflict	任务未暂停、仍在暂停过程中或不能恢复
500	internal_error	恢复任务失败

错误返回格式见[调用方式](#)。

重试导入任务

功能介绍

将 `failed` 或 `retry_waiting` 状态的任务重新加入队列。重试会清空旧的执行节点、进度、错误和起止时间，成功后返回 `status: pending`。

访问要求：有效 Token；需要当前项目的模型文件写入权限。

URI

```
POST /api/v1/model-artifacts/tasks/{taskId}/retry
```

参数	类型	必选	描述
taskId	string	是	导入任务 ID

请求消息

无请求体。

响应消息

返回重新进入队列的任务对象，`status` 为 `pending`。

请求示例

```
POST https://{endpoint}/api/v1/model-artifacts/tasks/task-91bc/retry
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "task-91bc",
  "modelArtifactId": "artifact-7f3a",
  "taskType": "import_and_publish",
  "sourceType": "huggingface",
  "sourceUri": "Qwen/Qwen2.5-0.5B-Instruct",
  "status": "pending",
  "totalBytes": 1073741824,
  "speedLimit": 20971520
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限
404	not_found	当前项目中不存在该任务

HTTP	error	说明
409	conflict	任务状态不是 <code>failed</code> 或 <code>retry_waiting</code>
500	internal_error	重试任务失败，包括基础镜像检查失败

错误返回格式见[调用方式](#)。

取消或删除导入任务

功能介绍

取消活动任务或删除非活动任务。对 `running` 、 `paused` 任务调用时，接口返回 204，任务进入 `deleting` ，工作节点随后停止下载并清理任务及关联模型文件；这是一项异步操作。

对于 `pending` 、 `retry_waiting` 、 `failed` 或终态任务，服务端可以直接删除：如果关联模型文件尚未就绪，会同时删除该模型文件；如果模型文件已经就绪，则只删除任务记录。 `deleteLinkedModels` 决定清理模型文件时如何处理引用它的模型。

访问要求：有效 Token；需要当前项目的模型文件写入权限。

URI

```
DELETE /api/v1/model-artifacts/tasks/{taskId}
```

参数	类型	必选	描述
taskId	string	是	导入任务 ID

查询参数

参数	类型	必选	描述
----	----	----	----

参数	类型	必选	描述
deleteLinkedModels	boolean	否	清理关联模型文件时是否连带删除引用它的模型；仅 <code>true</code> 启用，默认 <code>false</code>

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。活动任务返回 204 仅表示取消请求已接受。

请求示例

```
DELETE https://{endpoint}/api/v1/model-artifacts/tasks/task-91bc?
deleteLinkedModels=false
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型文件写入权限
404	not_found	当前项目中不存在该任务或关联模型文件
409	conflict	任务状态发生并发变化，或关联模型阻止清理
500	internal_error	取消或删除任务失败

错误返回格式见[调用方式](#)。

1.5 模型仓库

模型仓库

列表和详情会确保当前租户默认仓库存在，并可能初始化工作空间。

limitBytes >= -1; -1 表示不限。新容量不能低于 usedBytes + reservedBytes，否则返回 409 capacity_too_small。容量修改还要求仓库容量管理权限及 Harbor 管理服务可用。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/model-repositories	查询模型仓库
GET	/model-repositories/{id}	获取模型仓库详情
GET	/model-repositories/{id}/capacity	查询仓库容量
PUT	/model-repositories/{id}/capacity	设置仓库容量

查询模型仓库

功能介绍

查询模型仓库。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/model-repositories
```

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始

参数	类型	必选	描述
page_size	integer	否	每页数量；建议显式传入，例如 20；省略时部分列表返回全部

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<modelrepository.Response>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始
page_size	integer	每页数量

请求示例

```
GET https://{endpoint}/api/v1/model-repositories?page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "name": "示例资源",

```

```
{
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "summary": {},
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
},
"total": 1,
"page": 1,
"page_size": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to initialize public model repository
401	unauthorized	authentication required

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取模型仓库详情

功能介绍

获取模型仓库详情。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/model-repositories/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
backend	modelrepository.backendRef	后端资源标识
summary	modelrepository.Summary	摘要信息
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/model-repositories/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "summary": {},
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to initialize public model repository
401	unauthorized	authentication required
404	not_found	model repository not found

HTTP	error	说明
503	backend_unavailable	harbor admin client is not configured
502	workspace_unavailable	failed to initialize workspace

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询仓库容量

功能介绍

查询仓库容量。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/model-repositories/{id}/capacity
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
usedBytes	int64	已使用容量，字节
reservedBytes	int64	预留容量，字节

参数	参数类型	描述
limitBytes	int64	容量上限，字节；-1 表示不限

请求示例

```
GET https://{endpoint}/api/v1/model-repositories/resource-id/capacity
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "usedBytes": 0,
  "reservedBytes": 0,
  "limitBytes": 0
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
404	not_found	model repository not found
500	internal_error	failed to get model repository
503	backend_unavailable	harbor admin client is not configured
502	backend_unavailable	failed to query repository capacity

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

设置仓库容量

功能介绍

设置仓库容量。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
PUT /api/v1/model-repositories/{id}/capacity
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
limitBytes	int64	否	容量上限，字节；-1 表示不限

limitBytes >= -1；-1 表示不限。新容量不能低于 usedBytes + reservedBytes，否则返回 409 capacity_too_small。容量修改还要求仓库容量管理权限及 Harbor 管理服务可用。

响应消息

参数	参数类型	描述
usedBytes	int64	已使用容量，字节

参数	参数类型	描述
reservedBytes	int64	预留容量, 字节
limitBytes	int64	容量上限, 字节; -1 表示不限

请求示例

```
PUT https://{endpoint}/api/v1/model-repositories/resource-id/capacity
x-keystone-token:
Content-Type: application/json
```

```
{
  "limitBytes": 107374182400
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "usedBytes": 0,
  "reservedBytes": 0,
  "limitBytes": 0
}
```

状态码与错误码

成功状态码: 200。

HTTP	error	说明
400	invalid_request	limitBytes must be greater than or equal to -1
403	forbidden	repository capacity permission required

HTTP	error	说明
409	capacity_too_small	capacity cannot be lower than current usage and task reservations
503	backend_unavailable	harbor admin client is not configured
502	backend_unavailable	failed to update repository capacity
401	unauthorized	authentication required
404	not_found	model repository not found
500	internal_error	failed to get model repository

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.6 模型上传

本地模型上传

使用Keystone Token先创建上传任务，使用返回的 endpoint、files[].uploadToken 和文件标识上传，所有文件完成后调用 complete，等待打包发布完成。文件字节流由独立 tus 服务接收。

任务管理接口通过 x-keystone-token 请求头携带有效 Keystone Token。文件上传入口采用不同的校验方式：创建 tus 上传资源时，在 Upload-Metadata 中携带专用 uploadToken；仅携带 x-keystone-token 不能替代 uploadToken。

relativePath 必须以 rootName/ 开头，不允许绝对路径、目录穿越或重复路径；sizeBytes 不得为负。默认最多 20000 个文件、总计 2 TiB。

resume 的清单必须与原任务匹配；上传未完成调用 complete 返回 409 upload_incomplete。发布失败可调用 retry，具体允许状态由任务控制。

任务状态包含 created、uploading、paused、uploaded、packing、publishing、succeeded、failed、deleting、cancelled、expired。

方法	路径（产品接口统一前缀 /api/v1）	功能
POST	/model-upload-tasks	创建本地模型上传任务
GET	/model-upload-tasks/{id}	获取本地上传任务
GET	/model-upload-tasks/{id}/files	分页查询上传文件
POST	/model-upload-tasks/{id}/pause	暂停上传任务
POST	/model-upload-tasks/{id}/resume	恢复上传任务
POST	/model-upload-tasks/{id}/complete	提交上传完成
POST	/model-upload-tasks/{id}/retry	重试制品发布
DELETE	/model-upload-tasks/{id}	取消本地上传任务

创建本地模型上传任务

功能介绍

创建本地模型上传任务。

访问要求：通过 x-keystone-token 携带有效 Keystone Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/model-upload-tasks
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
repositoryId	string	是	模型仓库 ID
name	string	是	显示名称
description	string	否	描述；空值时可能省略
lora	boolean	否	LoRA 配置；空值时可能省略
rootName	string	是	上传根目录名称，从所选文件的相对路径提取。
files	array<modelartifactupload.ManifestFile>	是	非空文件清单，从所选模型目录生成。
files[].relativePath	string	是	以 rootName/ 开头的相对路径
files[].sizeBytes	int64	是	文件字节数，不得为负
files[].lastModified	int64	否	文件修改时间标记，恢复时保持原值

响应消息

参数	参数类型	描述
task	modelartifactupload.Task	模型任务类型或导入任务对象
files	array<modelartifactupload.File>	文件清单
endpoint	string	文件上传入口，默认 /beacon/uploads/files/；相对地址按服务源站解析

创建和恢复任务返回的文件对象除文件信息外，还包含以下上传相关字段：

参数	参数类型	描述
files[].id	string	文件 ID，用于创建 tus 上传资源
files[].relativePath	string	含根目录的相对路径
files[].sizeBytes	int64	文件总字节数
files[].uploadToken	string	专用上传凭证，创建 tus 上传资源时放入 Upload-Metadata
files[].uploadTokenExpiresAt	string(date-time)	专用上传凭证过期时间，默认有效期为 1800 秒，以服务配置和响应为准
files[].tusUploadId	string	已绑定的 tus 上传 ID，未建立时可省略
files[].tusUploadUrl	string	已记录的 tus 上传地址，尚未记录时可省略

uploadToken 并非 Keystone Token。创建任务或调用 resume 获取凭证时仍需通过普通 API 的 Keystone 认证。

请求示例

```
POST https://{endpoint}/api/v1/model-upload-tasks
x-keystone-token:
Content-Type: application/json
```

```
{
  "repositoryId": "repository-id",
  "name": "本地模型",
  "rootName": "model",
  "files": [
    {
      "relativePath": "model/config.json",
      "sizeBytes": 1024
    }
  ]
}
```

正常响应示例

```
HTTP/1.1 201
Content-Type: application/json
```

```
{
  "task": {
    "id": "id-example",
    "modelArtifactId": "modelArtifactId-example",
    "repositoryId": "repositoryId-example",
    "rootName": "",
    "status": "created",
    "totalFiles": 0,
    "completedFiles": 0,
    "totalBytes": 0,
    "completedBytes": 0,
    "progressPercent": 0,
    "retryCount": 0,
    "version": 1,
    "createdAt": "2026-09-21T00:00:00Z",
    "updatedAt": "2026-09-21T00:00:00Z"
  }
}
```

```
{,
  "files": [
    {
      "id": "id-example",
      "relativePath": "",
      "artifactPath": "",
      "sizeBytes": 0,
      "uploadedBytes": 0,
      "progressPercent": 0,
      "status": "created",
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:00:00Z"
    }
  ],
  "endpoint": "https://service.example.com"
}
```

创建任务后，使用响应中的 endpoint 创建每个文件的 tus 上传资源。以下示例使用当前默认网关前缀，文件传输接口不属于 /api/v1 路径。

创建文件上传请求：

```
POST https://{host}/beacon/uploads/files/
Tus-Resumable: 1.0.0
Upload-Length: 1024
Upload-Metadata: taskId <Base64任务ID>,fileId <Base64文件ID>,relativePath
<Base64相对路径>,uploadToken <Base64上传Token>
Content-Length: 0
```

元数据值使用 UTF-8 字节的标准 Base64 编码，尖括号是需要替换的占位符。taskId 使用 task.id，fileId 使用对应 files[].id，relativePath、uploadToken 使用同一文件的响应值；Upload-Length 必须等于该文件的 sizeBytes。Base64 编码不等于加密。

服务端通过 pre-create hook 校验 uploadToken 的签名、有效期、任务和文件绑定，并检查项目归属、文件路径、大小及任务状态。缺失或无效 uploadToken 返回 401 invalid upload token。此请求不要求额外携带 Keystone Token，单独携带有效 Keystone Token 仍不能通过上传创建校验。

创建成功返回：

```
HTTP/1.1 201 Created
Tus-Resumable: 1.0.0
Location: https://{host}/beacon/uploads/files/{upload-id}
```

使用 Location 查询已接收字节数：

```
HEAD https://{host}/beacon/uploads/files/{upload-id}
Tus-Resumable: 1.0.0
```

```
HTTP/1.1 200 OK
Tus-Resumable: 1.0.0
Upload-Length: 1024
Upload-Offset: 0
```

按服务端偏移发送文件分片：

```
PATCH https://{host}/beacon/uploads/files/{upload-id}
Tus-Resumable: 1.0.0
Content-Type: application/offset+octet-stream
Upload-Offset: 0
Content-Length: 512
```

<文件开始位置的 512 字节原始二进制内容>

```
HTTP/1.1 204 No Content
Tus-Resumable: 1.0.0
Upload-Offset: 512
```

继续从返回的偏移上传剩余内容，直到 Upload-Offset 等于 Upload-Length。PATCH 请求体是原始文件字节，不是 JSON 或 multipart/form-data。网络中断后先 HEAD 查询实际偏移，再续传。

当前 HEAD、PATCH 示例未携带认证头，与现有实现和实测一致。不要将请求成功理解为 Keystone Token 已被验证；当前即使携带错误的 x-keystone-token，续传仍可能成功。uploadToken 到期不能据此推断已创建上传 URL 的续传会被拒绝。

以上 201、200、204 为文件传输响应，和本节创建任务 API 的响应分开处理。文件传输错误可能来自网关或 tus 服务，不保证采用任务 API 的 JSON 错误封装。

状态码与错误码

成功状态码：201。

HTTP	error	说明
403	forbidden	write permission required
400	invalid_manifest	malformed upload manifest
500	internal_error	failed to create upload credentials
503	backend_unavailable	model upload service is not configured
502	backend_unavailable	详细原因由 message 返回。
404	not_found	upload task not found
409	capacity_exceeded	model repository capacity would be exceeded
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取本地上传任务

功能介绍

获取本地上传任务。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/model-upload-tasks/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
modelArtifactId	string	模型制品 ID
projectId	string	所属项目 ID；空值时可能省略
repositoryId	string	模型仓库 ID
ownerUserId	string	创建者用户 ID；空值时可能省略
rootName	string	上传根目录名称
status	string	资源状态；具体取值见业务规则
failureStage	string	失败阶段；空值时可能省略
totalFiles	integer	文件总数
completedFiles	integer	已完成文件数
totalBytes	int64	总字节数
completedBytes	int64	已完成字节数
lastProgressAt	string(date-time) / null	最近进度更新时间；空值时可能省略
progressPercent	integer	进度百分比
workerId	string	执行工作节点 ID；空值时可能省略

参数	参数类型	描述
retryCount	integer	重试次数
version	int64	并发控制版本
errorMessage	string	错误详情；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339
uploadCompletedAt	string(date-time) / null	文件上传完成时间；空值时可能省略
publishStartedAt	string(date-time) / null	发布开始时间；空值时可能省略
finishedAt	string(date-time) / null	结束时间；空值时可能省略
expiresAt	string(date-time) / null	到期时间；空值时可能省略

请求示例

```
GET https://{endpoint}/api/v1/model-upload-tasks/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "modelArtifactId": "modelArtifactId-example",
  "repositoryId": "repositoryId-example",
  "rootName": "",
  "status": "created",
  "totalFiles": 0,
  "completedFiles": 0,
  "totalBytes": 0,
  "completedBytes": 0,
```

```
"progressPercent": 0,  
"retryCount": 0,  
"version": 1,  
"createdAt": "2026-09-21T00:00:00Z",  
"updatedAt": "2026-09-21T00:00:00Z"  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
404	not_found	upload task not found
403	forbidden	repository is read-only
400	invalid_manifest	详细原因由 message 返回。
409	capacity_exceeded	model repository capacity would be exceeded
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。
500	internal_error	model upload operation failed

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

分页查询上传文件

功能介绍

分页查询上传文件。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI


```
GET /api/v1/model-upload-tasks/{id}/files
```

参数	类型	必选	描述
id	string	是	目标资源 ID

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始
page_size	integer	否	每页数量；默认 200，最多 1000

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<modelartifactupload.File>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始
page_size	integer	每页数量

请求示例

```
GET https://{endpoint}/api/v1/model-upload-tasks/resource-id/files?  
page=1&page_size=20
```

```
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "relativePath": "",
      "artifactPath": "",
      "sizeBytes": 0,
      "uploadedBytes": 0,
      "progressPercent": 0,
      "status": "created",
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:00:00Z"
    }
  ],
  "total": 1,
  "page": 1,
  "page_size": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
404	not_found	upload task not found
403	forbidden	repository is read-only
400	invalid_manifest	详细原因由 message 返回。
409	capacity_exceeded	model repository capacity would be exceeded

HTTP	error	说明
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。
500	internal_error	model upload operation failed

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

暂停上传任务

功能介绍

暂停上传任务。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/model-upload-tasks/{id}/pause
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
----	------	----

参数	参数类型	描述
id	string	资源唯一标识
modelArtifactId	string	模型制品 ID
projectId	string	所属项目 ID；空值时可能省略
repositoryId	string	模型仓库 ID
ownerUserId	string	创建者用户 ID；空值时可能省略
rootName	string	上传根目录名称
status	string	资源状态；具体取值见业务规则
failureStage	string	失败阶段；空值时可能省略
totalFiles	integer	文件总数
completedFiles	integer	已完成文件数
totalBytes	int64	总字节数
completedBytes	int64	已完成字节数
lastProgressAt	string(date-time) / null	最近进度更新时间；空值时可能省略
progressPercent	integer	进度百分比
workerId	string	执行工作节点 ID；空值时可能省略
retryCount	integer	重试次数
version	int64	并发控制版本
errorMessage	string	错误详情；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339
uploadCompletedAt	string(date-time) / null	文件上传完成时间；空值时可能省略
publishStartedAt	string(date-time) / null	发布开始时间；空值时可能省略

参数	参数类型	描述
finishedAt	string(date-time) / null	结束时间；空值时可能省略
expiresAt	string(date-time) / null	到期时间；空值时可能省略

请求示例

```
POST https://{endpoint}/api/v1/model-upload-tasks/resource-id/pause
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "modelArtifactId": "modelArtifactId-example",
  "repositoryId": "repositoryId-example",
  "rootName": "",
  "status": "created",
  "totalFiles": 0,
  "completedFiles": 0,
  "totalBytes": 0,
  "completedBytes": 0,
  "progressPercent": 0,
  "retryCount": 0,
  "version": 1,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
403	forbidden	write permission required
404	not_found	upload task not found
400	invalid_manifest	详细原因由 message 返回。
409	capacity_exceeded	model repository capacity would be exceeded
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。
500	internal_error	model upload operation failed

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

恢复上传任务

功能介绍

恢复上传任务。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/model-upload-tasks/{id}/resume
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
files	array<modelartifactupload.ManifestFile>	是	文件清单

响应消息

参数	参数类型	描述
task	modelartifactupload.Task	模型任务类型或导入任务对象
files	array<modelartifactupload.File>	文件清单
endpoint	string	上传服务地址或服务端点

请求示例

```
POST https://{endpoint}/api/v1/model-upload-tasks/resource-id/resume
x-keystone-token:
Content-Type: application/json
```

```
{
  "files": [
    {
      "relativePath": "model/config.json",
      "sizeBytes": 1024
    }
  ]
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "task": {
    "id": "id-example",
    "modelArtifactId": "modelArtifactId-example",
    "repositoryId": "repositoryId-example",
    "rootName": "",
    "status": "created",
    "totalFiles": 0,
    "completedFiles": 0,
    "totalBytes": 0,
    "completedBytes": 0,
    "progressPercent": 0,
    "retryCount": 0,
    "version": 1,
    "createdAt": "2026-09-21T00:00:00Z",
    "updatedAt": "2026-09-21T00:00:00Z"
  },
  "files": [
    {
      "id": "id-example",
      "relativePath": "",
      "artifactPath": "",
      "sizeBytes": 0,
      "uploadedBytes": 0,
      "progressPercent": 0,
      "status": "created",
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:00:00Z"
    }
  ],
  "endpoint": "https://service.example.com"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
------	-------	----

HTTP	error	说明
400	invalid_manifest	malformed upload manifest
500	internal_error	failed to create resume credentials
503	backend_unavailable	model upload service is not configured
502	backend_unavailable	详细原因由 message 返回。
404	not_found	upload task not found
403	forbidden	repository is read-only
409	capacity_exceeded	model repository capacity would be exceeded
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

提交上传完成

功能介绍

提交上传完成。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/model-upload-tasks/{id}/complete
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
modelArtifactId	string	模型制品 ID
projectId	string	所属项目 ID；空值时可能省略
repositoryId	string	模型仓库 ID
ownerUserId	string	创建者用户 ID；空值时可能省略
rootName	string	上传根目录名称
status	string	资源状态；具体取值见业务规则
failureStage	string	失败阶段；空值时可能省略
totalFiles	integer	文件总数
completedFiles	integer	已完成文件数
totalBytes	int64	总字节数
completedBytes	int64	已完成字节数
lastProgressAt	string(date-time) / null	最近进度更新时间；空值时可能省略
progressPercent	integer	进度百分比
workerId	string	执行工作节点 ID；空值时可能省略
retryCount	integer	重试次数
version	int64	并发控制版本
errorMessage	string	错误详情；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339

参数	参数类型	描述
updatedAt	string(date-time)	更新时间，RFC3339
uploadCompletedAt	string(date-time) / null	文件上传完成时间；空值时可能省略
publishStartedAt	string(date-time) / null	发布开始时间；空值时可能省略
finishedAt	string(date-time) / null	结束时间；空值时可能省略
expiresAt	string(date-time) / null	到期时间；空值时可能省略

请求示例

```
POST https://{endpoint}/api/v1/model-upload-tasks/resource-id/complete
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "modelArtifactId": "modelArtifactId-example",
  "repositoryId": "repositoryId-example",
  "rootName": "",
  "status": "created",
  "totalFiles": 0,
  "completedFiles": 0,
  "totalBytes": 0,
  "completedBytes": 0,
  "progressPercent": 0,
  "retryCount": 0,
  "version": 1,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
502	backend_unavailable	详细原因由 message 返回。
404	not_found	upload task not found
403	forbidden	repository is read-only
400	invalid_manifest	详细原因由 message 返回。
409	capacity_exceeded	model repository capacity would be exceeded
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。
500	internal_error	model upload operation failed

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

重试制品发布

功能介绍

重试制品发布。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/model-upload-tasks/{id}/retry
```

参数	类型	必选	描述
----	----	----	----

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
modelArtifactId	string	模型制品 ID
projectId	string	所属项目 ID；空值时可能省略
repositoryId	string	模型仓库 ID
ownerUserId	string	创建者用户 ID；空值时可能省略
rootName	string	上传根目录名称
status	string	资源状态；具体取值见业务规则
failureStage	string	失败阶段；空值时可能省略
totalFiles	integer	文件总数
completedFiles	integer	已完成文件数
totalBytes	int64	总字节数
completedBytes	int64	已完成字节数
lastProgressAt	string(date-time) / null	最近进度更新时间；空值时可能省略
progressPercent	integer	进度百分比
workerId	string	执行工作节点 ID；空值时可能省略

参数	参数类型	描述
retryCount	integer	重试次数
version	int64	并发控制版本
errorMessage	string	错误详情；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339
uploadCompletedAt	string(date-time) / null	文件上传完成时间；空值时可能省略
publishStartedAt	string(date-time) / null	发布开始时间；空值时可能省略
finishedAt	string(date-time) / null	结束时间；空值时可能省略
expiresAt	string(date-time) / null	到期时间；空值时可能省略

请求示例

```
POST https://{endpoint}/api/v1/model-upload-tasks/resource-id/retry
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "modelArtifactId": "modelArtifactId-example",
  "repositoryId": "repositoryId-example",
  "rootName": "",
  "status": "created",
  "totalFiles": 0,
  "completedFiles": 0,
  "totalBytes": 0,
  "completedBytes": 0,
```

```
"progressPercent": 0,  
"retryCount": 0,  
"version": 1,  
"createdAt": "2026-09-21T00:00:00Z",  
"updatedAt": "2026-09-21T00:00:00Z"  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
403	forbidden	write permission required
502	backend_unavailable	详细原因由 message 返回。
404	not_found	upload task not found
400	invalid_manifest	详细原因由 message 返回。
409	capacity_exceeded	model repository capacity would be exceeded
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。
500	internal_error	model upload operation failed

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

取消本地上传任务

功能介绍

取消本地上传任务。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
DELETE /api/v1/model-upload-tasks/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/model-upload-tasks/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
403	forbidden	write permission required
404	not_found	upload task not found
400	invalid_manifest	详细原因由 message 返回。

HTTP	error	说明
409	capacity_exceeded	model repository capacity would be exceeded
409	upload_not_resumable	详细原因由 message 返回。
409	upload_incomplete	not all files are uploaded
409	upload_task_conflict	详细原因由 message 返回。
500	internal_error	model upload operation failed

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.7 容器镜像

容器镜像

镜像列表要求 Harbor 配置启用；按当前租户及 public 筛选项目仓库。architecture、runtime、vendor 为可选镜像筛选条件。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/images/repositories	查询镜像仓库
GET	/images/tags	查询镜像标签

查询镜像仓库

功能介绍

查询镜像仓库。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/images/repositories
```

查询参数

参数	类型	必选	描述
public	string	否	是否为公共仓库
architecture	string	否	加速器架构
runtime	string	否	推理运行时
vendor	string	否	筛选参数 vendor

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<harbor.imageEntry>	结果列表
total	integer	匹配总数

请求示例

```
GET https://{endpoint}/api/v1/images/repositories
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "name": "示例资源",
      "harborProjectName": "",
      "public": true,
      "tagsCount": 0,
      "artifactCount": 0,
      "pullCount": 0,
      "description": "",
      "creationTime": "2026-09-21T00:00:00Z"
    }
  ],
}
```

```
"total": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
400	invalid_request	详细原因由 message 返回。
404	not_found	harbor resource not found
502	backend_unavailable	详细原因由 message 返回。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询镜像标签

功能介绍

查询镜像标签。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/images/tags
```

查询参数

参数	类型	必选	描述
project	string	是	Harbor 项目名称

参数	类型	必选	描述
repository	string	是	镜像仓库名称
architecture	string	否	加速器架构
runtime	string	否	推理运行时
vendor	string	否	筛选参数 vendor

请求消息

无请求体。

响应消息

参数	参数类型	描述
tags	array	镜像标签或模型标签列表

请求示例

```
GET https://{endpoint}/api/v1/images/tags?project=project-  
name&repository=repo-name  
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200  
Content-Type: application/json
```

```
{  
  "tags": [  
    ""  
  ]  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
400	invalid_request	project and repository are required
404	not_found	harbor resource not found
502	backend_unavailable	详细原因由 message 返回。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.8 推理引擎

推理引擎

资源未配置 Kubernetes 时可仅维护目录；配置 Kubernetes 时会创建或更新 KServe 引擎对象。201 并不保证运行环境已经就绪。

已被部署引用的引擎不允许修改受保护配置或删除；使用 inUse 判断引用状态，并处理 409 engine_in_use。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/engines	查询推理引擎
POST	/engines	创建推理引擎
GET	/engines/{id}	获取推理引擎详情
PATCH	/engines/{id}	更新推理引擎
DELETE	/engines/{id}	删除推理引擎

查询推理引擎

功能介绍

查询推理引擎。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/engines
```

查询参数

参数	类型	必选	描述
----	----	----	----

参数	类型	必选	描述
page	integer	否	页码，从 1 开始
page_size	integer	否	每页数量；建议显式传入，例如 20；省略时部分列表返回全部

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<engine.response>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始
page_size	integer	每页数量

请求示例

```
GET https://{endpoint}/api/v1/engines?page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
```



```
{
  "name": "示例资源",
  "description": "",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {},
  "inUse": true,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
],
"total": 1,
"page": 1,
"page_size": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to list engines
401	unauthorized	authentication required

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建推理引擎

功能介绍

创建推理引擎。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/engines
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	是	显示名称
description	string	否	描述
image	string	是	容器镜像完整地址，应包含仓库路径及标签。
engineImageId	string	否	镜像资源 ID；使用完整 image 地址创建时不需要提供。
architecture	string	是	运行架构，amd64 或 arm64；创建流程默认选择 amd64。
runtime	string	是	推理运行时，须选择有效类型。
command	array	否	容器启动命令数组
args	array	否	容器启动参数数组
env	array<engine.EnvVar>	否	环境变量数组
resources	engine.Resources / null	是	资源配置，须包含 cpu 和 memory；选择加速器时还需提供对应 GPU 配置。
nodeSelector	map<string, string>	否	节点选择标签
yaml	string	否	自定义 YAML 配置

参数	参数类型	是否必选	描述
labels	map<string, string>	是	须包含 visibility，取值 private 或 public。

参数	参数类型	是否必选	描述
resources.cpu	string	是	CPU 数量，不能为空。
resources.memory	string	是	内存容量，包含单位，例如 16Gi。
resources.gpu	string	条件必选	选择 GPU 产品时必选，为大于等于 1 的整数数量。
resources.gpuProduct	string	条件必选	使用 GPU 时指定 GPU 产品。
resources.gpuResource	string	条件必选	所选 GPU 产品提供资源名称时，需携带该资源名称。

上述资源子字段以创建或完整替换资源配置为前提。镜像仓库与镜像标签在提交时组合为 image，不是独立的请求参数。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略

参数	参数类型	描述
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	engine.backendRef	后端资源标识
spec	engine.Spec	资源配置快照
labels	map<string, string>	标签键值对；空值时可能省略
inUse	boolean	是否正被其他资源引用
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
POST https://{endpoint}/api/v1/engines
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "vllm-engine",
  "image": "registry.example.com/ai/vllm:approved",
  "runtime": "vllm",
  "resources": {
    "cpu": "4",
    "memory": "16Gi"
  },
  "architecture": "amd64",
  "labels": {
    "visibility": "private"
  }
}
```

正常响应示例

```
HTTP/1.1 201
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {},
  "inUse": true,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
400	invalid_request	malformed JSON body
403	forbidden	public engine is read-only for current tenant
500	internal_error	failed to encode engine spec
401	unauthorized	missing project-scoped token; re-authenticate with a project scope
404	not_found	详细原因由 message 返回。
409	conflict	an engine with that name already exists in the namespace

HTTP	error	说明
422	invalid_request	详细原因由 message 返回。
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取推理引擎详情

功能介绍

获取推理引擎详情。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/engines/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称

参数	参数类型	描述
description	string	描述
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	engine.backendRef	后端资源标识
spec	engine.Spec	资源配置快照
labels	map<string, string>	标签键值对；空值时可能省略
inUse	boolean	是否正被其他资源引用
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/engines/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
```

```
"description": "",
"projectId": "projectId-example",
"ownerUserId": "ownerUserId-example",
"backend": {
  "catalog": "",
  "type": "",
  "id": "id-example"
},
"spec": {},
"inUse": true,
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
404	not_found	engine not found
409	conflict	engine already exists
403	forbidden	public engine is read-only for current tenant
400	invalid_request	engine visibility cannot be changed; create a new engine instead
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

更新推理引擎

功能介绍

更新推理引擎。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
PATCH /api/v1/engines/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string / null	否	显示名称
description	string / null	否	描述
status	string / null	否	资源状态；具体取值见业务规则
image	string / null	否	修改时须提供包含仓库路径和标签的完整镜像地址。
architecture	string / null	否	加速器架构
runtime	string / null	否	修改运行时提供，不能为空。
command	array / null	否	容器启动命令数组
args	array / null	否	容器启动参数数组
env	array<engine.EnvVar> / null	否	环境变量数组

参数	参数类型	是否必选	描述
resources	engine.Resources / null	否	省略时不修改；替换资源配置时须提供 cpu、memory, GPU 子字段要求与创建一致。
yaml	string / null	否	自定义 YAML 配置
labels	map<string, string> / null	否	省略时不修改；visibility 不允许变更，若提供须与原值一致。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	engine.backendRef	后端资源标识
spec	engine.Spec	资源配置快照
labels	map<string, string>	标签键值对；空值时可能省略
inUse	boolean	是否正被其他资源引用
createdAt	string(date-time)	创建时间，RFC3339

参数	参数类型	描述
updatedAt	string(date-time)	更新时间, RFC3339

请求示例

```
PATCH https://{endpoint}/api/v1/engines/resource-id
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "更新后的名称"
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {},
  "inUse": true,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	malformed JSON body
500	internal_error	failed to resolve engine references
409	engine_in_use	engine is referenced by a deployment and cannot be modified; delete the referencing deployment(s) first
401	unauthorized	missing project-scoped token; re-authenticate with a project scope
422	invalid_request	engine has no valid backend reference
403	forbidden	write permission required
404	not_found	engine not found
409	conflict	engine already exists
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除推理引擎

功能介绍

删除推理引擎。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
DELETE /api/v1/engines/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/engines/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
500	internal_error	failed to resolve engine references
409	engine_in_use	engine is referenced by a deployment and cannot be deleted; delete the referencing deployment(s) first
401	unauthorized	missing project-scoped token; re-authenticate with a project scope
403	forbidden	write permission required
404	not_found	engine not found

HTTP	error	说明
409	conflict	engine already exists
400	invalid_request	engine visibility cannot be changed; create a new engine instead
422	invalid_request	详细原因由 message 返回。
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.9 模型部署

模型部署

模型部署将推理引擎与模型文件组合为可运行的推理服务。创建、更新、启动和停止接口在提交工作负载变更后即返回，实际状态由 KServe 异步收敛；调用方应通过详情接口中的 `status`、`ready`、`stopped` 和 `conditions` 判断部署是否最终可用。

部署是项目级资源。普通调用方只能访问当前项目的部署，具有全局读取权限的管理员可以查看所有项目的部署。创建部署时可以引用当前项目或公共的推理引擎、模型和模型文件，但这些关联资源的可见性必须一致。

创建部署至少需要 `name` 和 `engineId`，并且必须通过 `artifactId` 或 `modelUri` 解析出模型 OCI URI。`replicas` 省略或小于 1 时按 1 处理。更新接口不能替换已经部署的 `modelId`、`artifactId` 或 `modelUri`；如需更换模型文件，应创建新的部署。

部署响应中的 `spec.endpoint` 是集群内访问地址，仅在 KServe 已报告工作负载 Service 后返回，不能将其视为面向公网的推理地址。日志接口返回纯文本，`follow=true` 或 `follow=1` 时持续输出；终端接口使用 WebSocket。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/deployments	查询模型部署
POST	/deployments	创建模型部署
GET	/deployments/api-schemas	查询部署支持的 API 协议
GET	/deployments/{id}	获取模型部署详情
PATCH	/deployments/{id}	更新模型部署配置
POST	/deployments/{id}/stop	停止部署
POST	/deployments/{id}/start	启动部署
DELETE	/deployments/{id}	删除模型部署
GET	/deployments/{id}/pods	查询部署 Pod

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/deployments/{id}/logs	读取或持续跟踪容器日志
GET	/deployments/{id}/metrics	查询部署监控指标
GET	/deployments/{id}/exec	建立 WebSocket 终端会话

查询模型部署

功能介绍

分页查询调用方有权查看的模型部署。普通调用方返回当前项目的部署；具有全局读取权限的管理员可以返回所有项目的部署。

列表响应提供部署的目录状态和配置快照，不读取实时 `ready` 和 `conditions`。需要判断工作负载的实时状态时，应调用部署详情接口。

访问要求：有效 Token；需要模型部署读取权限。

URI

```
GET /api/v1/deployments
```

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始；省略或小于 1 时按 1 返回
page_size	integer	否	每页数量，最大 200；省略或传 0 时返回全部匹配项，响应中的 <code>page_size</code> 为 0

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array	部署列表；对象字段与部署详情基本一致，但不包含实时 <code>ready</code> 和 <code>conditions</code>
total	integer	匹配条件的部署总数
page	integer	当前页码
page_size	integer	请求的每页数量；未限制时为 0

请求示例

```
GET https://{endpoint}/api/v1/deployments?page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "deployment-7f3a",
      "name": "qwen-chat",
      "description": "Qwen 对话模型",
      "status": "ready",
      "projectId": "project-a",
      "ownerUserId": "user-a",
      "backend": {
        "catalog": "k8s",
        "type": "LLMInferenceService",
        "id": "project-a/qwen-chat"
      }
    }
  ]
}
```

```
    },
    "spec": {
      "architecture": "amd64",
      "k8sName": "qwen-chat",
      "engineId": "engine-vllm",
      "engineName": "vLLM",
      "engineConfigName": "vllm-config",
      "modelId": "model-qwen",
      "artifactId": "artifact-qwen",
      "artifactName": "Qwen2.5-7B-Instruct",
      "modelUri":
"oci://registry.example.com/models/qwen@sha256:0123456789abcdef",
      "modelName": "qwen2.5-7b-instruct",
      "replicas": 1,
      "endpoint": "http://qwen-chat-kserve-workload-svc.project-
a.svc.cluster.local:8000"
    },
    "stopped": false,
    "labels": {
      "beacon.io/runtime": "vllm"
    },
    "createdAt": "2026-09-21T00:00:00Z",
    "updatedAt": "2026-09-21T00:05:00Z"
  }
],
"total": 1,
"page": 1,
"page_size": 20
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型部署读取权限

HTTP	error	说明
500	internal_error	查询部署列表失败

错误返回格式见[调用方式](#)。

创建模型部署

功能介绍

在当前项目中创建模型部署，并提交对应的 KServe 工作负载。接口成功仅表示部署资源已经创建，初始 `status` 为 `pending`；应继续查询详情，直到 `status` 变为 `ready`，或者根据 `conditions` 处理异常。

访问要求：有效 Token；需要模型部署写入权限。平台已配置 Kubernetes 时，还需要项目作用域令牌。

URI

```
POST /api/v1/deployments
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	是	部署显示名称；去除首尾空白后不能为空，同时用于生成 Kubernetes 工作负载名称
description	string	否	部署说明
engineId	string	是	推理引擎 ID；可以引用当前项目或公共引擎
modelId	string	否	来源模型 ID，用于记录模型关联关系

参数	参数类型	是否必选	描述
artifactId	string	条件必选	模型文件 ID；与 <code>modelUri</code> 至少提供一种。提供后优先从该模型文件解析 OCI URI
modelUri	string	条件必选	模型 OCI URI；未提供可解析的 <code>artifactId</code> 时必选。省略 <code>oci://</code> 前缀时服务端会自动补全
modelName	string	否	推理服务对外标识的模型名称；引擎引用 <code>SERVED_MODEL_NAME</code> 时建议提供
replicas	integer	否	期望副本数；省略或小于 1 时按 1 处理
image	string	否	覆盖引擎默认值的容器镜像
command	array	否	覆盖引擎默认值的容器启动命令
args	array	否	覆盖引擎默认值的容器启动参数
env	array	否	覆盖或补充的容器环境变量
resources	object / null	否	覆盖引擎默认值的计算资源配置
kvCacheOffloading	object / null	否	KV Cache CPU 卸载配置，仅适用于使用加速器资源的 vLLM 引擎
labels	object	否	部署标签，键和值均为字符串；系统管理的标签可能被补充或覆盖
lora	object / null	否	LoRA 适配器配置，仅适用于 vLLM 引擎

参数	参数类型	是否必选	描述
env[].name	string	是	环境变量名称
env[].value	string	是	环境变量值

参数	参数类型	是否必选	描述
resources.cpu	string	否	Kubernetes CPU quantity, 例如 4 或 500m
resources.memory	string	否	Kubernetes 内存 quantity, 例如 16 Gi
resources.gpu	string	否	加速器数量; 非空值会覆盖引擎中的加速器配置
kvCacheOffloading.cpu	string	条件必选	用于卸载 KV Cache 的正数内存 quantity, 例如 8Gi
kvCacheOffloading.evictionPolicy	string	否	淘汰策略: lru 或 arc ; 省略时使用 lru
lora.adapters	array	否	LoRA 适配器列表; 省略或传空数组表示不加载适配器
lora.adapters[].name	string	是	适配器名称; 同一部署内必须唯一
lora.adapters[].modelId	string	否	适配器来源模型 ID
lora.adapters[].artifactId	string	是	LoRA 模型文件 ID; 文件必须已就绪、带有 LoRA 标识并使用摘要固定的 OCI URI

当同时提供 `engineId`、`modelId`、`artifactId` 或 LoRA 模型文件时, 这些资源必须全部为公共资源, 或全部为当前项目的私有资源。`kvCacheOffloading` 不能与命令、参数、环境变量或引擎 YAML 中手工配置的 `--kv-transfer-config` 同时使用。

响应消息

参数	参数类型	描述
----	------	----

参数	参数类型	描述
id	string	部署唯一标识
name	string	部署显示名称
description	string	部署说明
status	string	部署状态；创建成功时为 <code>pending</code>
projectId	string	所属项目 ID
projectName	string	项目名称；空值时省略
domainId	string	所属域 ID；空值时省略
domainName	string	所属域名称；空值时省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时省略
backend	object	后端工作负载引用，包含 <code>catalog</code> 、 <code>type</code> 和 <code>id</code>
spec	object	部署配置快照；字段与请求配置及服务端解析出的引擎、模型文件信息对应
stopped	boolean	是否已完全停止；创建成功时为 <code>false</code>
labels	object	部署标签；空值时省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

创建响应不包含实时 `ready` 和 `conditions` 。这些字段需要通过详情接口读取。

请求示例

```
POST https://{endpoint}/api/v1/deployments
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "qwen-chat",
  "description": "Qwen 对话模型",
  "engineId": "engine-vllm",
  "modelId": "model-qwen",
  "artifactId": "artifact-qwen",
  "modelName": "qwen2.5-7b-instruct",
  "replicas": 1
}
```

正常响应示例

```
HTTP/1.1 201 Created
Content-Type: application/json
```

```
{
  "id": "deployment-7f3a",
  "name": "qwen-chat",
  "description": "Qwen 对话模型",
  "status": "pending",
  "projectId": "project-a",
  "ownerUserId": "user-a",
  "backend": {
    "catalog": "k8s",
    "type": "LLMInferenceService",
    "id": "project-a/qwen-chat"
  },
  "spec": {
    "architecture": "amd64",
    "k8sName": "qwen-chat",
    "engineId": "engine-vllm",
    "engineName": "vLLM",
    "engineConfigName": "vllm-config",
    "modelId": "model-qwen",

```

```
"artifactId": "artifact-qwen",
"artifactName": "Qwen2.5-7B-Instruct",
"modelUri":
"oci://registry.example.com/models/qwen@sha256:0123456789abcdef",
"modelName": "qwen2.5-7b-instruct",
"replicas": 1
},
"stopped": false,
"labels": {
  "beacon.io/runtime": "vllm"
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
400	invalid_request	请求体格式错误、必填字段缺失、关联资源不存在或部署配置不符合要求；具体原因见 <code>message</code>
401	unauthorized	未认证，或缺少项目作用域令牌
403	forbidden	无模型部署写入权限，或 Kubernetes 拒绝操作
404	not_found	目标 Kubernetes 命名空间或对象不存在
409	conflict	同名工作负载或部署资源已存在
422	invalid_request	引擎后端引用无效，或 Kubernetes 拒绝工作负载配置
502	backend_unavailable	Kubernetes API 不可用
500	internal_error	创建部署失败

错误返回格式见[调用方式](#)。

查询部署支持的 API 协议

功能介绍

一次查询调用方可见的所有部署分别支持哪些推理 API 协议，可用于创建 AI 服务提供者时筛选部署和选择协议。当前协议值为 `OpenAI` 或 `Cohere`。

服务端根据部署工作负载实际暴露的接口确定协议。如果工作负载尚未就绪、已停止或无法访问，则返回兼容性默认值 `OpenAI`，并将 `unreachable` 标记为 `true`。因此该结果用于辅助配置，不代表部署当前已经可以处理推理请求。

访问要求：有效 Token；需要模型部署读取权限。

URI

```
GET /api/v1/deployments/api-schemas
```

请求消息

无请求体。

响应消息

参数	类型	描述
items	object	以部署 ID 为键的协议查询结果
items.{deploymentId}.schemas	array	部署支持的协议列表，取值为 <code>OpenAI</code> 、 <code>Cohere</code>
items.{deploymentId}.recommended	string	推荐使用的协议；支持 OpenAI 时优先推荐 <code>OpenAI</code>
items.{deploymentId}.unreachable	boolean	是否因工作负载不可访问而使用默认结果；值为 <code>false</code> 时省略

请求示例

```
GET https://{endpoint}/api/v1/deployments/api-schemas
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "items": {
    "deployment-7f3a": {
      "schemas": [
        "OpenAI"
      ],
      "recommended": "OpenAI"
    },
    "deployment-8b2c": {
      "schemas": [
        "OpenAI"
      ],
      "recommended": "OpenAI",
      "unreachable": true
    }
  }
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型部署读取权限
500	internal_error	查询部署列表失败

错误返回格式见[调用方式](#)。

获取模型部署详情

功能介绍

获取指定部署的配置及实时运行状态。服务端会读取 KServe 工作负载，更新 `ready`、`conditions`、`status` 和集群内访问地址；工作负载暂时无法读取时，仍可能返回目录中保存的部署信息。

访问要求：有效 Token；需要模型部署读取权限。普通调用方只能读取当前项目的部署，具有全局读取权限的管理员可以读取其他项目的部署。

URI

```
GET /api/v1/deployments/{id}
```

参数	类型	必选	描述
id	string	是	部署 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	部署唯一标识
name	string	部署显示名称
description	string	部署说明
status	string	部署状态： <code>pending</code> 、 <code>ready</code> 、 <code>stopping</code> 或 <code>stoppe d</code>

参数	参数类型	描述
projectId	string	所属项目 ID
projectName	string	项目名称；空值时省略
domainId	string	所属域 ID；空值时省略
domainName	string	所属域名称；空值时省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时省略
backend	object	后端工作负载引用，包含 <code>catalog</code> 、 <code>type</code> 和 <code>id</code>
spec	object	部署配置快照，主要字段见下表
ready	string	KServe Ready 条件值： <code>True</code> 、 <code>False</code> 或 <code>Unknown</code> ；无法读取时可能省略
stopped	boolean	工作负载是否已完全停止； <code>stopping</code> 阶段仍为 <code>false</code>
conditions	array	KServe 状态条件；无条件信息时省略
labels	object	部署标签；空值时省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

参数	参数类型	描述
spec.architecture	string	工作负载架构，当前为 <code>amd64</code> 或 <code>arm64</code>
spec.k8sName	string	Kubernetes 工作负载名称
spec.engineId	string	推理引擎 ID
spec.engineName	string	推理引擎名称

参数	参数类型	描述
spec.engineConfigName	string	引擎对应的 KServe 配置名称
spec.modelId	string	来源模型 ID；未关联时省略
spec.artifactId	string	模型文件 ID；未关联时省略
spec.artifactName	string	模型文件名称；未关联时省略
spec.modelUri	string	实际部署的模型 OCI URI
spec.modelName	string	推理服务使用的模型名称；空值时省略
spec.lora	object	已解析的 LoRA 适配器配置；未配置时省略
spec.lora.adapters[].name	string	LoRA 适配器名称
spec.lora.adapters[].modelId	string	适配器来源模型 ID；未关联时省略
spec.lora.adapters[].artifactId	string	LoRA 模型文件 ID
spec.lora.adapters[].uri	string	服务端从模型文件解析出的、使用摘要固定的 OCI URI
spec.replicas	integer	期望副本数
spec.image	string	部署级镜像覆盖；未配置时省略
spec.command	array	部署级启动命令覆盖；未配置时省略
spec.args	array	部署级启动参数覆盖；未配置时省略
spec.env	array	部署级环境变量；未配置时省略
spec.resources	object	部署级计算资源覆盖；未配置时省略
spec.kvCacheOffloading	object	KV Cache CPU 卸载配置；未配置时省略
spec.endpoint	string	集群内工作负载地址；尚未就绪或已停止时省略
conditions[].type	string	条件类型

参数	参数类型	描述
conditions[].status	string	条件状态
conditions[].reason	string	条件原因；空值时省略
conditions[].message	string	条件说明；空值时省略
conditions[].lastTransitionTime	string(date-time)	最近转换时间；空值时省略

请求示例

```
GET https://{endpoint}/api/v1/deployments/deployment-7f3a
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "deployment-7f3a",
  "name": "qwen-chat",
  "description": "Qwen 对话模型",
  "status": "ready",
  "projectId": "project-a",
  "ownerUserId": "user-a",
  "backend": {
    "catalog": "k8s",
    "type": "LLMInferenceService",
    "id": "project-a/qwen-chat"
  },
  "spec": {
    "architecture": "amd64",
    "k8sName": "qwen-chat",
    "engineId": "engine-vllm",
    "engineName": "vLLM",
```

```
"engineConfigName": "vllm-config",
"modelId": "model-qwen",
"artifactId": "artifact-qwen",
"artifactName": "Qwen2.5-7B-Instruct",
"modelUri":
"oci://registry.example.com/models/qwen@sha256:0123456789abcdef",
"modelName": "qwen2.5-7b-instruct",
"replicas": 1,
"endpoint": "http://qwen-chat-kserve-workload-svc.project-
a.svc.cluster.local:8000"
},
"ready": "True",
"stopped": false,
"conditions": [
{
"type": "Ready",
"status": "True",
"lastTransitionTime": "2026-09-21T00:05:00Z"
}
],
"labels": {
"beacon.io/runtime": "vllm"
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:05:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型部署读取权限
404	not_found	部署不存在或无权访问该项目的部署
500	internal_error	读取部署失败

错误返回格式见[调用方式](#)。

更新模型部署配置

功能介绍

按字段更新部署。省略的字段保持不变；修改运行配置后，KServe 会异步更新工作负载，响应中的 `status` 变为 `pending`。仅修改 `name` 或 `description` 时不会重命名或重启已有 Kubernetes 工作负载。

此接口不能替换 `modelId`、`artifactId`、`modelUri`、`resources` 或 `labels`。如需更换模型文件，应创建新的部署。

访问要求：有效 Token；需要当前项目的模型部署写入权限。修改运行配置且平台已配置 Kubernetes 时，还需要项目作用域令牌。

URI

```
PATCH /api/v1/deployments/{id}
```

参数	类型	必选	描述
id	string	是	部署 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	否	新的显示名称；去除首尾空白后不能为空，不会修改 Kubernetes 工作负载名称
description	string	否	新的部署说明
engineId	string	否	新的推理引擎 ID；必须与已部署的模型文件具有相同可见性

参数	参数类型	是否必选	描述
modelName	string	否	新的服务模型名称；空字符串表示清空
replicas	integer	否	新的期望副本数，必须大于等于 1
image	string	否	新的容器镜像覆盖；空字符串表示恢复使用引擎默认值
command	array	否	新的启动命令覆盖；空数组表示清除覆盖
args	array	否	新的启动参数覆盖；空数组表示清除覆盖
env	array	否	新的环境变量数组；空数组表示清除部署级环境变量
kvCacheOffloading	object / null	否	新的 KV Cache CPU 卸载配置； <code>null</code> 表示关闭，省略表示不修改
lora	object / null	否	新的 LoRA 配置； <code>null</code> 或空 <code>adapters</code> 数组表示清除，省略表示不修改

`env`、`kvCacheOffloading` 和 `lora` 的子字段及约束与创建接口相同。切换到非 vLLM 引擎前，必须先移除已配置的 LoRA 适配器。

响应消息

返回更新后的部署对象，字段与“获取模型部署详情”相同，但该响应不读取实时工作负载，因此通常不包含 `ready` 和 `conditions`。运行配置发生变化时，`status` 为 `pending`。

请求示例

```
PATCH https://{endpoint}/api/v1/deployments/deployment-7f3a
x-keystone-token:
Content-Type: application/json
```

```
{
  "replicas": 2,
```

```
"modelName": "qwen-chat-v2"  
}
```

正常响应示例

```
HTTP/1.1 200 OK  
Content-Type: application/json
```

```
{  
  "id": "deployment-7f3a",  
  "name": "qwen-chat",  
  "description": "Qwen 对话模型",  
  "status": "pending",  
  "projectId": "project-a",  
  "ownerUserId": "user-a",  
  "backend": {  
    "catalog": "k8s",  
    "type": "LLMInferenceService",  
    "id": "project-a/qwen-chat"  
  },  
  "spec": {  
    "architecture": "amd64",  
    "k8sName": "qwen-chat",  
    "engineId": "engine-vllm",  
    "engineName": "vLLM",  
    "engineConfigName": "vllm-config",  
    "modelId": "model-qwen",  
    "artifactId": "artifact-qwen",  
    "artifactName": "Qwen2.5-7B-Instruct",  
    "modelUri":  
"oci://registry.example.com/models/qwen@sha256:0123456789abcdef",  
    "modelName": "qwen-chat-v2",  
    "replicas": 2  
  },  
  "stopped": false,  
  "labels": {  
    "beacon.io/runtime": "vllm"  
  },  
  "createdAt": "2026-09-21T00:00:00Z",  
}
```

```
"updatedAt": "2026-09-21T01:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	请求体格式错误、字段取值无效或关联资源不符合要求；具体原因见 <code>message</code>
401	unauthorized	未认证，或修改工作负载时缺少项目作用域令牌
403	forbidden	无模型部署写入权限，或 Kubernetes 拒绝操作
404	not_found	部署、引擎或目标 Kubernetes 对象不存在
409	conflict	更新与当前资源状态冲突
422	invalid_request	后端引用无效，或 Kubernetes 拒绝工作负载配置
502	backend_unavailable	Kubernetes API 不可用
500	internal_error	更新部署失败

错误返回格式见[调用方式](#)。

停止部署

功能介绍

请求 KServe 停止指定部署。服务端保留部署配置和期望副本数，以便后续按原配置启动。

停止是异步过程。请求后如果 Pod 尚未全部退出，响应中的 `status` 为 `stopping` 且 `stopped` 为 `false`；Pod 全部退出后，详情接口返回 `status: stopped` 和 `stopped: true`。

访问要求：有效 Token 和项目作用域令牌；需要当前项目的模型部署写入权限。

URI

```
POST /api/v1/deployments/{id}/stop
```

参数	类型	必选	描述
id	string	是	部署 ID

请求消息

无请求体。

响应消息

返回已提交停止请求的部署对象，字段与“获取模型部署详情”相同，但不包含实时 `ready` 和 `conditions`。

请求示例

```
POST https://{endpoint}/api/v1/deployments/deployment-7f3a/stop
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "deployment-7f3a",
  "name": "qwen-chat",
  "description": "Qwen 对话模型",
  "status": "stopping",
  "projectId": "project-a",
  "ownerUserId": "user-a",
  "backend": {
    "catalog": "k8s",
    "type": "LLMInferenceService",
```

```
"id": "project-a/qwen-chat"
},
"spec": {
  "architecture": "amd64",
  "k8sName": "qwen-chat",
  "engineId": "engine-vllm",
  "engineName": "vLLM",
  "modelUri":
"oci://registry.example.com/models/qwen@sha256:0123456789abcdef",
  "replicas": 1,
  "endpoint": "http://qwen-chat-kserve-workload-svc.project-
a.svc.cluster.local:8000"
},
"stopped": false,
"labels": {
  "beacon.io/runtime": "vllm"
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T02:00:00Z"
}
```

如果停止请求提交时工作负载 Pod 已经全部退出，接口可直接返回 `status: stopped` 和 `stopped: true`。

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或缺少项目作用域令牌
403	forbidden	无模型部署写入权限，或 Kubernetes 拒绝操作
404	not_found	部署或目标 Kubernetes 对象不存在
409	conflict	Kubernetes 更新发生冲突
422	invalid_backend / invalid_request	部署没有有效工作负载，或 Kubernetes 拒绝请求

HTTP	error	说明
502	backend_unavailable	Kubernetes API 不可用
503	unavailable	Kubernetes 未配置
500	internal_error	停止部署失败

错误返回格式见[调用方式](#)。

启动部署

功能介绍

移除部署的停止标记，请求 KServe 按原有配置重新创建工作负载。接口返回时 `status` 为 `pending` 、`stopped` 为 `false` ，实际就绪仍需通过详情接口确认。

访问要求：有效 Token 和项目作用域令牌；需要当前项目的模型部署写入权限。

URI

```
POST /api/v1/deployments/{id}/start
```

参数	类型	必选	描述
id	string	是	部署 ID

请求消息

无请求体。

响应消息

返回已提交启动请求的部署对象，字段与“获取模型部署详情”相同，但不包含实时 `ready` 和 `conditions` 。

请求示例

```
POST https://{endpoint}/api/v1/deployments/deployment-7f3a/start
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "id": "deployment-7f3a",
  "name": "qwen-chat",
  "description": "Qwen 对话模型",
  "status": "pending",
  "projectId": "project-a",
  "ownerUserId": "user-a",
  "backend": {
    "catalog": "k8s",
    "type": "LLMInferenceService",
    "id": "project-a/qwen-chat"
  },
  "spec": {
    "architecture": "amd64",
    "k8sName": "qwen-chat",
    "engineId": "engine-vllm",
    "engineName": "vLLM",
    "modelUri":
"oci://registry.example.com/models/qwen@sha256:0123456789abcdef",
    "replicas": 1
  },
  "stopped": false,
  "labels": {
    "beacon.io/runtime": "vllm"
  },
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T03:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或缺少项目作用域令牌
403	forbidden	无模型部署写入权限，或 Kubernetes 拒绝操作
404	not_found	部署或目标 Kubernetes 对象不存在
409	conflict	Kubernetes 更新发生冲突
422	invalid_backend / invalid_request	部署没有有效工作负载，或 Kubernetes 拒绝请求
502	backend_unavailable	Kubernetes API 不可用
503	unavailable	Kubernetes 未配置
500	internal_error	启动部署失败

错误返回格式见[调用方式](#)。

删除模型部署

功能介绍

删除当前项目中的模型部署。配置 Kubernetes 时，服务端先删除对应的 KServe 工作负载，再将部署记录标记为已删除。该操作不会删除关联的推理引擎、模型或模型文件。

访问要求：有效 Token；需要当前项目的模型部署写入权限。配置 Kubernetes 时还需要项目作用域令牌。

URI

```
DELETE /api/v1/deployments/{id}
```

参数	类型	必选	描述
----	----	----	----

参数	类型	必选	描述
id	string	是	部署 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/deployments/deployment-7f3a
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
401	unauthorized	未认证，或配置 Kubernetes 时缺少项目作用域令牌
403	forbidden	无模型部署写入权限，或 Kubernetes 拒绝操作
404	not_found	部署或目标 Kubernetes 对象不存在
409	conflict	Kubernetes 删除操作发生冲突
422	invalid_request	Kubernetes 拒绝删除请求
502	backend_unavailable	Kubernetes API 不可用

HTTP	error	说明
500	internal_error	删除部署失败

错误返回格式见[调用方式](#)。

查询部署 Pod

功能介绍

查询指定部署当前的工作负载 Pod 及其容器状态，可用于选择日志或终端会话的目标 Pod 和容器。已成功结束的历史 Pod 不在结果中。

访问要求：有效 Token；需要模型部署读取权限。

URI

```
GET /api/v1/deployments/{id}/pods
```

参数	类型	必选	描述
id	string	是	部署 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array	当前工作负载 Pod 列表，按 Pod 名称排序
items[].name	string	Pod 名称

参数	参数类型	描述
items[].phase	string	Kubernetes Pod phase, 例如 <code>Pending</code> 、 <code>Running</code> 或 <code>Failed</code>
items[].status	string	便于展示的详细状态, 例如 <code>Running</code> 、 <code>PodInitializing</code> 或 <code>CrashLoopBackOff</code>
items[].nodeName	string	所在节点名称; 未调度时省略
items[].startTime	string(date-time)	Pod 启动时间; 尚未启动时省略
items[].containers	array	容器列表
items[].containers[].name	string	容器名称
items[].containers[].image	string	容器镜像; 空值时省略
items[].containers[].imageId	string	已运行镜像 ID; 空值时省略
items[].containers[].ready	boolean	容器是否就绪
items[].containers[].restartCount	integer	容器重启次数
items[].containers[].state	string	<code>running</code> 、 <code>waiting</code> 或 <code>terminated</code> ; 状态未知时省略

请求示例

```
GET https://{endpoint}/api/v1/deployments/deployment-7f3a/pods
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "items": [
    {
      "name": "qwen-chat-kserve-6d7c9f8b6d-abcde",
      "phase": "Running",
      "status": "Running",
      "nodeName": "worker-01",
      "startTime": "2026-09-21T00:04:00Z",
      "containers": [
        {
          "name": "main",
          "image": "registry.example.com/vllm:v1",
          "ready": true,
          "restartCount": 0,
          "state": "running"
        }
      ]
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型部署读取权限，或 Kubernetes 拒绝操作
404	not_found	部署或目标 Kubernetes 对象不存在
422	invalid_backend / invalid_request	部署没有有效工作负载，或 Kubernetes 拒绝请求

HTTP	error	说明
502	backend_unavailable	Kubernetes API 不可用
503	unavailable	Kubernetes 未配置
500	internal_error	查询 Pod 失败

错误返回格式见[调用方式](#)。

读取或持续跟踪容器日志

功能介绍

读取部署容器最近的日志，或保持 HTTP 连接持续接收新增日志。单 Pod 部署可以省略 `pod`；存在多个 Pod 时必须先通过 Pod 查询接口选择目标。建议始终显式传入 `container`。

访问要求：有效 Token；需要模型部署读取权限。

URI

```
GET /api/v1/deployments/{id}/logs
```

参数	类型	必选	描述
id	string	是	部署 ID

查询参数

参数	类型	必选	描述
pod	string	条件必选	Pod 名称；部署当前存在多个 Pod 时必须选，且必须属于该部署

参数	类型	必选	描述
container	string	否	容器名称；建议使用 Pod 查询接口返回的容器名称
follow	boolean / integer	否	<code>true</code> 或 <code>1</code> 表示持续跟踪；其他值表示仅读取当前日志
tailLines	integer	否	返回末尾日志行数；省略、无法解析或小于等于 0 时为 500

请求消息

无请求体。

响应消息

成功响应的 Content-Type 为 `text/plain; charset=utf-8`。未开启 `follow` 时连接在当前日志返回完毕后结束；开启后服务端逐块写入新日志，直到客户端断开或日志流结束。

请求示例

```
GET https://{endpoint}/api/v1/deployments/deployment-7f3a/logs?pod=qwen-  
chat-kserve-6d7c9f8b6d-abcde&container=main&tailLines=200  
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK  
Content-Type: text/plain; charset=utf-8  
X-Accel-Buffering: no  
  
INFO model loaded  
INFO server listening on :8000
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	pod_required	部署存在多个 Pod，但未指定 <code>pod</code>
400	invalid_pod	指定的 Pod 不属于该部署
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型部署读取权限，或 Kubernetes 拒绝操作
404	no_pods	部署当前没有工作负载 Pod
404	not_found	部署、Pod、容器或目标 Kubernetes 对象不存在
422	invalid_backend / invalid_request	部署没有有效工作负载，或 Kubernetes 拒绝请求
502	backend_unavailable	Kubernetes API 不可用
503	unavailable	Kubernetes 未配置
500	internal_error	读取日志失败

错误返回格式见[调用方式](#)。

查询部署监控指标

功能介绍

查询指定部署在最近时间窗口内的 CPU、内存和网络时间序列。传入 `runtime=vllm` 时还会查询 vLLM 的延迟与吞吐量指标；如果部署尚未报告工作负载 Service，则仍返回通用资源指标，但省略 `vllm`。

访问要求：有效 Token；需要模型部署读取权限。

URI

```
GET /api/v1/deployments/{id}/metrics
```

参数	类型	必选	描述
id	string	是	部署 ID

查询参数

参数	类型	必选	描述
range	string	否	时间窗口： 5m 、 15m 、 1h 或 3h ；省略或传入其他值时为 30 分钟
runtime	string	否	传 vllm 时附加查询 vLLM 指标；不区分大小写，其他值仅返回通用资源指标

请求消息

无请求体。

响应消息

每个时间序列由 [Unix 时间戳（秒），数值] 组成。

参数	参数类型	描述
cpu	array<[number, number]>	CPU 使用量，单位为核
memory	array<[number, number]>	内存工作集，单位为字节
netRx	array<[number, number]>	网络接收速率，单位为 bytes/s
netTx	array<[number, number]>	网络发送速率，单位为 bytes/s

参数	参数类型	描述
stepSeconds	integer	时间序列采样步长，单位为秒；服务端按窗口生成约 60 个点，最小 15 秒
vllm	object	vLLM 指标；未请求、无工作负载 Service 或无可用数据时可能省略
vllm.ttftP50 / ttftP95 / ttftP99 / ttftAvg	array<[number, number]>	首 Token 延迟，单位为秒
vllm.tpotP50 / tpotP95 / tpotP99 / tpotAvg	array<[number, number]>	每输出 Token 延迟，单位为秒
vllm.promptThroughput	array<[number, number]>	输入 Token 吞吐量，单位为 tokens/s
vllm.generationThroughput	array<[number, number]>	输出 Token 吞吐量，单位为 tokens/s
vllm.totalThroughput	array<[number, number]>	总 Token 吞吐量，单位为 tokens/s

请求示例

```
GET https://{endpoint}/api/v1/deployments/deployment-7f3a/metrics?
range=1h&runtime=vllm
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json
```

```
{
  "cpu": [
    [1789948800, 0.5]
  ],
```

```
"memory": [
  [1789948800, 8589934592]
],
"netRx": [
  [1789948800, 1024]
],
"netTx": [
  [1789948800, 2048]
],
"stepSeconds": 60,
"vllm": {
  "ttftP50": [
    [1789948800, 0.12]
  ],
  "ttftP95": [
    [1789948800, 0.21]
  ],
  "ttftP99": [
    [1789948800, 0.28]
  ],
  "ttftAvg": [
    [1789948800, 0.14]
  ],
  "tpotP50": [
    [1789948800, 0.018]
  ],
  "tpotP95": [
    [1789948800, 0.026]
  ],
  "tpotP99": [
    [1789948800, 0.032]
  ],
  "tpotAvg": [
    [1789948800, 0.02]
  ],
  "promptThroughput": [
    [1789948800, 180.5]
  ],
  "generationThroughput": [
    [1789948800, 72.3]
  ],
}
```

```
"totalThroughput": [  
  [1789948800, 252.8]  
]  
}  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型部署读取权限
404	not_found	部署不存在或无权访问该项目的部署
422	invalid_backend	部署没有有效工作负载
502	prometheus_error	Prometheus/Thanos 查询失败，具体原因见 <code>message</code>
503	unavailable	Kubernetes 或监控服务未配置
500	internal_error	查询部署信息失败

错误返回格式见[调用方式](#)。

建立 WebSocket 终端会话

功能介绍

将 HTTP 连接升级为 WebSocket，并在指定部署容器中建立交互式终端。单 Pod 部署可以省略 `pod`；存在多个 Pod 时必须指定目标。该接口可以在容器中执行命令，权限要求高于普通读取接口。

访问要求：有效 Token；需要模型部署写入权限。

URI

```
GET /api/v1/deployments/{id}/exec
```

参数	类型	必选	描述
id	string	是	部署 ID

查询参数

参数	类型	必选	描述
pod	string	条件必选	Pod 名称；部署当前存在多个 Pod 时必选，且必须属于该部署
container	string	否	容器名称；建议使用 Pod 查询接口返回的容器名称

请求消息

客户端通过文本帧发送 JSON 消息：

字段	类型	描述
type	string	消息类型： <code>input</code> 或 <code>resize</code>
data	string	<code>type=input</code> 时发送到终端的输入内容
cols	integer	<code>type=resize</code> 时的终端列数
rows	integer	<code>type=resize</code> 时的终端行数

输入示例：`{"type": "input", "data": "ls\n"}`。调整终端大小示例：`{"type": "resize", "cols": 120, "rows": 30}`。

响应消息

握手成功返回 101 Switching Protocols。服务端将终端标准输出作为 WebSocket 文本帧发送；会话结束前会发送一条包含结束原因的文本帧，随后发送正常关闭帧。

请求示例

```
GET wss://{endpoint}/api/v1/deployments/deployment-7f3a/exec?pod=qwen-chat-
kserve-6d7c9f8b6d-abcde&container=main
x-keystone-token:
Upgrade: websocket
Connection: Upgrade
```

正常响应示例

```
HTTP/1.1 101 Switching Protocols
Upgrade: websocket
Connection: Upgrade
```

状态码与错误码

成功状态码：101。

以下错误发生在 WebSocket 升级前；升级后的执行错误通过终端文本帧和关闭帧返回。

HTTP	error	说明
400	pod_required	部署存在多个 Pod，但未指定 <code>pod</code>
400	invalid_pod	指定的 Pod 不属于该部署
401	unauthorized	未认证或 Token 无效
403	forbidden	无模型部署写入权限，或 Kubernetes 拒绝操作
404	no_pods	部署当前没有工作负载 Pod
404	not_found	部署或目标 Kubernetes 对象不存在

HTTP	error	说明
422	invalid_backend / invalid_request	部署没有有效工作负载，或 Kubernetes 拒绝请求
502	backend_unavailable	Kubernetes API 不可用
503	unavailable	Kubernetes 未配置
500	internal_error	建立终端会话失败

错误返回格式见[调用方式](#)。

1.10 服务提供者

服务提供者

AI 服务提供者用于将可调用的上游服务接入 Beacon，供模型网关路由使用。服务来源包括平台外部的 AI 服务和集群内部的模型部署。

schema 支持 OpenAI、AzureOpenAI、AWSBedrock、GCPVertexAI、Anthropic、Cohere；端点通过 endpointUrl 指定。visibility 是历史 API 字段，表示服务接入类型而非资源可见性：public 表示外部 AI 服务，private 表示集群内部模型部署，缺省为 public。外部服务要求 credential.value 非空，内部服务的凭据可省略。

凭据内容不在详情中返回。PUT credential 轮换凭据成功返回 204。被路由引用的服务删除受到引用检查。

models 为尽力探测；后端不可达时可能返回 200、models=[]、unreachable=true。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/ai-service-providers	查询 AI 服务提供者
POST	/ai-service-providers	创建 AI 服务提供者
GET	/ai-service-providers/{id}	获取 AI 服务提供者详情
GET	/ai-service-providers/{id}/models	探测服务端模型列表
PATCH	/ai-service-providers/{id}	更新 AI 服务提供者
DELETE	/ai-service-providers/{id}	删除 AI 服务提供者
PUT	/ai-service-providers/{id}/credential	轮换 AI 服务提供者凭据

查询 AI 服务提供者

功能介绍

查询 AI 服务提供者。结果可同时包含外部 AI 服务和集群内部模型部署。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/ai-service-providers
```

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始
page_size	integer	否	每页数量；建议显式传入，例如 20；省略时部分列表返回全部

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<provider.response>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始
page_size	integer	每页数量

请求示例

```
GET https://{endpoint}/api/v1/ai-service-providers?page=1&page_size=20
x-keystone-token:
```


正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "name": "示例资源",
      "description": "",
      "status": "ready",
      "projectId": "projectId-example",
      "ownerUserId": "ownerUserId-example",
      "backend": {
        "catalog": "",
        "type": "",
        "id": "id-example"
      },
      "spec": {
        "providerKey": "",
        "schema": "OpenAI",
        "endpoint": "https://service.example.com",
        "credentialRef": "",
        "status": "ready"
      },
      "inUse": true,
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:00:00Z"
    }
  ],
  "total": 1,
  "page": 1,
  "page_size": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to list providers
401	unauthorized	authentication required

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建 AI 服务提供者

功能介绍

创建 AI 服务提供者。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/ai-service-providers
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	否	显示名称
schema	string	是	服务协议类型；OpenAI / AzureOpenAI / AWSBedrock / GCPVertexAI / Anthropic / Cohere
visibility	string	否	服务接入类型，public 表示外部 AI 服务，private 表示集群内部模型部署；缺省为 public。该字段不表示资源的可见范围。

参数	参数类型	是否必选	描述
endpointUrl	string	是	服务端点 URL；内部服务可根据所选部署自动填充，但请求中仍须提供。
credential	provider.credential	条件必选	外部服务（visibility=public）必选，须提供非空密钥；内部服务（visibility=private）可不设置密钥。
labels	map<string, string>	否	标签键值对

参数	参数类型	是否必选	描述
credential.type	string	条件必选	提供 API Key 凭据时使用 APIKey。
credential.value	string	条件必选	外部服务（visibility=public）的 API Key，不能为空。

部署选择与厂商预设用于生成端点、协议等参数，不作为 deploymentId 或 vendor 字段发送。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略

参数	参数类型	描述
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	provider.backendRef	后端资源标识
spec	provider.Spec	资源配置快照
inUse	boolean	是否正被其他资源引用
usedBy	array	引用方信息；空值时可能省略
usedByRoutes	array<provider.routeRef>	引用该服务的路由；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
POST https://{endpoint}/api/v1/ai-service-providers
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "示例服务",
  "schema": "OpenAI",
  "visibility": "public",
  "endpointUrl": "https://provider.example.com/v1",
  "credential": {
    "type": "APIKey",
    "value": "REPLACE_WITH_PROVIDER_KEY"
  }
}
```

正常响应示例

```
HTTP/1.1 201
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {
    "providerKey": "",
    "schema": "OpenAI",
    "endpoint": "https://service.example.com",
    "credentialRef": "",
    "status": "ready"
  },
  "inUse": true,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
400	invalid_request	malformed JSON body
502	backend_unavailable	kubernetes not configured; cannot project provider

HTTP	error	说明
401	unauthorized	missing project-scoped token; re-authenticate with a project scope
409	conflict	详细原因由 message 返回。
500	internal_error	failed to persist provider
403	forbidden	write permission required
404	not_found	provider not found
422	invalid_request	详细原因由 message 返回。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取 AI 服务提供者详情

功能介绍

获取 AI 服务提供者详情。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/ai-service-providers/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	provider.backendRef	后端资源标识
spec	provider.Spec	资源配置快照
inUse	boolean	是否正被其他资源引用
usedBy	array	引用方信息；空值时可能省略
usedByRoutes	array<provider.routeRef>	引用该服务的路由；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/ai-service-providers/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {
    "providerKey": "",
    "schema": "OpenAI",
    "endpoint": "https://service.example.com",
    "credentialRef": "",
    "status": "ready"
  },
  "inUse": true,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
404	not_found	provider not found

HTTP	error	说明
409	conflict	provider already exists
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

探测服务端模型列表

功能介绍

探测服务端模型列表。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/ai-service-providers/{id}/models
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
models	array	模型路由列表或模型名称列表，按类型区分
unreachable	boolean / null	是否未能访问后端

请求示例

```
GET https://{endpoint}/api/v1/ai-service-providers/resource-id/models
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "models": [
    ""
  ],
  "unreachable": false
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
404	not_found	provider not found
409	conflict	provider already exists
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

更新 AI 服务提供者

功能介绍

更新 AI 服务提供者。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
PATCH /api/v1/ai-service-providers/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string / null	否	显示名称
schema	string / null	否	服务协议类型；OpenAI / AzureOpenAI / AWSBedrock / GCPVertexAI / Anthropic / Cohere
endpointUrl	string / null	否	服务端点 URL

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述

参数	参数类型	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	provider.backendRef	后端资源标识
spec	provider.Spec	资源配置快照
inUse	boolean	是否正被其他资源引用
usedBy	array	引用方信息；空值时可能省略
usedByRoutes	array<provider.routeRef>	引用该服务的路由；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
PATCH https://{endpoint}/api/v1/ai-service-providers/resource-id
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "更新后的名称"
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {
    "providerKey": "",
    "schema": "OpenAI",
    "endpoint": "https://service.example.com",
    "credentialRef": "",
    "status": "ready"
  },
  "inUse": true,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	malformed JSON body
500	internal_error	failed to resolve provider references

HTTP	error	说明
409	conflict	详细原因由 message 返回。
403	forbidden	write permission required
401	unauthorized	authentication required
404	not_found	provider not found
422	invalid_request	详细原因由 message 返回。
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除 AI 服务提供者

功能介绍

删除 AI 服务提供者。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
DELETE /api/v1/ai-service-providers/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/ai-service-providers/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
500	internal_error	failed to resolve provider references
409	conflict	详细原因由 message 返回。
403	forbidden	write permission required
401	unauthorized	authentication required
404	not_found	provider not found
422	invalid_request	详细原因由 message 返回。
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

轮换 AI 服务提供者凭据

功能介绍

轮换 AI 服务提供者凭据。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
PUT /api/v1/ai-service-providers/{id}/credential
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
credential	provider.credential	否	访问凭据对象

参数	参数类型	是否必选	描述
credential.type	string	是	凭据类型，APIKey。
credential.value	string	是	新的 API Key，不能为空。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
PUT https://{endpoint}/api/v1/ai-service-providers/resource-id/credential
x-keystone-token:
Content-Type: application/json
```



```
{
  "credential": {
    "type": "APIKey",
    "value": "REPLACE_WITH_PROVIDER_KEY"
  }
}
```

正常响应示例

HTTP/1.1 204 No Content

状态码与错误码

成功状态码：204。

HTTP	error	说明
400	invalid_request	malformed JSON body
502	backend_unavailable	kubernetes not configured
403	forbidden	write permission required
401	unauthorized	authentication required
404	not_found	provider not found
409	conflict	provider already exists
500	internal_error	internal server error
422	invalid_request	详细原因由 message 返回。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.11 模型网关

模型网关

PUT 替换路由配置；PATCH 仅修改 name / description。具体可用状态读取 ready / conditions，不能仅依赖创建成功。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/routes	查询网关路由
POST	/routes	创建网关路由
PATCH	/routes/{id}	更新网关路由显示信息
GET	/routes/{id}	获取网关路由详情
PUT	/routes/{id}	替换网关路由配置
DELETE	/routes/{id}	删除网关路由
GET	/service-backends	查询可选路由后端
GET	/routes/filter-options	查询路由筛选项

查询网关路由

功能介绍

查询网关路由。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/routes
```

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始
page_size	integer	否	每页数量；建议显式传入，例如 20；省略时部分列表返回全部
name	string	否	显示名称
hostname	string	否	筛选参数 hostname
model	string	否	对外模型名称
backend	string	否	后端资源标识
status	string	否	资源状态；具体取值见业务规则
domain_id	string	否	筛选参数 domain_id
project_id	string	否	筛选参数 project_id
ids	string	否	逗号分隔的路由 ID；按指定资源集合筛选
sort	string	否	排序表达式
sort_field	string	否	筛选参数 sort_field
sort_order	string	否	筛选参数 sort_order
created_from	string	否	创建时间下界，RFC3339，可只指定一端
created_to	string	否	创建时间上界，RFC3339，可只指定一端

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<route.response>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始
page_size	integer	每页数量

请求示例

```
GET https://{endpoint}/api/v1/routes?page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "name": "示例资源",
      "description": "",
      "status": "ready",
      "projectId": "projectId-example",
      "ownerUserId": "ownerUserId-example",
      "backend": {
        "catalog": "",
        "type": "",
        "id": "id-example"
      },
      "spec": {},
      "createdAt": "2026-09-21T00:00:00Z",
    }
  ]
}
```

```
"updatedAt": "2026-09-21T00:00:00Z"
  }
],
"total": 1,
"page": 1,
"page_size": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to list routes
401	unauthorized	authentication required
400	invalid_request	详细原因由 message 返回。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建网关路由

功能介绍

创建网关路由。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/routes
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	是	显示名称
description	string	否	描述
namespace	string	否	Kubernetes 命名空间
models	array<route.ModelRoute>	条件必选	使用多模型映射时必选，至少一项；每项须包含 model 和非空 backendRefs。
model	string	条件必选	未提供 models、使用单模型兼容格式时必选。
subdomains	array	条件必选	使用平台根域名时必选，与完整 hostnames 二选一。
hostnames	array	条件必选	直接指定完整域名时必选，与 subdomains 二选一。
backendRefs	array<route.BackendRef>	条件必选	未提供 models 时，与顶层 model 一起提供，至少一个有效后端。
ipAllow	array	条件必选	启用 IP 白名单时须提供非空有效 CIDR 列表；否则可省略或为空。
ipDeny	array	条件必选	启用 IP 黑名单时须提供非空有效 CIDR 列表；否则可省略或为空。
labels	map<string, string>	否	标签键值对

参数	参数类型	是否必选	描述
models[].model	string	是	对外模型名，不可为空且在该路由内不可重复。
models[].backendRefs	array	是	该模型的后端列表，至少一项。

参数	参数类型	是否必选	描述
models[].backendRefs[].name	string	是	关联服务名称。
models[].backendRefs[].priority	integer	是	后端优先级, 0 至 100 的整数。
models[].backendRefs[].weight	integer	是	后端权重, 1 至 100 的整数。
models[].backendRefs[].modelNameOverride	string	是	发送给后端的模型名称, 不能为空。

嵌套表适用于 models 中的每个映射；单模型兼容格式的 backendRefs 子字段遵循相同约束。IP 白名单和黑名单按所选模式配置。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	route.backendRef	后端资源标识

参数	参数类型	描述
spec	route.Spec	资源配置快照
ready	string	是否就绪；空值时可能省略
conditions	array<route.Condition>	后端状态条件；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
POST https://{endpoint}/api/v1/routes
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "示例路由",
  "models": [
    {
      "model": "demo-model",
      "backendRefs": [
        {
          "name": "prov-backend",
          "priority": 0,
          "weight": 1,
          "modelNameOverride": "backend-model"
        }
      ]
    }
  ],
  "subdomains": [
    "demo"
  ]
}
```


正常响应示例

```
HTTP/1.1 201
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {},
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
400	invalid_request	malformed JSON body
401	unauthorized	missing project-scoped token; re-authenticate with a project scope
403	forbidden	write permission required
500	internal_error	failed to validate hostnames
409	conflict	详细原因由 message 返回。
404	not_found	详细原因由 message 返回。

HTTP	error	说明
422	invalid_request	详细原因由 message 返回。
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

更新网关路由显示信息

功能介绍

更新网关路由显示信息。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
PATCH /api/v1/routes/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string / null	否	显示名称
description	string / null	否	描述

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	route.backendRef	后端资源标识
spec	route.Spec	资源配置快照
ready	string	是否就绪；空值时可能省略
conditions	array<route.Condition>	后端状态条件；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
PATCH https://{endpoint}/api/v1/routes/resource-id
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "更新后的名称"
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {},
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	malformed JSON body
403	forbidden	write permission required
401	unauthorized	authentication required

HTTP	error	说明
404	not_found	route not found
409	conflict	route already exists
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取网关路由详情

功能介绍

获取网关路由详情。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/routes/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识

参数	参数类型	描述
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	route.backendRef	后端资源标识
spec	route.Spec	资源配置快照
ready	string	是否就绪；空值时可能省略
conditions	array<route.Condition>	后端状态条件；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/routes/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {},
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
404	not_found	route not found
409	conflict	route already exists
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

替换网关路由配置

功能介绍

替换网关路由配置。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
PUT /api/v1/routes/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	是	显示名称
description	string	否	描述
namespace	string	否	Kubernetes 命名空间
models	array<route.ModelRoute>	条件必选	使用多模型映射时必选，至少一项；每项须包含 model 和非空 backendRefs。
model	string	条件必选	未提供 models、使用单模型兼容格式时必选。
subdomains	array	条件必选	使用平台根域名时必选，与完整 hostnames 二选一。
hostnames	array	条件必选	直接指定完整域名时必选，与 subdomains 二选一。

参数	参数类型	是否必选	描述
backendRefs	array<route.BackendRef>	条件必选	未提供 models 时，与顶层 model 一起提供，至少一个有效后端。
ipAllow	array	条件必选	启用 IP 白名单时须提供非空有效 CIDR 列表；否则可省略或为空。
ipDeny	array	条件必选	启用 IP 黑名单时须提供非空有效 CIDR 列表；否则可省略或为空。
labels	map<string, string>	否	标签键值对

使用完整路由配置替换模型映射等配置；不是局部 PATCH。

参数	参数类型	是否必选	描述
models[].model	string	是	对外模型名，不可为空且在该路由内不可重复。
models[].backendRefs	array	是	该模型的后端列表，至少一项。
models[].backendRefs[].name	string	是	关联服务名称。
models[].backendRefs[].priority	integer	是	后端优先级，0 至 100 的整数。
models[].backendRefs[].weight	integer	是	后端权重，1 至 100 的整数。
models[].backendRefs[].modelNameOverride	string	是	发送给后端的模型名称，不能为空。

嵌套表适用于 models 中的每个映射；单模型兼容格式的 backendRefs 子字段遵循相同约束。IP 白名单和黑名单按所选模式配置。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	资源状态；具体取值见业务规则
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	route.backendRef	后端资源标识
spec	route.Spec	资源配置快照
ready	string	是否就绪；空值时可能省略
conditions	array<route.Condition>	后端状态条件；空值时可能省略
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
PUT https://{endpoint}/api/v1/routes/resource-id
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "示例路由",
  "models": [
    {
      "model": "demo-model",
      "backendRefs": [
        {
          "name": "prov-backend",
          "priority": 0,
          "weight": 1,
          "modelNameOverride": "backend-model"
        }
      ]
    }
  ],
  "subdomains": [
    "demo"
  ]
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "description": "",
  "status": "ready",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",

```

```
{
  "type": "",
  "id": "id-example"
},
"spec": {},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	malformed JSON body
500	internal_error	route has no k8s backend id
401	unauthorized	missing project-scoped token; re-authenticate with a project scope
403	forbidden	write permission required
404	not_found	route not found
409	conflict	route already exists
422	invalid_request	详细原因由 message 返回。
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除网关路由

功能介绍

删除网关路由。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
DELETE /api/v1/routes/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/routes/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
401	unauthorized	missing project-scoped token; re-authenticate with a project scope

HTTP	error	说明
403	forbidden	write permission required
404	not_found	route not found
409	conflict	route already exists
500	internal_error	internal server error
422	invalid_request	详细原因由 message 返回。
502	backend_unavailable	kubernetes API unavailable

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询可选路由后端

功能介绍

查询可选路由后端。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/service-backends
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<route.backendView>	结果列表

请求示例

```
GET https://{endpoint}/api/v1/service-backends
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "name": "示例资源",
      "displayName": "",
      "namespace": "",
      "kind": ""
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
400	invalid_request	详细原因由 message 返回。
403	forbidden	详细原因由 message 返回。
404	not_found	详细原因由 message 返回。
409	conflict	a route with that name already exists in the namespace
422	invalid_request	详细原因由 message 返回。

HTTP	error	说明
502	backend_unavailable	kubernetes API unavailable
500	internal_error	failed to provision route

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询路由筛选项

功能介绍

查询路由筛选项。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/routes/filter-options
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
backends	array<route.filterOption>	后端筛选项
domains	array<route.filterOption>	部门筛选项
projects	array<route.filterOption>	项目筛选项

请求示例


```
GET https://{endpoint}/api/v1/routes/filter-options
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "backends": [
    {
      "value": "",
      "label": ""
    }
  ],
  "domains": [
    {
      "value": "",
      "label": ""
    }
  ],
  "projects": [
    {
      "value": "",
      "label": ""
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.12 模型体验

模型体验

体验请求使用 x-keystone-token: 通过控制面鉴权，并通过 X-Experience-Key 提供网关 API Key。X-Experience-Host 选择网关路由域名。

chat / images / speech / transcriptions 请求体原样转发，Beacon 不定义统一固定的上游字段表。请求体最大 32 MiB。支持字段与效果取决于所选后端服务。

上游状态 ≥ 400 时，Beacon 返回 HTTP 200 并保留错误 JSON；真实上游状态在 X-Upstream-Status 中。非 JSON 上游错误会包装为 error 对象。

chat 可返回 SSE；speech 返回音频二进制；transcriptions 使用 multipart/form-data。音色、语言等能力由 audio/capabilities 查询。

方法	路径（产品接口统一前缀 /api/v1）	功能
POST	/experience/chat	流式聊天补全
POST	/experience/images	生成图片
POST	/experience/audio/speech	生成语音
GET	/experience/audio/capabilities	查询语音模型能力
POST	/experience/audio/transcriptions	音频转写
GET	/experience/routes	查询体验中心可用路由

对话

功能介绍

模型对话。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/experience/chat
```

接口请求头

请求头	说明
X-Experience-Host	路由域名。
X-Experience-Key	必选，网关 API Key。

请求消息

请求体由 Beacon 原样转发给后端服务，字段约束以所选服务为准；以下为常见调用格式，非 Beacon 本地校验规则。

字段	类型	说明
model	string	路由提供的模型名称
messages	array	对话消息数组，如 role/content
stream	boolean	是否请求 SSE

响应消息

成功体沿用上游：chat 为 JSON 或 text/event-stream；images 为 JSON；speech 为音频字节；transcriptions 为上游转写结果。具体取决于当前接口与请求参数。上游错误见[调用方式](#)。

请求示例

```
POST https://{endpoint}/api/v1/experience/chat
x-keystone-token:
X-Experience-Key:
X-Experience-Host: demo.ai.example.com
Content-Type: application/json
```

```
{
  "model": "demo-model",
  "messages": [
    {
      "role": "user",
      "content": "你好"
    }
  ],
  "stream": true
}
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: text/event-stream

data: {"choices":[{"delta":{"content":"你好"}}]}

data: [DONE]
```

状态码与错误码

成功状态码：200。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

本地可能返回 400、413、500、502、503；上游 ≥ 400 被包装为 200，请检查 X-Upstream-Status 与 error。

生成图片

功能介绍

生成图片。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/experience/images
```

接口请求头

请求头	说明
X-Experience-Host	路由域名。
X-Experience-Key	必选，网关 API Key。

请求消息

请求体由 Beacon 原样转发给后端服务，字段约束以所选服务为准；以下为常见调用格式，非 Beacon 本地校验规则。

字段	类型	说明
model	string	路由提供的模型名称
prompt	string	图像描述

响应消息

成功体沿用上游：chat 为 JSON 或 text/event-stream；images 为 JSON；speech 为音频字节；transcriptions 为上游转写结果。具体取决于当前接口与请求参数。上游错误见[调用方式](#)。

请求示例

```
POST https://{endpoint}/api/v1/experience/images
x-keystone-token:
X-Experience-Key:
X-Experience-Host: demo.ai.example.com
Content-Type: application/json
```

```
{
  "model": "demo-model",
  "prompt": "一座山"
}
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json

{"data":[{"b64_json":""}]}
```

状态码与错误码

成功状态码：200。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

本地可能返回 400、413、500、502、503；上游 ≥ 400 被包装为 200，请检查 X-Upstream-Status 与 error。

生成语音

功能介绍

生成语音。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/experience/audio/speech
```

接口请求头

请求头	说明
-----	----

请求头	说明
X-Experience-Host	路由域名。
X-Experience-Key	必选，网关 API Key。

请求消息

请求体由 Beacon 原样转发给后端服务，字段约束以所选服务为准；以下为常见调用格式，非 Beacon 本地校验规则。

字段	类型	说明
model	string	路由提供的模型名称
input	string	待合成文本
voice	string	后端支持的音色
response_format	string	后端支持的音频格式

响应消息

成功体沿用上游：chat 为 JSON 或 text/event-stream；images 为 JSON；speech 为音频字节；transcriptions 为上游转写结果。具体取决于当前接口与请求参数。上游错误见[调用方式](#)。

请求示例

```
POST https://{endpoint}/api/v1/experience/audio/speech
x-keystone-token:
X-Experience-Key:
X-Experience-Host: demo.ai.example.com
Content-Type: application/json
```

```
{
  "model": "demo-model",
  "input": "你好",
```



```
"voice": ""  
}
```

正常响应示例

```
HTTP/1.1 200 OK  
Content-Type: audio/mpeg  
  
<音频字节>
```

状态码与错误码

成功状态码：200。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

本地可能返回 400、413、500、502、503；上游 ≥ 400 被包装为 200，请检查 X-Upstream-Status 与 error。

查询语音模型能力

功能介绍

查询语音模型能力。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/experience/audio/capabilities
```

查询参数

参数	类型	必选	描述
routeId	string	是	路由 ID
model	string	是	对外模型名称

请求消息

无请求体。

响应消息

参数	参数类型	描述
model	string	对外模型名称
servedModel	string	后端实际模型名；空值时可能省略
providerId	string	外部服务 ID
voices	array	可用音色列表
languages	array	可用语言列表
languageSupported	boolean	是否支持语言参数
defaultLanguage	string	默认语言；空值时可能省略
responseFormats	array	响应编码格式列表
defaultResponseFormat	string	默认响应格式；空值时可能省略
speed	experience.numberRange / null	语速范围；空值时可能省略
discoveryWarnings	array	能力探测提示；空值时可能省略

请求示例

```
GET https://{endpoint}/api/v1/experience/audio/capabilities?routeId=route-id&model=demo-model
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "model": "",
  "providerId": "providerId-example",
  "voices": [
    ""
  ],
  "languages": [
    ""
  ],
  "languageSupported": true,
  "responseFormats": [
    ""
  ]
}
```

状态码与错误码

成功状态码：200。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

音频转写

功能介绍

音频转写。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/experience/audio/transcriptions
```

接口请求头

请求头	说明
X-Experience-Host	路由域名。

请求头	说明
X-Experience-Key	必选，网关 API Key。

请求消息

请求体由 Beacon 原样转发给后端服务，字段约束以所选服务为准；以下为常见调用格式，非 Beacon 本地校验规则。

表单字段	类型	说明
file	binary	上传音频文件。
model	string	路由提供的模型名称。
language / response_format	string	可选参数，是否支持由后端决定。

响应消息

成功体沿用上游：chat 为 JSON 或 text/event-stream；images 为 JSON；speech 为音频字节；transcriptions 为上游转写结果。具体取决于当前接口与请求参数。上游错误见[调用方式](#)。

请求示例

```
POST https://{endpoint}/api/v1/experience/audio/transcriptions
x-keystone-token:
X-Experience-Key:
X-Experience-Host: demo.ai.example.com
Content-Type: multipart/form-data; boundary=example

--example
Content-Disposition: form-data; name="model"

demo-model
--example
Content-Disposition: form-data; name="file"; filename="audio.wav"
Content-Type: audio/wav
```

```
<音频文件二进制>
--example--
```

正常响应示例

```
HTTP/1.1 200 OK
Content-Type: application/json

{"text": "你好"}
```

状态码与错误码

成功状态码：200。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

本地可能返回 400、413、500、502、503；上游 ≥ 400 被包装为 200，请检查 X-Upstream-Status 与 error。

查询体验中心可用路由

功能介绍

查询体验中心可用路由。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/experience/routes
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
----	------	----

参数	参数类型	描述
items	array<experience.RouteView>	结果列表

请求示例

```
GET https://{endpoint}/api/v1/experience/routes
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "name": "示例资源",
      "status": "ready",
      "projectId": "projectId-example",
      "backend": {
        "catalog": "",
        "type": "",
        "id": "id-example"
      },
      "spec": {}
    }
  ]
}
```

状态码与错误码

成功状态码：200。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.13 套餐管理

套餐管理

套餐用于配置 API Key 可访问的模型网关路由，以及按周期生效的请求数和 Token 数限流。套餐可启用或禁用，禁用后绑定该套餐的 API Key 无法通过套餐访问路由。

rps 和 tps 是历史 API 字段名：rps 表示配置周期内的请求数上限，tps 表示配置周期内的 Token 数上限，二者并非固定按秒计数。rps 范围为 0..2147483647，tps 为非负整数；0 表示不限量。周期单位支持 minute、hour、day、month、year；空或未知单位归一为 minute。

rpsPeriod / tpsPeriod 缺省为 1；各单位最大倍数为 minute=10080、hour=2160、day=1095、month=60、year=5。

套餐共享允许其他项目查看套餐，并将本项目的 API Key 绑定到该套餐。共享项目列表通过 PUT shares 整体替换；如果被移除的项目仍有 API Key 绑定该套餐，请求返回 409。所有 shares / share-targets 操作都要求 plan_share/global/manage 权限。sharedProjectIds 仅向拥有套餐全局读权限的调用方返回。

套餐仍被 API Key 绑定时不能删除，需先解除绑定。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/plans	查询套餐
POST	/plans	创建套餐
GET	/plans/{id}	获取套餐详情
PATCH	/plans/{id}	更新套餐
DELETE	/plans/{id}	删除套餐
GET	/plans/{id}/shares	查询套餐共享项目
PUT	/plans/{id}/shares	替换套餐共享项目
GET	/plans/{id}/share-targets	查询套餐可共享项目

查询套餐

功能介绍

查询当前调用方可见的套餐，包括本项目套餐和共享给本项目的套餐。

访问要求：有效 Token 和套餐读权限。普通调用方可查看本项目套餐；具有共享读权限时还可查看共享给本项目的套餐；具有全局读权限时可查看全部项目的套餐。

URI

```
GET /api/v1/plans
```

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始
page_size	integer	否	每页数量；建议显式传入，例如 20；省略时部分列表返回全部

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<plan.response>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始
page_size	integer	每页数量

请求示例

```
GET https://{endpoint}/api/v1/plans?page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "id-example",
      "name": "标准套餐",
      "description": "",
      "status": "active",
      "projectId": "projectId-example",
      "ownerUserId": "ownerUserId-example",
      "backend": {
        "catalog": "",
        "type": "",
        "id": "id-example"
      },
      "spec": {
        "planKey": "pk-a1b2c3d4e5f6",
        "rps": 0,
        "rpsUnit": "minute",
        "rpsPeriod": 1,
        "tps": 0,
        "tpsUnit": "minute",
        "tpsPeriod": 1,
        "status": "active",
        "routeIds": []
      },
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:00:00Z"
    }
  ],
```

```
"total": 1,  
"page": 1,  
"page_size": 1  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to list plans
401	unauthorized	authentication required

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建套餐

功能介绍

创建套餐，配置请求限流、Token 限流和可访问路由。

访问要求：有效 Token 和本项目的套餐写权限。

URI

```
POST /api/v1/plans
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	是	套餐名称

参数	参数类型	是否必选	描述
description	string	否	描述
rps	integer	否	配置周期内的请求数上限，范围为 0..2147483647；0 表示不限量，缺省为 0。
rpsUnit	string	否	请求限流周期单位，minute、hour、day、month 或 year；缺省为 minute。
rpsPeriod	integer / null	否	请求限流周期倍数，为有效范围内的正整数；缺省为 1。
tps	int64	否	配置周期内的 Token 数上限，为非负整数；0 表示不限量，缺省为 0。
tpsUnit	string	否	Token 限流周期单位，minute、hour、day、month 或 year；缺省为 minute。
tpsPeriod	integer / null	否	Token 限流周期倍数，为有效范围内的正整数；缺省为 1。
status	string	否	套餐状态：active 表示启用，disabled 表示禁用；缺省为 active。
routelds	array	否	可访问路由 ID 数组，可为空；为空时套餐未授权任何路由。
labels	map<string, string>	否	标签键值对

建议创建时完整指定限流周期和上限；不限制某类用量时传入 0。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	套餐名称

参数	参数类型	描述
description	string	描述
status	string	套餐状态：active 或 disabled
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	plan.backendRef	后端资源标识
spec	plan.Spec	套餐配置快照，包含系统生成的 planKey、限流配置、状态和可访问路由 ID
labels	map<string, string>	标签键值对；空值时可能省略
sharedProjectIds	array	套餐已共享的项目 ID 数组；仅向拥有套餐全局读权限的调用方返回，共享接收方无法通过该字段查看其他接收项目。
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
POST https://{endpoint}/api/v1/plans
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "标准套餐",
  "rps": 60,
  "rpsUnit": "minute",
  "rpsPeriod": 1,
  "tps": 10000,
  "tpsUnit": "minute",
  "tpsPeriod": 1,
  "routeIds": [
    "route-id"
  ],
  "status": "active"
}
```

正常响应示例

```
HTTP/1.1 201
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "标准套餐",
  "description": "",
  "status": "active",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {
    "planKey": "pk-a1b2c3d4e5f6",
    "rps": 60,
    "rpsUnit": "minute",
    "rpsPeriod": 1,
    "tps": 10000,
    "tpsUnit": "minute",
    "tpsPeriod": 1,
  }
}
```

```
"status": "active",
"routeIds": [
  "route-id"
],
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
400	invalid_request	malformed JSON body
409	conflict	详细原因由 message 返回。
500	internal_error	failed to persist plan
403	forbidden	write permission required
401	unauthorized	authentication required

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取套餐详情

功能介绍

获取套餐详情。共享接收项目可查看共享给本项目的套餐，但不会获取其他共享接收项目的信息。

访问要求：有效 Token 和套餐读权限；本项目套餐、共享给本项目的套餐或全局读权限范围内的套餐可读。

URI

```
GET /api/v1/plans/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	套餐名称
description	string	描述
status	string	套餐状态：active 或 disabled
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	plan.backendRef	后端资源标识
spec	plan.Spec	套餐配置快照，包含系统生成的 planKey、限流配置、状态和可访问路由 ID
labels	map<string, string>	标签键值对；空值时可能省略

参数	参数类型	描述
sharedProjectIds	array	套餐已共享的项目 ID 数组；仅向拥有套餐全局读权限的调用方返回，共享接收方无法通过该字段查看其他接收项目。
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/plans/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "标准套餐",
  "description": "",
  "status": "active",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {
    "planKey": "pk-a1b2c3d4e5f6",
    "rps": 0,
    "rpsUnit": "minute",
    "rpsPeriod": 1,
```



```
{
  "tps": 0,
  "tpsUnit": "minute",
  "tpsPeriod": 1,
  "status": "active",
  "routeIds": [
    "route-id"
  ],
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to resolve shared plan
401	unauthorized	authentication required
404	not_found	plan not found

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

更新套餐

功能介绍

更新套餐的基本信息、状态、限流配置或可访问路由。

访问要求：有效 Token 和套餐写权限；只能更新本项目所属的套餐，共享接收项目不能修改共享套餐。

URI

```
PATCH /api/v1/plans/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string / null	否	套餐名称
description	string / null	否	描述
rps	integer / null	否	配置周期内的请求数上限；0 表示不限量
rpsUnit	string / null	否	请求限流周期单位
rpsPeriod	integer / null	否	请求限流周期倍数
tps	int64 / null	否	配置周期内的 Token 数上限；0 表示不限量
tpsUnit	string / null	否	Token 限流周期单位
tpsPeriod	integer / null	否	Token 限流周期倍数
status	string / null	否	套餐状态：active 表示启用，disabled 表示禁用
routeIds	array / null	否	可访问路由 ID 数组
labels	map<string, string> / null	否	标签键值对

PATCH 仅需发送待修改字段。rps/tps 可以为 0，表示对应用量不限；周期倍数必须是有效范围内的正整数。修改限流周期会重置当前周期的已使用量。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	套餐名称
description	string	描述
status	string	套餐状态：active 或 disabled
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
backend	plan.backendRef	后端资源标识
spec	plan.Spec	套餐配置快照，包含系统生成的 planKey、限流配置、状态和可访问路由 ID
labels	map<string, string>	标签键值对；空值时可能省略
sharedProjectIds	array	套餐已共享的项目 ID 数组；仅向拥有套餐全局读权限的调用方返回，共享接收方无法通过该字段查看其他接收项目。
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
PATCH https://{endpoint}/api/v1/plans/resource-id
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "更新后的名称"
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "更新后的名称",
  "description": "",
  "status": "active",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "backend": {
    "catalog": "",
    "type": "",
    "id": "id-example"
  },
  "spec": {
    "planKey": "pk-a1b2c3d4e5f6",
    "rps": 0,
    "rpsUnit": "minute",
    "rpsPeriod": 1,
    "tps": 0,
    "tpsUnit": "minute",
    "tpsPeriod": 1,
    "status": "active",
    "routeIds": [
      "route-id"
    ]
  }
}
```

```
{,
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	malformed JSON body
500	internal_error	failed to bind routes
403	forbidden	write permission required
401	unauthorized	authentication required
404	not_found	plan not found
409	conflict	plan already exists

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除套餐

功能介绍

删除套餐。套餐仍被 API Key 绑定时返回 409，需先解除绑定。

访问要求：有效 Token 和套餐写权限；只能删除本项目所属且未被 API Key 绑定的套餐。

URI

```
DELETE /api/v1/plans/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/plans/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
409	conflict	详细原因由 message 返回。
500	internal_error	failed to remove plan detail
403	forbidden	write permission required
401	unauthorized	authentication required
404	not_found	plan not found

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询套餐共享项目

功能介绍

查询已被授予该套餐使用权限的项目。

访问要求：Token + plan_share/global/manage。

URI

```
GET /api/v1/plans/{id}/shares
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
projectIds	array	套餐已共享的项目 ID 数组

请求示例

```
GET https://{endpoint}/api/v1/plans/resource-id/shares
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "projectIds": [
    "project-id"
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
403	forbidden	plan share management permission required
500	internal_error	failed to list plan shares
401	unauthorized	authentication required
404	not_found	plan not found

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

替换套餐共享项目

功能介绍

用请求中的项目 ID 数组整体替换套餐的共享项目。传入空数组表示撤销全部共享。

访问要求：Token + plan_share/global/manage。

URI

```
PUT /api/v1/plans/{id}/shares
```


参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
projectIds	array	否	共享项目 ID 数组；省略或传入空数组表示撤销全部共享

响应消息

参数	参数类型	描述
projectIds	array	更新后的共享项目 ID 数组

请求示例

```
PUT https://{endpoint}/api/v1/plans/resource-id/shares
x-keystone-token:
Content-Type: application/json
```

```
{
  "projectIds": [
    "project-id"
  ]
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "projectIds": [
    "project-id"
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
403	forbidden	plan share management permission required
400	invalid_request	malformed JSON body
503	backend_unavailable	failed to validate tenant projects
500	internal_error	failed to inspect current plan shares
409	conflict	详细原因由 message 返回。
401	unauthorized	authentication required
404	not_found	plan not found

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询套餐可共享项目

功能介绍

查询可作为套餐共享目标的项目。套餐所属项目不包含在结果中。

访问要求：Token + plan_share/global/manage。

URI

```
GET /api/v1/plans/{id}/share-targets
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<keystone.Project>	结果列表

请求示例

```
GET https://{endpoint}/api/v1/plans/resource-id/share-targets
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "project-id",
      "name": "示例项目",
      "domainId": "domainId-example"
    }
  ]
}
```

```
}  
]  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
403	forbidden	plan share management permission required
503	backend_unavailable	failed to list tenant projects
401	unauthorized	authentication required
404	not_found	plan not found
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.14 基准测试

基准测试

基准测试功能需要平台已启用基准测试服务。

创建要求 name、deploymentId、profile。profile 表示测试类型，仅支持 latency / throughput；测试方法由服务端按版本定义，客户端不传入任意目标 URL。

创建和停止返回 202，随后轮询 state；状态为 pending、running、completed、error、stopped。同一模型部署存在进行中的基准测试时返回 409 active_benchmark。

列表分页参数使用 size，缺省 20、最大 200。日志 tail 缺省 200，返回日志文本与截断标记。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/benchmark-profiles	查询基准测试方法
GET	/benchmarks	查询基准测试任务
POST	/benchmarks	创建基准测试任务
GET	/benchmarks/{id}	查询基准测试详情
GET	/benchmarks/{id}/logs	查询基准测试日志
POST	/benchmarks/{id}/stop	停止基准测试任务
DELETE	/benchmarks/{id}	删除基准测试任务

查询基准测试方法

功能介绍

查询平台预定义的基准测试方法。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/benchmark-profiles
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
methodologyVersion	string	基准测试方法版本
items	array<benchmark.Profile>	结果列表

请求示例

```
GET https://{endpoint}/api/v1/benchmark-profiles
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "methodologyVersion": "beacon-benchmark/v1",
  "items": [
    {
      "id": "id-example",
      "methodologyVersion": "beacon-benchmark/v1",
      "datasetVersion": "beacon-jsonl/v1",
      "description": "",
      "inputShape": "",
      "outputTokens": 0,
      "runnerProfile": "",
    }
  ]
}
```

```
"warmup": {
  "mode": "",
  "value": 0
},
"dataSamples": 0,
"maxRequests": 0,
"maxSeconds": 0,
"sampleRequests": 0,
"maxErrors": 0
}
]
```

状态码与错误码

成功状态码：200。

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询基准测试任务

功能介绍

查询基准测试任务。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/benchmarks
```

查询参数

参数	类型	必选	描述
state	string	否	任务状态
profile	string	否	测试类型：latency 或 throughput

参数	类型	必选	描述
deploymentId	string	否	目标模型部署 ID
page	integer	否	页码，从 1 开始；默认 1
size	integer	否	数量或大小，单位见上下文；默认 20

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array<benchmark.Item>	结果列表
total	integer	匹配总数
page	integer	页码，从 1 开始
size	integer	数量或大小，单位见上下文

请求示例

```
GET https://{endpoint}/api/v1/benchmarks?page=1&size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
```



```
{
  "id": "id-example",
  "name": "示例资源",
  "deploymentId": "deploymentId-example",
  "profile": "latency",
  "state": "pending",
  "spec": {
    "methodologyVersion": "beacon-benchmark/v1",
    "datasetVersion": "beacon-jsonl/v1",
    "seed": 0,
    "profile": {}
  },
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
},
"total": 1,
"page": 1,
"size": 1
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_state	unsupported benchmark state
400	invalid_profile	unsupported benchmark profile
404	not_found	benchmark or deployment not found
409	active_benchmark	deployment already has an active benchmark
409	invalid_state	operation is not allowed in current state
422	invalid_spec	benchmark specification is too large
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建基准测试任务

功能介绍

创建基准测试任务。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/benchmarks
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	是	显示名称
deploymentId	string	是	目标模型部署 ID
profile	string	是	测试类型，latency 表示延迟测试，throughput 表示吞吐测试；创建流程默认 latency。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
projectId	string	所属项目 ID；空值时可能省略

参数	参数类型	描述
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
deploymentId	string	目标模型部署 ID
profile	string	测试类型：latency 或 throughput
state	string	任务状态
cancelRequested	boolean	是否已请求停止；空值时可能省略
runnerPodPhase	string	执行 Pod 阶段；空值时可能省略
spec	benchmark.Spec	资源配置快照
result	JSON	基准测试结果；空值时可能省略
errorMessage	string	错误详情；空值时可能省略
startedAt	string(date-time) / null	执行开始时间；空值时可能省略
finishedAt	string(date-time) / null	结束时间；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
POST https://{endpoint}/api/v1/benchmarks
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "延迟基准",
  "deploymentId": "deployment-id",
```

```
"profile": "latency"
}
```

正常响应示例

```
HTTP/1.1 202
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "deploymentId": "deploymentId-example",
  "profile": "latency",
  "state": "pending",
  "spec": {
    "methodologyVersion": "beacon-benchmark/v1",
    "datasetVersion": "beacon-jsonl/v1",
    "seed": 0,
    "profile": {
      "id": "id-example",
      "methodologyVersion": "beacon-benchmark/v1",
      "datasetVersion": "beacon-jsonl/v1",
      "description": "",
      "inputShape": "",
      "outputTokens": 0,
      "runnerProfile": "",
      "warmup": {
        "mode": "",
        "value": 0
      },
      "dataSamples": 0,
      "maxRequests": 0,
      "maxSeconds": 0,
      "sampleRequests": 0,
      "maxErrors": 0
    }
  },
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：202。

HTTP	error	说明
400	invalid_request	invalid JSON body
422	invalid_request	name and deploymentId are required
422	invalid_profile	profile must be latency or throughput
503	benchmark_unavailable	benchmark target validation is not configured
403	forbidden	write permission required
404	not_found	benchmark or deployment not found
409	active_benchmark	deployment already has an active benchmark
409	invalid_state	operation is not allowed in current state
422	invalid_spec	benchmark specification is too large
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询基准测试详情

功能介绍

查询基准测试详情。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/benchmarks/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
projectId	string	所属项目 ID；空值时可能省略
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
deploymentId	string	目标模型部署 ID
profile	string	测试类型：latency 或 throughput
state	string	任务状态
cancelRequested	boolean	是否已请求停止；空值时可能省略
runnerPodPhase	string	执行 Pod 阶段；空值时可能省略
spec	benchmark.Spec	资源配置快照
result	JSON	基准测试结果；空值时可能省略
errorMessage	string	错误详情；空值时可能省略
startedAt	string(date-time) / null	执行开始时间；空值时可能省略

参数	参数类型	描述
finishedAt	string(date-time) / null	结束时间；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/benchmarks/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "deploymentId": "deploymentId-example",
  "profile": "latency",
  "state": "pending",
  "spec": {
    "methodologyVersion": "beacon-benchmark/v1",
    "datasetVersion": "beacon-jsonl/v1",
    "seed": 0,
    "profile": {
      "id": "id-example",
      "methodologyVersion": "beacon-benchmark/v1",
      "datasetVersion": "beacon-jsonl/v1",
      "description": "",
      "inputShape": "",
      "outputTokens": 0,
      "runnerProfile": "",
      "warmup": {
        "mode": "",

```

```
    "value": 0
  },
  "dataSamples": 0,
  "maxRequests": 0,
  "maxSeconds": 0,
  "sampleRequests": 0,
  "maxErrors": 0
}
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
404	not_found	benchmark or deployment not found
409	active_benchmark	deployment already has an active benchmark
409	invalid_state	operation is not allowed in current state
422	invalid_spec	benchmark specification is too large
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询基准测试日志

功能介绍

查询基准测试日志。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/benchmarks/{id}/logs
```

参数	类型	必选	描述
id	string	是	目标资源 ID

查询参数

参数	类型	必选	描述
tail	integer	否	日志末尾行数；默认 200

请求消息

无请求体。

响应消息

参数	参数类型	描述
logs	string	日志文本
truncated	boolean	日志是否被截断；空值时可能省略

请求示例

```
GET https://{endpoint}/api/v1/benchmarks/resource-id/logs
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "logs": ""
}
```

状态码与错误码

成功状态码：200, 202。

HTTP	error	说明
503	logs_unavailable	benchmark logs are not configured
404	not_found	benchmark or deployment not found
409	active_benchmark	deployment already has an active benchmark
409	invalid_state	operation is not allowed in current state
422	invalid_spec	benchmark specification is too large
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

停止基准测试任务

功能介绍

停止基准测试任务。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
POST /api/v1/benchmarks/{id}/stop
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
projectId	string	所属项目 ID；空值时可能省略
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
deploymentId	string	目标模型部署 ID
profile	string	测试类型：latency 或 throughput
state	string	任务状态
cancelRequested	boolean	是否已请求停止；空值时可能省略
runnerPodPhase	string	执行 Pod 阶段；空值时可能省略
spec	benchmark.Spec	资源配置快照
result	JSON	基准测试结果；空值时可能省略
errorMessage	string	错误详情；空值时可能省略
startedAt	string(date-time) / null	执行开始时间；空值时可能省略

参数	参数类型	描述
finishedAt	string(date-time) / null	结束时间；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
POST https://{endpoint}/api/v1/benchmarks/resource-id/stop
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 202
Content-Type: application/json
```

```
{
  "id": "id-example",
  "name": "示例资源",
  "deploymentId": "deploymentId-example",
  "profile": "latency",
  "state": "pending",
  "spec": {
    "methodologyVersion": "beacon-benchmark/v1",
    "datasetVersion": "beacon-jsonl/v1",
    "seed": 0,
    "profile": {
      "id": "id-example",
      "methodologyVersion": "beacon-benchmark/v1",
      "datasetVersion": "beacon-jsonl/v1",
      "description": "",
      "inputShape": "",
      "outputTokens": 0,
      "runnerProfile": "",
      "warmup": {
        "mode": "",

```

```
    "value": 0
  },
  "dataSamples": 0,
  "maxRequests": 0,
  "maxSeconds": 0,
  "sampleRequests": 0,
  "maxErrors": 0
}
},
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：202。

HTTP	error	说明
403	forbidden	write permission required
404	not_found	benchmark or deployment not found
409	active_benchmark	deployment already has an active benchmark
409	invalid_state	operation is not allowed in current state
422	invalid_spec	benchmark specification is too large
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除基准测试任务

功能介绍

删除基准测试任务。

访问要求：有效 Token；需要对应资源写权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
DELETE /api/v1/benchmarks/{id}
```

参数	类型	必选	描述
id	string	是	目标资源 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/benchmarks/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
403	forbidden	write permission required
404	not_found	benchmark or deployment not found
409	active_benchmark	deployment already has an active benchmark

HTTP	error	说明
409	invalid_state	operation is not allowed in current state
422	invalid_spec	benchmark specification is too large
500	internal_error	internal server error

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.15 租户和配额

租户和配额

租户对应用户登录身份中的项目。项目和部门上下文由 Token 确定，调用方不能通过查询参数或请求头切换资源所属项目。

租户命名空间是 Beacon 在 Kubernetes 中为当前项目使用的工作空间。创建操作具有幂等性：命名空间已存在时返回 200，首次创建成功时返回 201；提交的名称会归一为小写的 Kubernetes 命名空间名称。未配置 Kubernetes 时，查询接口返回 exists=true、ready=true，以避免非集群环境被初始化流程阻塞，但创建接口返回 503。

计算配额接口返回 Horizon 项目配额中 eks_managed 产品的资源配额和当前用量。该结果表示当前项目可使用的 CPU、内存和 AI 加速卡等计算资源，不等同于集群剩余资源或套餐的请求数、Token 数限流。

模型网关产生的请求数和 Token 统计请参见[Token 用量](#)。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/tenants/current	获取当前租户上下文
GET	/tenants/current/namespace	查询租户命名空间状态
POST	/tenants/current/namespace	创建租户命名空间（幂等）
GET	/tenants/current/quota	获取当前租户计算配额

获取当前租户上下文

功能介绍

获取 Token 所属项目和部门信息。

访问要求：有效 Token。

URI


```
GET /api/v1/tenants/current
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
projectId	string	当前项目 ID
projectName	string	当前项目名称
domainId	string	当前部门 ID
domainName	string	当前部门名称

请求示例

```
GET https://{endpoint}/api/v1/tenants/current
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "projectId": "project-a1b2c3d4",
  "projectName": "生产项目",
  "domainId": "domain-a1b2c3d4",
  "domainName": "研发部门"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required

后端错误与错误封装见[调用方式](#)。

查询租户命名空间状态

功能介绍

查询当前项目是否已经具有 Beacon 使用的 Kubernetes 命名空间。接口只返回是否存在和是否就绪，不返回实际命名空间名称。

访问要求：有效 Token。

URI

```
GET /api/v1/tenants/current/namespace
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
exists	boolean	当前项目的 Beacon 命名空间是否存在
ready	boolean	当前项目的 Beacon 命名空间是否可用；当前实现与 exists 保持一致

请求示例

```
GET https://{endpoint}/api/v1/tenants/current/namespace
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "exists": true,
  "ready": true
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
403	forbidden	无权读取当前项目的 Kubernetes 命名空间状态
502	backend_unavailable	查询 Kubernetes 命名空间失败，详细原因由 message 返回

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建租户命名空间（幂等）

功能介绍

为当前项目创建 Beacon 使用的 Kubernetes 命名空间，并初始化推理引擎权限和镜像拉取凭据。当前项目的命名空间已经存在时不重复创建，而是补齐所需初始化配置。

访问要求：有效 Token。当前项目尚未创建命名空间时，还需要项目作用域 Token 和 namespace/project/bootstrap 权限；命名空间已经存在时，本接口按幂等操作返回现有状态。

URI

```
POST /api/v1/tenants/current/namespace
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	是	期望的命名空间名称；服务端转换为小写，将空格、点和下划线转换为连字符，并截断为最长 63 个字符

响应消息

参数	参数类型	描述
exists	boolean	当前项目的 Beacon 命名空间是否存在
ready	boolean	当前项目的 Beacon 命名空间是否可用

请求示例

```
POST https://{endpoint}/api/v1/tenants/current/namespace
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "beacon-demo"
}
```

正常响应示例

首次创建成功：

```
HTTP/1.1 201 Created
Content-Type: application/json
```

命名空间已经存在：

```
HTTP/1.1 200 OK
Content-Type: application/json
```

两种情况的响应体相同：

```
{
  "exists": true,
  "ready": true
}
```

状态码与错误码

成功状态码：200, 201。

HTTP	error	说明
400	invalid_request	请求体格式错误，或 name 无法转换为有效的命名空间名称
401	unauthorized	authentication required，或缺少项目作用域 Token
403	forbidden	缺少命名空间初始化权限，或 Kubernetes 拒绝创建
409	conflict	指定名称已被其他命名空间占用
502	backend_unavailable	创建或初始化 Kubernetes 命名空间失败，详细原因由 message 返回
503	backend_unavailable	Kubernetes 未配置，不能创建命名空间

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取当前租户计算配额

功能介绍

获取当前项目在 Horizon 中的 eks_managed 计算资源配额、已用量和剩余量。项目和部门由 Token 确定，接口不接受其他项目 ID。

访问要求：有效 Token，并具有当前项目的计算配额读取权限。

URI

```
GET /api/v1/tenants/current/quota
```

接口请求头

请求头	说明
Cookie	可选；存在时转发给 Horizon 配额服务。使用Token认证时，仍需提供x-keystone-token；平台登录会话认证见 调用方式

请求消息

无请求体。

响应消息

参数	参数类型	描述
source	string	配额数据来源，当前为 horizon_projectquota
product	string	Horizon 云产品名称，当前为 eks_managed
quotas	map<string, object / null>	资源名称到配额信息的映射；只包含 eks_managed 产品返回的资源

参数	参数类型	描述
quotas.*.used	number	当前项目已使用数量
quotas.*.quota	number	当前项目配额上限；负数表示不限量
quotas.*.available	number	剩余数量，计算方式为 quota - used；不限量时省略
quotas.*.unlimited	boolean	是否不限量
quotas.*.unit	string	计量单位，例如 core、MiB、GiB 或 card；空值时可能省略
updatedAt	string(date-time)	配额数据获取时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/tenants/current/quota
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "source": "horizon_projectquota",
  "product": "eks_managed",
  "quotas": {
    "cpu": {
      "used": 12,
      "quota": 64,
      "available": 52,
      "unlimited": false,
      "unit": "core"
    },
    "memory": {
```

```
    "used": 32768,
    "quota": 131072,
    "available": 98304,
    "unlimited": false,
    "unit": "MiB"
  },
  "NVIDIA-H100-80GB-HBM3": {
    "used": 2,
    "quota": 8,
    "available": 6,
    "unlimited": false,
    "unit": "card"
  }
},
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
403	forbidden	当前用户无权读取项目配额
404	not_found	当前项目未配置计算配额
502	backend_unavailable	Horizon 配额查询失败，详细原因由 message 返回
503	backend_unavailable	Horizon 配额服务未配置

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.16 平台设置

平台设置

根域名变更会经过根域名校验器，并受授权检查限制；后续路由子域名展开使用该值。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/settings/route-root-domain	读取路由根域名
PUT	/settings/route-root-domain	更新路由根域名

读取路由根域名

功能介绍

读取路由根域名。

访问要求：有效 Token；需要对应资源读权限，并受项目、所有者、公共或共享资源规则限制。

URI

```
GET /api/v1/settings/route-root-domain
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
rootDomain	string	路由根域名

请求示例

```
GET https://{endpoint}/api/v1/settings/route-root-domain
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "rootDomain": "ai.example.com"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
500	internal_error	failed to read setting
401	unauthorized	missing principal

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

更新路由根域名

功能介绍

更新路由根域名。

访问要求：Token + 对应写权限，且要求 cloud_admin 角色。

URI

```
PUT /api/v1/settings/route-root-domain
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
rootDomain	string	是	路由根域名

响应消息

参数	参数类型	描述
rootDomain	string	路由根域名

请求示例

```
PUT https://{endpoint}/api/v1/settings/route-root-domain
x-keystone-token:
Content-Type: application/json
```

```
{
  "rootDomain": "ai.example.com"
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "rootDomain": "ai.example.com"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
403	forbidden	只有云管理员可以修改根域名
400	invalid_request	invalid request body
500	internal_error	setting has no validator
401	unauthorized	missing principal

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.17 消费者与API Key

消费者与 API Key

消费者是调用模型网关的逻辑主体。一个消费者可绑定一个套餐并持有多个 API Key；同一消费者下的 API Key 共享该消费者的套餐、路由授权和限流配额。

本页接口用于管理消费者和 API Key，调用平台网关调用时，使用x-keystone-token: 认证。签发的 API Key 用于调用模型网关，请求头格式为 `Authorization: bearer <API_KEY>`，不能替代本页管理接口所需的 Keystone Token。

创建消费者时同时签发首个 API Key。API Key 明文仅在签发响应中返回一次，后续查询只返回脱敏值 `masked`，请在收到响应后立即安全保存。

消费者可以暂不绑定套餐，但未绑定套餐时不能通过套餐获得任何模型网关路由访问权限。API Key 能否使用还取决于消费者状态、API Key 状态及有效期、套餐状态和套餐授权路由；禁用消费者会使其下所有 API Key 停止使用。

实时配额接口的 ids 为逗号分隔的 1 至 200 个消费者 ID，重复 ID 会被忽略，响应仅包含调用者可见的消费者。已绑定套餐但无法读取实时配额时仍返回 200，并以 `available=false` 表示配额信息暂不可用；未绑定套餐时 `available=true`，但不返回 `requests` 和 `tokens` 配额维度。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/consumers	查询 API 消费者
POST	/consumers	创建消费者并签发首个 API Key
GET	/consumers/quota	批量查询消费者实时配额
GET	/consumers/{id}	获取消费者详情
PATCH	/consumers/{id}	更新消费者
DELETE	/consumers/{id}	删除消费者
POST	/consumers/{id}/apikeys	签发新的 API Key
PATCH	/consumers/{id}/apikeys/{keyId}	启用或禁用 API Key

方法	路径（产品接口统一前缀 /api/v1）	功能
DELETE	/consumers/{id}/apikeys/{keyId}	删除 API Key

查询 API 消费者

功能介绍

查询 API 消费者。

访问要求：有效 Token 和消费者读权限。普通调用方只能查询当前项目中自己创建的消费者；具有消费者所有者管理权限时可查询项目内其他用户的消费者。全局读权限允许跨项目查询，但仍受所有者范围限制，除非同时具有消费者所有者管理权限。

URI

```
GET /api/v1/consumers
```

查询参数

参数	类型	必选	描述
page	integer	否	页码，从 1 开始
page_size	integer	否	每页数量；建议显式传入，例如 20；省略或传 0 时返回全部匹配项，响应中的 page_size 为 0

请求消息

无请求体。

响应消息

参数	参数类型	描述
----	------	----

参数	参数类型	描述
items	array	消费者列表，字段与消费者详情一致，但不包含 keys
total	integer	匹配总数
page	integer	页码，从 1 开始
page_size	integer	每页数量

请求示例

```
GET https://{endpoint}/api/v1/consumers?page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "id": "con-a1b2c3d4e5f6",
      "name": "示例消费者",
      "description": "",
      "status": "active",
      "projectId": "projectId-example",
      "ownerUserId": "ownerUserId-example",
      "spec": {
        "planId": "plan-id",
        "planKey": "pk-a1b2c3d4e5f6",
        "planName": "标准套餐",
        "status": "active",
        "keyCount": 1
      },
      "createdAt": "2026-09-21T00:00:00Z",
      "updatedAt": "2026-09-21T00:00:00Z"
    }
  ]
}
```

```
}  
],  
"total": 1,  
"page": 1,  
"page_size": 20  
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
500	internal_error	查询消费者列表失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

创建消费者并签发首个 API Key

功能介绍

创建消费者并签发首个 API Key。

访问要求：有效 Token 和当前项目的消费者写权限。绑定本项目套餐时需具有套餐使用权限；绑定其他项目共享的套餐时，还需具有共享套餐使用权限。

URI

```
POST /api/v1/consumers
```

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
----	------	------	----

参数	参数类型	是否必选	描述
name	string	是	消费者名称
description	string	否	描述
planId	string	否	绑定的套餐 ID，可暂不绑定；未绑定时不具备套餐授权的路由访问能力。
labels	map<string, string>	否	标签键值对

响应消息

参数	参数类型	描述
consumer	object	新创建的消费者对象，字段与消费者详情一致，但不包含 keys
apiKey	object	同时签发的首个 API Key；明文 key 仅在本次响应中返回
apiKey.id	string	API Key 资源 ID，用于启用、禁用和删除 API Key
apiKey.keyId	string	API Key 标识，可用于用量记录关联；不能代替 apiKey.id 作为管理接口路径参数
apiKey.key	string	API Key 明文，仅本次签发响应返回
apiKey.masked	string	API Key 脱敏值
apiKey.expiresAt	string(date-time) / null	到期时间，RFC3339；未设置有效期时省略

请求示例

```
POST https://{endpoint}/api/v1/consumers
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "示例消费者",
  "planId": "plan-id"
}
```

正常响应示例

```
HTTP/1.1 201
Content-Type: application/json
```

```
{
  "consumer": {
    "id": "con-a1b2c3d4e5f6",
    "name": "示例消费者",
    "description": "",
    "status": "active",
    "projectId": "projectId-example",
    "ownerUserId": "ownerUserId-example",
    "spec": {
      "planId": "plan-id",
      "planKey": "pk-a1b2c3d4e5f6",
      "planName": "标准套餐",
      "status": "active",
      "keyCount": 1
    },
    "createdAt": "2026-09-21T00:00:00Z",
    "updatedAt": "2026-09-21T00:00:00Z"
  },
  "apiKey": {
    "id": "ak-a1b2c3d4e5f6",
    "keyId": "key-a1b2c3d4e5f6",
    "key": "sk-beacon-EXAMPLE-NOT-A-REAL-KEY",
    "masked": "sk-beacon-...-KEY"
  }
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
401	unauthorized	authentication required
400	invalid_request	请求体格式错误、名称无效或指定的套餐不存在
403	forbidden	缺少消费者写权限或共享套餐使用权限
409	conflict	消费者资源冲突，或绑定套餐期间套餐共享关系被撤销
500	internal_error	消费者或首个 API Key 创建失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

批量查询消费者实时配额

功能介绍

批量查询消费者当前限流周期内的请求数和 Token 数配额。多个 API Key 属于同一消费者时，共享这里返回的用量和剩余额度。

访问要求：有效 Token 和消费者读权限，消费者可见范围与查询消费者接口一致。请求中不可见或不存在的消费者 ID 不会出现在响应中。

URI

```
GET /api/v1/consumers/quota
```

查询参数

参数	类型	必选	描述
ids	string	是	逗号分隔的消费者 ID，去重后须包含 1 至 200 个非空 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array	消费者实时配额列表
items[].consumerId	string	消费者 ID
items[].available	boolean	是否成功获取配额信息；未绑定套餐时也为 true，但不包含 requests 和 tokens
items[].requests	object / null	请求数配额；套餐未限制请求数、未绑定套餐或配额不可用时省略
items[].requests.limit	int64	当前配置周期内的请求数上限
items[].requests.used	int64	当前周期已使用请求数
items[].requests.remaining	int64	当前周期剩余请求数，最小为 0
items[].requests.unit	string	周期单位：minute、hour、day、month 或 year
items[].requests.period	integer	周期倍数
items[].requests.resetAt	string(date-time) / null	当前周期重置时间，RFC3339；尚未形成活跃周期时省略
items[].tokens	object / null	Token 数配额；字段含义与 requests 相同，套餐未限制 Token 数、未绑定套餐或配额不可用时省略

请求示例

```
GET https://{endpoint}/api/v1/consumers/quota?ids=consumer-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "consumerId": "con-a1b2c3d4e5f6",
      "available": true,
      "requests": {
        "limit": 100,
        "used": 12,
        "remaining": 88,
        "unit": "minute",
        "period": 1,
        "resetAt": "2026-09-21T00:01:00Z"
      },
      "tokens": {
        "limit": 100000,
        "used": 18000,
        "remaining": 82000,
        "unit": "minute",
        "period": 1,
        "resetAt": "2026-09-21T00:01:00Z"
      }
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_request	ids must contain 1 to 200 consumer ids
401	unauthorized	authentication required
500	internal_error	查询消费者或实时配额失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

获取消费者详情

功能介绍

获取消费者详情及其 API Key 脱敏列表。响应不包含任何 API Key 明文。

访问要求：有效 Token 和消费者读权限。普通调用方只能获取当前项目中自己创建的消费者；具有消费者所有者管理权限时可读取项目内其他用户的消费者。全局读权限允许跨项目读取，但仍受所有者范围限制，除非同时具有消费者所有者管理权限。

URI

```
GET /api/v1/consumers/{id}
```

参数	类型	必选	描述
id	string	是	消费者 ID

请求消息

无请求体。

响应消息

参数	参数类型	描述
----	------	----

参数	参数类型	描述
id	string	资源唯一标识
name	string	显示名称
description	string	描述
status	string	消费者状态：active 表示启用，disabled 表示禁用
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
spec	object	消费者配置快照
spec.planId	string	绑定的套餐 ID；未绑定套餐时为空字符串
spec.planKey	string	套餐内部标识；未绑定套餐时可能省略
spec.planName	string	套餐名称快照；未绑定套餐时可能省略
spec.status	string	消费者状态：active 或 disabled
spec.keyCount	integer	状态为 active 的 API Key 数量，不包含已禁用的 Key；到期但状态仍为 active 的 Key 仍可能计入
spec.ownerEmail	string	创建者登录邮箱；空值时可能省略
keys	array	API Key 脱敏列表；没有 API Key 时可能省略
keys[].id	string	API Key 资源 ID，用于启用、禁用和删除 API Key
keys[].keyId	string	API Key 标识，可用于用量记录关联；不能代替 id 作为管理接口路径参数

参数	参数类型	描述
keys[].masked	string	API Key 脱敏值，不可用于网关认证
keys[].status	string	API Key 状态：active 或 disabled
keys[].notBefore	string(date-time) / null	生效时间，RFC3339；未设置时省略
keys[].expiresAt	string(date-time) / null	到期时间，RFC3339；未设置时省略
keys[].createdAt	string(date-time)	创建时间，RFC3339
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
GET https://{endpoint}/api/v1/consumers/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "con-a1b2c3d4e5f6",
  "name": "示例消费者",
  "description": "",
  "status": "active",
```



```
"projectId": "projectId-example",
"ownerUserId": "ownerUserId-example",
"spec": {
  "planId": "plan-id",
  "planKey": "pk-a1b2c3d4e5f6",
  "planName": "标准套餐",
  "status": "active",
  "keyCount": 1
},
"keys": [
  {
    "id": "ak-a1b2c3d4e5f6",
    "keyId": "key-a1b2c3d4e5f6",
    "masked": "sk-beacon-...-KEY",
    "status": "active",
    "createdAt": "2026-09-21T00:00:00Z"
  }
],
"createdAt": "2026-09-21T00:00:00Z",
"updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
404	not_found	消费者不存在或调用方不可见
500	internal_error	查询消费者失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

更新消费者

功能介绍

更新消费者名称、绑定套餐或启用状态。变更套餐或消费者状态后，同一消费者下所有 API Key 的网关授权会按新配置重新生效。重新启用消费者不会同时启用此前被单独禁用的 API Key。

访问要求：有效 Token 和当前项目的消费者写权限。普通调用方只能修改自己创建的消费者；具有消费者所有者管理权限的调用方可修改项目内其他用户的消费者。绑定其他项目共享的套餐时，还需具有共享套餐使用权限；不能修改其他项目的消费者。

URI

```
PATCH /api/v1/consumers/{id}
```

参数	类型	必选	描述
id	string	是	消费者 ID

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
name	string	否	消费者名称；提供时不能为空
planId	string	否	绑定的套餐 ID；传空字符串解除套餐绑定，省略表示不修改；不要传 null，当前实现会将 null 视为未提供
status	string	否	消费者状态：active 表示启用，disabled 表示禁用

响应消息

参数	参数类型	描述
id	string	资源唯一标识

参数	参数类型	描述
name	string	显示名称
description	string	描述
status	string	消费者状态：active 或 disabled
projectId	string	所属项目 ID
projectName	string	项目名称；空值时可能省略
domainId	string	所属部门 ID；空值时可能省略
domainName	string	部门名称；空值时可能省略
ownerUserId	string	创建者用户 ID
ownerName	string	创建者名称；空值时可能省略
spec	object	消费者配置快照，字段含义见获取消费者详情
labels	map<string, string>	标签键值对；空值时可能省略
createdAt	string(date-time)	创建时间，RFC3339
updatedAt	string(date-time)	更新时间，RFC3339

请求示例

```
PATCH https://{endpoint}/api/v1/consumers/resource-id
x-keystone-token:
Content-Type: application/json
```

```
{
  "name": "更新后的名称"
}
```

正常响应示例

HTTP/1.1 200
Content-Type: application/json

```
{
  "id": "con-a1b2c3d4e5f6",
  "name": "更新后的名称",
  "description": "",
  "status": "active",
  "projectId": "projectId-example",
  "ownerUserId": "ownerUserId-example",
  "spec": {
    "planId": "plan-id",
    "planKey": "pk-a1b2c3d4e5f6",
    "planName": "标准套餐",
    "status": "active",
    "keyCount": 1
  },
  "createdAt": "2026-09-21T00:00:00Z",
  "updatedAt": "2026-09-21T00:00:00Z"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
400	invalid_request	请求体格式错误、名称为空或指定的套餐不存在
403	forbidden	缺少消费者写权限、尝试修改其他项目资源或缺少共享套餐使用权限
404	not_found	消费者不存在或调用方无权访问
409	conflict	消费者资源冲突，或绑定套餐期间套餐共享关系被撤销
500	internal_error	更新消费者失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除消费者

功能介绍

删除消费者及其全部 API Key。删除后，这些 API Key 不能再用于调用模型网关，且无法通过查询接口恢复。

访问要求：有效 Token 和当前项目的消费者写权限。普通调用方只能删除自己创建的消费者；具有消费者所有者管理权限的调用方可删除项目内其他用户的消费者；不能删除其他项目的消费者。

URI

```
DELETE /api/v1/consumers/{id}
```

参数	类型	必选	描述
id	string	是	消费者 ID

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/consumers/resource-id
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
401	unauthorized	authentication required
403	forbidden	缺少消费者写权限或尝试删除其他项目的消费者
404	not_found	消费者不存在或调用方无权访问
500	internal_error	删除消费者或其 API Key 失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

签发新的 API Key

功能介绍

为指定消费者签发新的 API Key。该操作只新增 Key，不会自动禁用或删除已有 Key；用于密钥轮换时，应在新 Key 验证可用后再禁用或删除旧 Key。

访问要求：有效 Token 和当前项目的消费者写权限。普通调用方只能为自己创建的消费者签发 Key；具有消费者所有者管理权限的调用方可管理项目内其他用户的消费者；不能为其他项目的消费者签发 Key。

URI

```
POST /api/v1/consumers/{id}/apikeys
```

参数	类型	必选	描述
id	string	是	消费者 ID

请求消息

请求体可选。提供请求体时，Content-Type 为 application/json。

参数	参数类型	是否必选	描述
expiresInHours	integer	否	密钥有效小时数；大于 0 时设置到期时间，省略、传 0 或负数时不设置到期时间

响应消息

参数	参数类型	描述
id	string	API Key 资源 ID，用于启用、禁用和删除 API Key
keyId	string	API Key 标识，可用于用量记录关联；不能代替 id 作为管理接口路径参数
key	string	API Key 明文，仅本次签发响应返回
masked	string	API Key 脱敏值
expiresAt	string(date-time) / null	到期时间；空值时可能省略

请求示例

```
POST https://{endpoint}/api/v1/consumers/resource-id/apikeys
x-keystone-token:
Content-Type: application/json
```

```
{
  "expiresInHours": 24
}
```

正常响应示例

```
HTTP/1.1 201
Content-Type: application/json
```

```
{
  "id": "ak-a1b2c3d4e5f6",
  "keyId": "key-a1b2c3d4e5f6",
  "key": "sk-beacon-EXAMPLE-NOT-A-REAL-KEY",
  "masked": "sk-beacon-...-KEY",
  "expiresAt": "2026-09-22T00:00:00Z"
}
```

状态码与错误码

成功状态码：201。

HTTP	error	说明
401	unauthorized	authentication required
403	forbidden	缺少消费者写权限或尝试操作其他项目的消费者
404	not_found	消费者不存在或调用方无权访问
500	internal_error	API Key 签发失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

启用或禁用 API Key

功能介绍

启用或禁用指定 API Key。禁用后，该 Key 不能再调用模型网关；重新启用不会更换明文，也不会绕过消费者、套餐和有效期状态检查。

访问要求：有效 Token 和当前项目的消费者写权限。普通调用方只能管理自己创建的消费者下的 Key；具有消费者所有者管理权限的调用方可管理项目内其他用户的消费者；不能操作其他项目的消费者或不属于该消费者的 Key。

URI


```
PATCH /api/v1/consumers/{id}/apikeys/{keyId}
```

参数	类型	必选	描述
id	string	是	消费者 ID
keyId	string	是	API Key 资源 ID，对应消费者详情中 keys[].id，而不是 keys[].keyId

请求消息

Content-Type: application/json。

参数	参数类型	是否必选	描述
status	string	是	API Key 状态：active 表示启用，disabled 表示禁用

响应消息

参数	参数类型	描述
id	string	API Key 资源 ID
status	string	API Key 状态：active 或 disabled

请求示例

```
PATCH https://{endpoint}/api/v1/consumers/con-a1b2c3d4e5f6/apikeys/ak-a1b2c3d4e5f6
x-keystone-token:
Content-Type: application/json
```

```
{
  "status": "disabled"
}
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "id": "ak-a1b2c3d4e5f6",
  "status": "disabled"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
400	invalid_request	请求体格式错误
403	forbidden	缺少消费者写权限或尝试操作其他项目的消费者
404	not_found	消费者不存在、调用方无权访问，或 Key 不属于该消费者
500	internal_error	更新 API Key 状态失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

删除 API Key

功能介绍

永久删除指定 API Key。删除后，该 Key 不能再用于调用模型网关，也不能重新启用；需要恢复访问时只能重新签发 Key。

访问要求：有效 Token 和当前项目的消费者写权限。普通调用方只能管理自己创建的消费者下的 Key；具有消费者所有者管理权限的调用方可管理项目内其他用户的消费者；不能操作其他项目的消费者或不属于该消费者的 Key。

URI

```
DELETE /api/v1/consumers/{id}/apikeys/{keyId}
```

参数	类型	必选	描述
id	string	是	消费者 ID
keyId	string	是	API Key 资源 ID，对应消费者详情中 keys[].id，而不是 keys[].keyId

请求消息

无请求体。

响应消息

成功返回 204 No Content，无响应体。

请求示例

```
DELETE https://{endpoint}/api/v1/consumers/con-a1b2c3d4e5f6/apikeys/ak-a1b2c3d4e5f6
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 204 No Content
```

状态码与错误码

成功状态码：204。

HTTP	error	说明
401	unauthorized	authentication required
403	forbidden	缺少消费者写权限或尝试操作其他项目的消费者
404	not_found	消费者不存在、调用方无权访问，或 Key 不属于该消费者
500	internal_error	删除 API Key 失败

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.18 AI 加速卡资源

AI 加速卡资源

查询集群中可用于模型推理的 AI 加速卡型号及容量概览。接口按加速卡型号、Kubernetes 扩展资源名称和节点架构汇总节点可分配数量与 Pod 已申请数量，可用于选择推理引擎的加速卡型号，并判断单个工作负载可申请的卡数。

接口数据来自 Kubernetes 节点和 Pod，不表示加速卡的实时计算利用率。`available` 是集群维度的剩余数量，`maxAvailablePerNode` 是单个节点上的最大剩余数量；二者均不等同于当前项目的资源配额，也不保证工作负载一定能够立即调度。

未配置 Kubernetes 客户端时返回空列表；读取 Kubernetes 资源失败时返回 502。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/gpu/inventory	查询集群 AI 加速卡资源

查询集群 AI 加速卡资源

功能介绍

按型号查询集群中的 AI 加速卡资源及可用数量。支持平台已配置或自动发现的 GPU、HCU、NPU 和 NVIDIA MIG 等 Kubernetes 扩展资源。

访问要求：有效 Token。接口仅返回集群级聚合数量和调度信息，不包含节点、Pod 或租户明细，所有已认证用户均可查询。

URI

```
GET /api/v1/gpu/inventory
```

请求消息

无请求体。

响应消息

参数	参数类型	描述
items	array	AI 加速卡资源列表；未配置 Kubernetes 或集群中未发现支持的加速卡时为空数组
items[].product	string	加速卡型号或资源显示名称，例如 NVIDIA-H100-80GB-HBM3
items[].architecture	string	加速卡所在节点的处理器架构，例如 amd64 或 arm64；无法识别时可能省略
items[].resourceName	string	Kubernetes 扩展资源名称，例如 nvidia.com/gpu、hygon.com/hcu 或 huawei.com/Ascend910
items[].nodeSelector	map<string, string>	用于将工作负载调度到对应型号和架构节点的标签选择器
items[].total	int64	集群中该型号加速卡的可分配总数
items[].used	int64	未终止 Pod 的普通容器已申请数量之和，不表示实时硬件利用率
items[].available	int64	集群剩余数量，计算方式为 $\max(\text{total} - \text{used}, 0)$
items[].maxAvailablePerNode	int64	任一单个节点上的最大剩余数量；单个 Pod 申请多张卡时应重点参考该值

请求示例

```
GET https://{endpoint}/api/v1/gpu/inventory
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "items": [
    {
      "product": "NVIDIA-H100-80GB-HBM3",
      "architecture": "amd64",
      "resourceName": "nvidia.com/gpu",
      "nodeSelector": {
        "nvidia.com/gpu.product": "NVIDIA-H100-80GB-HBM3",
        "kubernetes.io/arch": "amd64"
      },
      "total": 8,
      "used": 3,
      "available": 5,
      "maxAvailablePerNode": 4
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
401	unauthorized	authentication required
502	backend_unavailable	无法读取 Kubernetes 节点或 Pod 资源，详细原因由 message 返回

后端错误与错误封装见[调用方式](#)。

1.19 Token用量

Token 用量

Token 用量接口用于查询通过模型网关产生的请求数、输入 Token、输出 Token、总 Token 和缓存命中 Token。用量数据由网关异步写入统计后端，可能与刚完成的请求存在短暂延迟；它表示实际调用记录，不等于套餐中的实时限流剩余额度。

用量采集采用不阻塞模型请求的故障开放策略。统计通道或后端异常不会中断模型调用，但可能造成部分记录延迟或缺失，因此该接口适合运营分析和用量观察，不应作为唯一的财务计费或合规审计依据。

查询范围由 Token 身份和用量聚合权限确定，查询参数只能缩小范围，不能扩大权限：member 仅能查询当前项目中自己创建的消费者产生的用量；project_admin 可查询当前项目；domain_admin 可查询当前部门；cloud_admin 可查询全部部门和项目。

start 和 end 支持 RFC3339 或 YYYY-MM-DD，查询范围为 [start, end)；省略时默认查询截至当前时间的最近 7 天。eventTime 按北京时间自然日筛选，格式为 YYYY-MM-DD。分页默认第 1 页、每页 50 条，最多 200 条；page_size 优先于兼容参数 limit。

为兼容现有页面，字段 apiKeyId 和 apiKeyName 实际表示消费者 ID 和消费者名称，即同一消费者下多个 API Key 共享的调用身份；单次请求实际使用的 Key 标识由请求明细中的 keyId 返回。

当前项目的 CPU、内存和 AI 加速卡资源配额请参见[租户与计算配额](#)；消费者套餐的实时限流剩余额度请参见[消费者与 API Key](#)。

方法	路径（产品接口统一前缀 /api/v1）	功能
GET	/tenants/current/usage	查询 Token 用量概览
GET	/tenants/current/usage/recent	查询最近请求用量明细
GET	/tenants/current/usage/model-daily	查询模型每日用量趋势

查询 Token 用量概览

功能介绍

查询授权范围内的 Token 用量汇总，并按部门、项目、用户、消费者、路由域名、模型和请求类型组成的完整维度进行分组。相同模型由不同消费者、用户或路由调用时会返回多条记录。

访问要求：有效 Token。返回范围由 member、project_admin、domain_admin 或 cloud_admin 用量权限确定。

URI

```
GET /api/v1/tenants/current/usage
```

查询参数

参数	类型	必选	描述
start	string	否	开始时间，RFC3339 或 YYYY-MM-DD；包含该时刻
end	string	否	结束时间，RFC3339 或 YYYY-MM-DD；不包含该时刻
limit	integer	否	兼容的每页数量参数，必须为正整数；page_size 优先
page	integer	否	聚合结果页码，从 1 开始，缺省为 1
page_size	integer	否	聚合结果每页数量，必须为正整数，缺省为 limit 或 50，最大为 200
apiKeyId	string	否	消费者 ID 精确筛选；字段名为历史兼容名称
domainId	string	否	部门 ID 精确筛选
projectId	string	否	项目 ID 精确筛选
userId	string	否	消费者创建者用户 ID 精确筛选
model	string	否	对外模型名称精确筛选
domainLike	string	否	调用方可见部门名称的模糊筛选

参数	类型	必选	描述
projectLike	string	否	调用方可见项目名称的模糊筛选
userLike	string	否	消费者创建者名称的模糊筛选，不区分大小写
modelLike	string	否	对外模型名称的模糊筛选，不区分大小写
apiKeyName	string	否	消费者名称的模糊筛选，不区分大小写； 字段名为历史兼容名称
host	string	否	路由域名精确筛选
hostLike	string	否	路由域名模糊筛选，不区分大小写
requestType	string	否	请求类型精确筛选，例如 ai_chat、ai_completion、embeddings、responses、rerank、image、speech 或 transcription
status	integer	否	HTTP 响应状态码精确筛选，必须为正整数
eventTime	string	否	按北京时间自然日筛选，格式为 YYYY-MM-DD
sortField	string	否	排序字段；支持 totalTokens、requestCount、promptTokens、completionTokens、cacheHitTokens、domainId、projectId、userId、userName、apiKeyId、apiKeyName、host、model、requestType； 缺省为 totalTokens
sortValue	string	否	排序方向：ascend 或 descend，缺省为 descend

请求消息

无请求体。

响应消息

参数	参数类型	描述
start	string(date-time)	实际查询开始时间，RFC3339
end	string(date-time)	实际查询结束时间，RFC3339
accessLevel	string	实际权限范围：member、project_admin、domain_admin 或 cloud_admin
domainId	string	请求中的部门 ID 筛选；未指定时空字符串
projectId	string	请求中的项目 ID 筛选；未指定时空字符串
userId	string	请求中的用户 ID 筛选；未指定时空字符串
apiKeyId	string	请求中的消费者 ID 筛选；未指定时空字符串
apiKeyName	string	apiKeyId 对应的消费者名称；无法解析或未指定时空字符串
domainOptions	array	可用部门筛选项；仅 cloud_admin 返回可选项
projectOptions	array	可用项目筛选项；domain_admin 和 cloud_admin 返回可选项
userOptions	array	可用用户筛选项；project_admin、domain_admin 和 cloud_admin 返回可选项
apiKeyOptions	array	授权范围内可用的消费者筛选项

参数	参数类型	描述
domainOptions[].id / projectOptions[].id / userOptions[].id	string	筛选项 ID
domainOptions[].name / projectOptions[].name / userOptions[].name	string	筛选项名称
domainOptions[].displayName / projectOptions[].displayName / userOptions[].displayName	string	用于界面展示的名称；重名时可能包含 ID 后缀
apiKeyOptions[].id	string	消费者 ID
apiKeyOptions[].name	string	消费者名称
apiKeyOptions[].displayName	string	消费者显示名称
summary	object	当前筛选范围的用量汇总，不受聚合列表分页影响
summary.requestCount	int64	请求总数
summary.promptTokens	int64	输入 Token 总数
summary.completionTokens	int64	输出 Token 总数
summary.totalTokens	int64	Token 总数
summary.cacheHitTokens	int64	缓存命中 Token 总数
byModel	array	按完整调用维度聚合的用量列表
byModel[].domainId	string	部门 ID
byModel[].domainName	string	部门名称
byModel[].projectId	string	项目 ID
byModel[].projectName	string	项目名称
byModel[].userId	string	消费者创建者用户 ID
byModel[].userName	string	消费者创建者名称

参数	参数类型	描述
byModel[].apiKeyId	string	消费者 ID；字段名为历史兼容名称
byModel[].apiKeyName	string	消费者名称；字段名为历史兼容名称
byModel[].host	string	路由域名
byModel[].model	string	对外模型名称
byModel[].requestType	string	请求类型
byModel[].requestCount	int64	请求数
byModel[].promptTokens	int64	输入 Token 数
byModel[].completionTokens	int64	输出 Token 数
byModel[].totalTokens	int64	Token 总数
byModel[].cacheHitTokens	int64	缓存命中 Token 数
byModelTotal	int64	聚合结果总数
byModelPage	integer	聚合结果页码
byModelPageSize	integer	聚合结果每页数量
sortField	string	实际使用的排序字段
sortValue	string	实际使用的排序方向：ascend 或 descend

请求示例

```
GET https://{endpoint}/api/v1/tenants/current/usage?start=2026-09-15&end=2026-09-22&page=1&page_size=20
x-keystone-token:
```

正常响应示例

HTTP/1.1 200
Content-Type: application/json

```
{
  "start": "2026-09-15T00:00:00Z",
  "end": "2026-09-22T00:00:00Z",
  "accessLevel": "member",
  "domainId": "",
  "projectId": "",
  "userId": "",
  "apiKeyId": "",
  "apiKeyName": "",
  "domainOptions": [],
  "projectOptions": [],
  "userOptions": [],
  "apiKeyOptions": [
    {
      "id": "con-a1b2c3d4e5f6",
      "name": "生产应用",
      "displayName": "生产应用"
    }
  ],
  "summary": {
    "requestCount": 120,
    "promptTokens": 240000,
    "completionTokens": 60000,
    "totalTokens": 300000,
    "cacheHitTokens": 80000
  },
  "byModel": [
    {
      "domainId": "domain-a1b2c3d4",
      "domainName": "研发部门",
      "projectId": "project-a1b2c3d4",
      "projectName": "生产项目",
      "userId": "user-a1b2c3d4",
      "userName": "示例用户",
      "apiKeyId": "con-a1b2c3d4e5f6",
      "apiKeyName": "生产应用",
      "host": "api.example.com",
```

```
{
  "model": "qwen2.5-72b",
  "requestType": "ai_chat",
  "requestCount": 120,
  "promptTokens": 240000,
  "completionTokens": 60000,
  "totalTokens": 300000,
  "cacheHitTokens": 80000
},
{
  "byModelTotal": 1,
  "byModelPage": 1,
  "byModelPageSize": 20,
  "sortField": "totalTokens",
  "sortValue": "descend"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_query	时间、分页、状态码、排序方向或排序字段无效，详细原因由 message 返回
401	unauthorized	authentication required
502	backend_unavailable	查询 Token 用量失败，详细原因由 message 返回
503	backend_unavailable	Token 用量后端未配置

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询最近请求用量明细

功能介绍

分页查询授权范围内的模型网关请求明细，包括调用身份、模型、请求类型、HTTP 状态码和 Token 数量。

访问要求：有效 Token。查询范围与 Token 用量概览接口一致。

URI

```
GET /api/v1/tenants/current/usage/recent
```

查询参数

参数	类型	必选	描述
start	string	否	开始时间，RFC3339 或 YYYY-MM-DD；包含该时刻
end	string	否	结束时间，RFC3339 或 YYYY-MM-DD；不包含该时刻
limit	integer	否	兼容的每页数量参数，必须为正整数；page_size 优先
page	integer	否	请求明细页码，从 1 开始，缺省为 1
page_size	integer	否	请求明细每页数量，必须为正整数，缺省为 limit 或 50，最大为 200
apiKeyId	string	否	消费者 ID 精确筛选；字段名为历史兼容名称
domainId	string	否	部门 ID 精确筛选
projectId	string	否	项目 ID 精确筛选
userId	string	否	消费者创建者用户 ID 精确筛选
model	string	否	对外模型名称精确筛选
domainLike	string	否	调用方可见部门名称的模糊筛选
projectLike	string	否	调用方可见项目名称的模糊筛选

参数	类型	必选	描述
userLike	string	否	消费者创建者名称的模糊筛选，不区分大小写
modelLike	string	否	对外模型名称的模糊筛选，不区分大小写
apiKeyName	string	否	消费者名称的模糊筛选，不区分大小写；字段名为历史兼容名称
host	string	否	路由域名精确筛选
hostLike	string	否	路由域名模糊筛选，不区分大小写
requestType	string	否	请求类型精确筛选
status	integer	否	HTTP 响应状态码精确筛选，必须为正整数
eventTime	string	否	按北京时间自然日筛选，格式为 YYYY-MM-DD
sortField	string	否	排序字段；支持 eventTime、totalTokens、promptTokens、completionTokens、cacheHitTokens、status、requestId、domainId、projectId、userId、userName、apiKeyName、consumerId、keyId、planId、model、host、requestType；缺省为 eventTime
sortValue	string	否	排序方向：ascend 或 descend，缺省为 descend

请求消息

无请求体。

响应消息

参数	参数类型	描述
start	string(date-time)	实际查询开始时间, RFC3339
end	string(date-time)	实际查询结束时间, RFC3339
recent	array	请求明细列表
recent[].requestId	string	网关请求 ID
recent[].domainId	string	部门 ID
recent[].domainName	string	部门名称
recent[].projectId	string	项目 ID
recent[].projectName	string	项目名称
recent[].userId	string	消费者创建者用户 ID
recent[].userName	string	消费者创建者名称
recent[].apiKeyName	string	消费者名称; 字段名为历史兼容名称
recent[].consumerId	string	消费者 ID
recent[].keyId	string	本次请求使用的 API Key 标识, 对应签发响应中的 keyId, 不是 API Key 资源 ID
recent[].planId	string	请求发生时关联的套餐 ID
recent[].model	string	对外模型名称
recent[].host	string	路由域名
recent[].requestType	string	请求类型

参数	参数类型	描述
recent[].promptTokens	int64	输入 Token 数
recent[].completionTokens	int64	输出 Token 数
recent[].totalTokens	int64	Token 总数
recent[].cacheHitTokens	int64 / null	缓存命中 Token 数；上游未提供时为 null
recent[].status	integer	HTTP 响应状态码
recent[].eventTime	string(date-time)	请求发生时间，RFC3339
recentTotal	int64	匹配的请求明细总数
recentPage	integer	请求明细页码
recentPageSize	integer	请求明细每页数量
sortField	string	实际使用的排序字段
sortValue	string	实际使用的排序方向：ascend 或 descend

请求示例

```
GET https://{endpoint}/api/v1/tenants/current/usage/recent?start=2026-09-15&end=2026-09-22&page=1&page_size=20
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "start": "2026-09-15T00:00:00Z",
  "end": "2026-09-22T00:00:00Z",
  "recent": [
    {
      "requestId": "req-a1b2c3d4",
      "domainId": "domain-a1b2c3d4",
      "domainName": "研发部门",
      "projectId": "project-a1b2c3d4",
      "projectName": "生产项目",
      "userId": "user-a1b2c3d4",
      "userName": "示例用户",
      "apiKeyName": "生产应用",
      "consumerId": "con-a1b2c3d4e5f6",
      "keyId": "key-a1b2c3d4e5f6",
      "planId": "plan-a1b2c3d4",
      "model": "qwen2.5-72b",
      "host": "api.example.com",
      "requestType": "ai_chat",
      "promptTokens": 2000,
      "completionTokens": 500,
      "totalTokens": 2500,
      "cacheHitTokens": 800,
      "status": 200,
      "eventTime": "2026-09-21T08:30:00Z"
    }
  ],
  "recentTotal": 120,
  "recentPage": 1,
  "recentPageSize": 20,
  "sortField": "eventTime",
  "sortValue": "descend"
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
------	-------	----

HTTP	error	说明
400	invalid_query	时间、分页、状态码、排序方向或排序字段无效，详细原因由 message 返回
401	unauthorized	authentication required
502	backend_unavailable	查询请求用量明细失败，详细原因由 message 返回
503	backend_unavailable	Token 用量后端未配置

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

查询模型每日用量趋势

功能介绍

查询指定模型在授权范围内的每日请求数和 Token 用量。日期按北京时间划分，只返回存在用量记录的日期，不自动补充零值日期。

访问要求：有效 Token。查询范围与 Token 用量概览接口一致。

URI

```
GET /api/v1/tenants/current/usage/model-daily
```

查询参数

参数	类型	必选	描述
model	string	是	对外模型名称，精确匹配
start	string	否	开始时间，RFC3339 或 YYYY-MM-DD；包含该时刻

参数	类型	必选	描述
end	string	否	结束时间，RFC3339 或 YYYY-MM-DD；不包含该时刻
apiKeyId	string	否	消费者 ID 精确筛选；字段名为历史兼容名称
domainId	string	否	部门 ID 精确筛选
projectId	string	否	项目 ID 精确筛选
userId	string	否	消费者创建者用户 ID 精确筛选
domainLike	string	否	调用方可见部门名称的模糊筛选
projectLike	string	否	调用方可见项目名称的模糊筛选
userLike	string	否	消费者创建者名称的模糊筛选，不区分大小写
apiKeyName	string	否	消费者名称的模糊筛选，不区分大小写；字段名为历史兼容名称
host	string	否	路由域名精确筛选
hostLike	string	否	路由域名模糊筛选，不区分大小写
requestType	string	否	请求类型精确筛选
status	integer	否	HTTP 响应状态码精确筛选，必须为正整数
eventTime	string	否	按北京时间自然日进一步筛选，格式为 YYYY-MM-DD

请求消息

无请求体。

响应消息

参数	参数类型	描述
start	string(date-time)	实际查询开始时间，RFC3339
end	string(date-time)	实际查询结束时间，RFC3339
model	string	对外模型名称
series	array	按北京时间自然日升序排列的用量数据，只包含存在记录的日期
series[].date	string	日期，格式为 YYYY-MM-DD
series[].requestCount	int64	当日请求数
series[].promptTokens	int64	当日输入 Token 数
series[].cacheHitTokens	int64	当日缓存命中 Token 数
series[].completionTokens	int64	当日输出 Token 数
series[].totalTokens	int64	当日 Token 总数

请求示例

```
GET https://{endpoint}/api/v1/tenants/current/usage/model-daily?
model=qwen2.5-72b&start=2026-09-15&end=2026-09-22
x-keystone-token:
```

正常响应示例

```
HTTP/1.1 200
Content-Type: application/json
```

```
{
  "start": "2026-09-15T00:00:00Z",
  "end": "2026-09-22T00:00:00Z",
  "model": "qwen2.5-72b",
  "series": [
    {
      "date": "2026-09-20",
      "requestCount": 45,
      "promptTokens": 90000,
      "cacheHitTokens": 30000,
      "completionTokens": 22000,
      "totalTokens": 112000
    },
    {
      "date": "2026-09-21",
      "requestCount": 75,
      "promptTokens": 150000,
      "cacheHitTokens": 50000,
      "completionTokens": 38000,
      "totalTokens": 188000
    }
  ]
}
```

状态码与错误码

成功状态码：200。

HTTP	error	说明
400	invalid_query	model 缺失，或时间、状态码等查询参数无效，详细原因由 message 返回
401	unauthorized	authentication required
502	backend_unavailable	查询模型每日用量失败，详细原因由 message 返回
503	backend_unavailable	Token 用量后端未配置

此外，认证/授权中间件可返回 401 / 403；后端错误与错误封装见[调用方式](#)。

1.20 发布记录

01 <2026-09-23>

变更类型	变更说明
首次发布	* API 简介、调用方式：说明平台术语、API 接入方式。 * 租户和配额、Token 用量：提供项目上下文、资源配额和模型调用用量查询接口。 * 模型广场、模型仓库、模型导入、模型上传：提供模型登记、存储、远程导入及本地上传接口。 * 推理引擎、AI 加速卡资源、容器镜像、模型部署：提供部署所需资源的查询与管理接口。 * 服务提供者、模型网关、套餐管理、消费者与API Key：提供模型服务接入、路由配置和调用权限管理接口。 * 模型体验、基准测试：提供模型能力体验和部署性能基准测试接口。 * 平台设置：提供平台配置查询与更新接口。

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)