# 云原生智能网关

# AI 网关

腾讯云

# 文档目录

# AI 网关
# AI 网关概述

最近更新时间：2026-07-23 17:07:31

**AI 网关**是腾讯云智能网关推出的面向大模型与智能化场景的新一代网关产品。它专注于解决企业接入、调度和管理多种 AI 模型时面临的协议复杂、治理困难、成本不可控及存量业务改造门槛高等核心问题。
AI 网关作为企业智能化架构的流量入口与治理中枢，通过统一的协议适配、智能的路由调度与全方位的可观测能力，帮助企业高效、安全、经济地集成与使用 AI 能力，加速业务创新与智能化转型。

## 产品特色

- **智能模型治理：**统一接入并智能调度腾讯云混元、开源模型及第三方商业模型，通过负载均衡、熔断降级与成本优化策略，实现性能、成本与稳定性的最佳平衡。
- **存量业务快速 AI 化：**内置强大的协议转换引擎，支持 MCP、OpenAI 等 AI 生态协议与传统 HTTP/gRPC 等业务协议的双向转换，助力存量业务系统快速具备 AI 能力，有效保护企业现有 IT 投资。
- **全链路安全合规：**构建从接入认证、参数过滤到数据脱敏的多层次安全防护体系，集成 WAF、DDoS 防护等能力，保障 AI 应用合规、安全、可靠地运行。
- **企业级高可用保障：**采用多可用区高可用部署架构，支持自动故障转移与实例弹性扩缩容，服务可用性有保障。

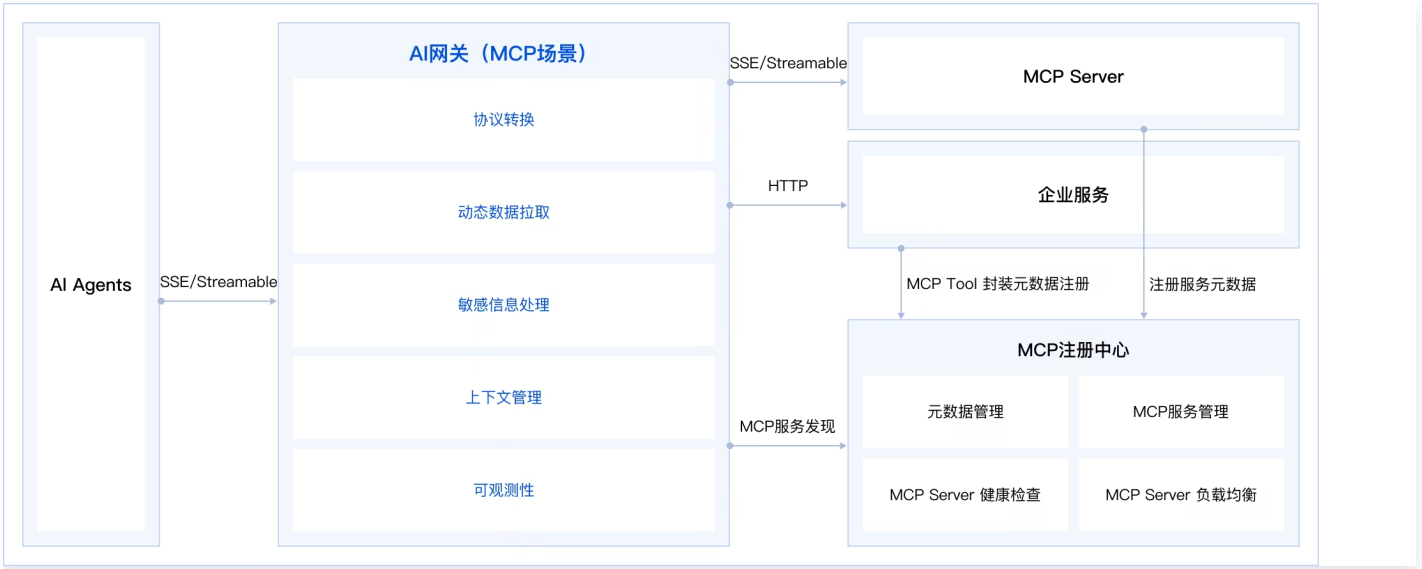
## 业务场景

### 场景一：多模型统一治理与智能调度

- **核心问题：**企业需同时使用多个 AI 模型，但缺乏统一平台，导致模型使用混乱、资源配置低效、成本不可控。
- **解决方案：**AI 网关作为统一入口，实现多模型的可视化接入与管理。通过智能路由策略，根据请求内容、模型性能、成本等因素自动选择最优模型，并配合限流、熔断机制保障服务稳定性，实现降本增效。


## 场景二：存量业务系统 AI 化改造

- **核心问题**：企业存量业务系统沉淀了大量核心数据与业务逻辑，是企业最有价值的资产。但这些系统普遍以 REST/HTTP 接口形式对外提供服务，无法直接被 AI Agent 以 MCP 协议调用，导致 AI 能力难以快速与现有业务能力打通，存量系统的价值无法在 AI 时代得到充分释放。
- **解决方案**：AI 网关通过协议转换引擎，将存量业务系统提供的标准 API 自动包装成 AI 应用可调用的标准化工具（MCP Tool），实现业务能力的“零代码改造” AI 化。开发人员无需关注底层集成细节，即可快速构建智能应用。




# 功能特性

| 功能点 | 说明 |
| --- | --- |
| 统一协议接入 | 100%兼容开源网关生态，并全面适配 AI 领域标准协议，支持 MCP、OpenAI 、SSE 等，提供传统 RESTful、gRPC 等协议的无缝转换，实现一套网关覆盖所有流量。 |
| 模型服务管理 | 提供模型服务的全生命周期管理，支持配置多模型供应商的密钥、API 端点。提供模型级流量控制、Fallback 容灾与精细化监控。 |
| 智能路由与编排 | 支持基于内容语义、成本、性能等策略的智能路由。可编排串联多个模型调用或业务 API，完成复杂任务。 |
| 精细化流量治理 | 提供从消费者、API 到模型等多个维度的限流、熔断、降级能力，保障后端服务与模型 API 的稳定性。 |
| 开箱即用的安全防护 | 集成认证鉴权、访问控制、敏感信息脱敏、防重放攻击等安全能力，提供企业级安全保障，满足合规要求。 |
| 全链路可观测性 | 提供从用户请求到模型响应的全链路追踪，监控 API 调用延迟、Token 消耗、模型费用等多维度指标，并支持智能诊断与告警，助力运维与成本优化。 |
| 细粒度的权限管理 | 通过消费者、消费者组的多级权限模型，实现 AI 能力在不同团队、项目间的安全隔离与便捷共享，支撑平台化运营。 |

# 快速入门
# 通过 AI 网关调用混元 API

最近更新时间：2026-07-23 17:07:31

AI 网关调用混元 API 是通过统一接口管理多个 AI 模型访问，实现高效、安全、可扩展的 AI 能力集成。它简化了混元 API 的调用流程，提供流量控制、监控和计费等一站式服务。本文将快速引导您完成首次配置，体验通过 AI 网关调用混元大模型服务的完整流程。您将依次完成模型密钥、模型服务、模型 API 的配置，并创建调用方身份（消费者与消费者组），最终通过网关成功发起调用。

## 前提条件

1. 已创建 AI 网关实例，具体操作请参见 新建 AI 网关 。
2. 已通过混元官方平台，获取混元大模型的 API 调用密钥。

## 操作概述

核心操作流程如下，您将依次完成以下六个步骤：

1. **创建模型密钥**：在网关中安全配置访问混元所需的 API 密钥。
2. **创建模型服务**：添加混元作为 AI 模型供应商，并关联上一步创建的密钥。
3. **创建模型 API**：创建一个对外提供文本生成能力的 API ，并绑定第2步创建的服务。
4. **创建消费者**：创建代表您自身应用的调用方，并为其添加身份凭证（ API Key）。
5. **创建消费者组并授权**：将消费者分组，并为该组授予访问第3步所创模型 API 的权限。
6. **获取地址并发起调用**：在控制台找到 API 的访问地址，使用消费者的凭证发起调用请求。

接下来，请按照下述详细说明完成每一步操作。

## 操作步骤

### 步骤一：创建模型密钥

模型密钥用于安全地存储和管理访问第三方大模型服务所需的凭证。

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**，选择需要使用的实例，进入实例详情。
2. 在左侧导航栏，单击**密钥管理**。
3. 在“密钥管理”页面，单击**新建**。
4. 在“新建密钥”窗口中，参考下表配置密钥信息。

| 参数 | 填写说明 |
| --- | --- |
| 密钥类型 | 选择**模型密钥**。 |

| | |
|---|---|
| 密钥名称 | 自定义名称，如 hunyuan-key。 |
| 生成方式 | 选择**自定义**。 |
| 凭证内容 | 此处填入您从混元官方获取的 API 密钥。 |
| 描述 | （可选）输入描述信息。 |

5. 单击 **确定**，完成密钥创建。


## 步骤二：创建模型服务

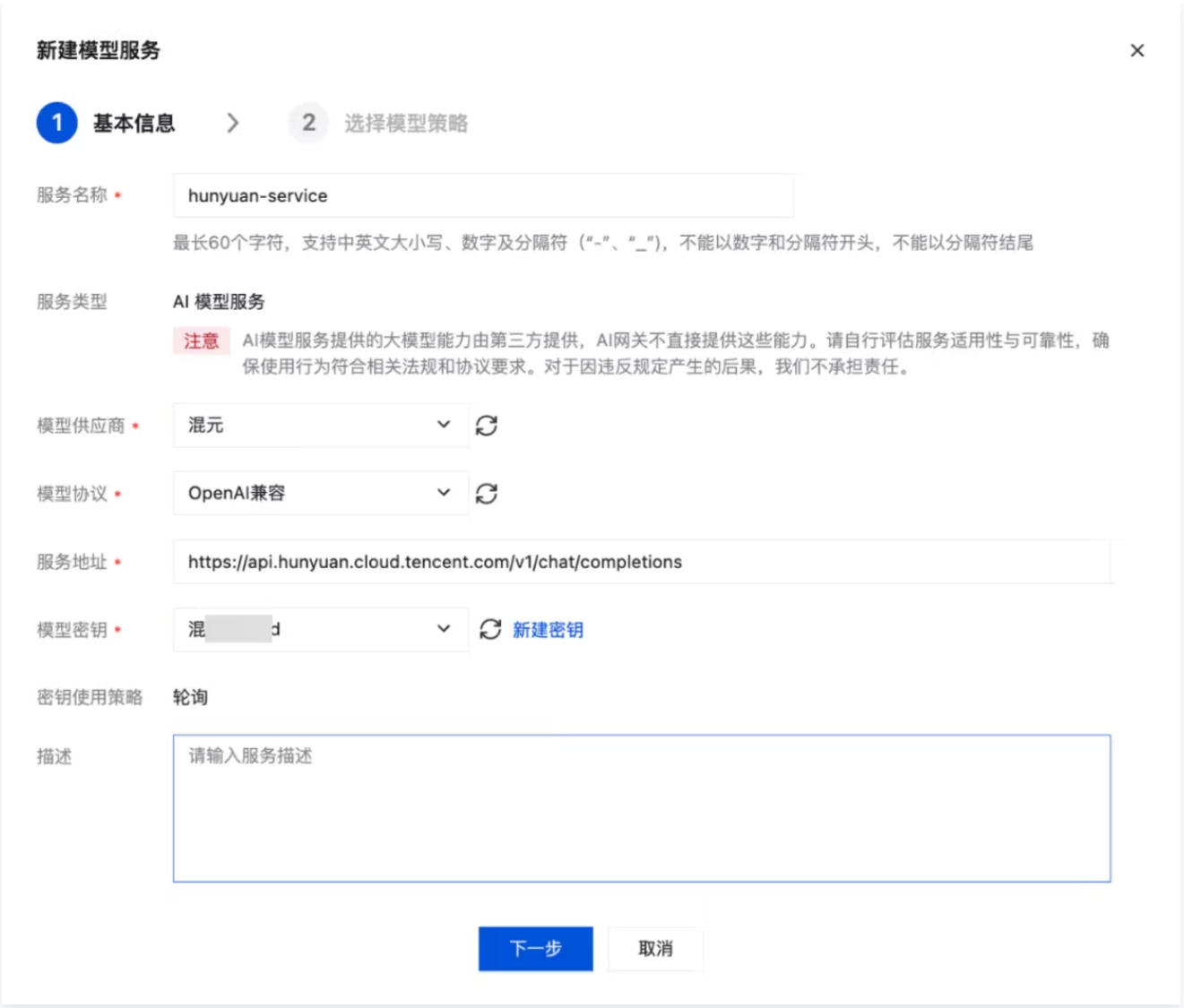
模型服务用于封装对一个特定大模型供应商（此处为混元）的调用配置。

1. 在左侧导航栏，单击**模型管理 > 模型服务**。
2. 在“模型服务”页面，单击**新建**。
3. 在“新建模型服务”窗口中，进入**基本信息**步骤，参考如下内容配置。

| 参数 | 填写说明 |
|---|---|
| 服务名称 | 自定义名称，如 hunyuan-service。 |
| 服务类型 | 固定为 **AI 模型服务**。 |

| | |
|---|---|
| 模型供应商 | 选择 hunyuan。 |
| 模型协议 | 选择 **OpenAI 兼容**。 |
| 服务地址 | 系统将自动填充混元的官方端点地址。 |
| 模型密钥 | 从下拉列表中选择您在 步骤一 创建的密钥（hunyuan-key）。 |
| 密钥使用策略 | 选择 **轮询**。 |

4. 单击**下一步**，进入**选择模型策略**步骤。
   - 模型选择方式：保持默认的指定模型。
   - 默认模型：从下拉列表中选择一个混元模型，如 HY-3。
5. 单击**确定**，完成模型服务创建。




## 步骤三：创建模型 API

模型 API 是网关对外暴露的调用端点，客户端通过访问此 API 来使用大模型能力。

1. 在左侧导航栏，单击**模型管理 > 模型 API** 。
2. 在“模型 API ”页面，单击**新建**。
3. 在“新建模型 API ”窗口中，进入**基本信息**步骤，参考下表配置。


| 参数 | 填写说明 |
| --- | --- |
| API 名称 | 自定义名称，如 hunyuan-api。 |
| 使用场景 | 选择 **文本生成**。 |

| | |
|---|---|
| 请求协议 | 选择 OpenAI。 |
| 路由 | 默认已勾选 /v1/chat/completions，请确认。 |
| Base Path | 输入一个路径前缀，作为 API 的唯一标识，如 /hunyuan。客户端将通过此路径访问。 |
| 路径简化 | 建议保持开启，使后端接收简洁路径。 |

4. 单击 **下一步**，进入**选择模型服务**步骤。
   - 服务类型：选择**单模型服务**。
   - 选择服务：从下拉列表中选择您在**步骤二**创建的服务（hunyuan-service）。
5. 单击**确定**，完成模型 API 创建。系统将自动生成访问路由。

## 步骤四：创建消费者

消费者代表调用 API 的客户端身份，需要配置认证凭证。

1. 在左侧导航栏，单击**消费者管理 > 消费者**。
2. 在“消费者”页面，单击**新建**。
3. 在“新建消费者”窗口中，参考下表配置。

| 参数 | 填写说明 |
|---|---|
| 消费者名称 | 自定义名称，如 my-app。 |
| 所属消费者组 | 此处可暂不选择，后续在消费者组中关联。 |
| 选择密钥 | 单击**新建密钥**，创建一个类型为 API Key 的密钥，用于此消费者调用网关时的身份认证。记下生成的 Key。 |

4. 单击**确定**，完成消费者创建。

## 步骤五：创建消费者组并授权

消费者组用于对消费者进行分组，并统一授权其访问模型 API 的权限。

1. 在左侧导航栏，单击**消费者管理 > 消费者组**。
2. 在“消费者组”页面，单击**新建**。
3. 配置消费者组基本信息，并在“消费者”选项中，关联您在**步骤四**创建的消费者（my-app）。
4. 创建完成后，找到该组，切换到 “已授权模型 API ” 卡片，单击**新增授权**按钮。
5. 在授权页面，将**步骤三**创建的模型 API（hunyuan-api）授权给此消费者组。

## 步骤六：获取访问地址并发起调用

所有配置完成后，即可通过网关地址调用大模型。

1. 获取网关入口地址：进入**基础信息 > 实例信息**，在**网络配置**卡片查看“公网负载均衡地址”或“内网私有连接地址”。
2. 获取路由访问地址：
   - 进入 **模型管理 > 模型 API**，单击您在步骤三创建的 API 名称。
   - 单击 **路由管理** 页签，复制“请求路径”。完整访问地址格式为：`协议://网关入口地址/请求路径`。
3. 发起调用：使用 curl 命令或任何 HTTP 客户端，参照以下示例发起请求。请将<访问地址>替换为上一步得到的完整 URL，将< API_KEY>替换为步骤四中为消费者创建的 API Key，将<MODEL_NAME>替换为步骤二选择的模型。

```
curl -i -X POST <访问地址> \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer < API _KEY>" \
  -d '{
    "model": "<MODEL_NAME>",
    "messages": [
      {"role": "user", "content": "你好，请介绍一下你自己。"}
    ],
    "stream": false
  }'
```

4. 查看结果：如果一切配置正确，您将收到来自混元大模型的 JSON 格式回复。

# 操作指南
# 网关管理
# 新建 AI 网关

最近更新时间：2026-07-23 17:07:33

## 操作场景

本文将介绍如何通过微服务平台控制台创建一个 AI 网关实例。

## 操作步骤

1. 登录微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击**新建**，进入云原生智能网关创建页。

| 参数 | 是否必选 | 说明 |
| --- | --- | --- |
| 网关类型 | 是 | 支持 **AI 网关**和**云原生网关**两种类型。 |
| 计费模式 | 是 | 支持**包年包月**和**按量计费**两种模式，如果您的网关使用时间在一个月以上，建议采用**预付费（包年包月）**模式。具体价格请参考 价格说明 文档。 |
| 地域 | 是 | 选择与您部署业务最靠近的地域。 |
| 产品版本 | 是 | 仅支持标准版。 |
| 名称 | 是 | 输入当前 AI 网关的名称，最长60个字符，支持中英文大小写，-，_。 |
| 描述 | 否 | 输入当前 AI 网关的描述信息。 |
| 资源标签 | 否 | 用于分类管理资源。 |
| 节点网络 | 是 | 选择当前 AI 网关所部署在的腾讯云 VPC 网络。 |
| 节点规格 | 是 | AI 网关集群由多个节点组成，此处选择每个节点的规格。 |
| 节点数量 | 是 | AI 网关集群由多个节点组成，此处选择节点的数量。<br>• 节点数量为2时，仅支持随机分配可用区。<br>• 节点数量大于等于3时，支持选择随机可用区和指定可用区（最少选择 2 个可用区，最多选择 3 个可用区）。 |
| 部署架构 | 是 | 2节点默认支持同城双可用区，2个以上支持同城多可用区 |

| | | |
|---|---|---|
| 公网负载均衡 | 否 | 开启公网后，将产生费用，具体价格参考 公网流量费用 。<br>• 计费模式：支持**按使用流量**和**按带宽计费**两种。<br>• 带宽上限：支持 1 - 2048Mbps 范围。<br>• 可用区类型：支持**单可用区**和**多可用区**。选择多可用区时，您需要选择一个主可用区和一个备可用区，当主可用区故障时，负载均衡可在较短时间内自动切换到备可用区并恢复服务，以保障网关的可用性。 |
| 日志采集 | 否 | • 开启实时日志服务：默认为您提供网关实时日志服务和简单搜索能力，免费使用。如需日志能力持久化存储，用于排障、审计等场景，建议开启 CLS 日志服务。<br>• 开启 CLS 日志服务：为您开启网关日志投递，提供日志分析。在勾选本选项前，请先确认您已开通 CLS 日志服务。日志服务由 CLS 日志 提供，会产生费用，具体计费项查看 费用详情 。 |

3. 单击 **立即购买**，耐心等待实例创建完成。

## 结果验证

返回网关列表页面，查看创建的网关信息和状态。网关信息和创建时一致，且状态为运行中，则说明网关新建成功。

## 注意事项

- 创建 AI 网关要求账号必须拥有 ApiGateWay_QCSRole 角色，并且角色下必须有 `QcloudAccessForApiGateWayRoleInCloudNativeAPIGateway` 策略。如无相关角色和策略，请在创建过程中弹出的授权弹窗中完成授权。
- AI 网关日志功能依赖于腾讯云日志服务（简称 CLS），请您确保已经开通 CLS 服务，否则在创建 AI 网关前请前往 CLS 控制台 开通服务。

# 升级网关版本

最近更新时间：2026-07-23 17:07:33

## 操作场景

如果当前的网关版本已进入下线阶段，为保证网关实例的稳定运行，建议您通过控制台将您的网关升级到稳定版本。

> **注意：**
> 为确保升级成功，请勿在升级前和升级过程中对网关实例执行变更类操作。升级过程中，网关控制台将禁止写入，请妥善安排您的升级计划。

## 版本升级

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 如果网关版本列显示**可升级**，在**操作 > 更多**，单击**升级网关版本**。
3. 单击**确定**，进入版本升级过程。
4. 升级前将执行升级环境检查，需确保所有检查项均通过后才能进行升级，如存在检查失败项，可修复后单击**重新检查**，重新执行环境检查。检查内容如下：

| 检查项 | 检查通过条件 | 检查失败处理措施 |
| --- | --- | --- |
| 实例状态 | 状态处于运行中 | 等待实例变更完成，状态处于运行中后执行重新检查 |
| 服务后端状态 | 不存在后端指向网关 admin 地址的服务 | 删除服务或修改服务地址后执行重新检查 |
| 弹性伸缩状态 | 所有弹性伸缩规则未启用 | 关闭启用状态的弹性伸缩规则后执行重新检查 |
| 子网 IP 数量 | 子网 IP 数量足够 | 前往 VPC，释放子网 IP 后执行重新检查 |
| 节点权重 | 节点权重一致 | 前往**实例详情 > 基础信息 > 部署架构**，手动调整节点权重，保持节点权重一致后执行重新检查，仅专业版支持修改节点权重。 |

5. 确认实例 ID、实例名称、选择升级目标版本，阅读并勾选网关升级相关事宜说明，单击**确定**，开启升级。升级整体耗时与实例当前节点数相关，如节点数较多可能会耗时较长，您可以在网关列表的状态栏查看任务状态。

## 查看升级任务进度

1. 在 AI 网关实例列表，找到目标网关，单击**状态**，单击**查看任务状态**。

2. 升级采用灰度策略，分为两个阶段，**版本升级**与**资源回收**。版本升级主要用于配置升级所用的组件资源，该过程中节点可能增加。完成版本升级后，需要进行资源回收，销毁升级过程中多余的节点，因此需要您手动确认，同意执行资源回收，之后网关会自动触发旧资源回收流程。

## 撤销升级

如在升级过程发现业务异常，可撤销升级。

1. 在 AI 网关实例列表，找到目标网关，单击**状态**，单击**查看任务状态**。
2. 单击**撤销升级**。网关将启动撤销升级任务，将流量回切以保证生产的稳定性。

# 查看网关详情

最近更新时间：2026-07-23 17:07:36

本文介绍在创建一个网关实例后，如何查看网关的基本信息以及访问地址等。

## 操作步骤

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击刚刚创建好的实例的“ID”，进入网关实例基本信息页面，可查看实例的基本信息。
   - **实例信息：**您可以在基本信息区域进行如下操作：
     - 单击基本信息模块右上角的**编辑**，可修改网关实例的名称和说明。
     - 单击公网计费模式的**编辑**，可以修改公网的计费方式。
     - 单击标签的**编辑**，可以修改网关绑定的标签信息。
   - **节点配置：**
     - 节点规格：网关实例每个节点的规格。
     - 节点数量：网关实例的节点数量。
   - **网络配置：**
     - 访问端口： AI 网关的访问端口。
     - 内网管理端口： AI 网关的管理端口。管理端口没有权限控制，仅支持通过内网私有连接访问。
     - 内网私有连接： AI 网关内网地址。客户端可以通过内网节点地址访问 AI 网关的网关节点。
     - 公网负载均衡：展示目前网关实例配置的公网负载均衡，支持修改公网带宽和配置访问策略，公网访问建议设置访问安全策略。

     > **说明：**
     > 内网和公网负载均衡支持选择负载均衡规格，单击**添加负载均衡**后，设置好负载均衡规格，单击**确定**即可。


3. 在实例详情页面，单击**部署架构**页签，展示网关节点的状态及可用区信息。

   单击切换视图，切换到部署架构图，展示接入层（CLB）和网关节点的部署情况。

# 升级网关规格

最近更新时间：2026-07-23 17:07:33

## 操作场景

如果当前的网关规格不能满足您的业务需求，可以通过控制台提升您的网关节点数量和节点规格。

> **说明**
> 如需降低网关节点数量和节点规格，请 提交工单 处理。

## 前提条件

只有**运行中/升级失败**状态的实例支持升级规格。

## 节点规格变更

1. 登录 微服务平台控制台，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，选择需要配置的网关实例“ID”，进入网关实例基本信息页面，可查看实例的基本信息。
3. 在**节点配置**模块，单击节点规格处的**变更**，进入节点变更页。
4. 在规格变更页，选择目标节点规格，单击**确定**，在确认弹窗中选择**确定**，等待大约3 – 5分钟完成实例升级，您可以在网关列表的状态栏查看任务状态。

## 节点数量变更

1. 登录 微服务平台控制台，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击刚刚创建好的实例的“ID”，进入网关实例基本信息页面，可查看实例的基本信息。
3. 在**节点配置**模块，单击节点数量处的**变更**，进入节点变更页。
4. 在变更页，选择目标节点数量，单击**确定**，在确认弹窗中选择**确定**，等待大约3 – 5分钟完成实例升级，您可以在网关列表的状态栏查看任务状态。

# 删除网关

最近更新时间：2026-07-23 17:07:33

当您不再需要使用某个 AI 网关实例时，可以将其删除以释放资源并停止计费。删除操作分为手动删除和到期/欠费自动释放两种场景。手动删除实例时，系统会提供风险确认和释放选项，请您谨慎操作。
本文档将指导您完成手动删除操作，并说明欠费到期后系统自动释放资源的规则。

## 操作步骤

### 手动删除

若您确认不再需要某个运行中的 AI 网关实例，可登录控制台手动将其删除。

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，找到目标实例，单击其操作列的更多，在更多操作中点击 删除。
3. 系统会弹出“确定删除”确认窗口，请您仔细阅读并确认相关信息与选项。
   - 释放资源选项
     - 立即释放：系统会即刻将相关配置、数据彻底释放，该操作不可逆，请谨慎操作。
     - 2小时后释放：系统会先将实例移入回收站并保留2小时（此期间仍会计费）。2小时内如需恢复实例，可尝试操作（但可能因资源不足导致恢复失败）；2小时后，系统将自动彻底释放相关配置与数据，此操作同样不可逆。
4. 完成上述信息确认并勾选复选框后，单击窗口底部的 **删除** 按钮。若需放弃操作，可单击 **取消**。

> ⚠ **注意：**
> - 只有**运行中/失败状态**的实例可以被销毁。
> - 实例释放后，其所有配置、路由、监控数据等均会被清除，**该操作不可逆，请谨慎操作**。

### 到期/欠费自动删除

如果您的 AI 网关实例到期未续费或账户处于欠费状态，系统将按以下规则自动处理：

1. 资源保留期（最多7天）：实例到期或欠费后，将在控制台中进入“已隔离”状态，并被保留最多7个自然日。此期间实例无法提供生产服务，但已保存的数据和配置不会被删除。详情请查看 欠费说明 。
2. 续费恢复：在7天保留期内，您可以在控制台“实例列表”中找到该实例，并通过操作列的 **续费** 来恢复服务。续费成功后，实例将恢复至“运行中”状态。
3. 自动释放：若实例在到期/欠费后第7天（含）仍未续费，系统将在第8天的0点（北京时间）开始自动释放所有相关计算与存储资源。资源释放后，实例内的所有数据将被永久清除且不可恢复。

# 模型管理
# 模型 API

最近更新时间：2026-07-23 17:07:33

## 操作场景

模型 API 是 AI 网关对外暴露的统一接口。客户端通过调用特定的模型 API 来使用大模型能力。其核心工作原理是：您创建一个模型 API 并为其关联后端模型服务，网关会根据配置自动生成对应的访问路由。客户端的请求通过匹配此路由进入网关，并由网关将其转发至关联的模型服务进行处理。
您需要在此创建和管理模型 API，以定义客户端如何访问、将请求路由到哪一个具体的模型服务。模型 API 支持绑定单个模型服务，也支持绑定多个模型服务并配置路由策略实现灵活的流量分发；同时支持配置 Header 匹配条件，在相同 Base Path 下创建多个模型 API 实现多租户隔离、灰度切流等场景；并可配置 QPM 限流及 Token 限流对请求进行精细化流量控制。本文介绍如何为 AI 网关添加模型 API 以及管理其关联的模型服务和自动生成的路由。

## 操作步骤

### 添加模型 API

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **模型管理**，然后单击**模型 API** 页签，在 API 列表页面单击**新建**。

> **说明：**
> 模型 API 列表展示的字段为：ID/名称、使用场景、请求协议、服务类型、描述、创建/修改时间、操作（编辑/删除）。其中"服务类型"在控制台显示为**单模型服务**或**多模型服务**。

4. 在“新建模型 API”窗口中，完成第一步“基本信息”的配置。

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| API 名称 | 是 | 输入此 API 的名称，用于标识。最长60个字符，支持中英文大小写、数字及分隔符（“-”、“_”），不能以数字和分隔符开头，不能以分隔符结尾。 |
| 使用场景 | 是 | 选择此 API 的用途，支持“文本生成”。系统将根据场景预置相关的默认路由。 |
| 请求协议 | 是 | 选择客户端调用此 API 时使用的协议，例如“OpenAI”。此选择将影响预置路由和网关对请求/响应格式的处理。 |

| | | |
|---|---|---|
| 路由 | 是 | 根据所选“使用场景”和“请求协议”自动预置的默认路由。请勾选需要为此 API 启用的路由。每个被勾选的路由都将与 Base Path 组合，生成一条独立的访问路径。 |
| Base Path | 否 | 为此 API 设置统一的路由前缀。客户端请求的完整路径为 `/{Base Path}/{路由路径}`。例如，Base Path 设为 `/qwen`，并勾选路由 `/v1/chat/completions`，则完整访问路径为 `/qwen/v1/chat/completions`。 |
| Header | 否 | 为此模型 API 配置 API 级别的入口过滤条件。填写后，只有请求携带满足所有条件的 Header 时，才能命中该模型 API。不填则表示无限制，所有请求均可进入。多条件为 AND 关系，每条规则包含：参数来源（Header）、参数键（如 X-User-Level）、匹配方式（等于 开头为 包含 / 存在）、参数值（匹配方式为"存在"时可不填）。同一 Base Path 下有 Header 条件的 API 优先于无条件 API 命中。 |
| 描述 | 否 | 该 API 的描述信息，便于后续管理。 |

在此步骤中，您定义的 Base Path 与所选的 **路由** 将组合成该 API 的最终访问路径。系统会根据您选择的“使用场景”和“请求协议”，自动预置一个或多个默认路由。例如，选择“文本生成”场景和“OpenAI”协议，系统会预置 `/v1/chat/completions` 路由。创建完成后，网关将基于此完整路径自动生成一条路由规则。

5. 完成基本信息配置后，单击 **下一步**，进入“选择模型服务”步骤。在此步骤中，您需要将此 API 与一个具体的模型服务（该服务已配置了供应商、密钥、模型 Fallback 等策略）进行绑定。

| 参数 | 是否必填 | 说明 |
|---|---|---|
| 服务类型 | 是 | 支持选择以下两种类型：<br>1. 单模型服务：此 API 固定路由到一个后端模型服务；<br>2. 多模型服务：此 API 将流量按路由策略分发到多个后端模型服务，适用于灰度发布、负载均衡、多供应商聚合等场景。 |
| 路由策略（多模型服务时显示） | 是 | 选择多模型服务时，需按照所选策略配置相关内容。 |
| 服务列表（多模型服务时显示） | 是 | 配置多个后端模型服务及其路由规则。每行包含：目标服务、路由规则等。 |
| 选择服务（单模型服务时显示） | 是 | 选择一个已创建的模型服务。您也可以通过“新建服务”链接跳转至模型服务页面快速创建 |

6. 单击**确定**，完成模型 API 的创建。此时，网关会为你在“基本信息”步骤中勾选的每一个路由，自动生成一条对应的路由规则。

## 查看与编辑模型 API

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **模型管理**，然后单击 **模型 API** 页签。
4. 单击 API 的“ID/名称”，可进入其详情页面。
5. 在“基本信息”页签下，可以查看 API 的完整配置信息。
6. 在详情页的“基本信息”页签右上角，单击 **编辑**，可修改其“基本信息”配置。修改后单击 **确定** 保存。

## 管理路由

路由是网关将客户端请求分发到对应模型 API 的规则。创建模型 API 时，系统已根据配置自动生成了路由。

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**模型管理**，然后单击**模型 API** 页签。
4. 单击 API 的“ID/名称”，可进入其详情页面。
5. 在 API 详情页面，单击**路由管理**页签，可以查看系统为此 API 自动生成的所有路由规则详情。这里展示了路由的 ID、名称、类型以及完整的匹配路径。网关正是通过匹配这些路由规则，来决定将进入的请求交给哪个模型 API 处理。

## 管理关联的模型服务

模型 API 需要关联模型服务才能实际工作。关联的模型服务、服务类型（单模型服务/多模型服务）及路由策略，均通过**编辑模型 API** 进行变更，而非在详情页的“基本信息”页签中直接管理。单模型服务类型的模型 API 最多绑定 1 个模型服务；多模型服务类型的模型 API 可绑定多个模型服务，并通过路由策略（权重路由、参数路由、模型名称路由）决定流量分发方式。如需变更服务类型或路由策略，请通过编辑 API 入口进行修改。

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**模型管理**，然后单击**模型 API** 页签。
4. 找到目标模型 API，单击其操作列的**编辑**（或进入 API 详情页后单击右上角的**编辑**）。
5. 在编辑向导中进入**选择模型服务与路由策略**步骤，可执行以下变更：
   - **变更关联的模型服务**：在服务列表中重新选择目标模型服务。
   - **变更服务类型**：在**单模型服务**与**多模型服务**之间切换。
   - **变更路由策略**：当服务类型为多模型服务时，可调整路由策略（权重路由、参数路由、模型名称路由）及各服务的路由规则。
6. 完成修改后，单击**确定**保存。此后，通过此模型 API 的请求将按新的模型服务与路由策略进行处理。

## 删除模型 API

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。

3. 在左侧导航栏单击**模型管理**，然后单击**模型 API**页签。
4. 在**模型 API** 列表页面，找到目标 API，单击其操作列下的**删除**，系统将进行删除前的依赖关系校验。
5. 系统会弹窗提示您确认删除，并自动检查该 API 是否存在被其他资源（如“消费者组”授权）绑定的情况。
   - 若无依赖：弹窗将直接显示 API 信息，单击**确定**即可删除。删除 API 的同时，其自动生成的所有路由规则也会被一并删除。
   - 若存在依赖：弹窗会显示“资源删除依赖关系检查结果”，并提示“存在未解除的依赖关系”，同时列出具体的依赖项。
6. 若存在依赖，您需要先行解除所有列出的依赖关系。解除依赖后，可单击弹窗内的**重新检查**链接，系统将再次进行校验。
7. 当校验通过，依赖提示消失后，单击**确定**即可最终删除该 API。若需放弃删除，可单击**取消**。

# 模型服务

最近更新时间：2026-07-24 16:45:00

## 操作场景

您需要将大模型服务添加到 AI 网关中，以便网关能代理请求至相应的模型供应商，实现统一接入、路由、降级与密钥管理。AI 网关支持添加混元、Google Gemini、DeepSeek、千问、OpenAI 等供应商的模型服务。本文介绍如何为 AI 网关添加、编辑和删除模型服务。

## 操作步骤

### 添加模型服务

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**模型管理**，然后单击**模型服务**页签，在服务列表中单击**新建**。
4. 在“新建模型服务”窗口中，完成第一步“基本信息”的配置。

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 服务名称 | 是 | 输入服务名称。最长60个字符，支持中英文大小写、数字及分隔符（“-”、“_”），不能以数字和分隔符开头，不能以分隔符结尾。 |
| 服务类型 | 是 | 固定为“AI 模型服务”。 |
| 模型供应商 | 是 | 选择模型供应商：<br>• 标准供应商：Hunyuan、OpenAI、Anthropic、Google-Gemini、DeepSeek、Qwen、月之暗面、智谱、百度千帆、腾讯云 TI-ONE<br>• MaaS 供应商：TokenHub、AWS Bedrock、Azure OpenAI、火山方舟、Google Vertex AI<br>• 自定义供应商 |
| 模型协议 | 是 | 根据模型供应商支持的模型协议，支持 OpenAI 兼容和 Anthropic 兼容。 |
| 服务地址 | 是 | 确认模型服务的服务地址。<br>若选择自定义供应商，此处可选手动填写或关联服务来源。<br>• 手动填写：请自行评估服务地址的有效性，确保服务地址符合相关法规和协议要求。提供**路径拼接**功能，在配置服务地址时单独指定路径前缀，系统将自动拼接完整请求地址。支持两种路径拼接模式：<br>  ○ 自动拼接（默认）：将模型 API 请求路径拼接到 Base URL 之后，适用于大多数标准 OpenAI 兼容协议服务。 |

| | | |
|---|---|---|
| | | ○ 使用固定路径：请求路径固定为 Base URL，不拼接，适用于 Bedrock、Azure 等固定 endpoint 场景。<br>• 关联服务来源：支持选择已创建的容器服务或北极星服务，并配置具体服务来源、命名空间、容器服务、请求协议。 |
| 模型密钥 | 否 | 选择已配置的该供应商 API 密钥，或单击“新建密钥”跳转至密钥管理页面进行添加。网关将使用此密钥调用对应模型 API。 |
| 密钥使用策略 | 否 | 当配置了多个密钥时，定义密钥的使用方式。默认为轮询，可在多个密钥间均衡负载。 |
| 密钥周期性轮换 | 否 | 开启后，网关仅从创建时间在周期内的密钥中遴选调度 |
| 描述 | 否 | 该服务的描述信息，便于后续管理。 |

**⚠ 注意：**
AI 模型服务提供的大模型能力由第三方提供，AI 网关不直接提供这些能力。请自行评估服务适用性与可靠性，确保使用行为符合相关法规和协议要求，否则请自行承担因违反规定产生的后果。

5. 配置高级配置(可选)

在 **高级配置** 区域,配置服务的网络超时、SNI、配额和标签等参数。

**超时配置说明：**

| 参数 | 默认值 | 范围 | 说明 |
|---|---|---|---|
| 连接超时时间 | 10000 ms | 1 ~ 3600000 | 建立与后端服务 TCP 连接的超时时间。建议根据模型推理时间设置，快速模型可适当降低超时时间 |
| 读取超时时间 | 60000 ms | 1 ~ 3600000 | 从后端服务读取响应数据的超时时间 |
| 写入超时时间 | 60000 ms | 1 ~ 3600000 | 向后端服务发送请求数据的超时时间 |

**超时配置建议值：**

| 模型类型 | 连接超时 | 读取超时 | 写入超时 | 说明 |
|---|---|---|---|---|
| 快速推理模型（< 5秒） | 5000 ms | 30000 ms | 30000 ms | 如 qwen-turbo、gpt-4o-mini |
| 标准推理模型（5 ~ 30秒） | 10000 ms | 60000 ms | 60000 ms | 默认值，适用于大多数模型 |

| | | | | |
|---|---|---|---|---|
| 慢速推理模型（> 30秒） | 15000 ms | 120000 ms | 120000 ms | 如深度推理模型、长文本生成 |
| 流式响应 | 10000 ms | 不限制 | 30000 ms | 流式模式下读取超时不限制 |

**SNI 配置说明：**

| 参数 | 是否必填 | 说明 |
|---|---|---|
| SNI | 否 | TLS 握手时发送的 Server Name Indication 值，留空时使用服务地址中的域名作为 SNI |

**配额配置说明：**

| 参数 | 是否必填 | 说明 | 示例 |
|---|---|---|---|
| RPM（每分钟请求数） | 否 | 供应商规定的模型每分钟允许的最大请求次数。不填写则无法在模型 API 中使用配额感知 Fallback | 1000 |
| TPM（每分钟 Token 数） | 否 | 供应商规定的模型每分钟允许的最大 Token 消耗量。不填写则无法在模型 API 中使用配额感知 Fallback | 100000 |

> **说明：**
> 配额配置用于配额感知 Fallback 场景。当消费者的剩余配额低于阈值时，网关自动将请求降级到备用服务。如果模型服务未配置 RPM/TPM，则无法参与配额感知 Fallback。

**服务标签说明：**

| 参数 | 是否必填 | 说明 | 示例 |
|---|---|---|---|
| 标签键 | 否 | 分类标识，建议使用有业务含义的键名 | region、env、provider |
| 标签值 | 否 | 对应键的值 | singapore、production、openai |

> **说明：**
> 服务标签用于在模型 API 中通过“基于标签路由”策略自动筛选服务。新增模型服务时只需配置标签，即可自动被匹配的 API 发现和使用，无需修改 API 配置。

6. 完成基本信息后，单击 下一步，进入“**选择模型策略**”步骤。

- 模型选择方式：此配置决定网关如何处理客户端请求中的模型（model）参数。

**指定模型**

网关将忽略客户端请求中的 model 参数，统一使用您在下方“默认模型”中指定的模型。此模式适合成本控制和高可用场景，便于统一路由和降级。

- 默认模型：当“模型选择方式”为“指定模型”时，必须在此处选择一个具体的模型名称
- 模型 Fallback：开启后，当请求“默认模型”失败时，网关可根据规则自动切换（Fallback）到其他可用模型，保障服务高可用。
- 备选规则：开启 Fallback 后，需在此选择或配置当主模型不可用时的备选模型列表及切换规则。

**透传请求模型**

网关将直接使用客户端请求中的 model 参数，并将其转发给供应商。此模式适合需要客户端灵活控制模型选择的场景（如评估请求延迟、统计 Token 用量等），但使用透传请求时网关无法明确识别用户实际请求的模型名称，请确保客户端传递正确的模型名称。

- 若用户请求的模型名称与后端供应商的模型名称一致或不需要对模型名称做任何转换处理，可直接透传用户请求中的 model name 至后端服务。
- 若需将用户请求中的 model name 替换为供应商中定义的模型名称，可配置**模型名称映射**信息。支持精确匹配和前缀匹配(*)，支持添加多条映射规则。
  - 请求模型名称（客户端请求的模型名称）
  - 目标模型名称（需要重写为的模型名称，即后端供应商实际使用的模型名称）

**高级配置（可选）**

- 模型参数校验：开启后，网关将校验客户端请求中的 model 参数是否在允许的列表内。
  - 允许的模型列表：定义客户端允许请求的模型名称白名单。
  - 校验失败处理：定义当模型校验失败时的处理策略，支持“返回404”或“使用默认模型降级”。

7. 配置完成后，单击**确定**即可创建模型服务。
8. 添加后，服务列表中会出现新增的服务，单击**服务 ID/名称**，查看详细的服务信息。

## 编辑服务

在**模型服务**列表页面，找到目标服务，单击其操作列下的**编辑**，即可修改服务配置信息，修改后单击**确定**保存。

## 删除服务

在**模型服务**列表页面，找到目标服务，单击其操作列下的**删除**，系统将进行删除前的依赖关系校验。

1. 系统会弹窗提示您确认删除，并自动检查该服务是否存在被其他资源（如“模型 API”）绑定的情况。

2. 确认结果：
   - 若无依赖：弹窗将直接显示服务 ID 和名称，单击**确定**即可删除。
   - 若存在依赖：弹窗会在服务信息下方显示“资源删除依赖关系检查结果”，并提示“存在未解除的依赖关系”，同时列出具体的依赖项。
3. 若存在依赖，您需要先行解除所有列出的依赖关系。解除依赖后，可单击弹窗内的**重新检查**操作，系统将再次进行校验。当校验通过，依赖提示消失后，单击**确定**即可最终删除该服务。若需放弃删除，可单击**取消**。

# 模型探测

最近更新时间：2026-07-23 17:07:33

## 操作场景

自定义供应商模型探测支持在创建自定义模型服务时，通过探测能力自动发现供应商提供的模型列表，避免手动逐个配置模型名称。适用于以下场景：

- **快速接入**：对接 OpenAPI 兼容协议的自定义供应商时，快速获取可用模型列表。
- **批量导入**：供应商提供多个模型时，一键导入所有模型名称，避免手动录入。
- **版本同步**：定期探测供应商的模型列表，自动同步新增或下线的模型。
- **兼容性验证**：验证自定义供应商的 API Endpoint 和协议兼容性，确保配置正确。

> **说明：**
> 模型探测仅支持 OpenAI 兼容协议的供应商，通过调用供应商的 `/v1/models` 接口获取模型列表。

## 前置条件

1. 已创建 AI 网关实例且处于运行中状态。
2. 准备好自定义供应商的 API Endpoint 和访问凭证。
3. 供应商需支持 OpenAI 兼容协议的模型列表查询接口（ `/v1/models` ）。

## 操作步骤

### 场景一：创建自定义模型服务时发起模型探测

#### 步骤 1：进入模型服务创建页

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**；
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 在左侧导航栏单击**模型管理**，然后单击**模型服务**页签，在服务列表中单击**新建**。

#### 步骤 2：选择自定义供应商

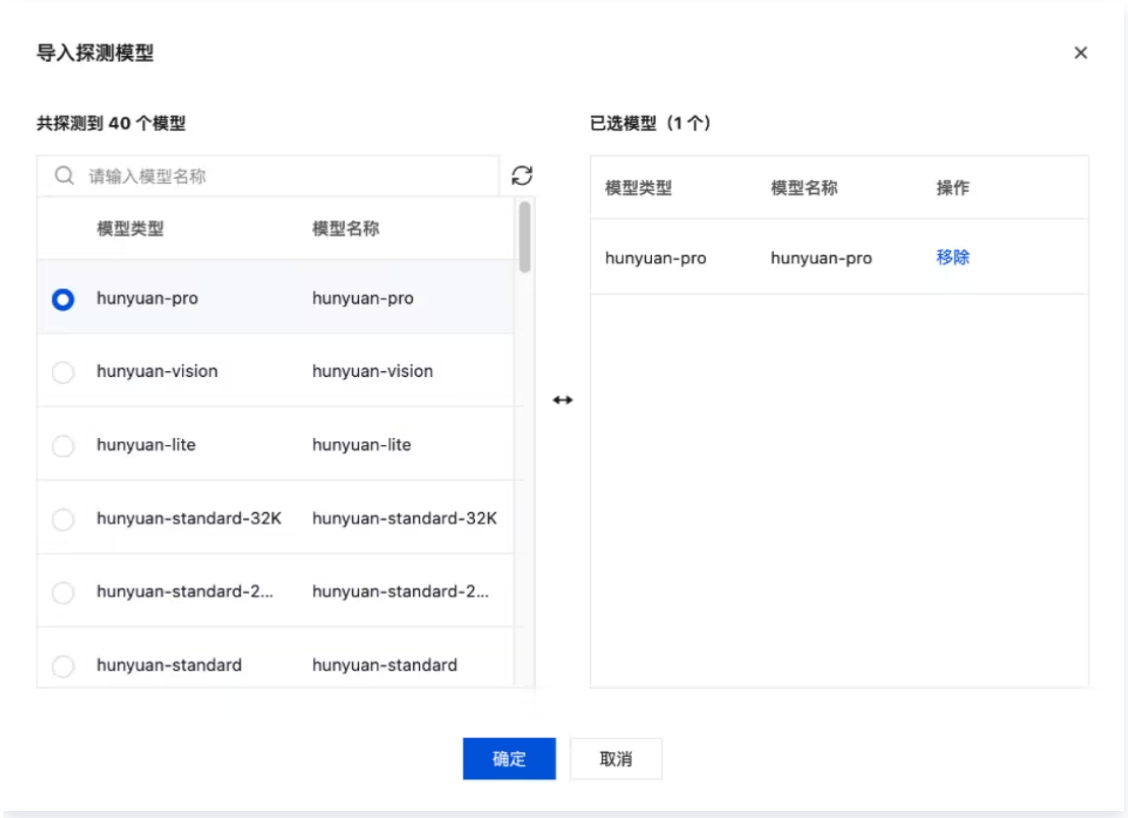
1. 在 **模型供应商** 区域，选择 **自定义供应商**；
2. 配置基本信息。

#### 步骤 3：发起模型探测

在配置完成基本信息后，在 **选择模型策略** 区域单击 **发起模型探测**：


### 步骤 4：确认并导入模型

1. 在探测结果列表中勾选需要导入的模型；
2. 单击 **确定** 或类似确认按钮；
3. 系统将选中的模型导入到模型配置中。




## 场景二：已有服务重新探测

对于已创建的自定义模型服务，可以重新发起模型探测以同步供应商的模型更新。

### 操作步骤

1. 进入目标模型服务的编辑页面；
2. 在 **选择模型策略** 区域单击 **发起模型探测**；

3. 查看探测结果，勾选新增的模型或取消勾选已下线的模型；
4. 保存配置。

## 注意事项

1. **API 接口要求**：模型探测通过调用供应商的 `/v1/models` 接口获取模型列表，请确保供应商支持该接口。
2. **访问凭证**：探测时需要使用供应商的 API Key 等访问凭证进行认证。
3. **探测失败处理**：探测失败时（如网络不可达、认证失败、接口不兼容），页面会显示具体错误信息，不影响后续手动配置模型。
4. **重新探测的频率**：建议定期（如每周）重新探测一次，同步供应商的模型更新。但注意不要过于频繁，避免对供应商造成不必要的负担。
5. **模型名称准确性**：导入的模型名称为供应商返回的原始名称，请确认与实际调用时使用的 model 参数一致。

# 认证策略

最近更新时间：2026-07-23 17:07:34

## 操作场景

认证策略用于控制客户端对 AI 网关 API 的访问权限。通过配置认证策略，您可以：

- 支持多种认证方式(API Key、JWT、OAuth 2.0、OIDC)。
- 为不同的消费者配置不同的访问权限。
- 单个消费者支持配置多种类型的凭证。
- 实现细粒度的访问控制和安全防护。

认证策略在创建模型 API 时配置，确保只有授权的客户端才能访问网关服务。
本文档指导您如何在 AI 网关中配置和管理认证策略。

## 前置条件

- 已创建 AI 网关实例。
- 已创建消费者。
- 已创建模型 API 或 Agent API。

## 操作步骤

### 创建消费者

消费者是 AI 网关的访问主体，必须先创建消费者才能配置认证策略。

#### 步骤1：进入消费者列表页

1. 登录 AI 网关控制台，选择对应实例。
2. 在左侧导航栏选择**消费者管理**，单击**消费者页签**。
3. 单击**新建**。

#### 步骤2：配置消费者信息

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 消费者名称 | 是 | 消费者的名称，支持中文、英文、数字、下划线，长度1-60字符。 |
| 所属消费者组 | 否 | 消费者所属的消费者组。 |
| 密钥 | 否 | 消费者绑定的密钥。 |

| | | |
|---|---|---|
| 描述 | 否 | 消费者的功能描述，最多200字符。 |

### 步骤3：完成创建

单击**确定**完成消费者创建。

## 为消费者添加认证凭证

创建消费者后，需要为其添加认证凭证，支持同时配置多种类型的凭证。

### 步骤1：进入消费者详情页

在消费者列表页，单击消费者名称或操作列的**详情**。

### 步骤2：添加凭证

在**凭证管理**区域，单击**添加凭证**，选择凭证类型。

## 认证方式1：API Key

API Key 是最简单的认证方式，适用于内部服务或可信客户端。

### 配置参数

| 参数 | 是否必填 | 说明 | 示例 |
|---|---|---|---|
| 凭证类型 | 是 | 选择 API Key | `API Key` |
| API Key | 自动生成 | 系统自动生成 API Key，也可自定义 | `ak-xxxxxxxxxx` |
| 有效期 | 否 | 设置凭证有效期,默认永久有效 | `2027-12-31` |

### 使用方式

客户端请求时，在请求头中携带 API Key：

```
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "Authorization: Bearer {API_Key}" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

**说明：**

- Authorization Header 格式为 `Bearer {API_Key}`。
- API Key 一旦生成，请妥善保管，不要泄露给无关人员。

## 认证方式2：JWT

JWT(JSON Web Token)适用于需要传递用户身份信息的场景，支持自定义 Claims。

### 配置参数

| 参数 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| 凭证类型 | 是 | 选择 **JWT** | `JWT` |
| 签名算法 | 是 | 选择 JWT 签名算法：<br>• **HS256**(HMAC SHA256)<br>• **RS256**(RSA SHA256)<br>• **ES256**(ECDSA SHA256) | `HS256` |
| 密钥/公钥 | 是 | 根据签名算法配置：<br>• HS256：配置密钥(Secret)<br>• RS256/ES256：配置公钥(Public Key) | `your-secret-key` |
| 有效期校验 | 否 | 是否校验 JWT 的 exp 字段 | 开启 |
| 必需 claims | 否 | 配置必须包含的 Claims 字段 | `sub, iss` |

### 使用方式

客户端请求时，在请求头中携带 JWT Token：

```
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "Authorization: Bearer {JWT_Token}" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

JWT Token 示例(HS256)：

```
Header:
{
```

```
  "alg": "HS256",
  "typ": "JWT"
}

Payload:
{
  "sub": "user123",
  "iss": "your-app",
  "exp": 1735660800
}

Signature:
HMACSHA256(
  base64UrlEncode(header) + "." + base64UrlEncode(payload),
  your-secret-key
)
```

## 认证方式3：OAuth 2.0

OAuth 2.0适用于需要接入第三方 OAuth 服务的场景，支持多种授权模式。

### 配置参数

| 参数 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| 凭证类型 | 是 | 选择 OAuth 2.0 | `OAuth 2.0` |
| 授权模式 | 是 | 选择 OAuth 2.0授权模式：<br>• Client Credentials(客户端凭证模式)<br>• Password(密码模式)<br>• Authorization Code(授权码模式) | `Client Credentials` |
| Token Endpoint | 是 | OAuth 服务的 Token 获取地址 | `https://oauth.example.com/token` |
| Client ID | 是 | OAuth 客户端 ID | `client-12345` |
| Client Secret | 是 | OAuth 客户端密钥 | `secret-xxxxxx` |
| Scope | 否 | 请求的权限范围 | `read write` |

| | | | |
|---|---|---|---|
| Token 缓存时长 | 否 | Token 缓存时长，默认根据 Token 的 expires_in 字段 | 3600秒 |

## 使用方式

### 方式1：客户端自行获取 Access Token

客户端先调用 OAuth 服务获取 Access Token，然后携带 Token 访问网关：

```
# 步骤1:获取Access Token
curl -X POST https://oauth.example.com/token \
  -d "grant_type=client_credentials" \
  -d "client_id=client-12345" \
  -d "client_secret=secret-xxxxxx"

# 响应示例:
{
  "access_token": "eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9...",
  "token_type": "Bearer",
  "expires_in": 3600
}

# 步骤2:携带Access Token访问网关
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "Authorization: Bearer eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..." \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

### 方式2：网关代为获取 Access Token

网关可以根据配置的 OAuth 信息，代为获取 Access Token 并缓存，客户端只需提供 Client ID 和 Client Secret：

```
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "X-Client-ID: client-12345" \
  -H "X-Client-Secret: secret-xxxxxx" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
```

```
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

## 认证方式4：OIDC

OIDC(OpenID Connect)是基于 OAuth 2.0的身份认证协议，适用于需要用户身份认证的场景。

### 配置参数

| 参数 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| 凭证类型 | 是 | 选择 OIDC | `OIDC` |
| Issuer URL | 是 | OIDC 服务的 Issuer 地址 | `https://oidc.example.com` |
| Client ID | 是 | OIDC 客户端 ID | `client-12345` |
| Client Secret | 否 | OIDC 客户端密钥(部分 Provider 需要) | `secret-xxxxxx` |
| Discovery Endpoint | 自动生成 | 根据 Issuer URL 自动生成 Discovery 地址 | `https://oidc.example.com/.well-known/openid-configuration` |
| 公钥获取方式 | 是 | 选择公钥获取方式：<br>• **自动发现**(从 Discovery Endpoint 获取)<br>• **手动配置**(手动输入 JWKS 或公钥) | 自动发现 |

### 使用方式

客户端请求时，在请求头中携带 ID Token：

```
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "Authorization: Bearer {ID_Token}" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

**说明：**

- OIDC 使用 ID Token 进行身份认证，ID Token 是一个 JWT。
- 网关会自动从 OIDC 服务的 Discovery Endpoint 获取公钥，验证 ID Token 签名。
- 网关会校验 ID Token 的 iss、aud、exp 等标准 Claims。

## 免认证方式

对于内网环境或测试场景，可以选择免认证方式。

**说明：**

- 免认证方式下，任何客户端都可以访问 API，不进行任何认证。
- **生产环境强烈不建议使用免认证方式。**

## 为 API 配置认证策略

创建消费者并添加凭证后，需要在 API 中配置认证策略，绑定消费者。

### 步骤1：进入 API 配置页

在创建或编辑模型 API/Agent API 时，找到**认证策略**区域。

### 步骤2：选择认证方式

| 参数 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| 认证方式 | 是 | 选择该 API 支持的认证方式(单选) | `API Key` |
| 绑定消费者 | 是(除免认证外) | 选择允许访问该 API 的消费者，支持多选 | 消费者 `A`<br>消费者 `B` |

**说明：**

- 单个 API 仅支持一种认证方式。。
- 单个 API 可以绑定多个消费者
- 单个消费者可以配置多种类型的凭证(如同时配置 API Key 和 JWT)，客户端可根据实际情况选择使用哪种凭证。

### 步骤3：保存配置

单击 **确定** 保存认证策略配置。

## 管理消费者凭证

### 查看凭证列表

在消费者详情页的 **凭证管理** 区域，可查看该消费者的所有凭证：

- 凭证类型(API Key、JWT、OAuth 2.0、OIDC)
- 凭证状态(启用、禁用、已过期)
- 创建时间和有效期

### 禁用凭证

如需临时禁用某个凭证，单击操作列的**禁用**。禁用后，使用该凭证的请求将被拒绝。

### 启用凭证

对于已禁用的凭证，单击操作列的**启用**即可重新启用。

### 删除凭证

单击操作列的**删除**，可永久删除该凭证。删除后，使用该凭证的请求将被拒绝。

> **说明：**
> 删除凭证是不可逆操作，请谨慎操作。

### 更新凭证

对于 JWT、OAuth 2.0、OIDC 凭证，单击操作列的 **编辑**，可更新凭证配置(如密钥、Token Endpoint 等)。

> **说明：**
> API Key 凭证创建后不可修改，如需更换请删除后重新创建。

## 结果验证

## 验证 API Key 认证

使用正确的 API Key 访问 API：

```
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "Authorization: Bearer {正确的API_Key}" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

**预期结果：返回200响应，成功调用模型。**

**使用错误的 API Key 访问 API：**

```
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "Authorization: Bearer wrong-api-key" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

**预期结果：返回401 Unauthorized 错误。**

```
{
  "error": {
    "code": "unauthorized",
    "message": "Invalid API Key"
  }
}
```

## 验证 JWT 认证

**生成 JWT Token（使用配置的密钥和签名算法）：**

```
import jwt
import time

# JWT配置
secret = os.getenv('SECRET_KEY')
algorithm = "HS256"

# 生成Token
payload = {
    "sub": "user123",
    "iss": "your-app",
    "exp": int(time.time()) + 3600  # 1小时后过期
}
token = jwt.encode(payload, secret, algorithm=algorithm)
```

```
print(f"JWT Token: {token}")
```

使用生成的 JWT Token 访问 API：

```
curl -X POST https://{网关域名}/{base_path}/v1/chat/completions \
  -H "Authorization: Bearer {JWT_Token}" \
  -H "Content-Type: application/json" \
  -d '{
    "model": "qwen-plus",
    "messages": [{"role": "user", "content": "你好"}]
  }'
```

**预期结果**：返回200响应，成功调用模型。

## 验证 OAuth 2.0认证

先获取 Access Token，再访问 API，可参考上文 使用方式 部分。

## 验证 OIDC 认证

使用 OIDC 服务颁发的 ID Token 访问 API，可参考上文 使用方式 部分。

## 注意事项

- **凭证安全**：API Key、密钥、Client Secret 等凭证信息务必妥善保管，不要提交到代码仓库或公开渠道
- **凭证轮换**：建议定期更换凭证，降低凭证泄露风险
- **有效期设置**：建议为凭证设置有效期，避免长期有效凭证泄露后持续被滥用
- **最小权限原则**：为不同的消费者配置不同的 API 访问权限，避免过度授权
- **免认证风险**：生产环境禁止使用免认证方式，仅适用于内网测试场景
- **OAuth Token 缓存**：网关会缓存 OAuth Access Token，如 Provider 侧 Token 失效，需等待缓存过期或手动刷新

## 常见问题

### 配置使用类

#### Q1：如何选择合适的认证方式?

A：根据您的实际场景选择：

- **API Key**：简单场景，适用于内部服务或可信客户端，配置简单，性能最优
- **JWT**：需要传递用户身份信息，支持自定义 Claims，适用于微服务间调用
- **OAuth 2.0**：需要接入第三方 OAuth 服务，支持标准 OAuth 流程

- **OIDC**：需要用户身份认证，支持 SSO(单点登录)
- **免认证**：仅适用于内网测试场景，生产环境禁用

### Q2：单个消费者可以配置多个凭证吗？

A：可以。单个消费者支持同时配置多种类型的凭证，例如：

- 凭证1：API Key 类型
- 凭证2：JWT 类型
- 凭证3：OAuth 2.0类型

客户端可以根据实际情况选择使用哪种凭证访问 API。

### Q3：单个 API 可以支持多种认证方式吗？

A：不可以。单个 API 仅支持一种认证方式，但可以绑定多个消费者，每个消费者可以配置多种凭证。
如需支持多种认证方式，请创建多个 API，分别配置不同的认证方式。

### Q4：JWT 的签名算法如何选择？

A：

- **HS256(对称加密)**：使用密钥(Secret)进行签名和验证，配置简单，性能最优，适用于内部服务
- **RS256/ES256(非对称加密)**：使用私钥签名、公钥验证，更安全，适用于需要分发公钥的场景

如 JWT 由第三方服务颁发，需根据第三方服务使用的签名算法进行配置。

## 功能限制类

### Q1：单个消费者最多可以配置多少个凭证？

A：单个消费者最多可以配置20个凭证，支持多种类型混合配置。

### Q2：单个 API 最多可以绑定多少个消费者？

A：单个 API 最多可以绑定100个消费者。

### Q3：OAuth 2.0支持哪些授权模式？

A：当前支持以下授权模式：

- **Client Credentials**(客户端凭证模式)：适用于机器对机器(M2M)场景
- **Password**(密码模式)：适用于可信客户端场景
- **Authorization Code**(授权码模式)：适用于 Web 应用场景

暂不支持 Implicit(隐式模式)和 Refresh Token 流程。

## 错误处理类

### Q1：调用 API 时返回401 Unauthorized 错误怎么办？

A：401错误表示认证失败，请检查：

1. 凭证是否正确(API Key、JWT Token 等)
2. 该消费者是否已绑定到 API 的认证策略中
3. 凭证是否已过期或被禁用
4. Authorization Header 格式是否正确(格式：`Bearer {凭证}`)
5. 如使用 JWT，检查签名算法和密钥是否与配置一致
6. 如使用 OAuth 2.0，检查 Access Token 是否有效

### Q2：配置 JWT 认证后，调用时返回"Invalid signature"错误?

A：这表示 JWT 签名验证失败，请检查：

1. 签名算法是否与配置一致(HS256、RS256、ES256)
2. 密钥/公钥是否正确：
   - HS256：密钥必须与签名时使用的密钥一致
   - RS256/ES256：公钥必须与签名时使用的私钥对应
3. JWT Token 是否被篡改或损坏

   建议使用 jwt.io 在线工具验证 JWT Token。

### Q3：配置 OAuth 2.0认证后，调用时返回"Failed to get access token"错误?

A：这表示网关无法从 OAuth 服务获取 Access Token，请检查：

1. token Endpoint 地址是否正确
2. Client ID 和 Client Secret 是否正确
3. OAuth 服务是否正常可用(网关能否访问到 token endpoint)
4. Scope 配置是否正确(某些 OAuth 服务要求 Scope 必须匹配)
5. 授权模式是否与 OAuth 服务支持的模式一致

建议先使用 curl 直接调用 OAuth 服务的 token endpoint，确认可以成功获取 Access Token。

## 实践教程类

### Q1：生产环境如何管理认证凭证?

A：建议的管理实践：

1. **凭证轮换**：定期更换凭证(如每季度更换一次)，降低凭证泄露风险
2. **有效期设置**：为凭证设置合理的有效期(如1年)，避免永久有效凭证泄露后持续被滥用
3. **权限分离**：为不同的业务系统创建不同的消费者，避免共用凭证
4. **监控告警**：监控认证失败率，异常流量及时告警
5. **凭证存储**：凭证信息存储在配置中心或密钥管理服务中，不要硬编码在代码中
6. **泄露应急**：如凭证泄露，立即禁用或删除该凭证，并创建新凭证

## Q2：如何实现多环境(开发、测试、生产)的认证隔离？

A：建议采用以下方案：

1. **方案1：多网关实例**
   - 为开发、测试、生产环境分别创建独立的网关实例
   - 每个实例配置独立的消费者和凭证
   - 环境间完全隔离，互不影响
2. **方案2：单网关实例+多消费者**
   - 使用单个网关实例
   - 为不同环境创建不同的消费者(如"开发消费者"、"生产消费者")
   - 创建不同的 API，分别绑定不同的消费者
   - 通过 API 访问地址区分环境(如 `/dev/api` 、 `/prod/api` )

建议使用方案1，环境隔离更彻底，更安全。

## Q3：如何设计多租户场景的认证方案？

A：对于 SaaS 等多租户场景，建议采用以下方案：

1. **租户隔离**：为每个租户创建独立的消费者
2. **认证方式**：
   - B 端租户：使用 API Key 或 OAuth 2.0，便于管理
   - C 端用户：使用 JWT 或 OIDC，支持用户身份信息传递
3. **租户识别**：
   - 方案 A：在 JWT 的 Claims 中包含租户 ID(如 `tenant_id` )
   - 方案 B：为每个租户分配独立的 API key
4. **访问控制**：在网关层识别租户，后端服务根据租户 ID 进行数据隔离
5. **计费与配额**：基于消费者维度进行流量统计和限流，实现租户级别的计费与配额管理

# 限流策略

最近更新时间：2026-07-24 16:45:00

## 操作场景

限流策略用于控制客户端对 AI 网关 API 的访问频率，保护后端模型服务免受突发流量冲击。AI 网关提供三层限流体系，您可以根据业务需求灵活选择：

| 限流层级 | 说明 | 适用场景 | 是否支持时间窗口策略 |
| --- | --- | --- | --- |
| 基础限流 | 全局限流，对当前模型 API 的所有请求进行整体限制，支持按并发数、请求数、总 Token 数等指标限流 | 保护后端服务的整体容量 | ✔ 支持 |
| 参数限流 | 按不同维度配置限流阈值，支持按消费者、API、路由等维度分别设置限流规则 | 多租户场景，不同消费者差异化配额 | ✖ 不支持 |
| 精细化限流 | 针对不同的参数值（如客户端 IP）设置精细化限流阈值，按顺序匹配 | 防止单一客户端异常占用 | ✖ 不支持 |

本文档将指导您如何在 AI 网关中配置和管理限流策略。

## 前置条件

- 已创建 AI 网关实例
- 已创建模型 API
- 已创建消费者（如需按消费者限流）

## 操作步骤

限流策略在创建或编辑模型 API 时配置，详情请参见 模型 API 。

### 步骤1：进入目标 API 限流策略配置页

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的"ID"，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**模型管理**，单击**模型 API** 页签。
4. 单击目标模型 API 名称（已创建的 API），进入其详情页，单击**限流策略**。


## 步骤2：开启限流

开启限流开关。

## 步骤3：选择限流层级并配置限流规则

在限流配置区域，首先选择限流层级，不同层级的配置规则如下。

### 基础限流

基础限流对整个 API 生效，对所有请求统一计数。

1. 开启基础限流。
2. 根据实际需求选择限流指标类型并配置对应阈值：
   - 总 Token（时间窗口内允许的总 Token 消耗量）
   - 请求数（时间窗口内允许的最大请求次数）
   - 并发数（允许的最大并发连接数）
3. 可单击添加限流规则，为当前 API 配置多层限流，例如：
   - 规则 1：并发数最多 1000 个并发
   - 规则 2：总 Token 每分钟请求数最多 1000 个

> **说明：**
> 基础限流的阈值在未匹配任何时间窗口规则时作为兜底值生效。如需按不同时段配置差异化阈值，请参见步骤4：启用时间窗口策略。

### 参数限流

参数限流允许按不同维度分别配置限流阈值，不同维度独立计数。

1. 开启参数限流。
2. 根据实际需求选择选择限流维度并配置对应阈值：
   - 消费者：按消费者限流，不同消费者独立计数，适用于多租户场景
   - API：按 API 限流，所有消费者共享配额，适用于保护后端服务
   - 路由：按路由限流，仅适用于 Agent API，可为不同的路由配置不同的限流策略
3. 可单击添加限流阈值，为消费者配置多层限流，例如：
   - 规则 1：消费者 A，并发数最多 1000 个并发

- 规则 2：消费者 A，每分钟请求数最多 5000 次

**精细化限流**

精细化限流支持按具体条件（如客户端 IP）设置限流规则，按配置顺序匹配。

1. 开启精细化限流。
2. 配置限流条件，根据实际需求选择条件类型、填写参数名和参数值。
3. 为该条件配置限流阈值。
4. 可单击添加限流阈值，为同一条件添加多条限流阈值。
5. 可单击添加规则，新增精细化限流的不同条件并配置对应限流阈值。

> **注意：**
> 在同一限流层级中，同一限流指标只能配置一条规则。若尝试重复添加，系统将弹出提示信息并拦截。
> 可同时配置基础限流、参数限流和精细化限流，三层规则组合生效。网关会依次检查所有规则，任意规则触发限流都会拒绝请求。

### 步骤4：启用时间窗口策略（可选）

时间窗口策略允许按不同时间段配置差异化的限流阈值，实现业务高峰期和低峰期的动态流量管控。仅基础限流支持配置时间窗口策略，参数限流和精细化限流暂不支持。

1. 在基础限流配置区域，勾选**启用时间窗口策略**复选框。
2. 勾选后，页面将自动展示时间窗口配置区域，包含快捷策略选项和时间窗口规则列表。
3. 通过以下三种方式添加时间窗口规则，并为每条规则的阈值配置相应限流阈值：
   - 快捷策略——工作日/非工作日
   - 快捷策略——白天/晚上
   - 自定义时段
4. 配置时间窗口策略并保存后，支持在限流策略页面单击规则生效预览，验证规则匹配结果。系统将展示所选时间点的基础限流生效规则详情，包括生效规则名称、优先级、限流指标与阈值。


**说明：**

**请求到达时，系统将按以下逻辑确定生效的基础限流阈值：**

| 场景 | 匹配结果 | 生效阈值 |
| --- | --- | --- |
| 当前时间匹配到某条时间窗口规则 | 匹配成功 | 使用该时间窗口规则的阈值 |
| 当前时间未匹配任何时间窗口规则 | 匹配失败 | 回退到基础限流配置的阈值（兜底值） |

**示例：基础限流配置 1 分钟最多 500 个请求，启用时间窗口策略后添加规则"工作日 09:00–18:00 1 分钟最多 100 个请求"。则工作日白天生效 100，其他时段生效 500。**

## 步骤5：保存配置

单击**确定**保存限流策略配置。

# 缓存策略

最近更新时间：2026-07-23 17:07:33

## 操作场景

AI 缓存能力基于双层缓存架构（精确缓存 L1 + 语义缓存 L2），帮助企业降低 LLM 调用成本、提升响应速度。适用于以下场景：

- **降低成本**：命中缓存的请求不调用 LLM，节省全部 Token 费用，典型场景可节省 30%～60%。
- **提升速度**：精确缓存命中延迟 < 2ms，语义缓存命中延迟 < 200ms，相比 LLM 推理（1–10 秒）大幅提升。
- **语义匹配**：语义相似但措辞不同的请求可命中缓存，突破传统精确匹配局限。
- **双层架构**：先精确后语义的两阶段查询，兼顾速度与命中率。

## 前置条件

### 精确缓存前置条件

1. 已创建 AI 网关实例，已创建模型 API 并配置至少一个路由；
2. 已配置 Redis 服务来源；
3. Redis 实例可正常访问。

### 语义缓存前置条件（额外）

已有向量数据库 VDB，并且在向量数据库中创建数据库和 Collection，要求：

- 索引类型：HNSW
- 相似性方法：**COSINE**（必须）
- 内置 Embedding：已开启并选择模型（如 `bge-base-zh`）
- 索引状态：`ready`

## 操作步骤

### 配置缓存

精确缓存基于 Redis 存储，对完全相同的 LLM 请求返回缓存响应，跳过 LLM 调用。语义缓存基于向量相似度匹配，对语义相似但措辞不同的请求返回缓存响应。

#### 步骤 1：进入模型 API 缓存策略页

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关 > 实例列表**；
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 选择左侧导航栏 **模型管理 > 模型 API**；
4. 单击目标模型 API 名称，进入详情页；

5. 单击 **缓存策略** Tab，在 **缓存策略** Tab 中，查看当前缓存配置状态。如果尚未配置，显示 **未配置**；
6. 单击 **配置缓存策略**，进入缓存配置对话框。

### 步骤 2：配置精确缓存

打开 **是否启用开关** 后，配置以下参数：
**配置参数说明：**

| 参数 | 是否必填 | 说明 | 推荐值 |
| --- | --- | --- | --- |
| 缓存键策略 | 是 | 缓存匹配范围 | 单轮问答选"最新用户消息"，多轮对话选"历史对话模式" |
| TTL（秒） | 是 | 缓存有效期，范围 60–604800 | 3600（1小时） |
| Redis 存储配置 | 是 | 缓存存储地址 | 服务地址，端口，用户名与密码 |

### 步骤 3：配置语义缓存（可选）

在缓存配置对话框中，打开 **是否启用开关** 以开启 **语义缓存**。
**配置参数说明：**

| 参数 | 是否必填 | 说明 | 推荐值 |
| --- | --- | --- | --- |
| VDB 服务来源 | 是 | 选择已配置的腾讯云 VDB 服务来源 | 同地域 VDB 实例 |
| 相似度阈值 | 是 | 范围 0–1，高于此值视为命中 | 0.85 |
| 最大 Token 距离 | 是 | 请求 Token 总数超过此值时不使用语义缓存 | 50 |
| TTL（秒） | 是 | 缓存有效期，范围 60–604800 | 3600（1小时） |
| 数据库 | 是 | VDB 实例下的数据库，系统自动拉取列表 | — |
| 向量存储配置 | 是 | 缓存存储地址 | 服务地址，端口，用户名与密码 |

高级配置：

| 参数 | 说明 |
| --- | --- |
| 按用户隔离缓存 | 默认关闭，不同用户的相同请求会生成独立缓存，提高隐私性 |
| 仅缓存成功响应 | 默认开启，只缓存 200 状态码的成功响应，错误请求不缓存 |
| 返回缓存标识 | 默认关闭，在响应 Header 中添加 X-Cache: HIT/MISS 标识 |

### 步骤 4：保存缓存配置

单击 **确定** 保存配置，缓存立即生效。

## 查看缓存策略列表

在左侧导航栏选择 **成本管理 > AI 缓存配置**，可查看缓存策略列表

## 查看缓存命中统计

在 **成本管理 > AI 缓存命中统计** 页面，可查看缓存命中率、节省成本等可观测指标，评估缓存效果。

### 步骤 1：进入 AI 缓存命中统计页面

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关 > 实例列表；**
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 在左侧导航栏选择 **成本管理 > AI 缓存命中统计**

### 步骤 2：配置筛选条件

| 筛选条件 | 说明 | 选项 |
| --- | --- | --- |
| 消费者组 | 按消费者组筛选（可选全部） | 全部 / 指定消费者组 |
| 模型 API | 按模型 API 筛选（可选全部） | 全部 / 指定模型 API |
| 时间范围 | 选择数据展示的时间窗口 | **今天/ 本周/ 本月/ 近7天/ 近30天/ 自定义** |

### 步骤 3：查看核心指标卡片

快速查看缓存命中的核心指标：

| 指标 | 说明 |
| --- | --- |
| **总请求次数** | 选定时间范围内网关收到的 LLM 请求总数 |

| | |
|---|---|
| 精确命中次数 | L1 精确缓存命中次数，显示实际命中数和占比 |
| 语义命中次数 | L2 语义缓存命中次数，显示实际命中数和占比 |
| 网关缓存命中率 | （精确命中 + 语义命中）/ 总请求数 × 100% |
| 网关缓存节省 | 命中请求未调用 LLM 节省的估算费用（基于模型单价计算） |

### 步骤 4：查看语义缓存相似度分布

展示语义缓存命中的相似度分数分布，帮助评估语义缓存阈值设置的合理性：
**相似度分布分析：**

| 相似度区间 | 说明 |
|---|---|
| 0.9 – 1.0 | 高相似度命中，表示用户请求与缓存内容非常接近 |
| 0.8 – 0.9 | 中等相似度命中，表示语义相似但措辞不同的请求 |
| 0.7 – 0.8 | 较低相似度命中，需要关注是否出现错误匹配 |
| 0.6 – 0.7 | 低相似度命中，建议检查是否误判 |
| 0 – 0.6 | 极低相似度命中，可能存在误匹配风险 |

**调优参考：**
如果大量命中集中在 0.8 以下区间（占比 > 30%），建议适当提高相似度阈值；如果 0.9 以上区间的命中数量过低（占比 < 20%），建议适当降低阈值以提高命中率。

### 步骤 5：查看缓存命中趋势与分布

趋势与分布面板说明：

| 面板名称 | 说明 |
|---|---|
| 缓存命中率趋势 | 展示选定时间范围内的缓存命中率变化趋势（按小时/天粒度） |
| 缓存类型分布 | 饼图展示精确缓存与语义缓存的命中比例 |
| 响应时间对比 | 对比缓存命中请求与未命中请求的平均响应时间，展示提升倍率 |

### 步骤 6：查看缓存命中明细

缓存命中明细表格展示按模型 API 和按消费者的详细命中数据。

| 列名 | 说明 |
| --- | --- |
| 消费者组 | 请求所属的消费者组 |
| 缓存命中数 | 缓存命中（精确 + 语义）的请求数量 |
| 缓存未命中数 | 未命中缓存、直接调用 LLM 的请求数量 |
| 命中率 | 缓存命中数 /（命中数 + 未命中数）× 100% |
| 节省成本 | 缓存命中节省的估算费用（￥） |
| 节省 Token | 缓存命中节省的估算 Token 数 |
| 命中平均响应时间 | 缓存命中请求的平均响应时间（ms） |
| 未命中平均响应时间 | 未命中请求的平均响应时间（ms） |
| 提升倍率 | 未命中平均响应时间 / 命中平均响应时间 |

### 步骤 7：导出统计报表

在缓存命中统计页面，可单击**导出 CSV** 按钮，下载当前筛选条件下的明细数据。

# 参数改写

最近更新时间：2026-07-23 17:07:33

## 操作场景

参数改写策略支持在请求转发给后端模型服务之前，对请求的 Body 参数和 Query 参数进行新增、编辑或删除操作。适用于以下场景：

- **参数注入**：为所有请求自动添加必要的参数（如固定的 `temperature` 、 `max_tokens` ）
- **参数覆盖**：覆盖客户端传入的参数值，统一控制模型行为
- **参数清理**：删除敏感或不必要的参数，防止参数泄露
- **协议适配**：适配不同供应商的参数差异（如将 `top_p` 改写为 `topP` ）
- **多版本兼容**：支持不同 API 版本的参数格式转换

> **说明：**
> 参数改写发生在请求路由之后、转发给后端模型服务之前，仅影响转发的请求，不影响客户端收到的响应。

## 前置条件

1. 已创建 AI 网关实例且处于运行中状态
2. 已创建模型 API
3. 了解后端模型服务支持的参数格式和要求

## 操作步骤

### 配置参数改写

#### 步骤 1：进入模型 API 详情页

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **模型管理**，然后单击**模型 API** 页签。
4. 单击目标模型 API 名称，进入 API 详情页。
5. 切换到 **参数改写** Tab。

#### 步骤 2：参数改写配置

1. 在 **参数改写** Tab 中，查看当前配置状态。
2. 单击 **配置参数改写规则** 进入配置对话框。
3. 在配置对话框中，开启 **是否启用** 开关。

4. 选择需要改写的参数，配置参数改写规则

**操作类型说明：**

| 操作类型 | 说明 | 使用场景 |
| --- | --- | --- |
| **新增** | 添加新的参数键值对 | 注入默认参数如 `temperature: 0.7` |
| **编辑** | 修改已有参数的值 | 覆盖客户端传入的参数 |
| **删除** | 删除指定的参数 | 清理敏感或不必要的参数 |

5. 单击**添加**按钮，可配置多条参数改写规则。规则按配置顺序依次执行。

### 步骤 3：保存配置

单击 **确定** 保存参数改写规则，配置立即生效。

## 典型配置示例

### 示例 1：统一温度参数

- **场景**：统一控制所有请求的 `temperature` 参数，避免客户端随意设置。
- **配置**：

| 阶段 | 操作类型 | 参数键 | 参数值 |
| --- | --- | --- | --- |
| Body 参数 | 编辑 | `temperature` | `0.7` |

- **效果**：无论客户端传入什么 temperature 值，最终都使用 0.7。

### 示例 2：清理敏感信息

- **场景**：删除请求中的用户标识信息，保护用户隐私。
- **配置**：

| 阶段 | 操作类型 | 参数键 | 参数值 |
| --- | --- | --- | --- |
| Body 参数 | 删除 | `user` | - |
| Body 参数 | 删除 | `metadata.user_id` | - |

- **效果**：请求中的用户标识字段被删除，不会转发给后端模型服务。

### 示例 3：适配 Azure OpenAI

- **场景**：适配 Azure OpenAI 的 `api-version` Query 参数要求。
- **配置**：

| 阶段 | 操作类型 | 参数键 | 参数值 |
| --- | --- | --- | --- |
| Query 参数 | 新增 | `api-version` | `2024-02-01` |

- **效果**：所有请求自动添加 `?api-version=2024-02-01` 。

## 注意事项

1. **执行时机**：参数改写发生在请求路由之后、转发给后端模型服务之前。
2. **执行顺序**：多条规则按配置顺序依次执行，建议将依赖性的规则按正确顺序排列。
3. **Body 参数限制**：仅支持对 JSON 格式的请求 Body 进行参数改写。
4. **参数键路径**：支持简单的参数键名（如 `temperature` ），暂不支持嵌套路径（如 `messages[0].content` ）。
5. **与缓存的配合**：参数改写发生在缓存查询之后，缓存的 Key 基于原始请求计算。
6. **与日志的关系**：LLM Log 记录的是改写后的请求包体（如已开启包体采集）。
7. **参数值类型**：当前版本参数值统一作为字符串处理，暂不支持数字、布尔等类型标记。

# 流量镜像

最近更新时间：2026-07-24 16:45:01

## 操作场景

流量镜像允许您在不影响线上用户的前提下，将模型 API 的部分生产请求异步复制到镜像模型服务，后台采集镜像模型的性能元数据（TTFT、TPOT、Token 数量、预估成本等），并与主模型进行并排对比。适用于以下场景：

- 评估新候选模型的性能，为模型切换提供数据支撑。
- 验证新模型的延迟和吞吐量是否满足生产要求。
- 用真实生产流量做性能对比，避免“沙盒好用、上线翻车”。

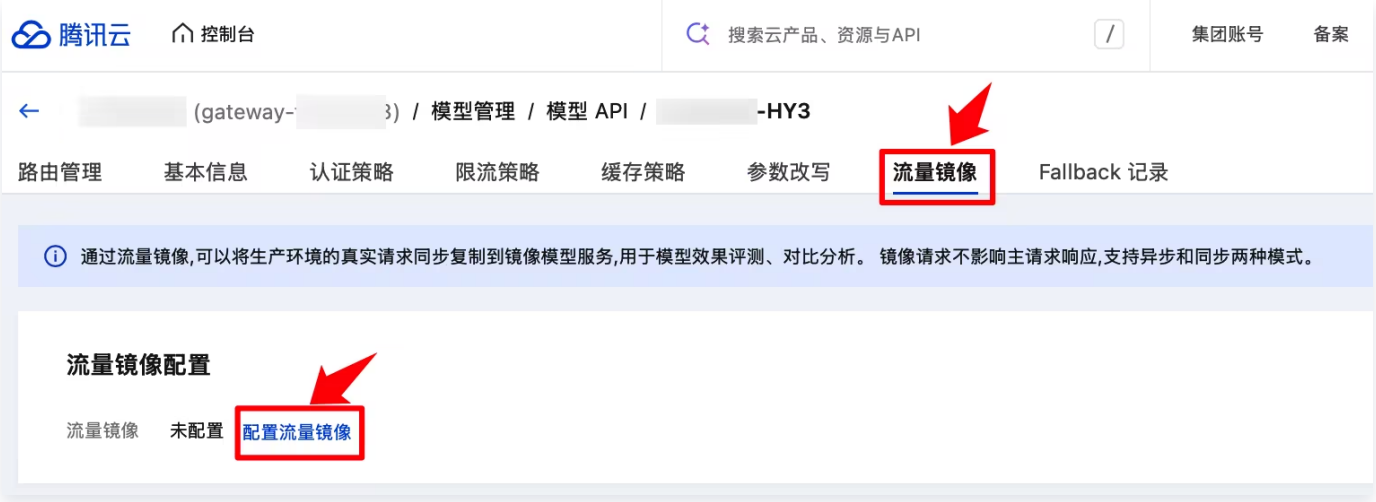
## 前提条件

- 已创建 AI 网关实例。
- 已创建模型 API。
- 已配置至少两个模型服务（一个作为主模型服务，一个作为镜像模型服务）。
- 镜像模型服务所使用的供应商 API Key 已配置且有效。
- 已开启 CLS 日志投递的 LLM 访问日志投递能力，详情请参见 CLS 日志投递。

## 操作步骤

### 步骤1：进入目标 API 流量镜像配置页

1. 登录 微服务平台控制台，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**模型管理**，单击**模型 API** 页签。
4. 单击目标模型 API 名称（已创建的 API），进入其详情页，单击**流量镜像 > 配置流量镜像**。




### 步骤2：配置流量镜像

1. 开启流量镜像开关，展开配置区域。
2. 在镜像目标下拉框中选择作为镜像目标的模型服务。
3. 设置镜像比例，默认值：10%，低于 100% 时系统按比例随机抽样决定是否镜像当前请求。

   建议值：

| 场景 | 建议比例 |
| --- | --- |
| 生产环境长期评估 | 10–30% |
| 临时调试、验证 | 100% |
| 首次开启测试 | 5–10% |

> **说明：**
> 镜像目标服务不能与主路径模型服务相同；
> 镜像目标服务独立于主路径，不参与主路径的 Fallback 和智能路由。

4. 在数据观测区域，勾选需要采集的性能指标（默认全选）：
   - 首 Token 时间 (TTFT)：从请求到收到第一个 Token 的时间。
   - 每 Token 耗时 (TPOT)：解码阶段平均每个输出 Token 的生成时间。
   - Token 数量 (Prompt + Completion)：输入和输出的 Token 数量。
   - 预估成本：基于 Token 数量和模型单价计算的预估费用。

> **说明：**
> 数据观测功能依赖 CLS 日志投递的 LLM 访问日志投递能力，请确保已开启。

## 步骤3：保存配置

单击确定保存配置，保存后配置立即生效。

> **注意：**
> 1. 镜像会产生额外费用：每次镜像请求会对镜像模型服务产生实际调用及对应 Token 费用，请提前评估成本。
> 2. 建议先小比例测试：首次开启建议设置镜像比例为 5–10%，确认镜像运行正常后再调高。
> 3. 镜像模型服务需足够容量：若镜像比例较高（>50%），请确保镜像模型服务有足够的并发承载能力。
> 4. 镜像数据不含质量评估：流量镜像仅采集性能元数据（TTFT/TPOT/Token/成本），响应质量评估需结合人工评审或外部评测工具。

# 降级策略

最近更新时间：2026-07-23 17:07:33

## 操作场景

降级策略用于在后端模型服务出现异常时，自动降级服务质量，保障系统可用性。通过配置降级策略，您可以：

- 当服务异常时自动触发降级
- 查看降级触发记录。

降级策略是保障系统稳定性的重要手段，与跨服务 Fallback 配合使用可实现完整的容灾方案。
本文档指导您如何在 AI 网关中配置和管理降级策略。

## 前置条件

- 已创建 AI 网关实例
- 已创建模型 API
- 已创建模型服务

## 操作步骤

降级策略在创建或编辑模型 API 时配置，详情请参见 模型 API 。

### 步骤1：进入 API 详情页

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的"ID"，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**模型管理**，单击**模型 API** 页签。
4. 单击目标模型 API 名称（已创建的 API），进入其详情页。
5. 在 API 详情页中找到**降级策略**区域，单击右上角**配置降级策略**按钮，进入降级策略配置对话框。

### 步骤2：配置降级触发条件

在配置对话框中，**开启降级开关**，配置以下参数：

| 参数 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| 触发条件 | 是 | 选择降级触发条件，支持多选：<br>• **服务不可用**：主模型服务返回 500 / 502 / 503 等错误时触发<br>• **连接超时**：请求超过超时时间无响应时触发<br>• **速率限制（HTTP 429）**：主模型服务触发限流时触发 | 服务不可用<br>速率限制 |

### 步骤3：配置备用模型服务

1. 依次添加备用模型服务（按顺序排列）。
2. 降级触发后，网关将按备用模型服务的顺序依次尝试调用。
3. 配置完成后，单击**确定**保存。

## 查看降级状态

在 API 详情页的**降级记录**区域，可查看历史降级记录：

- 降级触发时间
- 降级原因(具体的触发条件和阈值)
- 主服务
- 最终使用服务
- 降级结果

# 数据脱敏

最近更新时间：2026-07-23 17:07:33

## 操作场景

AI 网关支持在日志记录和请求转发两个环节对敏感数据进行脱敏处理，帮助您满足数据安全和合规要求：

- **日志脱敏**：在将 Prompt/Completion 内容写入 LLM Log 之前，对敏感信息进行掩码处理。适用于防止日志泄露场景，不影响 LLM 正常接收原始数据。
- **转发脱敏**：在将请求转发给后端 LLM 供应商之前，将敏感数据替换为占位符。适用于数据主权合规要求严格的场景，阻止敏感数据进入第三方模型供应商。

| 对比项 | 日志脱敏 | 转发脱敏 |
| --- | --- | --- |
| 处理时机 | 写入 LLM Log 之前 | 转发给 LLM 供应商之前 |
| 是否影响 LLM | 不影响，LLM 仍收到原始数据 | 影响，LLM 收到占位符 |
| 依赖包体采集 | 必须开启包体采集 | 不依赖 |
| 处理结果 | 掩码（如 `138****8888`） | 占位符（如 `[手机号]`） |
| 适用场景 | 大多数客户，既保合规又不影响模型效果 | 数据主权合规要求严格的客户 |

> **说明：**
> 数据脱敏功能需配合 **请求包体采集** 开关使用。若未开启包体采集，日志脱敏不会生效；转发脱敏不依赖包体采集开关。

## 前置条件

1. 已创建模型 API 且处于运行中状态；
2. （日志脱敏）已在模型 API 详情页开启 **请求包体采集（Prompt）** 或 **响应包体采集（Completion）**；
3. 已明确需要脱敏的敏感数据类型（如手机号、身份证号等）。

## 场景一：配置日志脱敏规则

日志脱敏在写入 LLM Log 前对敏感数据进行掩码处理，适用于需要记录日志内容同时保护敏感信息的场景。

### 步骤 1：进入模型 API 详情页

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。

3. 在左侧导航栏单击 **模型管理**，然后单击**模型 API** 页签。
4. 单击目标模型 API 名称，进入 API 详情页，切换至 **基本信息** Tab。

## 步骤 2：开启包体采集

在 **日志配置** 区域，确认已开启包体采集

> **注意：**
> 开启包体采集后，请求和响应的文本内容将写入日志存储。请确认内容不含用户个人隐私数据，且符合企业合规要求。

## 步骤 3：配置日志脱敏规则

在 **日志脱敏规则（包体采集开启时生效）** 区域，单击 **配置规则**，进入配置对话框。

**配置参数说明：**

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 是否启用 | 是 | 日志脱敏的总开关，关闭时所有脱敏规则不生效 |
| 内置脱敏类型 | 否 | 勾选需要脱敏的内置数据类型 |
| 脱敏范围 | 否 | 选择脱敏作用范围：仅 Prompt、仅 Completion、或两者 |
| 自定义脱敏规则 | 否 | 添加业务特有的敏感数据识别规则 |

**内置脱敏类型说明：**

| 脱敏类型 | 识别规则 | 掩码示例 |
| --- | --- | --- |
| 手机号 | 大陆 11 位手机号（ `1[3-9]\d{9}` ） | `138****8888` |
| 身份证号 | 18 位或 15 位身份证号 | `110101****1234` |
| 银行卡号 | 16–19 位数字（主流银行卡格式） | `6222****0001` |
| 邮箱 | 标准邮箱格式 | `abc***@***.com` |
| IP 地址 | IPv4 格式 | `192.168.*.*` |
| 姓名 | 2–4 位中文字符 | `张**` |

## 步骤 4：配置自定义脱敏规则（可选）

如果内置脱敏类型无法满足业务需求，可添加自定义规则：

1. 单击 **添加规则**，填写规则参数：

| 参数 | 说明 | 示例 |
| --- | --- | --- |
| 规则名称 | 规则标识名称，最长 60 个字符 | 员工工号 |
| 匹配正则 | 用于匹配敏感数据的正则表达式 | `E\d{6}` |
| 掩码格式 | 脱敏后的掩码格式 | `E***` |
| 启用 | 是否启用该规则 | 勾选启用 |

**规则名称限制：**

- 最长 60 个字符
- 支持中英文大小写、数字及分隔符（ `-` 、 `_` ）
- 不能以数字和分隔符开头
- 不能以分隔符结尾

2. 单击 **确定** 保存自定义规则

### 步骤 5：保存配置

在配置对话框中单击 **确定**，日志脱敏规则立即生效。

## 场景二：配置转发脱敏规则

转发脱敏在请求转发给 LLM 供应商之前，将敏感数据替换为占位符。适用于数据主权合规要求严格的场景。

> **注意：**
> 开启转发脱敏后，LLM 收到的是占位符而非真实数据，模型对上下文的理解可能受影响。建议仅在有明确数据主权合规要求时启用。

### 步骤 1：进入模型 API 详情页

同场景一的步骤 1。

### 步骤 2：配置转发脱敏规则

在 **数据安全配置（转发脱敏）** 区域，单击 **配置规则**，进入配置对话框。

**配置参数说明：**

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 是否启用 | 是 | 转发脱敏的总开关 |

| | | |
|---|---|---|
| 内置脱敏类型 | 否 | 勾选需要在转发时脱敏的数据类型 |
| 占位符格式 | 是 | 替换后的占位符格式，默认自动（`[{type}]`），支持自定义，系统会自动将 `{type}` 替换为对应敏感类型的名称。 |
| 脱敏失败处理 | 是 | 脱敏规则处理异常时的行为 |
| 自定义脱敏规则 | 否 | 添加业务特有的敏感数据识别规则 |

**脱敏失败处理策略：**

| 策略 | 说明 | 适用场景 |
|---|---|---|
| **拒绝请求并返回 500（推荐）** | 脱敏处理异常时直接拒绝请求，保证数据不泄露 | 强合规场景 |
| **跳过脱敏直接转发** | 脱敏失败时原样转发，不阻断请求 | 不推荐，存在数据泄露风险 |

### 步骤 3：保存配置

在配置对话框中单击 **确定**，转发脱敏规则立即生效。

## 注意事项

1. **包体采集依赖**：日志脱敏仅在开启包体采集时生效。若未开启包体采集，日志中不记录 Prompt/Completion 内容，无需脱敏
2. **转发脱敏权衡**：转发脱敏会改变 LLM 收到的实际内容，可能影响模型输出质量。建议仅在强合规场景下启用
3. **性能影响**：脱敏处理会增加网关的 CPU 开销，高并发场景下建议进行性能测试
4. **自定义正则安全**：自定义脱敏规则的正则表达式请谨慎编写，避免过于宽泛的规则导致正常内容被误脱敏
5. **脱敏范围选择**：建议根据实际需求选择脱敏范围：
   - 仅需审计请求输入 → 仅 Prompt
   - 仅需审计模型输出 → 仅 Completion
   - 需要完整审计 → Prompt + Completion
6. **姓名识别谨慎**：姓名脱敏的误判率较高，普通中文词可能被误识别为姓名，建议评估后谨慎启用
7. **占位符对模型的影响**：转发脱敏后的占位符（如 `[手机号]`）会被 LLM 当作普通文本处理，模型无法理解其真实含义

# 智能路由
# 概述

最近更新时间：2026-07-23 17:07:35

## 操作场景

智能路由策略是 AI 网关在多模型服务场景下的流量分发能力。通过配置智能路由策略，您可以：

- 实现多模型服务的负载均衡和流量分配。
- 根据模型名称自动匹配对应的后端服务。
- 基于权重配置灵活控制流量比例。
- 根据用户请求的语义内容自动分发到最优模型。
- 根据请求参数（如 Header）将不同租户/环境的流量隔离路由。
- 通过服务标签筛选可用服务，实现模型过滤。
- 根据实时延迟情况自动选择响应最快的服务。
- 支持 A/B 测试、灰度发布等场景。
- 配合 Fallback 机制实现高可用。

本文档指导您如何在 AI 网关控制台配置和使用智能路由策略。

## 前置条件

- 已创建 AI 网关实例；
- 已创建至少一个模型服务(内置供应商或自定义模型服务)；
- 已创建模型 API。

## 路由策略概述

AI 网关支持多种智能路由策略：

| 路由策略 | 使用场景 | 决策依据 | 适用范围 |
| --- | --- | --- | --- |
| 按模型名称路由 | 多模型服务精确匹配 | 请求中的 model 字段与模型服务配置的模型名称匹配 | 单模型 API 需访问不同供应商的相同模型 |
| 权重路由 | 负载均衡、A/B 测试、灰度发布 | 按配置的权重比例随机分配流量 | 同一模型的多实例部署或多供应商混合部署 |
| 意图识别路由 | 按语义内容分类分发 | 调用意图识别模型分析请求语义，识别意图后路由 | 多任务模型分发、按问题复杂度分级服务 |

| | | | |
|---|---|---|---|
| 参数路由 | 多租户隔离、环境隔离、用户分流 | 根据请求 Header 参数值匹配路由规则 | 需按租户/环境/用户类型将流量路由到不同服务 |
| 基于标签路由 | 模型标签筛选 | 通过模型服务标签（Key-Value）自动筛选可用服务 | 模型众多，需要通过标签筛选 |
| 延迟优先路由 | 跨地域多线路选路、响应体验优化 | 根据模型服务实时延迟监测数据选择最优路径 | 多地域部署、多 IDC 接入场景 |

# 模型名称路由

最近更新时间：2026-07-23 17:07:33

## 功能说明

按模型名称路由会根据客户端请求中的 `model` 字段自动匹配对应的模型服务。这种路由策略适用于以下场景：

- 同一个 API 需要支持多个不同的模型(如 `gpt-4o` 、 `gpt-4o-mini` 、 `claude-3-5-sonnet` )
- 不同模型由不同的供应商或服务实例提供
- 需要精确控制每个模型的路由目标
- 需要使用通配符匹配一组模型(如 `gpt-4-*` 匹配所有 gpt-4系列模型)

## 配置步骤

### 步骤1：创建模型服务

在配置路由策略前，需要先创建模型服务：

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关 > 实例列表；**
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 在左侧导航栏单击**模型管理**，然后单击**模型服务**页签；
4. 在服务列表中单击**新建**，配置供应商、访问凭证等信息；
5. 保存模型服务。

### 步骤2：在模型 API 中配置模型名称路由

1. 在**模型管理 > 模型 API** 页签单击**新建；**
2. 完成**基本信息**配置后，在**第二步：选择模型服务** 页面，配置服务类型和路由策略；
3. 选择**服务类型**为**多模型服务；**
4. 选择**路由策略**为**模型名称路由；**
5. 在模型服务表格中添加服务并配置路由规则。

   **配置参数说明：**

| 参数 | 说明 | 示例 | 是否必填 |
| --- | --- | --- | --- |
| 模型服务 | 从已创建的模型服务列表中选择 | `gpt-4o-openai` | 是 |
| 匹配模型名 | 客户端请求的 model 参数匹配规则，支持精确匹配和通配符匹配( `*` ) | `gpt-4o` 或 `gpt-4-*` | 是 |

| | | | |
|---|---|---|---|
| 重写模型名 | 转发到后端服务时，将请求中的 model 参数重写为指定值。留空则透传原始值 | `gpt-4o` 或留空 | 否 |

### 步骤3：保存并发布

配置完成后，单击 **确定** 保存配置。路由规则会立即生效。

## 配置示例

### 示例1：精确匹配多个模型

- **场景**：单个 API 需要支持3个不同的模型，分别由不同的服务提供
- **配置**：

| 模型服务 | 匹配模型名 | 重写模型名 |
|---|---|---|
| gpt-4o-openai | gpt-4o | (留空) |
| gpt-4o-mini-openai | gpt-4o-mini | (留空) |
| claude-3-5-sonnet-anthropic | claude-3-5-sonnet | (留空) |

- **测试请求**：

```
# 请求1:路由到 gpt-4o-openai
curl -X POST https://{网关域名}/ai/llm/v1/chat/completions \
  -H "Authorization:Bearer {API_Key}" \
  -H "Content-Type:application/json" \
  -d '{"model":"gpt-4o", "messages":[{"role":"user", "content":"你好"}]}'

# 请求2:路由到 claude-3-5-sonnet-anthropic
curl -X POST https://{网关域名}/ai/llm/v1/chat/completions \
  -H "Authorization:Bearer {API_Key}" \
  -H "Content-Type:application/json" \
  -d '{"model":"claude-3-5-sonnet", "messages": [{"role": "user", "content": "你好"}]}'
```

### 示例2：通配符匹配模型系列

- **场景**：希望所有 `gpt-4-*` 系列的模型都路由到同一个服务

- **配置：**

| 模型服务 | 匹配模型名 | 重写模型名 |
| --- | --- | --- |
| gpt-4-family-openai | gpt-4-* | (留空) |
| gpt-3.5-turbo-openai | gpt-3.5-turbo | (留空) |

- **测试请求：**

```
# 请求1：匹配通配符规则，路由到 gpt-4-family-openai，透传model=gpt-4o
curl -X POST https://{网关域名}/ai/llm/v1/chat/completions \
  -H "Authorization: Bearer {API_Key}" \
  -H "Content-Type: application/json" \
  -d '{"model": "gpt-4o", "messages": [{"role": "user", "content": "你好"}]}'

# 请求2：匹配通配符规则，路由到 gpt-4-family-openai，透传model=gpt-4-turbo
curl -X POST https://{网关域名}/ai/llm/v1/chat/completions \
  -H "Authorization: Bearer {API_Key}" \
  -H "Content-Type: application/json" \
  -d '{"model": "gpt-4-turbo", "messages": [{"role": "user", "content": "你好"}]}'

# 请求3：精确匹配，路由到 gpt-3.5-turbo-openai
curl -X POST https://{网关域名}/ai/llm/v1/chat/completions \
  -H "Authorization: Bearer {API_Key}" \
  -H "Content-Type: application/json" \
  -d '{"model": "gpt-3.5-turbo", "messages": [{"role": "user", "content": "你好"}]}'
```

## 示例3：模型名称重写

- **场景：客户端使用自定义模型名称，但后端服务只识别标准模型名**
- **配置：**

| 模型服务 | 匹配模型名 | 重写模型名 |
| --- | --- | --- |
| gpt-4o-openai | my-custom-gpt4 | gpt-4o |
| claude-3-5-sonnet-anthropic | my-custom-claude | claude-3-5-sonnet-20241022 |

- **测试请求：**

```
# 客户端请求model=my-custom-gpt4，网关转发时重写为model=gpt-4o
curl -X POST https://{网关域名}/ai/llm/v1/chat/completions \
  -H "Authorization: Bearer {API_Key}" \
  -H "Content-Type: application/json" \
  -d '{"model": "my-custom-gpt4", "messages": [{"role": "user",
"content": "你好"}]}'

# 转发到后端OpenAI服务的请求body：
# {"model": "gpt-4o", "messages": [{"role": "user", "content": "你
好"}]}
```

# 路由匹配规则

## 匹配优先级

当同一个请求可以匹配多个路由规则时，按以下优先级匹配：

1. **精确匹配优先于通配符匹配**

   如果配置了 `gpt-4o`（精确）和 `gpt-4-*`（通配符），请求 `model=gpt-4o` 会匹配精确规则
2. **匹配顺序按配置顺序**

   当多个通配符规则都匹配时，选择配置表格中最先添加的规则
3. **未匹配时返回404**

   如果请求的 model 参数无法匹配任何规则，返回错误响应

## 通配符规则

**支持的通配符：**

| 通配符 | 说明 | 示例 | 匹配结果 |
| --- | --- | --- | --- |
| `*` | 匹配任意字符(0个或多个) | `gpt-4-*` | 匹配 `gpt-4-turbo`、`gpt-4o`、`gpt-4-0125-preview` |
| `gpt-*` | 前缀匹配 | `gpt-*` | 匹配所有以 `gpt-` 开头的模型 |

**注意事项：**

- 通配符 `*` 只能在匹配模型名字段使用，不支持在重写模型名字段使用
- 单个匹配规则只能包含一个通配符 `*`
- 通配符大小写敏感，`GPT-*` 不会匹配 `gpt-4o`

# 模型名称重写逻辑

**重写时机：**

- 网关在转发请求到后端模型服务之前，会将请求 body 中的 `model` 字段替换为"重写模型名"配置的值
- 如果"重写模型名"留空，则透传客户端请求中的原始 model 值

**典型场景：**

| 场景 | 匹配模型名 | 重写模型名 | 客户端请求 model | 转发到后端的 model |
| --- | --- | --- | --- | --- |
| 透传原始 model | `gpt-4o` | (留空) | `gpt-4o` | `gpt-4o` |
| 统一模型标识 | `gpt-4-*` | `gpt-4-turbo-2024-04-09` | `gpt-4-turbo` | `gpt-4-turbo-2024-04-09` |
| 自定义别名 | `my-gpt4` | `gpt-4o` | `my-gpt4` | `gpt-4o` |

**未匹配处理**

如果请求的模型名称无法匹配任何路由规则，网关会返回：

```
{
  "error": {
    "code": "model_not_found",
    "message": "模型 'gpt-5' 不存在，请检查模型名称是否正确。可用模型：gpt-4o, gpt-4-*, claude-3-5-sonnet",
    "type": "invalid_request_error"
  }
}
```

错误信息中会列出当前 API 支持的所有匹配规则，方便客户端排查问题。

# 权重路由

最近更新时间：2026-07-23 17:07:33

## 功能说明

权重路由按照配置的权重比例将流量分配到多个模型服务。这种路由策略适用于以下场景：

- **负载均衡**：为同一模型的多个实例分配流量，避免单点过载
- **A/B 测试**：按比例测试不同供应商或模型版本的效果
- **灰度发布**：小流量验证新服务，逐步放大流量比例
- **成本优化**：将大部分流量导向成本更低的服务，保留少量高质量服务作为补充

## 配置步骤

### 步骤1：在模型 API 中配置权重路由

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 在左侧导航栏单击**模型管理 > 模型 API**；
4. 单击**新建**或编辑已有 API；
5. 选择服务类型为多模型服务；
6. 完成**基本信息**配置后，在**第二步：选择模型服务与路由策略**页面；
7. 选择路由策略为**权重路由**；
8. 为每个模型服务配置权重。

   **配置参数**：

| 参数 | 说明 | 示例 | 约束 |
| --- | --- | --- | --- |
| 模型服务 | 从已创建的模型服务列表中选择 | `qwen-plus-aliyun` | - |
| 权重 | 该服务的流量权重，范围1-100 | `70` | 必填，整数 |

### 步骤2：保存并发布

配置完成后，单击 **确定** 保存配置。权重路由策略会立即生效。

### 权重分配逻辑

- **权重总和不限制**：

  权重总和不要求等于100，网关会自动按比例计算。

服务A权重50，服务B权重30：

- 服务A流量占比 = 50 / (50+30) = 62.5%
- 服务B流量占比 = 30 / (50+30) = 37.5%

- **动态调整权重：**

  您可以随时编辑模型 API，调整权重配置，变更会在保存后立即生效(无需重启服务)。

# 典型场景示例

## 场景1：负载均衡 – 多实例部署

- **需求：**自建了3个 vLLM 实例，希望均匀分配流量
- **配置：**

| 模型服务 | 权重 | 流量占比 |
| --- | --- | --- |
| vllm-instance-1 | 33 | 33% |
| vllm-instance-2 | 33 | 33% |
| vllm-instance-3 | 34 | 34% |

- **效果：**每个实例接收约1/3的流量，实现负载均衡。

## 场景2：A/B 测试 – 新旧服务对比

- **需求：**测试新部署的 qwen-max-2.5模型，与现有 qwen-max 对比效果
- **配置：**

| 模型服务 | 权重 | 流量占比 | 说明 |
| --- | --- | --- | --- |
| qwen-max(旧版本) | 80 | 80% | 主流量 |
| qwen-max-2.5(新版本) | 20 | 20% | 测试流量 |

- **效果：**20%流量路由到新版本，收集效果数据后决定是否全量切换。

## 场景3：灰度发布 – 逐步放量

- **需求：**新模型服务需要灰度发布，从5%流量逐步放大到100%
- **阶段1：初始灰度(5%流量)**

| 模型服务 | 权重 |
| --- | --- |
| 旧服务 | 95 |

| | |
|---|---|
| 新服务 | 5 |

- **阶段2：扩大灰度(30%流量)**

| 模型服务 | 权重 |
|---|---|
| 旧服务 | 70 |
| 新服务 | 30 |

- **阶段3：全量切换(100%流量)**

| 模型服务 | 权重 |
|---|---|
| 新服务 | 100 |

删除旧服务或权重设为0

## 场景4：成本优化 – 混合供应商

- **需求**：大部分流量使用成本较低的国产模型，保留少量 Open AI 服务作为备份
- **配置**：

| 模型服务 | 权重 | 流量占比 | 成本 |
|---|---|---|---|
| qwen-max-aliyun | 85 | 85% | 低 |
| gpt-4o-openai | 15 | 15% | 高 |

- **效果**：85%流量使用成本更低的 qwen-max，15%流量使用 gpt-4o 保障质量。

# 意图识别路由

最近更新时间：2026-07-23 17:07:33

## 操作场景

意图识别路由策略基于用户请求的语义内容，自动识别用户的意图类别，并将请求路由到对应的模型服务。适用于以下场景：

- **多任务模型分发**：根据用户意图将翻译、代码生成、对话等不同任务路由到最优模型
- **成本优化**：简单意图（如闲聊）路由到低成本模型，复杂意图（如代码审查）路由到高性能模型
- **分级服务**：VIP 用户意图路由到专有服务，普通用户路由到共享服务
- **场景隔离**：将不同业务场景的请求隔离到不同的模型服务，避免相互影响

> **说明：**
> 意图识别路由需选择一个**意图识别模型服务**来执行语义分析。该模型用于理解用户请求的意图，不直接处理用户的业务请求。

## 前置条件

- 已创建 AI 网关实例且处于运行中状态
- 已创建至少两个模型服务（意图识别模型 + 业务模型）
- 已明确需要识别的意图类别及其对应的路由目标

## 操作步骤

### 步骤 1：配置模型 API

1. 登录 微服务平台控制台，在左侧导航栏单击 **AI 网关 > 实例列表**；
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 选择左侧导航栏 **模型管理 > 模型 API**；
4. 单击**新建**或编辑已有 API；
5. 完成基本信息配置后，进入 **第二步：选择模型服务**；
6. 选择 **服务类型** 为 **多模型服务**；
7. 在 **路由策略** 区域，选择 **意图识别路由**。

### 步骤 2：配置意图识别模型

在 **意图识别模型配置** 区域，选择用于意图识别的模型：
**配置参数说明：**

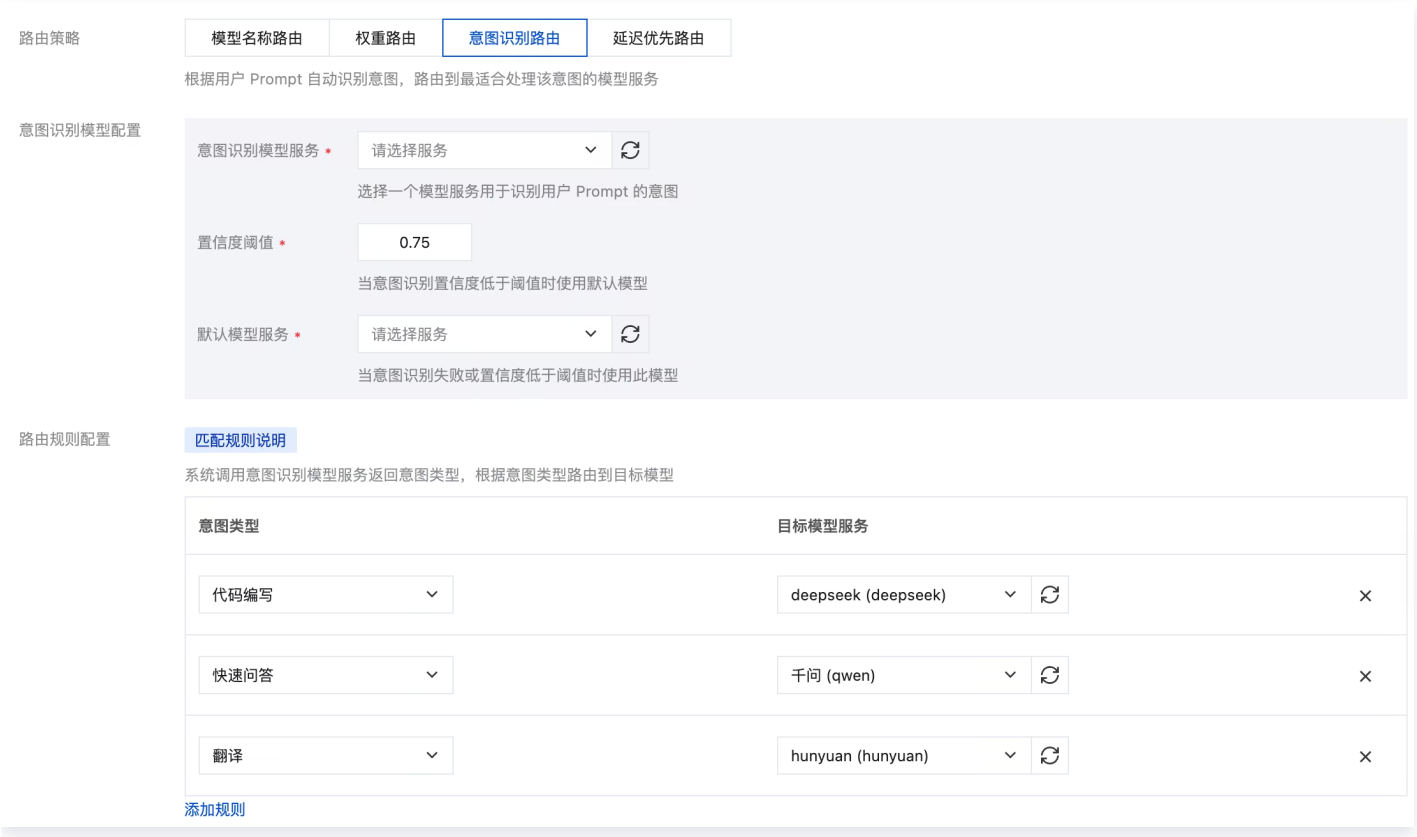
| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 意图识别模型 | 是 | 用于执行意图识别的模型服务 |
| 置信度阈值 | 是 | 意图识别置信度低于此值时，使用默认服务 |
| 默认模型服务 | 是 | 未匹配任何意图或置信度不足时使用的服务 |

**置信度阈值建议值：**

| 阈值 | 说明 | 适用场景 |
| --- | --- | --- |
| 0.50 | 较低的置信度要求，更多请求按识别意图路由 | 意图类别较少且区分度高时 |
| 0.70 | 推荐的置信度阈值，兼顾准确率和覆盖率 | 大多数场景 |
| 0.85 | 较高的置信度要求，仅置信度高的意图才按类别路由 | 对路由准确性要求高的场景 |

## 步骤 3：配置意图类型与模型映射

在**路由规则配置**区域，配置各意图对应的模型服务，支持代码编写、数学计算、翻译、快速问答和复杂推理场景。




## 步骤 4：保存配置

单击**确定**保存模型 API 配置，意图识别路由立即生效。

## 注意事项

- **意图识别模型选择**：意图识别模型需要具备良好的自然语言理解能力（NLU）。
- **额外 Token 消耗**：每个请求在路由前会先调用意图识别模型进行语义分析，产生额外的 Token 消耗和时间开销。
- **意图请求不记录**：意图识别调用的请求和响应不写入 LLM Log（仅为路由决策服务）。
- **置信度阈值调优**：建议先使用较低的阈值（0.5）观察识别效果，根据实际表现逐步调整。
- **路由延迟增加**：由于需要先调用意图识别模型，路由延迟可能会增加（取决于模型响应速度）。

# 延迟优先路由

最近更新时间：2026-07-23 17:07:34

## 操作场景

延迟优先智能路由策略根据模型服务的实时延迟情况，自动优先路由到延迟较低的模型服务。适用于以下场景：

- **跨地域多线路选路**：多 IDC、多云专线场景下，根据网络延迟选择最优线路。
- **端到端响应体验优化**：多个模型后端性能差异显著，优先选择响应快的后端。
- **容量充足场景**：第三方 API 等容量充足服务，追求最低延迟。
- **容量受限场景**：自建集群或混合场景，需要均衡负载与延迟。

## 前置条件

1. 已创建 AI 网关实例且处于运行中状态；
2. 已创建多个模型服务，且服务间存在延迟差异；
3. 了解各模型服务的网络延迟和性能特征。

## 操作步骤

### 步骤 1：进入模型 API 配置页

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 选择左侧导航栏**模型管理 > 模型 API**；
4. 单击**新建**或编辑已有 API；
5. 完成**基本信息**配置后，进入**第二步：选择模型服务**。

### 步骤 2：选择服务类型和路由策略

1. 选择**服务类型**为**多模型服务**；
2. 在**路由策略**区域，选择**延迟优先路由**。

### 步骤 3：选择延迟场景

在**延迟场景**区域，选择适合的延迟场景

场景一：网络延迟最优

**网络延迟最优** 适用于跨地域多线路选路场景，优先选择网络链路质量最好的服务。包括以下场景：

- 跨地域多线路：多个 IDC、多云专线场景，需要选择网络最优路径

- 机房直连线路：后端服务本身性能相近，差异主要来自网络延迟
- 关注链路质量：业务对网络延迟敏感，对模型业务属性无特殊要求

场景二：模型延迟最优

**模型延迟最优**适用于端到端响应体验优化场景，综合考虑后端模型服务的处理延迟。包括以下场景：

- 后端性能差异显著：不同模型服务（如不同供应商、不同规格）的处理速度差异大 。
- 关注端到端体验：业务对整体响应时间敏感，需要优化用户体验。
- 负载影响延迟：自建集群场景，后端负载会直接影响响应延迟。

根据服务容量特征，选择合适的路由策略：

| 路由策略 | 说明 | 适用场景 |
| --- | --- | --- |
| **快速策略** | 始终选择延迟最低的服务，不考虑负载均衡 | 第三方 API 等容量充足、延迟稳定的服务 |
| **均衡策略** | 在延迟和负载间均衡，避免单点过载 | 自建集群或混合场景，容量有限 |

## 步骤 4：配置目标模型服务

1. 在**目标模型服务**区域，单击**请选择服务**下拉框；
2. 选择需要参与延迟优先路由的模型服务（可多选）；
3. 系统将自动监测这些服务的网络延迟，并优先路由到延迟最低的服务。

## 步骤 5：保存配置

单击**确定**保存模型 API 配置，延迟优先路由立即生效。

## 注意事项

1. **网络延迟监测机制**：网关通过周期性探测监测各服务的延迟，实际路由决策基于最近一次的探测数据。
2. **网络延迟数据时效性**：延迟数据存在一定的更新滞后，不适合应对秒级的网络抖动。
3. **快速策略的容量风险**：使用快速策略时，所有流量可能集中到延迟最低的服务，需确保该服务容量充足。
4. **均衡策略的延迟权衡**：均衡策略会牺牲部分延迟性能换取负载均衡，适用于对延迟要求不高的场景。
5. **与 Fallback 的配合**：延迟优先路由可与全局跨服务 Fallback 配合使用，当延迟最低的服务不可用时自动切换到备用服务。

# Token 长度路由

最近更新时间：2026-07-24 16:45:01

## 操作场景

基于 Token 长度的智能路由是一种根据请求 Token 数量自动选择更加适合当前模型服务的路由策略。系统接收用户请求后，自动估算请求的 Token 长度，根据预设的分段规则匹配目标模型，实现成本与性能的最优平衡。
典型场景：

- 多模型成本优化：企业同时接入 GPT-4（短上下文，成本高）、Claude（中长上下文，成本中）、Gemini（超长上下文，成本低），根据请求 Token 长度自动选择最经济的模型
- 长短文本分流：短文本请求（如 0-4K Token）路由到经济型模型，长文本请求（如 32K+）路由到长上下文模型，避免截断和成本浪费

## 前提条件

- 已创建 AI 网关实例且处于运行中状态。
- 已配置多个模型服务（作为候选服务池）。
- 已明确预期进行分段的 Token 长度值。

## 操作步骤

### 步骤 1：进入模型 API 配置页

1. 登录 微服务平台控制台，在左侧导航栏单击 **AI 网关 > 实例列表**；
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 选择左侧导航栏**模型管理 > 模型 API**；
4. 单击**新建**或编辑已有 API；
5. 完成基本信息配置后，进入**第二步：选择模型服务**。

### 步骤 2：选择服务类型和路由策略

1. 选择**服务类型**为**多模型服务**；
2. 若您需要先筛选候选服务池，可启用**标签过滤**（可选）。开启后，将根据标签筛选符合条件的模型服务；不启用标签过滤时，候选服务池为全部模型服务列表。
   - 过滤模式：选择 AND（服务必须同时满足所有标签）或 OR（服务满足任意一个标签即可）。推荐使用 AND。
   - 标签配置：单击添加标签，填写标签的 Key 和 Value。
3. 在**路由策略**区域，选择 **Token 长度路由**。

### 步骤 3：选择 Tokenizer 编码

选择模型对应的分词规则，不同的编码会直接影响 Token 数量、API 费用以及中文处理效率。请根据你实际使用的模型进行选择：

| 编码 | 推荐场景 | 特点 |
| --- | --- | --- |
| o200k_base（推荐） | GPT-4o / o1 / o3 / GPT-5 系列 | 最新、最省成本。对中文支持最好，相同的中文内容消耗的 Token 更少。 |
| cl100k_base | GPT-4 / GPT-3.5 / Claude 系列 | 通用兼容。适合旧版 OpenAI 模型或与 Claude 对齐 Token 计数。 |
| p50k_base | 早期 GPT-3 / Codex | 偏向代码处理，中文效率低，仅用于旧系统维护。 |
| r50k_base | 早期 GPT-2 / GPT-3 | 最古老的编码，中文消耗极大，一般不建议选择。 |

## 步骤4：配置默认目标

Token 计数失败、规则为空或未命中任何分段规则时，使用此默认目标。

1. 选择**二级路由策略**：单模型服务、权重路由、延迟优先路由。
2. 根据所选二级路由策略完成后续配置。

## 步骤5：配置分段规则

1. 在分段规则区域，按 Token 长度区间配置多个分段，为每个分段设置以下参数：
   - Token 长度区间： [分段包含的最小 Token 数，分段包含的最大 Token 数]。前一分段的上界即为下一分段的下界，分段必须连续，不允许区间重叠
     - 典型配置示例：
     - 短文本：[1, 100] / [101, 500] / [501, 2000]
     - 中等文本：[1, 1000] / [1001, 8000] / [8001, 32000]
     - 长文本：[1, 4000] / [4001, 32000] / [32001, 200000]
   - 二级路由策略：单模型服务、权重路由、延迟优先路由。
   - 根据所选二级路由策略完成后续配置。分段内配置多个模型时，自动按所选策略在多模型间分发。
2. 支持单击**添加分段规则**增加新分段。

## 步骤6：确认并保存

单击**确定**保存配置，通过配置合法性检验后智能路由策略生效。

# 服务标签筛选

最近更新时间：2026-07-23 17:07:33

## 操作场景

基于标签路由通过为模型服务绑定标签（Key-Value 键值对），在模型 API 中按标签条件动态筛选可用的模型服务，实现灵活的服务选择。适用于以下场景：

- **按需筛选**：从大批量模型中根据标签条件自动筛选符合条件的服务，无需手动逐个选择
- **多云混合管理**：为不同云服务商、不同地域的模型服务打上标签，按业务需求灵活组合
- **灰度发布与版本管理**：通过 `version`、`env` 等标签筛选特定版本或环境的服务

## 前置条件

1. 已创建 AI 网关实例且处于运行中状态
2. 已创建多个模型服务
3. 已在模型服务上配置了标签（Key-Value 对）

## 操作步骤

### 步骤一：为模型服务配置标签

标签需要在创建或编辑模型服务时配置，标签是后续路由筛选的基础。

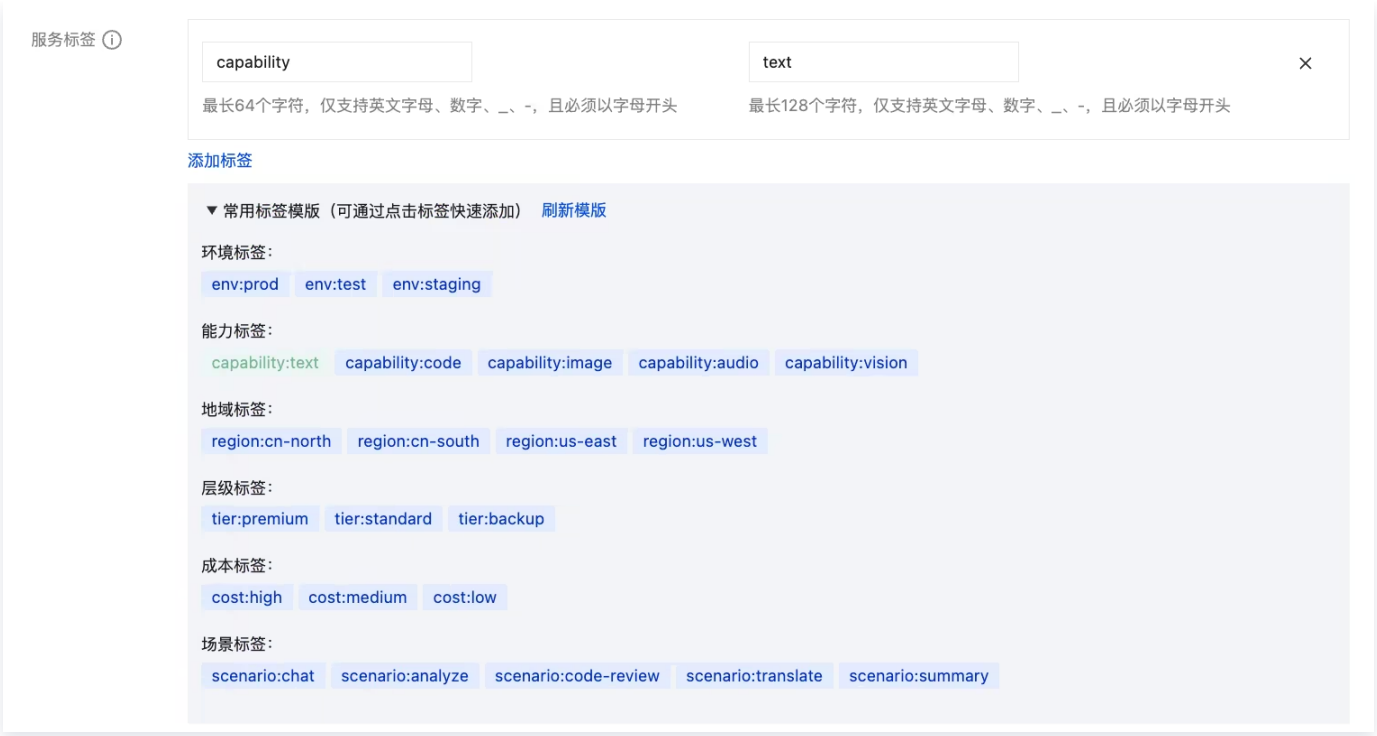
#### 创建或编辑模型服务时配置标签

1. 登录 微服务平台控制台，在左侧导航栏单击 **AI 网关 > 实例列表**；
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 在左侧导航栏单击**模型管理**，然后单击**模型服务**页签；
4. 单击**新建**或编辑已有服务；
5. 在**基本信息 > 高级配置**的**服务标签**区域，配置标签键值对。

   **标签配置说明**：

| 参数 | 说明 | 示例 |
| --- | --- | --- |
| 标签键 | 分类标识，建议使用有业务含义的键名 | `region`、`provider`、`env`、`tier`、`team` |
| 标签值 | 对应键的值 | `singapore`、`openai`、`production`、`vip` |

6. 内置常用标签模板，可单击标签快速添加




## 步骤二：在模型 API 中配置标签筛选

完成模型服务的标签配置后，在创建或编辑模型 API 时通过标签条件筛选服务。

### 1. 进入模型 API 配置页

1.1 选择左侧导航栏 **模型管理 > 模型 API**

1.2 单击 **新建模型 API** 或编辑已有 API

1.3 完成基本信息配置后，进入 **第二步：选择模型服务**

### 2. 配置标签筛选条件

在 **标签筛选** 区域，配置筛选条件和匹配模式：


**配置参数说明：**

| 参数 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |

| | | | |
|---|---|---|---|
| 匹配模式 | 是 | 标签条件的组合逻辑 | `AND` 或 `OR` |
| 标签键 | 是 | 模型服务标签的键名 | `region` |
| 标签值 | 是 | 期望匹配的标签值 | `singapore` |
| 已匹配服务 | 自动 | 根据条件自动筛选出的服务列表 | `OpenAI-Singapore-Prod` 等 |

### 3. 配置路由后续策略

标签筛选完成后，选中已匹配的服务，然后可选择后续的路由策略（如模型名称路由、权重路由等）对这些服务进行更细粒度的流量分配。

### 4. 保存配置

单击 **确定** 保存模型 API 配置，标签路由规则立即生效。

# 配额感知降级

最近更新时间：2026-07-23 17:07:33

## 操作场景

配额感知降级在模型 API 级别配置，当消费者的 Token 配额或请求配额即将耗尽时，自动将请求降级到备用服务，防止因配额不足导致请求失败。适用于以下场景：

- **成本控制**：当高成本模型服务的 Token 配额即将用尽时，自动降级到低成本备用服务，避免超预算。
- **配额共享**：多消费者共享同一配额池时，配额耗尽后自动降级而非直接报错。
- **分级服务**：VIP 消费者使用高配额的服务，普通消费者配额不足时降级到基础服务。
- **预算管理**：按月度/季度预算控制模型调用量，防止超支。

> **说明：**
> 配额感知降级是全局跨服务 Fallback 的增强特性，需在模型 API 的 Fallback 配置中开启。

## 前置条件

1. 已创建 AI 网关实例且处于运行中状态；
2. 已配置消费者，并已为其分配配额（如 Token 限额或请求限额）；
3. 已创建至少两个模型服务（主服务和备用服务）；
4. 已开启全局跨服务 Fallback 并配置备用服务链。

## 操作步骤

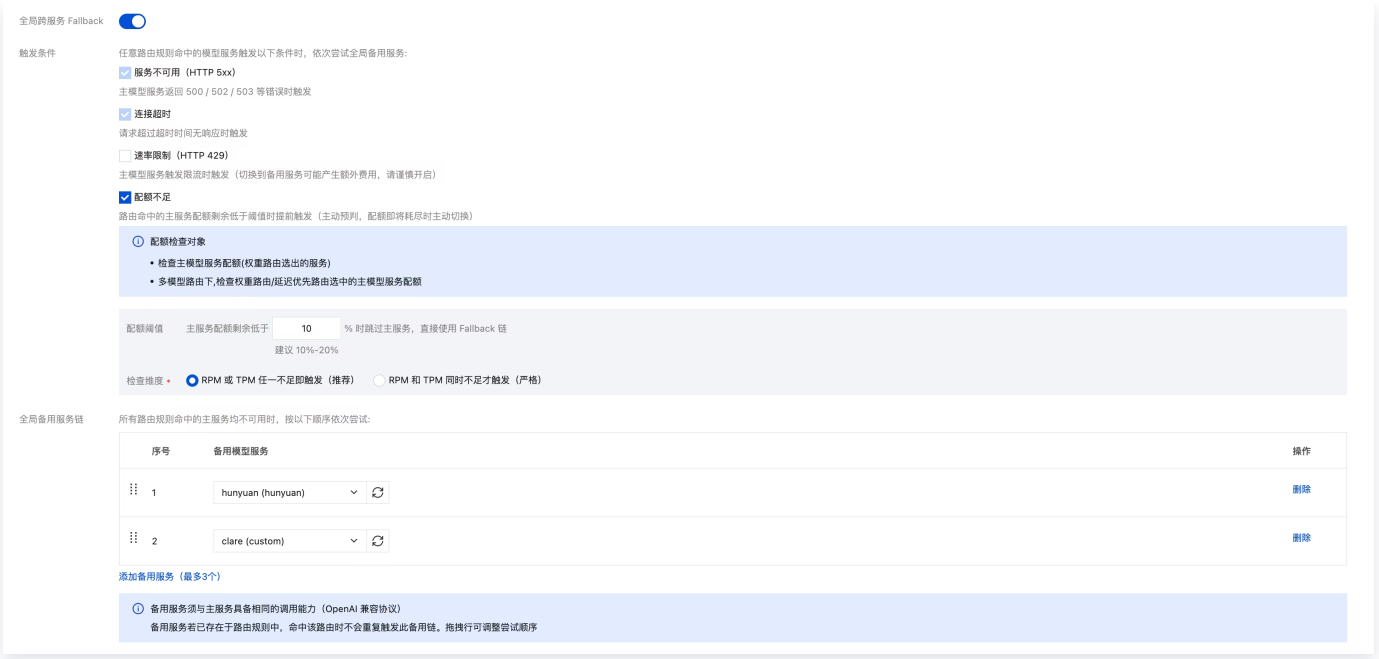
### 模型 API 中配置配额感知降级

#### 步骤 1：进入模型 API 配置页

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**；
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面；
3. 在左侧导航栏单击 **模型管理**，然后单击**模型 API** 页签；
4. 单击**新建**或编辑已有 API；
5. 完成基本信息配置后，进入 **第二步：选择模型服务**。

#### 步骤 2：开启全局跨服务 Fallback

1. 在 **全局跨服务 Fallback** 区域，开启 **全局跨服务 Fallback** 开关
2. 配置备用服务链




## 步骤 3：配置配额不足触发条件

**当 配额不足 触发条件启用后，需要配置具体的配额阈值：**

**配置参数说明：**

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 配额阈值 | 是 | 剩余配额低于此百分比时触发降级，默认 10%，建议10%–20% |
| 检查维度 | 是 | 支持两种：<br>• 推荐策略：RPM 或 TPM 任一不足即触发<br>• 严格策略：RPM 和 TPM 同时不足才触发 |

**配额阈值建议值：**

| 阈值 | 说明 | 适用场景 |
| --- | --- | --- |
| 5% | 激进的降级触发时机，尽可能用完配额 | 成本敏感场景，希望最大化利用已购配额 |
| 10% | 推荐的降级触发时机 | 大多数场景的推荐值 |
| 20% | 保守的降级触发时机，提前准备 | 高可用优先场景，需要预留缓冲 |
| 50% | 非常保守的降级触发时机 | 关键业务，配额充足时即开始降级 |

## 步骤 4：保存配置

**单击确定 保存模型 API 配置，配额感知降级规则立即生效。**

## 查看 Fallback 记录

在模型 API 详情页的 **Fallback 记录** Tab 中，可以查看所有 Fallback 事件，包括配额感知降级触发的记录。

## 注意事项

1. **配额预配置**：配额感知降级依赖消费者或 API 的配额配置。如果未配置配额，配额感知降级不会触发。
2. **阈值设定**：建议将配额阈值设置为 5%-20%，过低可能导致大量请求因配额用尽而失败，过高则可能太早切换到备用服务。
3. **降级与限流的区别**：
   - **限流**：超出配额直接拒绝请求（返回 429）。
   - **配额感知降级**：配额即将用尽时，将请求迁移到备用服务（返回正常响应）。
4. **双重降级组合**：配额感知降级可以与服务不可用、连接超时等触发条件组合使用，实现更完整的降级策略。
5. **成本控制**：配额感知降级特别适合成本敏感场景，可在高成本服务配额用尽时自动切换到低成本服务，避免超预算。

# MCP 管理
# MCP 服务管理

最近更新时间：2026-07-23 17:07:35

## 操作场景

随着大模型应用的普及，越来越多的业务系统希望以 Model Context Protocol(MCP) 的方式将自身能力开放给 AI Agent 调用。AI 网关作为统一的 MCP 接入层，为您屏蔽底层差异，提供统一的接入点、协议转换、Tool 管理与多后端聚合能力，使您能够安全、便捷地把后端服务以标准 MCP 形态对外暴露。
AI 网关支持以下两种 MCP 服务类型，以适配不同的接入场景：

- **标准 MCP 服务**：适用于后端已基于官方 MCP SDK 实现 MCP 协议、或已注册到 TSF Nacos MCP Registry 的场景。网关作为透传代理，将客户端的 MCP 请求原样转发到后端，Tool 元数据由后端自动同步，在控制台呈现为只读。
- **HTTP 转 MCP 服务**：适用于将存量 REST/HTTP 接口快速升级为 MCP 能力的场景。网关负责 HTTP 与 MCP 协议之间的转换，Tool 由您按需定义，支持手动创建或基于 OpenAPI 规范文件批量导入。该类型同时包含 **虚拟 MCP Server** 模式，可将多个不同系统的接口聚合为同一个 MCP Server 对外暴露。

本文介绍如何在 AI 网关中配置服务来源、创建后端服务，以及如何添加、查看、编辑 MCP 服务。

## 操作步骤

### 配置服务来源

服务来源用于批量接入已有的注册中心或私有域名解析服务。完成配置后，在创建 MCP 服务或后端服务时即可直接引用，避免重复填写接入参数。AI 网关支持 **MCP 注册中心** 与 **私有 DNS** 两种类型的服务来源：

- **MCP 注册中心**：对接 TSF Nacos 3.x，自动同步注册中心中的原生 MCP Server，实现 MCP Tool 的自动发现。适用于标准 MCP 服务接入。
- **私有 DNS**：对接腾讯云 Private DNS 进行内网域名解析。适用于通过私有域名访问的后端 HTTP 服务。

#### 添加服务来源

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**服务管理**，然后单击**服务来源**页签，在列表上方单击**新建**。
4. 在“新建服务来源”抽屉中，根据您的接入场景选择**来源类型**，并参考下表完成配置。
   - **来源类型 1：MCP 注册中心**

| 参数 | 是否必选 | 说明 |
| --- | --- | --- |

| | | |
|---|---|---|
| 来源类型 | 是 | 选择 **MCP 注册中心**，用于同步原生 MCP Server，实现 Tool 自动发现。 |
| 来源产品 | 是 | 默认为 **TSF Nacos**，用于对接注册在 TSF Nacos 3.x 上的服务。 |
| 服务来源名称 | 是 | 输入服务来源名称。最长 60 个字符，支持中英文大小写、数字及分隔符("-"、"_")，不能以数字和分隔符开头，不能以分隔符结尾。同网关实例下唯一。 |
| 选择实例 | 是 | 从下拉列表中选择已有的 Nacos 实例。仅展示与当前网关实例位于同一 VPC 内的实例，确保网关可正常访问注册中心。 |
| 用户名 | 是 | 输入 Nacos 访问用户名，用于网关连接注册中心时的身份认证。 |
| 密码 | 是 | 输入 Nacos 访问密码，以加密方式存储。 |
| 描述 | 否 | 输入该服务来源的描述信息，便于后续管理。最长 200 个字符。 |

- **来源类型 2：私有 DNS**

| 参数 | 是否必选 | 说明 |
|---|---|---|
| 来源类型 | 是 | 选择 **私有 DNS**，用于通过腾讯云 Private DNS 进行内网域名解析。 |
| 来源产品 | 是 | 固定为 **腾讯云私有解析**。 |
| 描述 | 否 | 输入该服务来源的描述信息，便于后续管理。最长 200 个字符。 |

5. 单击 **确定** 完成创建。创建成功后，服务来源将出现在 **服务来源** 列表中，后续在新建 MCP 服务或后端服务时可直接选择。

## 配置后端服务

若创建 **HTTP 转 MCP 服务(域名/IP 模式)** 或 **标准 MCP 服务(域名/IP 模式)**，需要先在 **服务管理 > 服务** 页签下创建服务，作为 MCP 服务关联的实例。

### 添加服务

1. 在网关实例详情页的左侧导航栏单击 **服务管理**，切换到 **服务** 页签，在列表上方单击 **新建**。
2. 在“新建服务”抽屉中，参考下表完成配置。其中部分字段会根据 **服务类型** 的选择动态展示。

| 参数 | 是否必选 | 说明 |
|---|---|---|
| 服务名称 | 是 | 输入服务名称，用于在 MCP 服务的“后端服务”下拉中标识此实例。最长 60 个字符，支持中英文大小写、数字及分隔符("-"、"_")，不能以数字和分 |

| | | |
|---|---|---|
| | | 隔符开头，不能以分隔符结尾。 |
| 服务类型 | 是 | 选择后端的接入方式：<br>• **私有 DNS(默认)**：通过已创建的私有 DNS 服务来源解析私有域名，适用于内网访问的后端服务。<br>• **域名/IP**：直接填写 IP 或公网域名，适用于公网或固定地址的后端服务。 |
| 服务来源 | 是(私有 DNS) | 当 **服务类型** 选择 **私有 DNS** 时必填，从下拉列表中选择已创建的 **私有 DNS** 类型服务来源。 |
| 私有域名 | 是(私有 DNS) | 当 **服务类型** 选择 **私有 DNS** 时必填，输入需要解析的私有域名，例如 `order.internal.example.com` 。 |
| 服务地址 | 是(域名/IP) | 当 **服务类型** 选择 **域名/IP** 时必填，输入后端服务的 IP 或公网域名，例如 `10.0.0.1` 或 `www.tencent.com` 。 |
| 服务协议 | 是 | 选择后端服务的协议，默认为 **HTTP**，可选 **HTTPS**。 |
| 服务端口 | 是 | 输入后端服务的端口，支持 1~65535 的任意端口。如 SRV 记录已填写端口，控制台输入端口会覆盖 SRV 记录的端口内容。 |
| 服务路径 | 是 | 输入后端服务的基础路径(如 `/api/v1` )，后续 Tool 的相对路径将拼接在此路径之后。 |
| 超时时间 | 是 | 网关请求后端的超时时间，默认 60000 毫秒，取值范围 0-60000。 |
| 重试次数 | 是 | 请求失败时的重试次数，默认 5 次，取值范围 0-10。 |

3. 单击 **确定** 完成创建。创建后的后端服务将出现在服务列表中，供后续新建 MCP 服务时引用。

## 添加 MCP 服务

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **云原生智能网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **MCP 管理**，在 MCP 服务列表上方单击 **新建**。
4. 在“新建 MCP 服务”窗口中，完成第一步“基本信息”的配置。

| 参数 | 是否必选 | 说明 |
|---|---|---|
| MCP 服务名称 | 是 | 输入 MCP 服务的唯一标识，该名称将作为接入路径的一部分( `/mcpservers/<MCP 服务名称>/mcp` )。最长 60 个字符，支持英文大小写、数字及分隔 |

| | | |
|---|---|---|
| | | 符("-"、"_")，不能以数字和分隔符开头，不能以分隔符结尾。同网关实例下唯一，创建后不可修改。 |
| 服务展示名称 | 是 | 输入服务展示名称，用于在控制台列表与详情页展示。最长 60 个字符，支持中英文大小写、数字及分隔符("-"、"_")，不能以数字和分隔符开头，不能以分隔符结尾。 |
| 服务类型 | 是 | 选择 MCP 服务类型：<br>• **标准 MCP 服务**：后端已实现 MCP 协议，网关透传代理。Tool 自动同步，只读。<br>• **HTTP 转 MCP 服务**：存量 HTTP 服务，网关负责协议转换，包含虚拟 MCP Server 模式。Tool 由用户配置。 |
| 请求协议 | 是 | 客户端调用网关时使用的 MCP 传输协议，当前版本固定为 **Streamable HTTP**。 |
| 描述 | 否 | 该服务的描述信息，便于后续管理。最长 200 个字符。 |

5. **完成基本信息后，单击 下一步，进入“后端服务配置”步骤。系统将根据“基本信息”中所选的 服务类型 展示不同的后端配置项，共四种场景。**
   - **场景一：标准 MCP 服务 + MCP 注册中心**

     **适用于后端 MCP Server 已注册到 TSF Nacos，由网关从注册中心自动同步服务地址与 Tool 元数据。在 后端类型 区选择 MCP 注册中心，参考下表完成配置。**

| 参数 | 是否必选 | 说明 |
|---|---|---|
| 服务协议 | 是 | 选择后端 MCP Server 的协议，默认为 **HTTP**，可选 **HTTPS**。 |
| 服务来源 | 是 | 从下拉列表选择已创建的 **MCP 注册中心** 类型服务来源。如未创建，可单击 **新建服务来源** 快速跳转至创建流程。 |
| 命名空间 | 是 | 输入 Nacos 命名空间。最长 128 个字符，仅支持英文字母、数字、`_`、`-`。 |
| MCP 服务 | 是 | 输入目标 MCP Server 的服务标识。最长 128 个字符，仅支持英文字母、数字、`_`、`-`。 |
| MCP Endpoint | 是 | 输入 MCP 协议的 Endpoint 路径。最长 64 个字符，以 `/` 开头，仅支持英文字母、数字、`/`、`_`、`-`。 |
| 服务分组 | 是 | 输入 Nacos 服务分组。最长 128 个字符，仅支持英文字母、数字、`.`、`-`。 |

| | | |
|---|---|---|
| 服务名称 | 是 | 输入 Nacos 服务名称。最长 128 个字符，仅支持英文字母、数字、`.`、`-`。 |
| 超时时间 | 是 | 网关请求后端的超时时间，默认 3000 毫秒，取值范围 0-60000。 |
| 重试次数 | 是 | 请求失败时的重试次数，默认 3 次，取值范围 0-10。 |
| 健康检查 | 否 | 是否开启 MCP Server 健康检查，如需进行健康检查则打开按钮并配置健康检查参数：<br>• 检查类型：HTTP/HTTPS<br>• 检查间隔：5秒以上（必填）<br>• 超时时间：1秒以上（必填）<br>• 失败阈值：1次以上（必填）<br>• 恢复阈值：1次以上（必填）<br>• 探测路径：请自行配置，默认：/health（必填）<br>• 探测方法：GET/POST |

- **场景二：标准 MCP 服务 + 域名/IP**

适用于后端 MCP Server 未使用注册中心，需要手动指定服务地址的场景。在 **后端类型** 区选择 **域名/IP**

| 参数 | 是否必选 | 说明 |
|---|---|---|
| 服务协议 | 是 | 选择 HTTP 或 HTTPS。 |
| 服务地址 | 是 | 输入后端 MCP Server 的 IP 或域名，例如 `10.0.0.1` 或 `mcp-server.example.com`。 |
| 服务端口 | 是 | 输入后端服务的端口，支持 1~65535 的任意端口。 |
| MCP Endpoint | 是 | 输入 MCP 协议的 Endpoint 路径。最长 64 个字符，以 `/` 开头，仅支持英文字母、数字、`/`、`_`、`-`。 |
| 超时时间 | 是 | 默认 3000 毫秒，取值范围 0-60000。 |
| 重试次数 | 是 | 默认 3 次，取值范围 0-10。 |
| 健康检查 | 否 | 是否开启 MCP Server 健康检查，如需进行健康检查则打开按钮并配置健康检查参数<br>• 检查类型：HTTP/HTTPS<br>• 检查间隔：5秒以上（必填）<br>• 超时时间：1秒以上（必填）<br>• 失败阈值：1次以上（必填）<br>• 恢复阈值：1次以上（必填） |

| | | |
|---|---|---|
| | | • 探测路径：请自行配置，默认：/health（必填）<br>• 探测方法：GET/POST |

- **场景三：HTTP 转 MCP 服务 + 域名/IP**

  适用于单一后端 HTTP 服务的接入，所有 Tool 共享同一个后端服务实例。在 **后端类型** 区选择 **域名/IP**，参考下表完成配置。

| 参数 | 是否必选 | 说明 |
|---|---|---|
| 后端服务 | 是 | 从下拉列表选择已在 **服务管理** > **服务** 下创建的后端服务。如未创建，可单击 **新建服务** 快速跳转至创建流程，或单击 **刷新** 重新加载下拉数据。 |

- **场景四：HTTP 转 MCP 服务 + 虚拟 MCP Server**

  适用于将多个不同的 HTTP 接口聚合为同一个 MCP Server 对外暴露的场景。网关自身充当 MCP Server，每个 Tool 独立配置完整后端 URL。

6. 配置完成后，单击 **确定** 即可创建 MCP 服务。创建成功后，服务列表中会出现新增的 MCP 服务，单击其 ID 可查看详细的服务信息。
   - 对于 **标准 MCP 服务**，Tool 由后端 MCP Server 自动同步，可在服务详情页的 **Tools 管理** 页签查看。
   - 对于 **HTTP 转 MCP 服务**，需进入服务详情页的 **Tools 管理** 页签手动创建 Tool 或通过 OpenAPI 批量导入，详见 MCP Tools 管理 。

## 查看 MCP 服务详情

在 MCP 服务列表中，单击目标服务的 **ID/名称**，进入服务详情页。详情页包含 **基本信息** 与 **Tools 管理** 两个页签。

**基本信息** 页签展示以下三个区块：

- **基本信息区**：展示 MCP 服务 ID、MCP 服务名称、服务展示名称、服务类型、请求协议、创建时间、修改时间、描述等元数据。单击右上角的 **编辑** 可修改服务展示名称与描述等可变字段。
- **调用方式区**：展示供 AI Agent 使用的 `mcpServers` JSON 配置模板，支持一键复制后粘贴到 Cursor、CodeBuddy、Claude Desktop 等 MCP 客户端的配置文件中。内网调用地址格式如下：

```
{
  "mcpServers": {
    "<MCP 服务名称>": {
      "url": "http://<网关 IP>/mcpservers/<MCP 服务名称>/mcp"
    }
  },
  "transportType": "streamable-http"
}
```

网关接入路径固定为 `/mcpservers/<MCP 服务名称>/mcp`，由网关根据实例 IP 与服务名自动生成。如需通过公网访问，请在 AI 网关实例的 **基础信息** > **网络配置** 中开启公网负载均衡，开启后调用方式区会同步展示公网接入地址。

- **后端服务配置区**：展示后端类型与对应的各字段信息。**虚拟 MCP Server** 模式下，服务协议显示为 `-`，表示无统一服务级后端协议，各 Tool 在创建时独立指定。

**Tools 管理** 页签用于管理该服务下的 Tool，详见 MCP Tools 管理 。

## 编辑 MCP 服务

在 MCP 服务列表中，找到目标服务，单击其操作列下的 **编辑**，进入“编辑 MCP 服务”窗口。
第一步“基本信息”中，**MCP 服务 ID**、**MCP 服务名称**、**服务类型**、**请求协议** 为只读，仅 **服务展示名称** 与 **描述** 可编辑。单击 **下一步** 后，可修改后端服务配置中的各字段(如服务地址、服务端口、超时时间、重试次数、后端服务等)。修改完成后，单击 **确定** 保存即可生效。

## 注意事项

- **标准 MCP 服务(MCP 注册中心模式)** 依赖 TSF Nacos 3.x 的版本特性，请确认所选 Nacos 实例的版本。
- **HTTP 转 MCP 服务(域名/IP 模式)** 必须先在 **服务管理** > **服务** 页签下创建服务，再在新建 MCP 服务时引用。**虚拟 MCP Server 模式** 下创建 MCP 服务时无需配置后端地址。

# MCP Server 上下线与健康检查管理

最近更新时间：2026-07-23 17:07:33

## 操作场景

在 MCP Server 的日常运维中，您可能需要临时停止某个服务对外提供能力（如发布、维护期间），或在恢复后重新对外提供服务。通过 **上下线管理**，您可以手动控制 MCP Server 及其工具的可用状态；同时可配置 **健康检查**，由网关自动探测后端健康状况并联动服务状态。

## 前提条件

- 已创建云原生 AI 网关实例。
- 已创建 MCP Server。

## 操作步骤

### 手动上线 / 下线 MCP Server

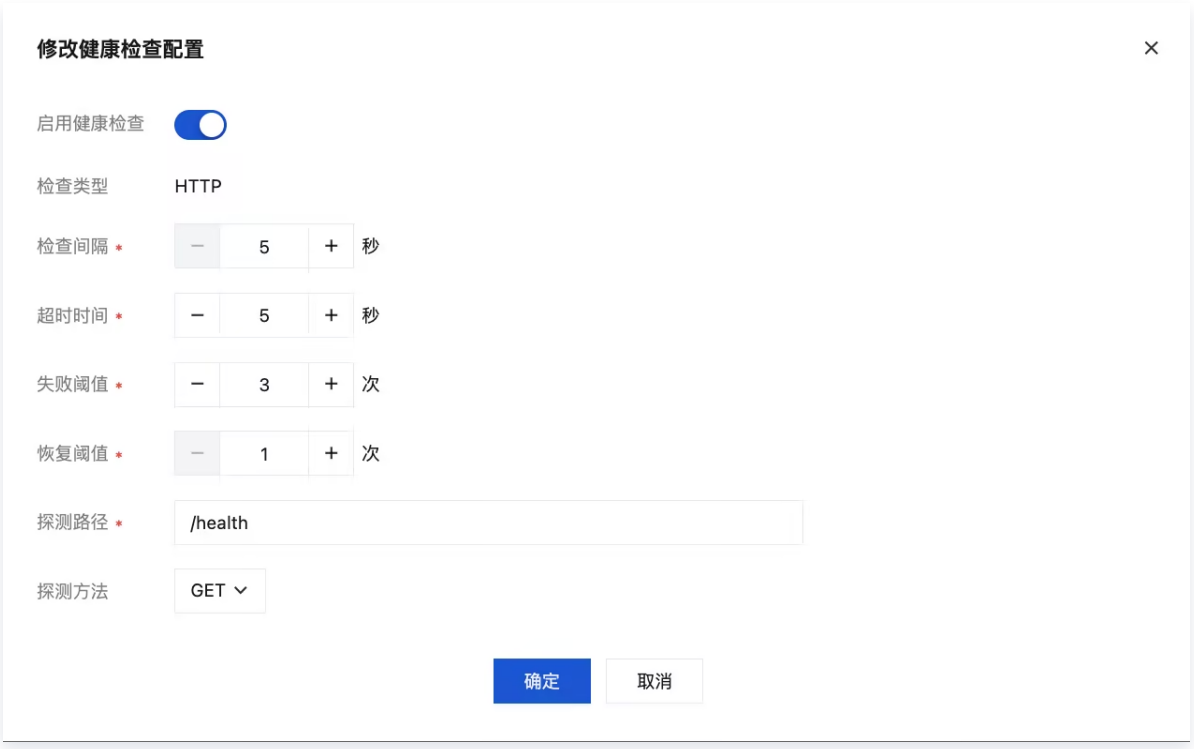
1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **MCP 管理**，然后单击 **MCP 服务**页签，找到目标 MCP 服务后单击进入 MCP 服务详情页面。
4. 单击“运行状态”操作列的 **上线（Online）** 或 **下线（Offline）**：
   - **下线**：该 Server 暂停对外提供服务。
   - **上线**：恢复对外提供服务。


### 启用 / 禁用单个工具

切换到“Tools 管理”标签页，选择目标工具，单击**启用（Enable）** 或 **禁用（Disable）**，控制单个工具是否对外可用。


# 配置健康检查

## 1. 在 MCP 服务详情页，找到**健康检查**配置区域，单击**编辑**。




## 2. 配置以下参数：

| 参数 | 取值范围 / 默认值 | 说明 |
| --- | --- | --- |
| 检查类型（Type） | MCP / HTTP | 探测协议类型 |
| 检查间隔（Interval） | 默认 30（秒） | 探测周期 |
| 超时时间（Timeout） | 默认 5（秒） | 单次探测超时 |
| 失败阈值（FailThreshold） | 默认 3（次） | 连续失败判定为不健康 |
| 恢复阈值（RecoverThreshold） | 默认 1（次） | 连续成功判定为恢复 |
| 检查路径（Path） | 默认 `/health` | 仅 HTTP 类型需配置 |
| 请求方法（Method） | 默认 GET | 仅 HTTP 类型需配置 |

3. 单击**保存**使健康检查生效。

## 相关说明

- 手动下线后，服务状态不受健康检查自动恢复影响，需手动重新上线。
- 健康检查类型为 HTTP 时才需配置检查路径与请求方法；类型为 MCP 时按 MCP 协议探测。
- 上下线、启停操作即时生效，建议在变更窗口内操作以避免影响线上调用。

# 分布式会话缓存

最近更新时间：2026-07-23 17:07:33

## 操作场景

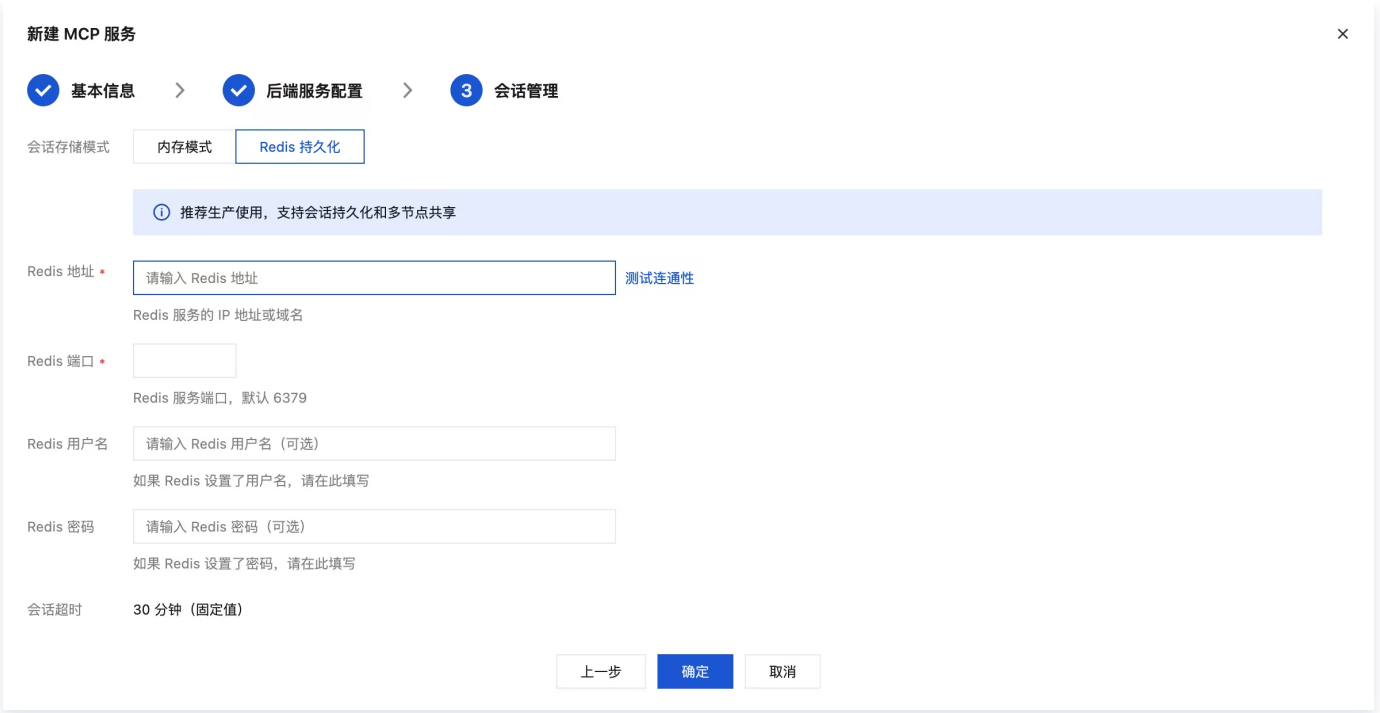
当您的 MCP 网关使用多个数据面节点（多副本）对外提供服务时，MCP 客户端通过 **SSE** 或 **Streamable HTTP** 协议建立的会话（Session）默认仅保存在单个节点的内存中。一旦后续请求被负载均衡路由到其他节点，会话信息会丢失，导致连接中断或会话失效。
通过为 MCP 服务开启**分布式会话缓存**，网关会将会话状态统一存储到外部 Redis，使任意节点都能识别并处理同一会话的请求，从而保障多节点环境下 SSE / Streamable HTTP 长连接的稳定性。

## 前提条件

- 已创建云原生 AI 网关实例。
- 已创建 MCP Server，且传输协议为 **SSE** 或 **Streamable HTTP**。
- 已准备一个网关可访问的 Redis 实例（建议与网关同 VPC），并获取其连接地址、端口、用户名及密码。

## 操作步骤

1. 登录 云原生 API 网关控制台，进入目标网关实例。
2. 在左侧菜单选择 **MCP 管理 > MCP 服务**，进入 MCP 服务列表页面。
3. 单击**新建**按钮，完成“基本信息”和“后端服务配置”后，进入到“会话管理”配置页面。
4. 会话存储模式：内存模式/Redis 持久化




| 参数 | 是否必填 | 取值范围 / 默认值 | 说明 |
| --- | --- | --- | --- |
| 地址（Host） | 是 | 有效的 IP 或域名 | Redis 实例连接地址 |
| 端口（Port） | 是 | 1-65535，默认 `6379` | Redis 服务端口 |
| 用户名（Username） | 否 | 字符串，可留空 | Redis 6.0+ ACL 用户名 |
| 密码（Password） | 否 | 字符串 | Redis 访问密码 |

5. 单击 **保存**，配置即时生效。

## 相关说明

- **会话有效期（TTL）**：会话在 Redis 中默认保留 **30 分钟（1800 秒）**，超时后自动失效，当前为固定值。
- **适用协议版本**：SSE 协议采用 `2024-11-05` 版本，双端点为 `GET /sse`（建立连接）与 `POST /message`（消息传输）。
- 切换为 Redis 存储后，已建立的旧会话不会自动迁移，建议在业务低峰期操作。
- 若 Redis 不可达，新会话将无法建立，请确保网络连通性与鉴权信息正确。

# MCP Tools 管理
# Tools 管理

最近更新时间：2026-07-23 17:07:36

## 操作场景

Tool 是 MCP 服务的最小能力单元，每个 Tool 对应后端的一个 HTTP 接口，代表 AI Agent 可以通过 MCP 协议直接调用的一项具体功能。Tool 的元数据(如英文名、中文名、描述、参数定义)将以标准 MCP 形式暴露给 AI Agent，AI 模型据此理解 Tool 的用途、参数与调用方式，从而做出合理的工具选用决策。因此，Tool 配置的清晰度与准确性，直接影响 AI Agent 的调用效果。
针对 **HTTP 转 MCP 服务**，AI 网关支持以下两种 Tool 创建方式：

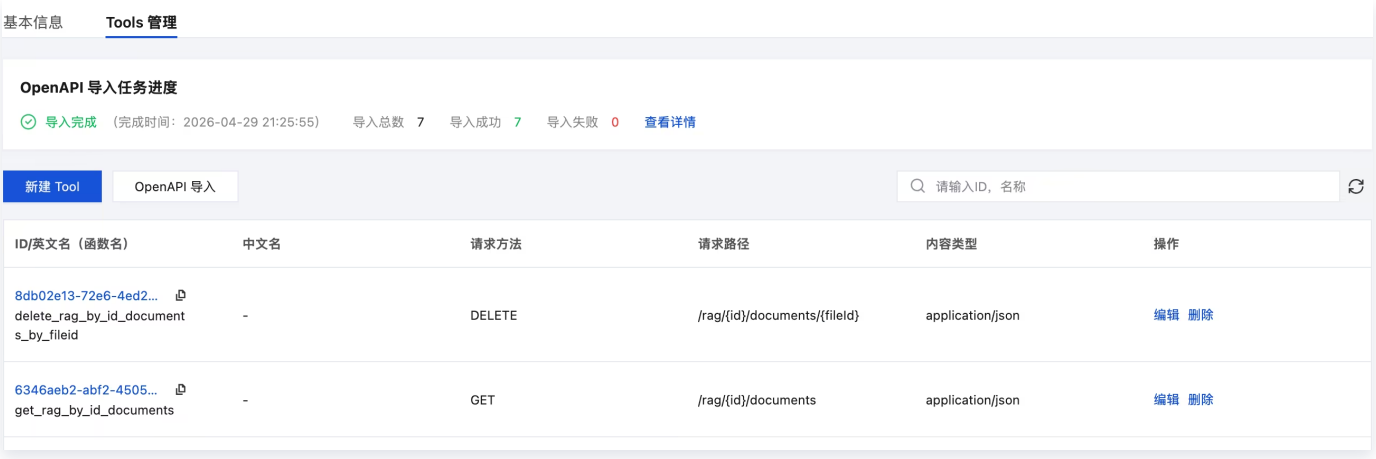
- **手动创建 Tool**：逐条配置 Tool 的基本信息、后端 HTTP 请求与参数定义。适合 Tool 数量较少、需要精细化控制每个字段的场景。
- **OpenAPI 批量导入**：基于已有的 OpenAPI 3.0 / Swagger 2.0 规范文件，一键解析并批量生成 Tool。适合后端已维护标准规范文档、希望快速完成大量 Tool 配置的场景。

本文介绍如何在 AI 网关中为 HTTP 转 MCP 服务添加、编辑 Tool，以及如何通过 OpenAPI 文件批量导入 Tool 与配置常用的协议转换。

## 操作步骤

### 手动创建 Tool

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **MCP 管理**，在服务列表中单击目标 HTTP 转 MCP 服务的 **ID/名称**，进入服务详情页。
4. 在服务详情页顶部单击 **Tools 管理** 页签，在列表上方单击 **新建 Tool**，打开“新建 MCP Tool”页面。




5. 在“新建 MCP Tool”中，完成 **基本信息** 区域的配置。

| 参数名称 | 是否必选 | 说明 |
| --- | --- | --- |
| 英文名(函数名) | 是 | 输入 Tool 的英文名，作为 AI Agent 调用时的唯一标识。最长 60 个字符，支持英文大小写、数字及分隔符("-"、"")，不能以数字和分隔符开头，不能以分隔符结尾。同服务下唯一，创建后不可修改。建议采用 动作对象的命名规范，如 create_order、query_stock。 |
| 中文名 | 否 | 输入 Tool 的中文名称，用于在控制台列表与详情页中展示，便于人工识别。最长 60 个字符，支持中英文大小写、数字及分隔符("-"、"_")，不能以数字和分隔符开头，不能以分隔符结尾。 |
| 描述 | 否 | 用自然语言描述该 Tool 的功能。 |

6. 继续完成 **后端配置** 区域的配置。

| 参数名称 | 是否必选 | 说明 |
| --- | --- | --- |
| 后端服务 | 是 | 从下拉列表选择已在 **服务管理** > **服务** 下创建的后端服务，Tool 的请求路径将与后端服务的协议、地址、端口、基础路径自动拼接。如未创建，可单击 **新建服务** 快速跳转至创建流程，或单击 **刷新** 重新加载下拉数据。<br>在虚拟 MCP Server 模式下，不同 Tool 可指向不同的后端服务，从而实现多后端聚合。 |
| 请求方法 | 是 | 选择后端 HTTP 接口的请求方法，默认为 **GET**，可选 **POST**、**PUT**、**DELETE**、**PATCH**。 |
| 请求路径 | 是 | 输入后端 HTTP 接口的相对路径，以 `/` 开头，例如 `/api/v1/orders/{order_id}`。路径中可使用 `{参数名}` 占位符，网关会自动用 `path` 类型参数的值替换。 |
| 内容类型 | 是 | 选择 HTTP 请求的 `Content-Type`，默认为 `application/json`，可选 `multipart/form-data`、`application/x-www-form-urlencoded` 等。 |

7. 继续完成 **参数配置** 区域的配置。

| 配置项 | 说明 |
| --- | --- |
| 参数名 | 输入参数名，与后端接口的实际参数名保持一致。 |
| 参数类型 | 选择参数的数据类型，默认 `string`，可选 `number`、`integer`、`boolean`、`object`、`array`。 |

| | |
|---|---|
| 参数位置 | 选择参数在后端 HTTP 请求中的位置：<br>• **path**：URL 路径参数，需在"请求路径"中存在对应的 `{参数名}` 占位符。<br>• **query**(默认)：URL 查询参数。<br>• **header**：HTTP 请求头。<br>• **body**：HTTP 请求体，仅适用于 POST、PUT、PATCH 请求方法。 |
| 默认值 | 输入客户端未传该参数时使用的默认值。配合"是否必填"勾选，可实现常量注入(例如固定鉴权 Token、租户 ID 等)。 |
| 是否必填 | 勾选后，AI Agent 调用 Tool 时必须提供该参数，否则网关将直接返回校验失败错误。 |
| 描述 | 输入参数的自然语言描述，影响 LLM 对参数含义与取值范围的理解。 |

- **协议转换说明**：AI 网关通过参数配置实现常用的协议转换能力，无需独立配置。
- **请求参数映射**：通过"参数位置"将 MCP 调用入参映射到后端 HTTP 请求的不同位置(path / query / header / body)。
- **请求头注入(常量注入)**：将固定的请求头(如 `Authorization: Bearer <token>`、`X-Tenant-ID`)以 `header` 类型参数的形式定义，填写"默认值"并勾选"必填"，所有调用即可自动注入。
- **请求体组装**：多个 `body` 类型参数将被网关自动组装成后端期望的 JSON 请求体，支持 `object`、`array` 等嵌套结构。
- **请求方法与内容类型转换**：客户端通过 MCP 协议统一调用，网关根据 Tool 配置的"请求方法"和"内容类型"转换为后端实际所需的 HTTP 请求形式。

8. 配置完成后，单击 **确定** 即可创建 Tool。创建后的 Tool 将出现在 Tools 列表中，AI Agent 可立即通过 MCP 协议进行调用。


## 通过 OpenAPI 文件批量导入 Tool

如果您的后端接口已维护标准的 OpenAPI 3.0 或 Swagger 2.0 规范文档，可通过 OpenAPI 导入功能一键批量生成 Tool。

1. 进入目标 HTTP 转 MCP 服务的详情页，单击 **Tools 管理** 页签，在列表上方单击 **OpenAPI 导入**，打开"从 OpenAPI 文件导入 Tool"对话框。


2. 在对话框中，单击上传区域或拖拽文件选择本地的 OpenAPI 规范文件，然后单击 **上传并解析**。当前支持 `.json`、`.yaml`、`.yml` 三种格式，文件大小不超过 40 KB。

3. 文件解析完成后，系统会展示识别到的所有 HTTP 接口列表，字段包括接口方法、接口路径、接口描述、Tool 名称。检查所有 Tool 信息无误后，勾选需要导入的接口，然后单击 确认导入。系统将批量创建 Tool，完成后返回 **Tools 管理** 列表。列表顶部会展示"OpenAPI 导入任务进度"卡片，显示导入状态(导入完成/导入失败)、完成时间、导入总数、导入成功数与导入失败数，单击 **查看详情** 可查看每条接口的导入结果与失败原因。

## 编辑 Tool

在 Tools 管理列表中，找到目标 Tool，单击其操作列下的 **编辑**，即可在选项中修改除 **英文名(函数名)** 以外的所有字段，包括中文名、描述、后端服务、请求方法、请求路径、内容类型、参数配置等。修改后单击 **确定** 保存，变更将立即生效。

# 版本管理

最近更新时间：2026-07-23 17:07:36

## 功能简介

MCP Tool 支持配置级版本管理，帮助你追溯历史变更、对比版本差异、一键回滚误改，缩短线上误变更的恢复时间。其核心行为：

- **自动记录版本**：创建 Tool 时自动生成初始版本 v1；后续每次编辑触发版本化字段变更时，按你填写的版本号生成一条新版本快照。
- **运行时无感**：版本快照仅记录配置历史，不影响 Tool 的实际调用链路；运行时配置始终以当前生效版本为准。

## 版本号规则

| 规则项 | 说明 |
| --- | --- |
| 格式 | 由字母、数字、下划线组成，长度 1~32 位，由用户自定义命名，如 v1、v2_0、release_20260425。 |
| 唯一性 | 同一个 Tool 下版本号不可重复；编辑弹窗会实时校验并提示版本号是否已存在。 |
| 保留上限 | 每个 Tool 默认最多保留 100 个版本，超出后自动裁剪最早的非生效版本；当前生效版本永不被裁剪。 |

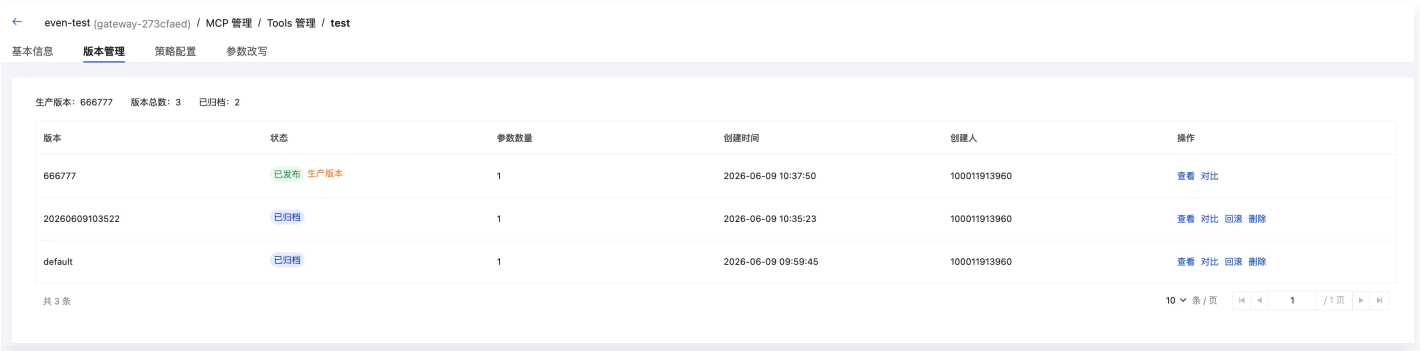
## 编辑 Tool 并生成新版本

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 选择 **MCP 管理**，找到目标 **HTTP 转 MCP 服务**，进入其 **Tool 列表**。
4. 在 MCP Tool 列表中找到目标 Tool，单击**编辑**。
5. 修改 Tool 的展示名称、描述或入参等配置。
6. 当修改涉及版本化字段时，需在弹窗中填写本次变更的版本号（必填）。
7. 单击**确定**保存。系统会生成一条新版本快照并将其置为当前生效版本。


> **说明：**若仅修改非版本化字段（不影响展示名称/描述/入参），则不会要求填写版本号，也不会生成新版本。

# 查看版本列表

在 Tool 详情中切换到“版本管理”页签，可分页查看该 Tool 的所有历史版本，列表字段包括：

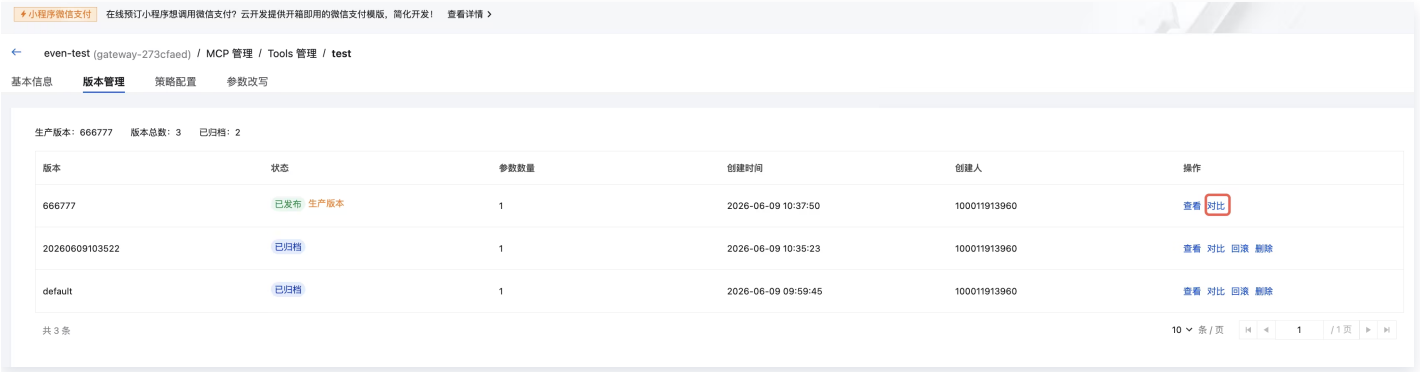


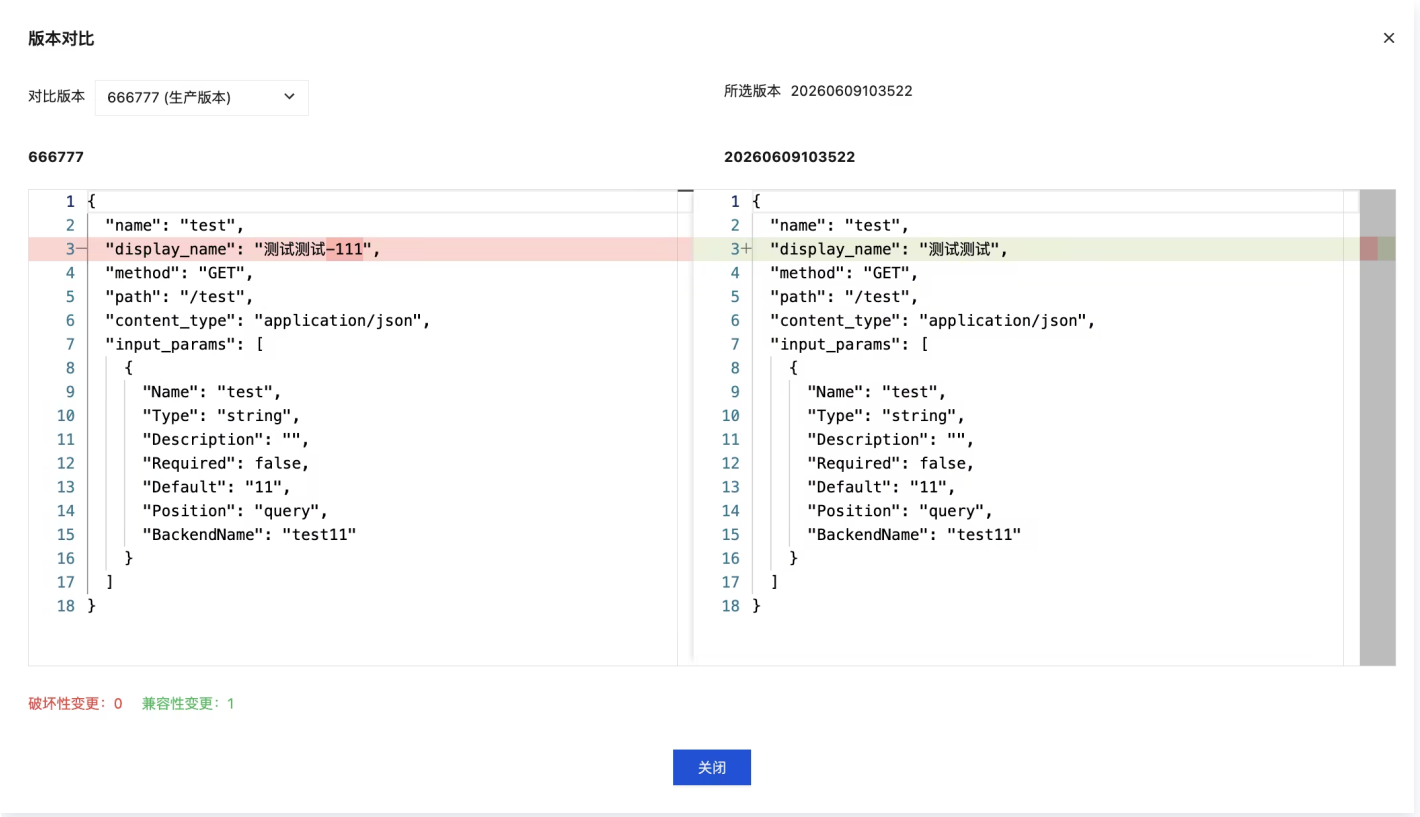

| 字段 | 说明 |
| --- | --- |
| 版本号 | 用户自定义的版本标识。 |

| | |
|---|---|
| 状态 | 标识该版本是否为已发布版本/归档版本。 |
| 参数数量 | 该版本的参数数量 |
| 创建人 | 发起该次变更的账号。 |
| 创建时间 | 版本生成时间。 |

# 对比版本差异

在版本列表中选择任一版本，单击对比，系统按字段展示二者差异，每条差异包含：




# 回滚到历史版本

1. 在版本列表中选择目标历史版本，单击**回滚**。
2. 在二次确认弹窗中确认操作。

3. 系统将该历史版本的配置快照重新下发为当前生效配置，并将其标记为生效版本。

> **说明：**
> - 回滚操作即时生效于运行时配置。
> - 若目标版本已是当前生效版本，回滚为幂等操作，不产生变化

# 限流策略

最近更新时间：2026-07-23 17:07:36

## 功能简介

MCP 网关在 Tool 维度提供的精细化流量控制能力，支持 QPS（每秒/分/时/天调用次数）和并发数两个维度的限速。用于防止 Agent 死循环打垮后端、避免单个 Tool 独占连接池、匹配第三方 MCP Server 的调用配额。仅对 HTTP 转 MCP 服务的 Tools 生效，默认关闭。

> **说明：**重试仅对 **HTTP 转 MCP 服务（Rest2MCP）**的 Tool 生效；标准 MCP 服务（ServerType = MCP）的 Tool 由后端 MCP Server 自身负责容错，网关不提供重试配置。

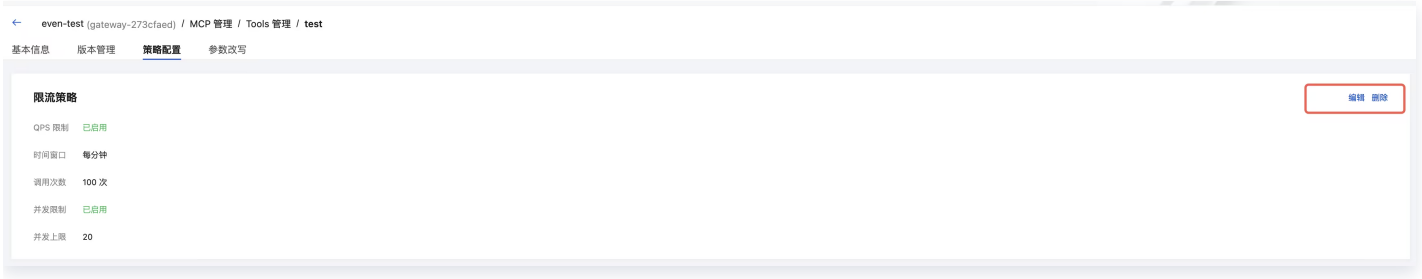
## 操作步骤

### 配置限速

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关 > 实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 选择 **MCP 管理**，找到目标 **HTTP 转 MCP 服务**，进入目标 MCP Server 的 Tool 列表。
4. 找到需要的 Tool（其所属 Server 须为 HTTP 转 MCP 服务），在 Tool 操作中选择**限流限速**，打开限流配置弹窗。
5. 按需配置以下一项或多项：
   - **QPS 限制**：选择时间窗口（秒/分/时/天），填写调用次数上限；
   - **并发限制**：填写并发上限（1～1000）。
6. 单击**确定**保存。配置将下发到数据面并生效。


### 修改与关闭限流

- **修改**：再次打开该 Tool 的 策略配置–限流策略 弹窗，调整阈值后保存即可，约 5 秒内生效。
- **删除**：删除限速配置后，该 Tool 不再受限速约束；删除该 Tool 时会自动级联清理其限速配置。



# 重试策略

最近更新时间：2026-07-23 17:07:36

## 功能简介

MCP 网关支持为单个 MCP Tool 配置应用层调用重试，按 Tool 粒度提升调用成功率，自动容错临时性后端故障。其核心能力：

- **Tool 级独立配置**：每个 Tool 单独一份重试配置，不做 Server 级 / Route 级继承或降级。
- **三类重试条件**：支持按“后端 5xx”“连接失败”“请求超时”三种条件触发重试，可任选一项或多项组合。
- **幂等安全保护**：对 `POST / PUT / PATCH` 等非幂等方法，必须显式确认后端接口具备幂等性后才允许开启重试，避免重复写入造成数据副作用。
- **默认关闭**：存量 Tool 和新建 Tool 均默认不启用重试，行为与原先完全一致，按需开启。

> **说明：**
> 重试仅对 **HTTP 转 MCP 服务**的 Tool 生效；标准 MCP 服务的 Tool 由后端 MCP Server 自身负责容错，网关不提供重试配置。

## 重试参数说明

| 参数 | 取值范围 | 说明 |
| --- | --- | --- |
| 启用重试 | 开 / 关 | 重试总开关；关闭时下述配置均不生效 |
| 重试次数 | 1 ~ 5 | 启用重试时必填；非幂等方法建议设为 1 |
| 重试条件 | 5xx / 连接失败 / 请求超时 | 启用重试时至少勾选一项 |
| 已确认后端接口具备幂等性 | 勾选 / 不勾选 | Tool 的 HTTP Method 为 POST/PUT/PATCH 时必须勾选，否则无法保存；GET/HEAD/OPTIONS/DELETE 忽略此项 |

## 三类重试条件的覆盖范围

| 重试条件 | 触发场景 |
| --- | --- |
| **后端 5xx** | 后端返回 500 / 502 / 503 / 504 状态码 |
| **连接失败** | 网关在收到后端首字节响应前的错误，包括 TCP 建连失败、connection refused、DNS 解析失败，以及建连后发送请求 / 读响应头阶段对端 RST |
| **请求超时** | 读写超时 / 上游处理超时 |

## 配置重试

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表；
2. 在**实例列表**页面，单击需要配置的网关实例的“ID”，进入该网关实例的**基本信息**页面；
3. 进入目标 MCP Server 的 **Tool 列表 / Tool 详情页**，找到需要重试的 Tool（其所属 Server 须为 HTTP 转 MCP 服务）；
4. 展开 **配置重试策略** 折叠面板；
5. 填写 **重试次数**（1 ~ 5）；
6. 勾选 **重试条件**（5xx / 连接失败 / 请求超时，至少一项）；
7. 若该 Tool 的 HTTP Method 为 POST / PUT / PATCH，需勾选“**已确认后端接口具备幂等性**”，否则保存会被拒绝；
8. 单击保存。配置将在下发到数据面并生效。


## 修改与关闭重试

- **修改**：再次展开该 Tool 的“重试策略”面板，调整参数后保存即可，约 5 秒内生效。
- **关闭 / 清除**：将“启用重试”总开关关闭并保存，或删除重试配置，该 Tool 恢复为不重试（行为与存量一致）。
- 删除该 Tool 时会自动级联清除其重试配置，无需单独操作。

# 熔断策略

最近更新时间：2026-07-23 17:07:36

## 操作场景

当 MCP 网关接入的某个后端工具（Tool）出现**持续性故障**（如响应持续变慢、持续返回错误、第三方接口持续限频）时，普通的“调用重试”反而会放大无效请求、拖垮整体性能。为此，MCP 网关提供**工具熔断与降级**能力：当某个 Tool 的后端持续故障达到设定阈值时，网关会自动“熔断”该 Tool，让后续请求**快速失败并返回降级响应**，不再打向已故障的后端；待后端恢复后，网关会自动探测并恢复正常调用。每个 Tool 拥有**独立的熔断器**，单个 Tool 故障不会影响同一服务下的其他 Tool。本文档介绍如何在控制台为 MCP Tool 配置熔断与降级策略。

## 前提条件

- 已创建 AI 网关（MCP 网关）实例。
- 已创建 **HTTP 转 MCP 服务**并完成 Tool 的配置。

## 熔断工作原理

熔断器有三种状态，并在三者之间自动切换，无需人工干预：

| 状态 | 说明 |
| --- | --- |
| **关闭（正常）** | 请求正常转发到后端，同时持续统计错误率 / 慢调用率 |
| **打开（熔断中）** | 触发熔断后，所有请求立即返回降级响应，不再打向后端；持续一段时间后进入“半开” |
| **半开（探测中）** | 放行少量探测请求到后端：若探测成功达到阈值，则恢复为“关闭”；若探测失败，则重新熔断 |

## 操作步骤

### 步骤一：进入 Tool 熔断策略配置

1. 登录 微服务平台控制台 ，左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 选择 **MCP 管理**，找到目标 **HTTP 转 MCP 服务**，进入其 **Tool 列表**。
4. 单击目标 Tool 进入 **Tool 详情**，选择**策略配置**页签，找到**熔断策略**区域。

### 步骤二：开启并配置熔断策略

1. 打开**熔断策略**，展开配置表单。
2. 配置**熔断触发条件**（以下两类条件可单独或同时启用，**至少启用一项**，任一条件满足即触发熔断）：

**基础统计设置**

| 配置项 | 默认值 | 取值范围 | 说明 |
| --- | --- | --- | --- |
| 统计时长 | 10 秒 | 1~60 秒 | 统计错误率 / 慢调用率的滑动时间窗口 |
| 最小请求数 | 10 次 | 1~100 次 | 窗口内请求数不足时不触发熔断，避免低流量误判 |

**错误比例触发**（默认开启）

| 配置项 | 默认值 | 取值范围 | 说明 |
| --- | --- | --- | --- |
| 触发状态码 | 500、502、503、504 | 400~599，最多 32 个 | 哪些 HTTP 状态码计入“错误”（可按需加入如 429） |
| 错误比例阈值 | 50% | 1~100 | 错误率达到该比例时触发熔断 |

**慢调用触发**（默认关闭）

| 配置项 | 默认值 | 取值范围 | 说明 |
| --- | --- | --- | --- |
| 最大响应时间 | 3000 毫秒 | 100~60000 毫秒 | 超过该时间的调用视为“慢调用”。**该值必须小于后端服务的请求超时时间**，否则慢调用无法被统计到 |
| 慢调用比例阈值 | 80% | 1~100 | 慢调用率达到该比例时触发熔断 |

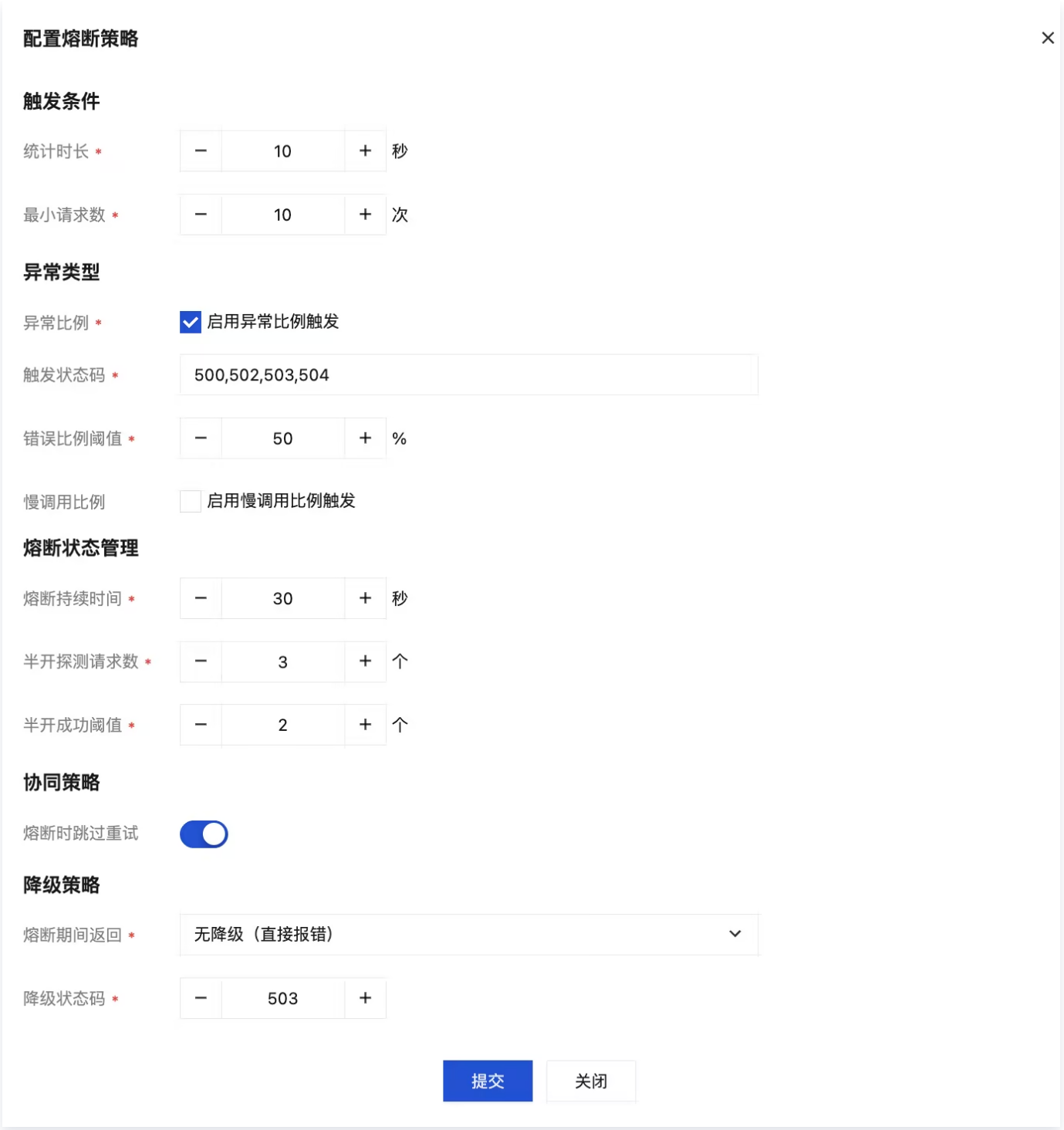
3. 配置**熔断器恢复行为**：

| 配置项 | 默认值 | 取值范围 | 说明 |
| --- | --- | --- | --- |
| 熔断持续时间 | 30 秒 | 5~600 秒 | 熔断打开后持续该时长，到期后进入半开探测 |
| 半开探测请求数 | 3 个 | 1~10 个 | 半开状态下放行到后端的探测请求数量 |
| 半开成功阈值 | 2 个 | 1~探测请求数 | 探测成功数达到该值则恢复正常 |
| 熔断时跳过重试 | 开启 | 开 / 关 | 建议保持开启。熔断期间跳过该 Tool 的重试，避免无效循环 |

4. 配置**降级策略**（熔断期间向调用方返回的内容）：

| 降级策略 | 默认返回码 | 返回内容 | 适用场景 |
| --- | --- | --- | --- |
| 无降级 | 503 | 返回标准错误信息（Circuit breaker open） | 调用方（Agent）能识别错误并自行处理，不需要“假成功”响应 |
| 固定 JSON 响应 | 200 | 返回您预设的固定 JSON 内容 | 调用方可接受空结果或兜底提示，无需真实数据 |

5. 确认配置无误后，单击 提交。




## 步骤三：查看与修改

- **查看已启用配置**：开启后，熔断策略区域会显示 **“已开启”** 及当前配置摘要（触发条件、统计窗口、熔断持续时间、降级策略等），可单击**编辑**修改。
- **修改配置**：调整参数后重新保存即可，约 5 秒内生效。
- **关闭熔断**：将**熔断策略**开关关闭并保存，该 Tool 立即恢复为正常转发（不熔断）。

## 应用场景示例

| 场景 | 现象 | 推荐配置 |
| --- | --- | --- |
| **后端持续变慢** | 后端工具响应从毫秒级劣化到秒级，重试导致请求大量堆积 | 开启“慢调用触发”（最大响应时间 3000ms、比例 80%），降级返回固定 JSON |
| **后端持续报错** | 后端依赖故障持续返回 5xx，重试放大无效请求 | 开启“错误比例触发”（状态码 500/502/503/504、阈值 50%），保持“熔断时跳过重试” |
| **第三方接口限频** | 第三方 API 持续返回 429，Agent 陷入无效调用循环 | 将 429 加入“触发状态码”，开启错误比例触发，降级返回固定 JSON 兜底 |

## 相关说明

- 熔断与降级**默认关闭**，开启后才生效，不影响存量 Tool 的现有行为。
- 每个 Tool 的熔断器相互独立，互不影响。
- 建议将熔断与“工具调用重试”配合使用，并保持“熔断时跳过重试”开启，以获得最佳容错效果。

# 配置改写

最近更新时间：2026-07-23 17:07:36

## 操作场景

当您使用 HTTP 转 MCP 方式将已有的 RESTful 接口封装为 MCP 工具时，工具对外暴露的参数结构与后端接口的实际报文格式往往并不一致。通过 **报文动态调整** 功能，您可以在不改动后端服务的前提下，对工具的请求参数和响应内容进行重命名、定位、固定注入和模板化改写，使 MCP 工具的输入输出更符合大模型调用习惯。

## 前提条件

- 已创建云原生 AI 网关实例。
- 已创建 HTTP 转 MCP 类型的 MCP Server 及其工具。

## 操作步骤

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **MCP 管理**，在服务列表中单击目标 HTTP 转 MCP 服务的 **ID/名称**，进入服务详情页。
4. 在服务详情页顶部单击 **Tools 管理** 页签，在 Tools 列表页面选择目标 Tool 后进入到 Tool 基本信息页面。
5. 在 tool 详情页面选择“参数改写”tab 页面后，单击**配置参数改写**。

   请求参数改写：
   - **参数重命名**：为对外参数设置映射到后端的真实字段名（后端字段名）。
   - **参数位置（Position）**：指定参数在后端请求中的位置（如 Query、Header、Body 等）。
   - **固定注入字段（InjectFields）**：为后端请求注入固定值字段（如固定的鉴权头、渠道标识等）。

   响应模板改写：
   - **响应模板（Response Body）**：使用 GJSON 语法从后端响应中提取并重组返回内容。
   - **错误响应模板**：自定义错误场景下返回给客户端的内容。
6. 单击**保存**使配置生效。

## 相关说明

配置时请遵守以下限制：

| 配置项 | 取值范围 / 限制 |
| --- | --- |
| 工具参数数量 | ≤ 20 个 |
| 标签键长度 | ≤ 128 字符 |

| | |
|---|---|
| 标签值长度 | ≤ 4096 字符 |
| 请求 Body 模板长度 | ≤ 16384 字符 |
| 响应 Body 模板长度 | ≤ 16384 字符 |
| 错误响应模板长度 | ≤ 2048 字符 |

**响应模板采用 GJSON 语法提取字段，配置后建议在调试中验证输出是否符合预期。**

# 访问控制

最近更新时间：2026-07-23 17:07:36

## 操作场景

为保障 MCP Server 的访问安全，您可以为 MCP Server 开启 **调用方鉴权**，要求客户端携带有效凭证才能访问；同时还可以基于调用方（Consumer）配置 **Server 级** 和 **Tool 级** 的黑白名单（ACL），实现“谁能调用整个服务”“谁能调用某个具体工具”的精细化访问控制。

## 前提条件

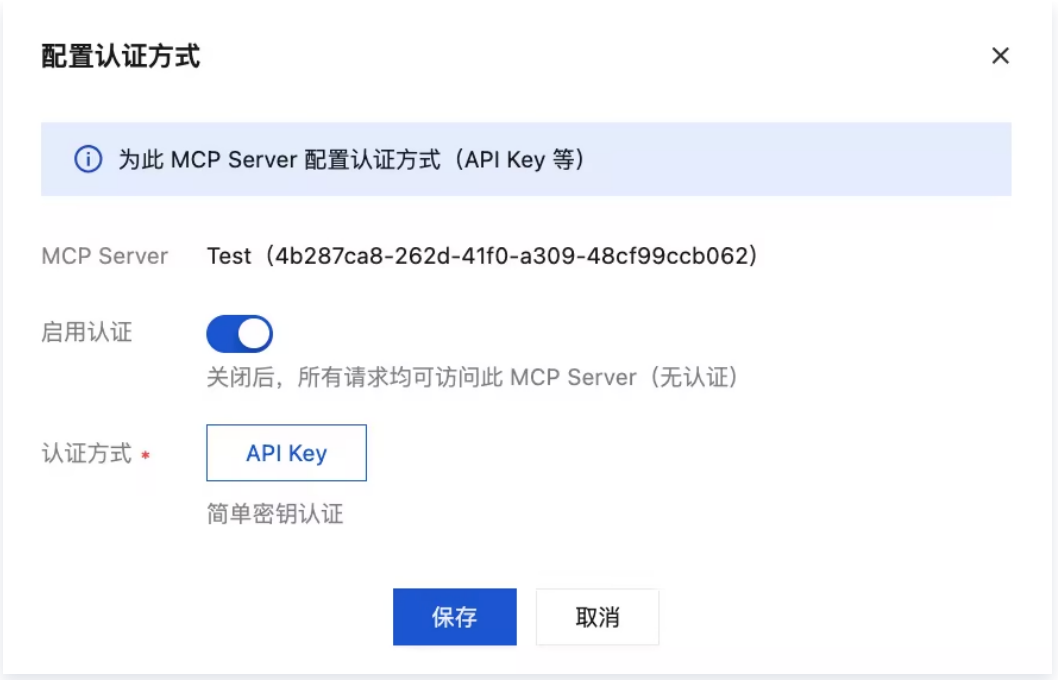
- 已创建云原生 AI 网关实例。
- 已创建 MCP Server。
- 如使用密钥鉴权，请提前创建好调用方（Consumer）及其凭证。

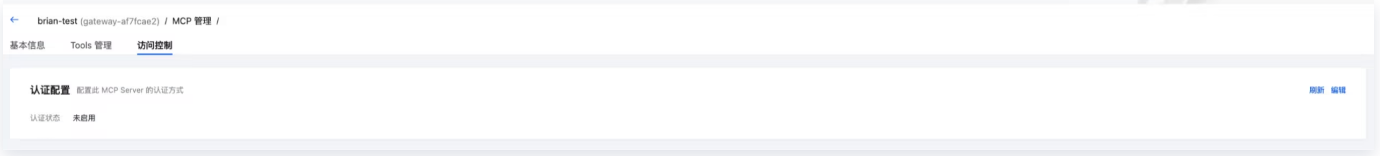
## 操作步骤

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关 > 实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧菜单选择 **MCP 管理 > MCP 服务**，选择目标 MCP 服务，进入目标 MCP 服务详情页。
4. 切换到 **访问控制** Tab 页面。

## 配置调用方鉴权方式

1. 设置 **鉴权方式（auth_type）**
2. 单击“认证配置”右侧的**编辑**按钮。
3. 配置认证方式：API Key。并保存。




| 取值 | 说明 |
| --- | --- |
| None | 无鉴权，不校验调用方身份 |
| ApiKey | 密钥鉴权，要求调用方携带有效 API Key |

## 配置 Server 级黑白名单

1. 设置 Server 访问控制（acl_type）：

| 取值 | 说明 |
| --- | --- |
| None | 不限制，所有调用方均可访问 |
| Allow | 白名单，仅名单内调用方可访问 |
| Deny | 黑名单，名单内调用方禁止访问 |


2. 根据所选类型，添加对应的调用方（Consumer）列表。

## 配置 Tool 级黑白名单

1. 在工具列表中选择目标工具，单独为其设置 **Tool 访问控制**（None / Allow / Deny）及对应调用方名单。
2. 单击**保存**使配置生效。


# 相关说明

- 鉴权与黑白名单分为 **Server** 和 **Tool** 两层，Tool 级规则在 Server 级规则的基础上进一步收敛权限。
- 配置变更后即时生效，请确认调用方凭证已正确下发，避免误拦截正常业务。

# MCP 日志管理

最近更新时间：2026-07-23 17:07:34

## 操作场景

MCP 网关支持记录 MCP 协议的调用日志（**MCP 访问日志**），帮助您审计工具调用情况、排查问题和分析用量。您可以在控制台查看日志，也可以将日志投递到 **腾讯云日志服务 CLS** 进行长期存储与检索。

## 前提条件

- 已创建云原生 API 网关实例。
- 已通过网关创建并对外提供 MCP Server 服务。
- 如需投递到 CLS，请提前开通日志服务 CLS 并准备好日志集 / 日志主题。

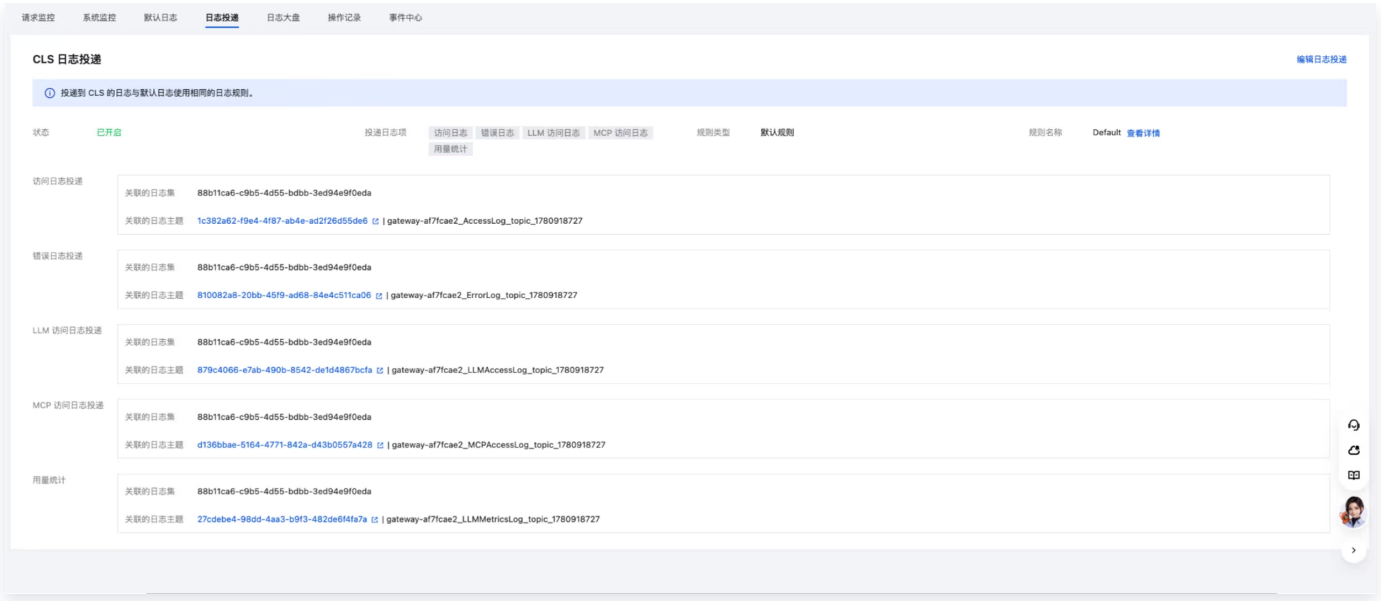
## 操作步骤

### 查看 MCP 访问日志

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。
3. 在左侧菜单选择**数据观测**，然后选择 **日志大盘 > MCP 监控**。
4. 通过时间范围、关键字等条件检索，查看具体的 MCP 调用记录。

### 配置日志投递到 CLS

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。
3. 在左侧菜单选择**数据观测**，然后选择**日志投递页面**。
4. 点击右上角的**编辑**按钮后，投递日志项勾选 **MCP 访问日志**。
5. 选择目标 **CLS 日志集 / 日志主题**。
6. 单击**保存**，日志将按配置投递到 CLS。


请求监控
系统监控
默认日志
日志投递
日志大盘
操作记录
事件中心
CLS 日志投递
编辑日志投递
投递到 CLS 的日志与默认日志使用相同的日志规则。
状态
已开启
投递日志项
访问日志
错误日志
LLM 访问日志
MCP 访问日志
用量统计
规则类型
默认规则
规则名称
Default
查看详情
访问日志投递
关联的日志集
88b11ca6-c9b5-4d55-bdbb-3ed94e9f0eda
关联的日志主题
1c382a62-f9e4-4f87-ab4e-ad2f26d55de6 | gateway-af7fcae2_AccessLog_topic_1780918727
错误日志投递
关联的日志集
88b11ca6-c9b5-4d55-bdbb-3ed94e9f0eda
关联的日志主题
810082a8-20bb-45f9-ad68-84e4c511ca06 | gateway-af7fcae2_ErrorLog_topic_1780918727
LLM 访问日志投递
关联的日志集
88b11ca6-c9b5-4d55-bdbb-3ed94e9f0eda
关联的日志主题
879c4066-e7ab-490b-8542-de1d4867bcfa | gateway-af7fcae2_LLMAccessLog_topic_1780918727
MCP 访问日志投递
关联的日志集
88b11ca6-c9b5-4d55-bdbb-3ed94e9f0eda
关联的日志主题
d136bbae-5164-4771-842a-d43b0557a428 | gateway-af7fcae2_MCPAccessLog_topic_1780918727
用量统计
关联的日志集
88b11ca6-c9b5-4d55-bdbb-3ed94e9f0eda
关联的日志主题
27cdebe4-98dd-4aa3-b9f3-482de6f4fa7a | gateway-af7fcae2_LLMMetricsLog_topic_1780918727


## 相关说明

- MCP 访问日志包含一组以 `aigw.mcp.*` 为前缀的专用字段（共 13 个，如工具名称、会话标识、调用结果等），便于精细化分析。
- 日志以 `mcp-access.log` 形式记录，采用 OTel 规范字段格式。
- 投递到 CLS 后，可结合 CLS 的检索分析、仪表盘、告警等能力做进一步运营。

# Tools 分组

最近更新时间：2026-07-23 17:07:36

## 功能简介

当接入的 MCP Server 与 Tool 数量增长后，扁平化的工具管理会带来三类问题：Agent 上下文 Token 膨胀、多来源工具同名冲突、工具权限无法分级。Tools 分组通过“虚拟 MCP Server 端点”解决上述问题：

- 减少上下文消耗：按业务场景（如“库存管理”“供应链查询”）组织 Tool 子集，Agent 仅加载所需子集。
- 分组隔离与冲突自动处理：不同分组的 Tool 彼此不可见；同名工具在添加时自动检测冲突并提供解决策略。
- 统一入口：每个分组对外是一个标准 MCP Server 端点，Agent 零配置接入。

> **说明：**
> Tools 分组为可选增强能力，不使用分组时网关行为与原先完全一致；该功能要求网关实例版本 ≥ 3.9.4。

## 前提条件

- 已创建 MCP 网关实例。
- 已存在至少一个 MCP Server 并完成 Tool 注册（分组通过“引用”已有 Tool 构建，不重复创建 Tool）。

## 操作步骤

### 新建/编辑 Tools 分组

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **MCP 管理**，进入服务列表页。
4. 在服务列表页顶部单击 **Tools 分组** 页签，在列表上方单击 **新建分组**，打开“新建 Tools 分组”页面。
5. 在弹窗中填写以下配置：

| 配置项 | 是否必填 | 说明 |
| --- | --- | --- |
| 分组名称（Name） | 是 | 英文标识，最大 60 字符，同网关实例内唯一，作为接入端点路径的一部分。 |
| 分组展示名称（DisplayName） | 是 | 分组中文/展示名，最大 60 字符。 |
| 描述（Description | 否 | 最大 200 字符。 |

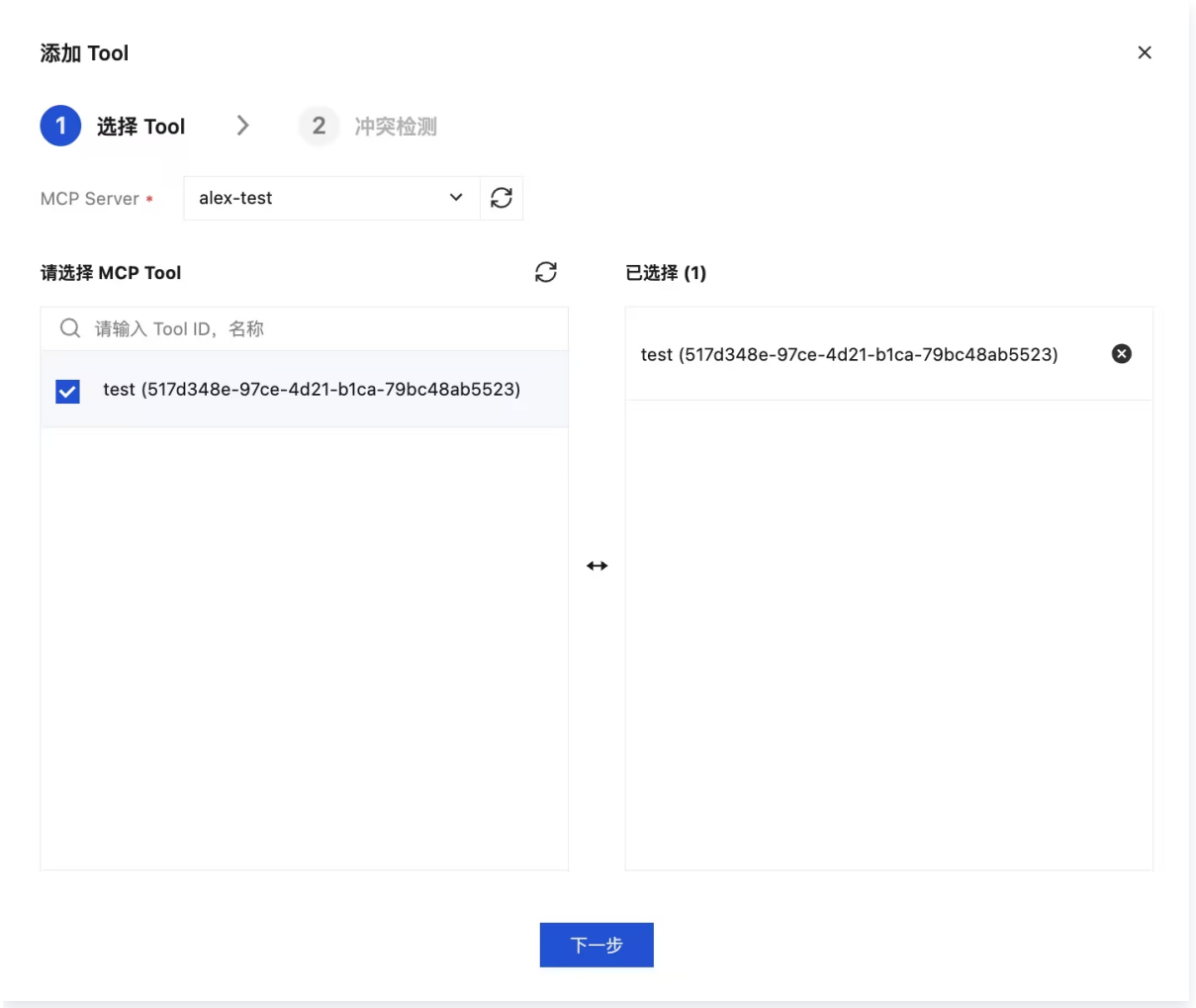
| | | |
|---|---|---|
| ） | | |
| Tools 总数上限（ToolCountLimit） | 否 | 范围 1 ~ 100，默认值 10。建议单分组 1 ~ 20个 Tool。 |
| 冲突解决策略（ConflictStrategy） | 否 | Reject（拒绝注册）/ AutoPrefix（自动加前缀，默认，自动为 Tool 名称添加前缀。例：toolname -> mcpserver_toolname）。 |

6. 单击确定，网关将自动创建该分组对应的虚拟 MCP Server 端点。


# 添加 MCP Tools

## 步骤一：选择 Tools

从已有 MCP Server 中按需勾选要加入分组的 Tool（支持跨 Server 多选）。同一 Tool 可被多个分组引用，添加时仅写入引用关系，不复制 Tool 数据。




## 步骤二：Tools 名称冲突检测

系统会将待添加 Tool 的名称与分组内已有 Tool 进行比对：

- 无冲突：直接添加成功。
- 存在冲突：根据分组的名称冲突策略处理。
- Reject：阻断本次添加，提示冲突的 Tool 名称及其所属原 MCP Server。
- AutoPrefix：自动为冲突 Tool 添加前缀以区分归属，添加成功。

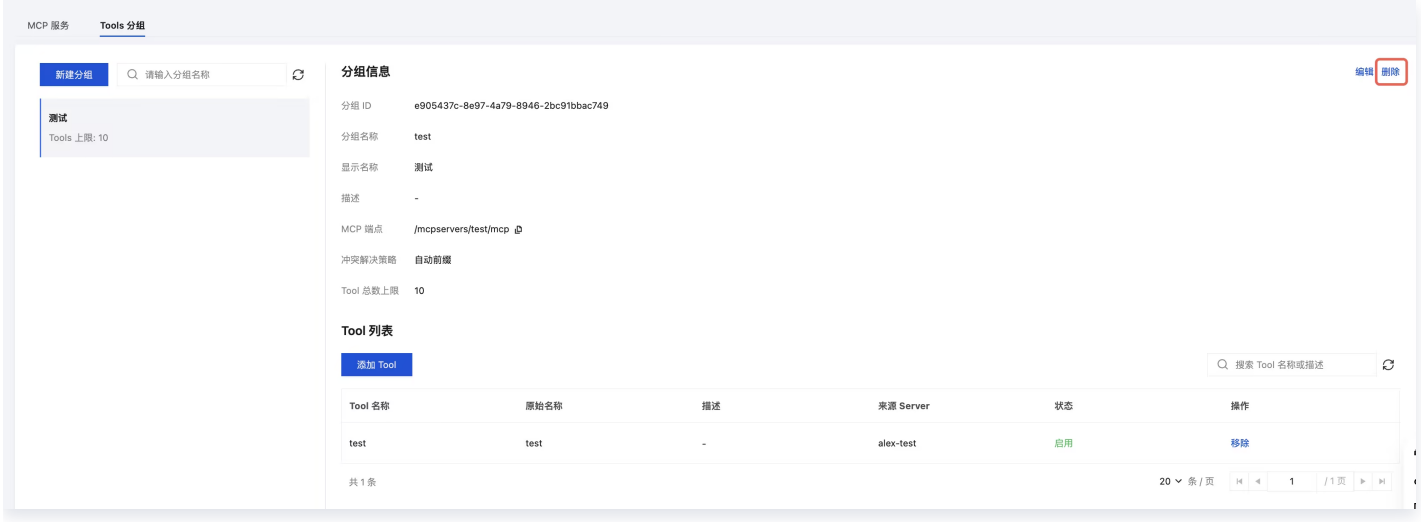
> **注意：**
> 命名规范（如“域–名词–动词”）为建议级别，不强制。MCP 协议本身仅要求 Tool 名称在同一端点内唯一。


## 移除 Tool

在分组内 Tool 列表中单击目标 Tool 的移除。移除仅解除引用关系，不影响该 Tool 在其原 MCP Server 中的定义。


## 删除分组

删除分组将自动清理分组内所有 Tool 引用关系并回收对应端点资源，无需先手动移除分组内 Tool。



# Agent 管理
# Agent API

最近更新时间：2026-07-23 17:07:35

## 操作场景

企业同时使用多个 Agent 平台（如 Coze、Dify、自建 Agent Runtime），各平台独立部署、各自暴露 HTTP 接口，业务侧需分别对接不同地址与鉴权方式，管理成本高且无法统一审计。通过 AI 网关 Agent API，您可以将所有 Agent 服务统一接入至单一入口，实现：

- 多源 Agent 统一接入：将不同平台的 Agent 服务（Coze / Dify / 自建 Runtime / MCP Server）通过路由规则分发至对应后端，业务系统只需对接网关一个入口。
- 协议无关透明代理：不依赖任何特定厂商协议，后端只要是 HTTP/S 服务即可接入，无需适配 SDK 或改造现有 Agent。
- 统一治理与可观测：所有 Agent 请求经过网关统一鉴权、限流、审计，调用链路全量可追踪。

## 前置条件

- 已创建 AI 网关实例。
- 已创建后端服务(在**服务管理** > **服务**中配置)。

## 操作步骤

### 创建 Agent API

#### 步骤1：进入 Agent API 列表页

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。
3. 在左侧导航栏选择 **Agent 管理**；
4. 单击 **新建**。

#### 步骤2：配置基本信息

在弹出的创建弹窗中，配置以下信息：

| 参数 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| API 名称 | 是 | Agent API 的唯一标识，支持英文、数字、下划线、短横线，长度1-64字符 | `agent-backend-api` |

| | | | |
|---|---|---|---|
| 协议 | 是 | 固定为"自定义"(区别于模型 API 的 OpenAI 兼容协议) | 自定义 |
| Base Path | 是 | API 的路径前缀，所有路由路径都会以此为前缀 | /agent |

**Base Path 说明:**
假设配置 Base Path 为 `/agent`，创建路由路径为 `/api/users`， 客户端访问地址： `https://{网关域名}/agent/api/users`

### 步骤3：完成创建

确认配置无误后，单击 **确定** 完成 Agent API 创建。创建成功后，会自动跳转到 Agent API 详情页。

## 管理 Agent API

### 查看 Agent API 列表

在 **Agent 管理** > **Agent API** 列表页，可查看所有已创建的 Agent API:

### 查看 Agent API 详情

单击 API 名称进入详情页，查看 API 的基本信息。

### 编辑 Agent API

1. 在 Agent API 列表页，单击操作列的 **编辑**;
2. 修改 API 基本信息(注意：API 名称创建后不可修改);
3. 单击 **确定** 保存修改。

### 删除 Agent API

1. 在 Agent API 列表页，单击操作列的 **删除**;
2. 在弹出的确认对话框中，单击 **确定**。

> **注意:**
> 删除 Agent API 后，该 API 下的所有请求将无法路由，请谨慎操作。

## 创建路由规则

Agent API 通过路由规则将请求分发到不同的后端服务。一个 Agent API 可以创建多条路由规则。

### 步骤1：进入路由列表

1. 在 Agent API 详情页，选择 **路由列表** Tab;
2. 单击 **创建路由**。

### 步骤2：配置基本信息

#### 基本信息

| 字段 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| 路由名称 | 否 | 路由的名称，支持中文、英文、数字、`.`、`-`、`_` | 用户服务路由 |
| 请求协议 | 是 | 选择支持的协议，默认"HTTP&HTTPS" | `HTTP&HTTPS` |

#### 匹配规则

| 字段 | 是否必填 | 说明 | 示例 |
| --- | --- | --- | --- |
| 请求方法 | 否 | 支持的 HTTP 方法，可多选 | `GET`、`POST` |
| 请求路径 | 否 | 路径匹配规则，支持精确匹配、前缀匹配、正则匹配 | 前缀匹配：`/api/users` |
| Host | 否 | 主机名匹配，支持大小写敏感配置，可添加多个 | `api.example.com` |
| Header | 否 | 请求头匹配，键支持数字、字母、`_`、`-`，可添加多个 | `X-Service: user-service` |

### 步骤3：配置路由后端

在第二步中，配置路由的后端服务，从"服务管理>服务"列表中选择后端服务。

### 步骤4：完成创建

单击 **完成创建**，路由规则创建成功。

## 编辑和删除路由

### 编辑路由

1. 在 Agent API 详情页 > 路由列表 Tab，选择需要编辑的路由；
2. 单击右侧路由详情区域的 **编辑路由** 按钮；
3. 修改路由配置(两步向导，与创建流程相同)；
4. 单击 **确定** 保存修改；
5. 修改后的路由需要重新发布才会生效。

### 删除路由

1. 在 Agent API 详情页 > 路由列表 Tab，选择需要删除的路由；
2. 单击右侧路由详情区域的 **更多操作 > 删除；**
3. 在确认对话框中单击 **确定。**

> **注意：**
> 删除路由后，该路由的所有请求将无法路由，请谨慎操作。

## 查看路由详情

在 Agent API 详情页 > 路由列表 Tab，左侧路由列表选择路由后，右侧会展示路由详情。

# 服务管理
# 服务来源

最近更新时间：2026-07-23 17:07:35

## 操作场景

当您需要将已有的注册中心或域名服务中的服务批量接入 AI 网关时，可以通过配置服务来源，实现 服务的自动同步和统一管理。服务来源分为三种类型：

- **MCP 注册中心**：对接注册中心（如 TSF Nacos），同步原生 MCP Server，实现 MCP Tool 自动发现。
- **普通注册中心**：对接注册中心（如 TSF Nacos 2.x、TSF PolarisMesh），自动同步后端服务地址和元数据，适用于将后端服务批量接入 AI 网关。
- **私有 DNS**：对接腾讯云 Private DNS 进行域名解析，适用于通过域名访问的服务场景。

## 前置条件

- 已开通 腾讯云 AI 网关服务
- 已创建 AI 网关实例
- 注册中心实例与网关实例在同一 VPC 网络

## 操作步骤

### 步骤1：进入服务来源配置页面

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。
3. 在左侧导航栏，选择**服务管理 > 服务来源**；
4. 单击**新建**按钮。

### 步骤2：选择服务来源类型

系统提供三种服务来源类型，请根据您的实际场景选择。

### 场景一：配置 MCP 注册中心（同步原生 MCP Server）

**适用场景**：您的 MCP Server 已注册到 TSF Nacos，需要 AI 网关自动发现和同步 Tool。
**参数配置**：

| 参数名称 | 是否必选 | 说明 |
| --- | --- | --- |

| | | |
|---|---|---|
| 来源类型 | 是 | 选择 **MCP 注册中心** |
| 来源产品 | 是 | **TSF Nacos**（推荐）<br>直接对接注册在 TSF Nacos 3.x 上的服务 |
| 服务来源名称 | 是 | 自定义名称，最长60个字符，支持中英文、数字及 `-` 、`_` 符号，不能以隔符开头和结尾，同网关实例下唯一 |
| 选择实例 | 是 | 选择已有的 Nacos 实例（仅显示与网关实例同 VPC 的实例） |
| 用户名 | 是 | Nacos 鉴权用户名 |
| 密码 | 是 | Nacos 鉴权密码，支持英文字母、数字、空格及常见符号 |
| 描述 | 否 | 服务来源描述信息，最长200个字符 |

**操作步骤：**

1. 在新建服务来源弹窗中，选择来源类型为 **MCP 注册中心**；
2. 选择来源产品为 **TSF Nacos**；
3. 填写服务来源名称（如：`mcp-nacos-prod`）；
4. 从下拉列表中选择 Nacos 实例；
5. 填写鉴权用户名和密码；
6. （可选）填写描述信息；
7. 单击 **确定** 完成创建。

**同步效果：**

- AI 网关将自动同步 Nacos 中注册的 MCP Server 服务。
- MCP Server 的 Tool 定义将自动发现并在控制台展示。
- 服务实例变更时，网关自动感知并更新。

### 场景二：配置普通注册中心（批量接入后端服务）

**适用场景**：您的后端服务已注册到 TSF Nacos 2.x 或 TSF PolarisMesh，需要 AI 网关自动发现并接入这些服务。

**参数配置：**

| 参数名称 | 是否必选 | 说明 |
|---|---|---|
| 来源类型 | 是 | 选择 **普通注册中心** |

| | | |
|---|---|---|
| 来源产品 | 是 | **TSF Nacos**（推荐）：直接对接注册在 TSF Nacos 2.x 上的服务；<br>**TSF PolarisMesh**：直接对接注册在 TSF PolarisMesh 上的服务 |
| 服务来源名称 | 是 | 自定义名称，最长60个字符，支持中英文、数字及 `-`、`_` 符号，不能以分隔符开头和结尾，同网关实例下唯一 |
| 选择实例 | 是 | 选择已有的 Nacos / PolarisMesh 实例（仅显示与网关实例同 VPC 的实例） |
| 用户名 | 是 | 注册中心鉴权用户名。支持英文字母、数字、空格及常见符号 |
| 密码 | 是 | 注册中心鉴权密码。支持英文字母、数字、空格及常见符号 |
| 描述 | 否 | 服务来源描述信息，最长200个字符 |

**操作步骤：**

1. 在新建服务来源弹窗中，选择来源类型为**普通注册中心**；
2. 选择来源产品为 **TSF Nacos**（推荐）或 **TSF PolarisMesh**；
3. 填写服务来源名称（如：`nacos-prod`、`polaris-prod`）；
4. 从下拉列表中选择对应的注册中心实例；
5. 填写鉴权用户名和密码；
6. （可选）填写描述信息；
7. 单击 **确定** 完成创建。

**同步效果：**

- AI 网关将自动同步注册中心中已注册的后端服务（包含服务地址、端口等元数据）。
- 服务实例变更时，网关自动感知并更新，无需手动维护服务列表。
- 同步的服务可在“**服务管理** > **服务**”列表中查看，可直接用于创建模型服务的后端来源。

## 场景三：配置私有 DNS（通过域名解析访问服务）

**适用场景：**您的后端服务通过域名访问，需要使用腾讯云 Private DNS 进行域名解析。

**参数配置：**

| 参数名称 | 是否必选 | 说明 |
|---|---|---|
| 来源类型 | 是 | 选择 **私有 DNS** |
| 来源产品 | 是 | **腾讯云私有解析**<br>使用 Private DNS 进行域名解析 |
| 描述 | 否 | 服务来源描述信息，最长200个字符 |

**操作步骤：**

1. 在新建服务来源弹窗中，选择来源类型为 **私有 DNS**；
2. 选择来源产品为 **腾讯云私有解析**；
3. （可选）填写描述信息；
4. 单击 **确定** 完成创建。

**配置说明：**

- 配置完成后，在创建服务时可选择该来源。
- 后端地址支持填写域名（如：`server.internal.example.com`）。
- 网关将通过 Private DNS 解析域名并转发请求。

### 步骤3：保存配置

1. 确认所有参数填写正确；
2. 单击 **确定** 按钮完成创建；
3. 系统自动跳转到服务来源列表页。

## 结果验证

配置完成后，可通过以下方式验证：

- 在服务来源列表页，查看新建的来源记录。
- **MCP 注册中心类型**：在新建 MCP 服务时，可在服务来源下拉列表中选择该来源。
- **普通注册中心类型**：在新建服务时，可在服务来源下拉列表中选择该来源，同步的注册中心服务将作为可选后端。
- **私有 DNS 类型**：在新建服务时，选择该来源，后端地址可填写域名。
- 在 **MCP 管理 > MCP 服务** 页面，新建服务时可在服务来源下拉列表中选择该来源。

## 常见问题

### Q1：为什么选择实例时看不到我的 Nacos 实例？

**A：** 请检查以下几点：

1. Nacos 实例与 AI 网关实例是否在同一 VPC 网络；
2. 您的账号是否有 Nacos 实例的查看权限；
3. Nacos 实例是否处于运行中状态。

### Q2：一个服务来源可以关联多个 MCP 服务吗？

**A：** 可以。单个服务来源可以关联多个 MCP 服务，适用于以下场景：

- 同一个 Nacos 实例中注册了多个 MCP Server。
- 同一个 Nacos / PolarisMesh 实例中注册了多个后端服务。
- 同一个 Private DNS 解析多个不同的域名。

## Q3： 服务来源配置错误导致同步失败，如何排查？

A： 请按以下步骤排查：

1. 进入服务来源详情页，查看错误信息；
2. **MCP 注册中心**：检查用户名/密码是否正确，Nacos 实例是否可访问；
3. **私有 DNS**：检查域名解析记录是否正确配置；
4. 查看 AI 网关日志获取详细错误信息。

# 服务

最近更新时间：2026-07-23 17:07:35

## 操作场景

服务管理是 AI 网关中统一管理后端服务地址的核心模块。通过在此模块注册后端服务，可以在创建 MCP 服务、配置路由、设置限流等场景中快速引用，避免重复填写服务地址，便于统一维护。
根据后端服务的接入方式，服务分为以下两种类型：

- **域名/IP**：后端服务通过固定域名或 IP 地址访问，网关直接连接后端服务。
- **私有 DNS**：后端服务通过私有域名访问，网关使用腾讯云 Private DNS 进行域名解析。

本文档介绍如何创建和管理这两种类型的服务。

## 前置条件

- 已开通腾讯云 AI 网关服务。
- 已创建 AI 网关实例。
- 后端服务已部署并可访问(网络连通)。
- **(私有 DNS 类型)** 已在腾讯云 Private DNS 中配置相应域名的解析记录。
- **(私有 DNS 类型)** 已创建私有 DNS 类型的服务来源。

## 操作步骤

## 场景一：创建域名/IP 类型的服务

### 适用场景

后端服务通过固定的域名或 IP 地址访问，无需使用 Private DNS 进行域名解析。适用于：

- 后端服务地址固定不变。
- 使用公网域名或公网 IP 访问。
- 使用内网 IP 地址直接访问。

### 前置条件

- 后端服务已部署并可通过域名或 IP 访问。
- 已确认后端服务的完整地址(协议 + 域名/IP + 端口 + 路径)。

### 操作步骤

**步骤1：进入服务管理页面**

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。

3. 在左侧导航栏，选择**服务管理 > 服务**；
4. 单击**新建**按钮。

**步骤2：选择服务类型并填写参数**

在新建服务弹窗中，填写以下参数：

| 参数名称 | 是否必选 | 说明 |
| --- | --- | --- |
| 服务名称 | 是 | 自定义名称，最长60个字符，支持中英文、数字及 `-` 、`_` 符号，不能以数字和分隔符开头，不能以分隔符结尾 |
| 服务类型 | 是 | 选择 **域名/IP** |
| 服务协议 | 是 | **HTTP**： 标准 HTTP 协议<br>**HTTPS**： 加密 HTTPS 协议 |
| 服务地址 | 是 | 输入后端服务的域名或 IP 地址(不含协议头)<br>示例(域名)： `backend-server.example.com`<br>示例(IP)： `192.168.1.100` |
| 服务端口 | 是 | 后端服务的端口号，支持 1-65535 的任意端口 |
| 服务路径 | 是 | 后端服务的路径(如： `/api/v1` )<br>如无路径，填写 `/` |
| 超时时间 | 是 | 网关调用后端的超时时间，默认 60000 毫秒，支持调整 |
| 重试次数 | 是 | 请求失败时的重试次数，默认 5 次，支持调整 |

**步骤3：保存配置**

1. 确认所有参数填写正确；
2. 单击 **确定** 按钮完成创建；
3. 系统自动跳转到服务列表页。

## 结果验证

- 在服务列表页，查看新建的服务记录。
- 服务类型列显示“域名/IP”。
- 服务地址列显示完整的后端地址(协议 + 域名/IP + 端口 + 路径)。
- 在创建 MCP 服务、配置路由时，可从服务下拉列表中选择该服务。

### 注意事项

- 服务地址填写时不需要包含协议头(如 `http://` 或 `https://`)，协议通过“服务协议”字段单独选择。
- 服务地址可以填写域名或 IP，但必须确保网关实例可以访问该地址。
- 使用公网域名时，请确保网关实例具备公网访问能力。
- 使用内网 IP 地址时，请确保网关实例与后端服务在同一 VPC 或已配置网络互通。
- 建议使用 HTTPS 协议，确保数据传输安全。
- 超时时间和重试次数可根据实际后端服务响应情况调整。

## 场景二：创建私有 DNS 类型的服务

### 适用场景

后端服务通过内网私有域名访问，且域名解析由腾讯云 Private DNS 服务管理。适用于：

- 企业内网服务通过私有域名访问。
- 需要与腾讯云 Private DNS 集成的场景。
- 后端服务地址需要动态解析的场景。

### 前置条件

- 已在腾讯云 Private DNS 中配置相应域名的解析记录。
- 已创建私有 DNS 类型的服务来源(参考 服务来源管理 )。
- 后端服务已部署并可通过私有域名访问。

### 操作步骤

**步骤1：进入服务管理页面**

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。
3. 在左侧导航栏，选择**服务管理 > 服务**；
4. 单击**新建**按钮，

**步骤2：选择服务类型并填写参数**

在新建服务弹窗中，填写以下参数：

| 参数名称 | 是否必选 | 说明 |
| --- | --- | --- |
| 服务名称 | 是 | 自定义名称，最长60个字符，支持中英文、数字及 `-` 、`_` 符号，不能以数字和分隔符开头，不能以分隔符结尾 |
| 服务类型 | 是 | 选择 **私有 DNS** |

| | | |
|---|---|---|
| 服务来源 | 是 | 从下拉列表中选择已配置的私有 DNS 类型服务来源<br>(单击刷新按钮可更新服务来源列表) |
| 服务协议 | 是 | **HTTP**：标准 HTTP 协议<br>**HTTPS**：加密 HTTPS 协议 |
| 私有域名 | 是 | 输入后端服务的私有域名(如：`backend-server.internal.example.com`) |
| 服务端口 | 是 | 后端服务的端口号，支持 1-65535 的任意端口<br>**注意**：如 SRV 记录已填写端口，控制台填写的端口内容会覆盖 SRV 记录的端口内容 |
| 服务路径 | 是 | 后端服务的路径(如：`/api/v1`)<br>如无路径，填写 `/` |
| 超时时间 | 是 | 网关调用后端的超时时间，默认 60000 毫秒，支持调整 |
| 重试次数 | 是 | 请求失败时的重试次数，默认 5 次，支持调整 |

**步骤3：保存配置**

1. 确认所有参数填写正确；
2. 单击 **确定** 按钮完成创建；
3. 系统自动跳转到服务列表页。

### 结果验证

- 在服务列表页，查看新建的服务记录。
- 服务类型列显示“私有 DNS”。
- 服务地址列显示私有域名和完整地址。
- 在创建 MCP 服务、配置路由时，可从服务下拉列表中选择该服务。

### 注意事项

- 私有域名必须在腾讯云 Private DNS 中已配置解析记录，否则服务调用将返回域名解析失败错误。
- 如 SRV 记录已填写端口，控制台填写的服务端口将覆盖 SRV 记录中的端口。
- 私有 DNS 服务来源必须先创建，才能在新建服务时选择。
- 请确保私有域名的解析记录关联的 VPC 包含网关实例所在的 VPC。
- 建议定期检查 Private DNS 的解析记录是否正确，避免服务调用失败。

## 服务管理操作

## 查看服务列表

在服务列表页，可以查看以下信息：

- **ID/名称**：服务的唯一标识和名称
- **服务类型**：域名/IP 或 私有 DNS
- **服务地址**：完整的后端地址(协议 + 域名/IP + 端口 + 路径)
- **限流**：是否配置了限流策略
- **创建时间**：服务的创建时间
- **修改时间**：服务的最后修改时间
- **操作**：编辑、删除等操作按钮

## 编辑服务

1. 在服务列表页，单击目标服务的**编辑**按钮；
2. 在编辑弹窗中，修改服务参数(服务类型不可修改)；
3. 单击**确定**按钮保存修改。

> **说明：**
> 服务类型(域名/IP / 私有 DNS)一旦创建后不可修改，如需变更请重新创建服务。

## 删除服务

1. 在服务列表页，单击目标服务的**删除**按钮；
2. 在确认弹窗中，单击**确定**按钮。

# 证书管理

最近更新时间：2026-07-23 17:07:32

## 操作场景

本文档介绍如何管理您网关需要使用的 SSL 证书，包括新建、删除、编辑证书等操作。

## 前提条件

- 已经完成域名注册、实名认证、备案等流程，若您尚未完成，可以使用腾讯云提供的域名服务，详情可参见 快速注册域名及实名认证 。
- 已经拥有 SSL 证书，若您暂无证书，可以使用腾讯云提供的证书服务，详情可参见 快速注册域名及实名认证 。

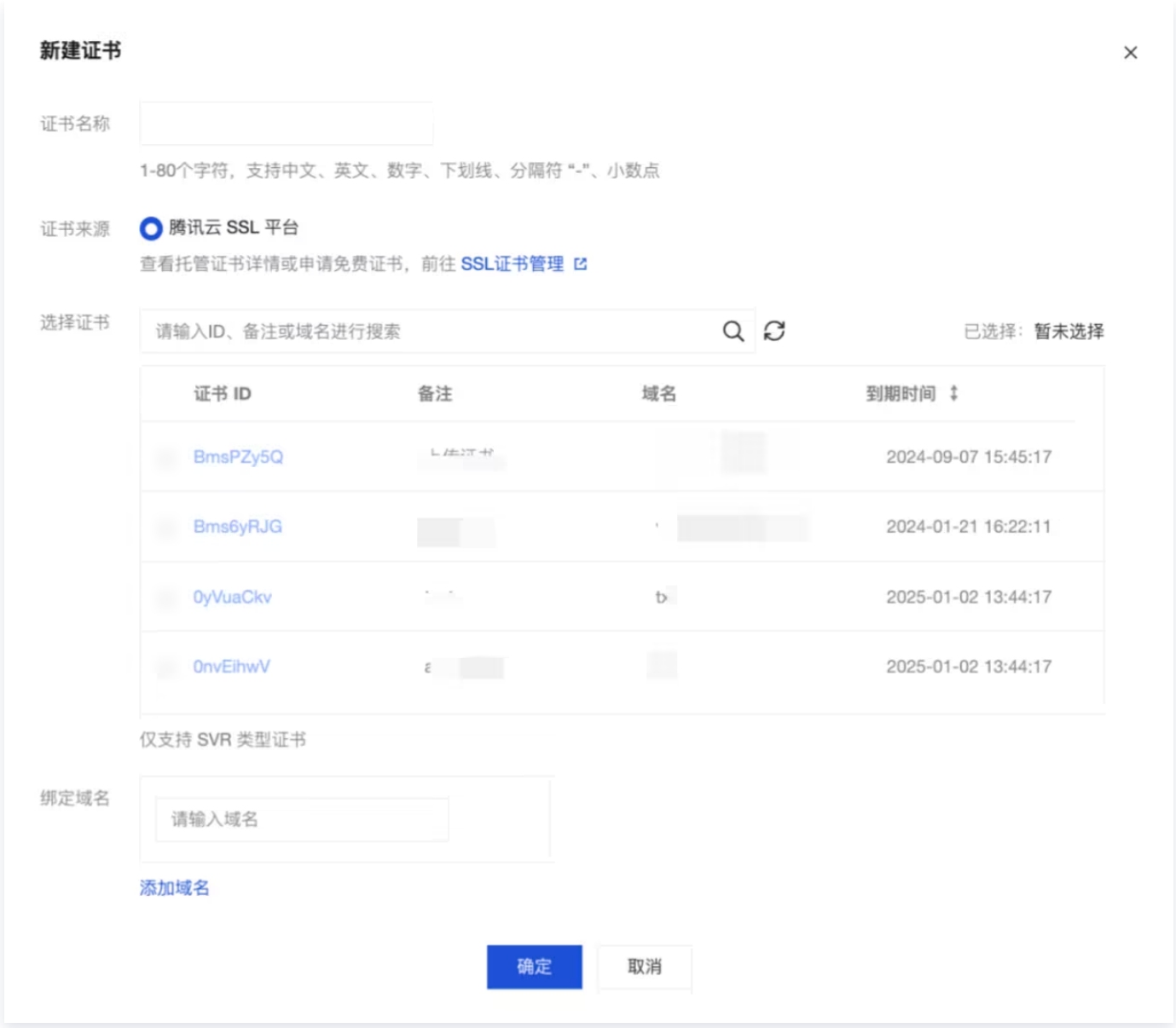
## 新建证书

### 步骤1：配置 DNS 解析

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。
3. 查看**网络配置 > 公网负载均衡**，获取访问 IP。
4. 根据 快速添加域名解析 指引，将您的域名解析至网关公网代理地址中的 IP。

### 步骤2：新建证书

1. 在左侧导航栏选择**证书管理**，单击**新建**。
2. 填写证书名称，选择证书并填写您的域名。**仅支持 SVR 类型证书。**




3. 单击**确定**，完成证书新建，返回证书列表查看创建好的证书。

## 步骤3：验证证书是否生效

访问 `https://域名` 验证是否生效。

## 更新证书

在证书过期前，您可以更新证书。

1. 在证书列表页面，选择您要更新的证书，单击操作栏的**更新证书**。
2. 在弹窗中选择更新的证书，确认更新后的证书信息，单击**确定**，完成更新。

## 编辑证书

您可以修改证书的名称和绑定的域名。

1. 在证书列表页面，选择您要更新的证书，单击操作栏的**编辑**。

2. 在弹窗中输入证书名称和绑定域名，单击**确定**，完成修改。


## 删除证书

> **注意：**
> 删除后，该证书绑定的域名将无法使用 HTTPS 访问网关。

1. 在证书列表页面，选择您要删除的证书，单击操作栏的**删除**。
2. 在弹窗中单击**确定**，完成删除。

# 域名管理

最近更新时间：2026-07-23 17:07:31

本文将快速引导您如何在 AI 网关中，使用自定义的域名访问网关。 AI 网关提供了 IP 访问的方式，并没有提供域名接入点。因此您需要在自己的域名 DNS 添加一个 A 记录，指向 AI 网关提供的 IP。

## 概述

您需要依次完成以下步骤：

1. 在 DNS 中 添加 A 记录，绑定 AI 网关提供的 IP。
2. 绑定证书（如果有使用 HTTPS 访问网关的需要）。

## 操作步骤

### 步骤1：在 DNS 中添加 A 记录，绑定 AI 网关提供的 IP

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 单击目标实例 ID/名称进入实例详情页，进入目标网关实例。
3. 查看**网络配置 > 公网负载均衡**，获取访问 IP。
4. 在 DNS 中添加 A 记录，将域名绑定该公网 IP。

### 步骤2：绑定证书（如果有 HTTPS 访问的需要）

如果您有使用 HTTPS 访问网关资源的需要，请参考 证书管理 步骤绑定证书。

> **注意：**
> 如果没有绑定证书，使用 HTTPS 协议，访问自定义域名时候，会出现证书跟自定义域名不一致的情况。如果您是使用 curl 工具进行测试，加上 `-k` 的参数跳过对服务端证书的校验。在生产的环境，使用 HTTPS 需要绑定证书。

# 密钥管理
# 模型密钥

最近更新时间：2026-07-23 17:07:34

## 操作场景

模型密钥 AI 网关安全调用大模型服务的核心凭证。企业/开发者通过 AI 网关对接多家 AI 厂商的模型 API 时，存在关键密钥场景：**模型密钥**用于存储各 AI 厂商的认证密钥（如厂商 AccessKey、API Secret），网关凭借这些密钥代表用户向厂商发起模型 API 调用。
为确保敏感密钥信息的安全，微服务 TSF AI 网关与腾讯云密钥管理系统（KMS）深度集成，实现密钥的全生命周期加密存储。KMS 使用经过第三方认证的硬件安全模块（HSM）来生成和保护密钥，确保包括腾讯云在内的任何人都无法获取您的明文主密钥，满足严格的合规性要求。通过集中化管理，本功能旨在强化安全管控，杜绝明文泄露和未授权访问风险，同时简化密钥的创建、更新、禁用、删除等运维流程。

## 前提条件

如果生成方式使用密钥管理系统（KMS 凭据）则需要创建凭据，详情请参见凭据管理系统-快速入门 。

## 操作步骤

### 查看密钥列表

1. 登录微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 列表页将展示所有已创建的模型密钥，包括密钥名称、类型、状态和生成方式等信息。您可以在此进行新建、编辑或删除等操作。
5. 当密钥状态为“**已启用**”时，删除操作将置灰不可用，系统会提示“请先禁用该密钥”。

### 新建密钥

1. 登录微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**模型密钥列表页**，单击**新建**。
5. 在“新建密钥”窗口中，配置以下参数：

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |

| | | |
|---|---|---|
| **密钥名称** | 是 | 最长60个字符，支持中英文大小写、数字及分隔符（“-”、“_”），不能以数字和分隔符开头，不能以分隔符结尾 |
| **生成方式** | 是 | **密钥管理系统（KMS 凭据）**：关联腾讯云 KMS 中的凭据。需填写“凭据名称”和“凭据版本”，若无 KMS 凭证，则可以点击“**新建凭证**”跳转新建<br>**自定义**：手动输入密钥值（模型密钥为 API-KEY 值，消费者密钥为凭证内容） |
| **描述** | 否 | 密钥的标识描述信息，最长可输入200个字符。 |

6. 单击**确定**，完成密钥创建。网关会通过集成 KMS 服务（若生成方式选择 KMS 凭据），确保密钥的加密安全存储。

> ⚠ **注意：**
> - 生成方式选择“**密钥管理系统（KMS 凭据）**”，若要管理 KMS 凭据，需跳转至 KMS 控制台 进行操作；
> - 生成方式为“**自定义**”时，不支持修改，但支持复制和查看，默认显示为“***”以保护敏感信息。

## 查看密钥详情

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**模型密钥列表页**，单击目标密钥的“ID/名称”。
5. 进入密钥详情页，您可以查看：
   - **基本信息**：包括密钥名称、类型、状态、创建时间等。
   - **已绑定模型资源**：展示当前密钥已关联的模型服务资源信息。

## 编辑密钥

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**模型密钥列表页**，找到目标密钥，单击其操作列下的**编辑**；或在密钥**详情页**右上角单击**编辑**。
5. 在编辑窗口中，支持修改密钥的**名称**和**描述**（备注）信息。
6. 单击**确定**保存修改。

## 密钥解除绑定关联模型服务

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。

3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 进入密钥管理页面，找到目标密钥。
5. 在"关联模型服务"列表中，单击目标服务右侧的解除关联。
6. 在弹出的确认窗口中，确认待解绑信息，单击确定完成解绑。

## 密钥绑定模型服务

模型服务与密钥为多对多关系，一个模型服务可以绑定多个密钥，一个密钥也可以绑定多个模型服务。您可以为密钥绑定多个模型服务。

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**模型密钥列表页**，单击目标密钥的“ID/名称”，进入**详情页**
5. 点击**添加资源**，在“新增资源”弹窗中，左侧“请选择模型服务”区域会列出所有可选的模型服务，您可以通过搜索框快速查找。
6. 在左侧列表中，勾选一个或多个需要绑定到该密钥的模型服务，被选中的模型服务会出现在右侧“已选择”列表中。
7. 如需移除模型服务，可在右侧“已选择”列表中，点击某个模型服务条目右侧的 × 图标，可将其从本组关联关系中移除。
8. 调整完毕后，点击 **确定** 保存关联关系。

## 启用/禁用密钥

模型密钥只有在启用状态下才可以生效，这意味着当密钥处于关闭或未激活状态时，AI 网关将无法识别或使用该密钥进行任何操作，因此在使用前务必确认密钥是否已正确启用。

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**模型密钥列表**页，找到目标密钥，单击其操作列下**禁用**后，密钥处于“已禁用”状态， AI 网关将无法识别或使用该密钥进行任何操作。
5. **启用操作**需要在目标密钥处于“已禁用”情况下：单击**启用**，密钥将处于“启用”状态。

## 删除密钥

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**模型密钥列表页**，找到目标密钥，单击其操作列下**禁用**后，才可以进行删除操作，等待密钥禁用后，点击**删除**即可。

5. 系统将进行删除前的依赖关系校验：
   - 若密钥已解除所有关联（模型密钥需解除所有模型服务关联），则弹窗直接显示密钥信息，点击**确定**即可删除。
   - 若密钥仍存在关联资源，弹窗会提示“存在未解除的依赖关系”并列出具体依赖项。您需要先解除所有依赖关系后，点击**重新检查**，校验通过后方可删除。

**注意事项：**
- KMS 凭据状态变化：如果您在腾讯云 KMS 控制台中修改了某个凭据，AI 网关为保障业务连续性，会短暂继续使用缓存的旧凭证内容（默认缓存时间约5分钟）。建议您在 KMS 上创建新版本的凭据，并在网关中关联新凭据后，再删除旧的 API Key 版本，以确保变更及时生效。
- 同时为了增加密钥高可用性，建议您为模型服务配置多个凭证，避免某个凭证被禁用后，模型服务不可用导致事故产生。

# 消费者密钥

最近更新时间：2026-07-23 17:07:34

## 操作场景

消费者密钥是客户端应用安全调用 AI 网关服务的核心凭证。当企业或开发者通过微服务 TSF AI 网关整合与调度后端 AI 能力时，消费者密钥为不同的客户端（如业务应用、移动 App 或第三方服务）创建唯一的身份认证凭据，用于这些客户端在调用网关 API 时的身份鉴权与访问控制。
为确保敏感密钥信息的安全，微服务 TSF AI 网关与腾讯云密钥管理系统（KMS）深度集成，实现密钥的全生命周期加密存储。KMS 使用经过第三方认证的硬件安全模块（HSM）来生成和保护密钥，确保包括腾讯云在内的任何人都无法获取您的明文主密钥，满足严格的合规性要求。通过集中化管理，本功能旨在强化安全管控，杜绝明文泄露和未授权访问风险，同时简化密钥的创建、更新、禁用、删除等运维流程。

## 前提条件

如果生成方式使用密钥管理系统（KMS 凭据）则需要创建凭据，详情请参见凭据管理系统-快速入门。

## 操作步骤

### 查看密钥列表

1. 登录微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 列表页将展示所有已创建的消费者密钥，包括密钥名称、类型、状态和生成方式等信息。您可以在此进行新建、编辑或删除等操作。
5. 当密钥状态为“**已启用**”时，删除操作将置灰不可用，系统会提示“请先禁用该密钥”。

### 新建密钥

1. 登录微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**消费者密钥列表页**，单击**新建**。
5. 在“新建密钥”窗口中，配置以下参数：

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| **密钥名称** | 是 | 最长60个字符，支持中英文大小写、数字及分隔符（“-”、“_”），不能以数字和分隔符开头，不能以分隔符结尾 |

| | | |
|---|---|---|
| **生成方式** | 是 | **密钥管理系统（KMS凭据）**：关联腾讯云KMS中的凭据。需填写“凭据名称”和“凭据版本”，若无 KMS 凭证，则可以单击“**新建凭证**”跳转新建 |
| | | **自动生成**：网关自动生成随机的 API key |
| | | **自定义**：手动输入密钥值（消费者密钥为凭证内容） |
| **描述** | 否 | 密钥的标识描述信息，最长可输入200个字符。 |

6. 单击**确定**，完成密钥创建。网关会通过集成KMS服务（若生成方式选择KMS 凭据），确保密钥的加密安全存储。

> **注意：**
> - 生成方式选择“**密钥管理系统（KMS 凭据）**”，若要管理 KMS 凭据，需跳转至 KMS 控制台 进行操作。
> - 生成方式为“**自定义**”和“**自动生成**”时，不支持修改，但支持复制和查看，默认显示为***以保护敏感信息；

## 查看密钥详情

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**消费者密钥列表页**，单击目标密钥的“ID/名称”。
5. 进入密钥详情页，您可以查看：
   - **基本信息**：包括密钥名称、类型、状态、创建时间等。
   - **已绑定消费者**：展示当前密钥已关联的消费者信息。

## 编辑密钥

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**消费者密钥列表页**，找到目标密钥，单击其操作列下的**编辑**；或在密钥**详情页**右上角单击**编辑**。
5. 在编辑窗口中，支持修改密钥的**名称**和**描述**（备注）信息。
6. 单击**确定**保存修改。

## 密钥绑定消费者

消费者与密钥为一对多关系，一个消费者可以绑定多个密钥，一个密钥只可以绑定一个消费者。您可以为密钥绑定一个消费者。

1. 登录 微服务平台控制台 ，在左侧导航栏选择 AI 网关进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**消费者密钥列表页**，单击目标密钥的“ID/名称”，进入**详情页**
5. 单击**添加资源**，在“新增资源”弹窗中，左侧“请选择消费者”区域会列出所有可选的消费者，您可以通过搜索框快速查找。
6. 在左侧列表中，勾选一个需要绑定到该密钥的消费者，被选中的消费者会出现在右侧“已选择”列表中。
7. 如需移除消费者，可在右侧“已选择”列表中，单击某个消费者条目右侧的 × 图标，可将其从本组关联关系中移除。
8. 调整完毕后，单击**确定**保存关联关系。

## 启用/禁用密钥

消费者密钥只有在启用状态下才可以生效，这意味着当密钥处于关闭或未激活状态时，AI 网关将无法识别或使用该密钥进行任何操作，因此在使用前务必确认密钥是否已正确启用。

1. 登录 微服务平台控制台 ，在左侧导航栏选择 AI 网关进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**消费者密钥列表**页，找到目标密钥，单击其操作列下的**禁用**按钮后，密钥处于“已禁用”状态， AI 网关将无法识别或使用该密钥进行任何操作。
5. **启用操作**需要在目标密钥处于“已禁用”情况下：单击**启用**，密钥将处于“启用”状态。

## 删除密钥

1. 登录 微服务平台控制台 ，在左侧导航栏选择 AI 网关进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**密钥管理**，进入密钥列表页。
4. 在**消费者密钥列表页**，找到目标密钥，单击其操作列下的**禁用**按钮后，才可以进行删除操作，等待密钥禁用后，单击**删除**即可。
5. 系统将进行删除前的依赖关系校验：
   - 若密钥已解除所有关联（消费者密钥需解除所有消费者关联），则弹窗直接显示密钥信息，单击**确定**即可删除。
   - 若密钥仍存在关联资源，弹窗会提示“存在未解除的依赖关系”并列出具体依赖项。您需要先解除所有依赖关系后，单击**重新检查**，校验通过后方可删除。

> ⚠ **注意事项：**
> - KMS 凭据状态变化：如果您在腾讯云 KMS 控制台中修改了某个凭据，AI 网关为保障业务连续性，会短暂继续使用缓存的旧凭证内容（默认缓存时间约5分钟）。建议您在 KMS 上创建新版本的凭据，并

在网关中关联新凭据后，再删除旧的 API Key 版本，以确保变更及时生效。

- 同时为了增加密钥的高可用性，建议您配置多个凭证，避免某个凭证被禁用后，消费者不可用导致事故发生。

# 消费者管理
# 消费者组

最近更新时间：2026-07-23 17:07:35

## 操作场景

消费者组用于对消费者（即 API 调用方）进行逻辑分组，通常对应于不同的业务团队、项目或应用。通过对消费者组进行授权，可以批量管理其下所有消费者对模型 API 的访问权限。本文介绍如何创建、查看、编辑、删除消费者组，以及管理消费者组与消费者之间的关联关系。

## 操作步骤

### 创建消费者组

1. 登录微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **消费者管理 > 消费者组**，进入消费者组列表页。
4. 在“新建消费者组”弹窗中，填写以下信息。

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 消费者组名称 | 是 | 输入消费者组名称。最长60个字符，支持中英文大小写、数字及分隔符（“-”、“_”），不能以数字和分隔符开头，不能以分隔符结尾。 |
| 状态 | 是 | 控制此消费者组是否生效。默认为“开启”状态。关闭后，组内所有消费者将无法通过网关访问任何API。 |
| 消费者 | 否 | 可从列表中选择一个或多个已存在的消费者，将其直接加入本组。创建后可随时增删。 |
| 描述 | 否 | 该消费者组的描述信息，便于后续管理。最长200个字符。 |

5. 单击 **确定**，完成消费者组的创建。

### 查看消费者组详情

1. 登录微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **消费者管理 > 消费者组**，进入消费者组列表页。
4. 在 **消费者组** 列表页面，单击任意一个消费者组的“ID/名称”，页面右侧将展开该消费者组的详情面板。

5. 详情面板分为上下两部分：
   - 消费者组信息：展示此消费者组的基础信息，包括消费者组ID、名称、状态、创建时间、修改时间和描述。
   - 消费者：以列表形式展示已关联到此消费者组的所有消费者，包括其ID/名称和描述。您可以在此对已关联的消费者执行“解除关联”操作。

## 编辑消费者组

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**消费者管理 > 消费者组**，进入消费者组列表页。
4. 在消费者组列表页面，找到目标消费者组，单击其右上角的编辑。
5. 在“编辑消费者组”弹窗中，您可以修改该组的名称、状态和描述。
6. 修改完成后，单击**确定**保存。

> ⚠ **注意：**
> 编辑操作不能直接修改组内关联的消费者。如需管理关联消费者，请参见下文 “管理关联消费者” 操作。

## 管理关联的消费者

消费者与消费者组为多对多关系，一个消费者可以属于多个组，一个组也可以包含多个消费者。您可以为消费者组添加或移除消费者。

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**消费者管理 > 消费者组**，进入消费者组列表页。
4. 在消费者组列表页面，找到目标消费者组，单击其操作列下的**关联消费者**。
5. 在“关联消费者”弹窗中，左侧“请选择消费者”区域会列出所有可选的消费者列表，您可以通过搜索框快速查找。
6. 在左侧列表中，勾选一个或多个需要添加到本组的消费者，被选中的消费者会出现在右侧“已选择”列表中。
7. 如需移除消费者，可在右侧“已选择”列表中，单击某个消费者条目右侧的 × 图标，可将其从本组关联关系中移除。
8. 调整完毕后，单击**确定**保存关联关系。

## 删除消费者组

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **消费者管理 > 消费者组**，进入消费者组列表页。
4. 在 **消费者组** 列表页面，找到目标消费者组，单击其右上角的 **删除**，系统将进行删除前的依赖关系校验。
5. 系统会弹窗提示您确认删除，并自动检查该消费者组是否存在被其他资源（如“模型API”授权）绑定的情况。

- 若无依赖：弹窗将提示“可安全删除”，单击 确定 即可删除。
- 若存在依赖：弹窗会显示“资源删除依赖关系检查结果: 存在未解除的依赖关系”，并列出具体的依赖项。

6. 若存在依赖，您需要先行解除所有列出的依赖关系。解除依赖后，可单击弹窗内的 **重新检查** 链接，系统将再次进行校验。
7. 当校验通过，提示变为“可安全删除”后，单击 **确定** 即可删除。删除后，该组下的所有消费者将立即失去通过本组获取的模型API访问权限。 若需放弃删除，可单击 **取消**。

# 消费者

最近更新时间：2026-07-23 17:07:35

## 操作场景

消费者代表最终的 API 调用方，例如一个独立的应用程序、服务或子系统。每个消费者需要配置其身份凭证（如 API Key），用于在访问网关时进行认证。消费者必须加入至少一个消费者组，通过该组获得模型 API 的访问授权。本文介绍如何创建、查看、编辑、删除消费者，以及管理消费者的访问凭证。

## 操作步骤

### 创建消费者并添加模型 API

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **消费者管理 > 消费者**，进入消费者列表页。
4. 在“新建消费者”弹窗中，填写以下信息。

| 参数 | 是否必填 | 说明 |
| --- | --- | --- |
| 消费者名称 | 是 | 输入消费者名称。最长60个字符，支持中英文大小写、数字及分隔符（“-”、“_”），不能以数字和分隔符开头，不能以分隔符结尾。 |
| 所属消费者组 | 否 | 从下拉列表中选择一个或多个已有的消费者组。消费者必须属于至少一个组，以获得 API 访问权限。 |
| 选择密钥 | 否 | 从下拉列表中选择一个或多个已创建的密钥，作为此消费者的身份凭证。您也可以通过“新建密钥”链接跳转至密钥管理页面快速创建。 |
| 描述 | 否 | 该消费者的描述信息，便于后续管理。最长200个字符。 |

5. 单击**确定**，完成消费者的创建。系统会为消费者生成一个唯一的消费者 ID。

### 查看消费者详情与管理凭证

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **消费者管理 > 消费者**，进入消费者列表页。
4. 在 **消费者** 列表页面，单击任意一个消费者的“ID/名称”，页面右侧将展开该消费者的详情面板。详情面板分为两部分：
   - 基础信息：展示此消费者的 ID、名称、所属消费者组、创建时间、修改时间和描述。

- 凭证管理：
  - 认证方式：显示当前使用的认证方式，如“API Key”。
  - 管理凭证：此处列表展示该消费者已绑定的所有密钥（凭证）。您可以通过上方的**添加凭证**按钮，为消费者绑定更多密钥。
  - 添加凭证操作：单击“添加凭证”后，在弹窗中选择认证方式和具体的密钥，单击**确定**即可完成绑定。
  - 移除凭证操作：单击“解除关联”后，在弹窗中输入密钥名称确认解除关联，单击**确定**即可完成移除凭证操作。

## 编辑消费者

1. 登录微服务平台控制台，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**消费者管理 > 消费者**，进入消费者列表页。
4. 在消费者列表页面，找到目标消费者，单击其右上角的编辑。
5. 在“编辑消费者”弹窗中，您可以修改该消费者的名称和描述。
6. 修改完成后，单击**确定**保存。

> **注意：**
> 编辑操作不能直接修改消费者的“所属消费者组”和已绑定的密钥。如需修改，请在消费者详情页的“基础信息”区域查看所属组，在“凭证管理”区域管理密钥。

## 删除消费者

1. 登录微服务平台控制台，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**消费者管理 > 消费者**，进入消费者列表页。
4. 在“消费者”列表页面，找到目标消费者，单击其右上角的删除，系统将进行删除前的依赖关系校验。
5. 系统会弹窗提示您确认删除，并自动检查该消费者是否存在被其他资源绑定的情况。
   - 若无依赖：弹窗将提示“可安全删除”，单击**确定**即可删除。
   - 若存在依赖：弹窗会显示“存在未解除的依赖关系”，并列出依赖项。
6. 若存在依赖，您需要先行解除所有列出的依赖关系。解除依赖后，可单击弹窗内的**重新检查**链接，系统将再次进行校验。
7. 当校验通过，依赖提示消失后，单击**确定**即可最终删除该消费者。若需放弃删除，可单击**取消**。

# 配额管理
# 消费者配额管理

最近更新时间：2026-07-23 17:07:35

消费者配额管理用于为 AI 网关中的消费者设置调用配额，包括 Token 用量配额和成本预算配额。通过配额管理，您可以控制消费者的模型调用量，避免资源滥用，保障业务连续性。

## 功能概述

消费者配额管理支持以下核心能力：

- **Token 用量配额**：限制消费者在指定时间周期内的 Token 消耗总量（输入 Token + 输出 Token）。
- **成本预算配额**：限制消费者在指定时间周期内的调用成本总额。
- **配额周期**：支持按日、按周、按月设置配额周期。
- **超配策略**：支持配置配额超限后的处理策略（拒绝请求 / 降级路由）。
- **配额状态**：实时展示各消费者的配额使用情况（正常 / 预警 / 超限）。

## 前置条件

- 已创建 AI 网关实例，且实例版本 ⩾ 3.9.4。
- 已创建消费者（Consumer）。
- 如需使用成本预算配额，需先配置模型单价（参见 模型单价配置 文档）。

## 操作详情

### 进入消费者配额管理页面

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏选择 **配额管理**。

配额管理页面展示该网关实例的配额配置列表和当前配额使用情况。

### 添加消费者配额规则

您可以选择消费者来添加配额规则。

1. 在配额管理页面，单击**新建配额规则**。
2. 在弹出的对话框中，配置以下参数：

| 参数 | 必填 | 说明 |
| --- | --- | --- |
| 消费者 | 是 | 选择当前实例的消费者，如果需要新建，可单击新建消费者 |

| | | |
|---|---|---|
| 配额类型 | 是 | 选择 Token 用量或请求次数 |
| 统计周期 | 是 | 配额统计周期：按日 / 按周 / 按月 |
| 限额 | 是 | Token 用量配额：输入 Token 用量上限（单位：Token）请求次数配额：输入请求次数上限（单位：次） |
| 启用状态 | 开关 | 当配额使用率达到指定百分比时触发预警，取值范围：50% ~ 95% |

3. 配额类型说明：
   - 配额限制可以按日、按周、按月统计
   - 超出配额后，系统将拒绝新的请求
   - 可以为同一对象设置多种配额类型（Token、请求次数）
   - 配额使用情况每小时更新一次
4. 单击 **确定**，完成配额配置添加。

> ⚠ **注意：**
> - 一个消费者可以同时配置多个配额（如同时配置 Token 用量配额和成本预算配额）。
> - 当多个配额同时超限时，以先触发的超配策略为准。
> - 降级路由的目标模型服务必须已创建且状态正常。

## 查看配额使用情况

配额配置完成后，系统会实时统计消费者的配额使用情况。
在配额管理页面，配额列表展示以下信息：

| 字段 | 说明 |
|---|---|
| 资源 ID/名称 | 配额规则对应的资源 ID/名称 |
| 配额类型 | Token 数量配额 / 请求次数配额 |
| 统计周期 | 按日 / 按周 / 按月 |
| 配额上限 | 配置的配额上限值 |
| 使用量 | 当前周期内已使用的配额量 |
| 使用率 | 当前使用量 / 配额上限 × 100% |
| 启用状态 | 当前配额规则情况 |
| 创建/修改时间 | 最新的创建和修改时间 |

| 操作 | 编辑 / 删除 |
| --- | --- |

## 编辑消费者配额

您可以根据需要修改已有配额配置。

1. 在配额管理页面，找到目标配额配置。
2. 单击操作列的**编辑**。
3. 修改需要变更的参数（消费者不可修改）。
4. 单击**确定**保存修改。

> **说明：**
> 编辑配额后，当前周期的配额使用量不会重置，将继续累计。

## 删除消费者配额

当某个配额不再需要时，可删除配额配置。

1. 在配额管理页面，找到目标配额配置。
2. 单击操作列的**删除**。
3. 在确认对话框中单击**确定**。

## 配额管理 Redis 配置

您在使用配额管理前，需要手动配置 Redis，以保存当前配额规则信息，若未配置 Redis，则**无法展示配额数据**。

1. 在配额管理页面，单击 Redis 配置 进入 Redis 配置页面。
2. 在 Redis 配置页，单击编辑 Redis 配置，填写 Redis 信息，根据您的 Redis 服务地址等信息完成填写。
3. 单击保存后完成 Redis 配置，可以开始配额规则配置

## Redis 配置编辑

您在使用配额管理时，若需要修改 Redis 配置，可以单击编辑 Redis 配置按钮进入 Redis 编辑页面：

1. 在配额管理页面，单击编辑 Redis 配置，进入 Redis 编辑页面。
2. 在 Redis 编辑页，修改 Redis 服务信息。
3. 单击保存后完成修改 Redis 配置，可以继续完成新配额规则配置

## 常见问题

### Q1：配额周期如何计算?

- **按日**：从当日 00:00:00 到 23:59:59。
- **按周**：从周一 00:00:00 到周日 23:59:59。
- **按月**：从当月第1日 00:00:00 到当月最后1日 23:59:59。

## Q2：一个消费者可以配置多个配额吗?

可以。一个消费者可以同时配置多个配额（如同时配置 Token 用量配额 和 成本预算配额），也可以配置多个同类型的配额（如两个 Token 用量配额，分别按日和按月统计）。

# 成本管理
# 模型单价配置

最近更新时间：2026-07-24 16:45:01

## 功能说明

模型单价配置是 AI 网关成本管理的基础数据模块，用于配置模型调用的计费单价。配置完成后，AI 网关在每次大模型调用结束时，会按照您配置的输入、缓存命中输入、输出三段单价，结合本次请求实际消耗的 Token 数实时计算成本，并驱动后续 **Token 消耗统计**与**成本分析报表**的成本数据呈现。
支持 10 家供应商官方最新价格自动同步，可选展示官方现价，同步周期为 24 小时，支持批量使用官方现价更新模型单价。

## 页面入口

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 单击左侧菜单成本管理模块中的**模型单价配置**。

## 配置列表字段说明

| 字段 | 说明 | |
|---|---|---|
| 模型供应商 | 模型所属供应商，支持选择标准供应商（如 Hunyuan、Google-Gemini 等）、MaaS 供应商（如 TokenHub、AWS Bedrock 等）以及自定义供应商（需选择或手动输入自定义供应商名称） | |
| 模型名称 | 模型名称，如 hunyuan-t1、gemini-2.5-flash | |
| 定价类型 | 固定价 | • 输入单价：输入 Token 的单价，格式如 ￥1.0000/1M Tokens<br>• 输出单价：输出 Token 的单价，格式如 ￥4.0000/1M Tokens<br>• 缓存输入单价：缓存命中输入单价，未配置时显示 – |
| | 阶梯价 | • 最小 Token（含）：该档包含的最小 Token 数<br>• 最大 Token（含）：该档包含的最大 Token 数（最后一档固定为“无上限”）<br>• 输入单价：输入 Token 的单价，格式如 ￥1.0000/1M Tokens |

| | | |
|---|---|---|
| | | • 输出单价：输出 Token 的单价，格式如 ￥4.0000/1M Tokens<br>• 缓存输入单价：缓存命中输入单价，未配置时显示 –<br>• 至少配置 2 个阶梯，点击 **添加阶梯** 最多可配置 5 档阶梯价格<br>• 阶梯区间必须连续且不重叠 |
| 计量单位 | 可选1M Tokens/1K Tokens | |
| 货币 | 货币单位，固定为 CNY | |

# 操作步骤

## 新增模型单价配置

1. 在模型单价配置页面，单击右上角**新增配置**，弹出**新增模型单价配置**对话框。
2. 配置以下参数：

| 参数 | 必填 | 说明 |
|---|---|---|
| 模型供应商 | 是 | 下拉选择，选项包括系统预置的 OpenAI、DeepSeek、Qwen、Google-Gemini、Hunyuan 等主流供应商，以及用户在“模型管理”中创建的自定义供应商 |
| 模型名称 | 是 | 下拉选择，选项根据所选供应商动态加载 |
| 定价类型 | 是 | 点选，固定价/阶梯价 |
| 计量单位 | 是 | 当前为 1M Tokens（每百万 Token） |
| 货币 | 是 | 固定为 CNY（人民币） |
| 输入单价 | 是 | 输入 Token 的单价，单位为 元/1M Tokens，建议保留 4 位小数 |
| 输出单价 | 是 | 输出 Token 的单价，单位为 元/1M Tokens，建议保留 4 位小数 |
| 缓存输入单价 | 否 | 缓存命中输入单价，单位为 元/1M Tokens；不填时成本计算自动按**输入单价**兜底 |
| 最小 Token（含） | 是（仅阶梯价） | 该档包含的最小 Token 数 |
| 最大 Token（含） | 是（仅阶梯价） | 该档包含的最大 Token 数（最后一档固定为“无上限”） |

3. 单击**确定**，新增的配置会出现在列表中。

## 批量导入模型单价配置

1. 单击**批量导入**，若首次进行此操作需在弹窗中先单击**下载导入模板**，完成填写后进行上传。
2. 模板包含表头和示例数据，模板需填字段与新增单个模型单价配置一致，可在模板中查看具体填写示例。
3. 单击**开始上传**。
4. 单击**解析文件**，将根据所填写模板解析出对应的模型单价配置，并以列表形式展示供确认。
5. 单击**确认导入**，导入成功后配置列表将自动新增以上批量配置。若有失败项，可单击**查看详情**查看批量导入任务详情。


## 编辑模型单价配置

1. 在配置列表中，找到目标配置。
2. 单击操作列的**编辑**，弹出编辑对话框。
3. 单击**确定**，保存修改。

## 导出模型单价配置

单击列表上方的下载按钮，可一键导出当前所有模型单价配置。


## 删除模型单价配置

1. 在配置列表中，找到目标配置。
2. 单击操作列的**删除**。
3. 在弹出的确认对话框中查看提示：**“确定删除模型单价配置？删除后将无法恢复，请确认操作”**。
4. 单击**确定**完成删除。

> **注意：**
> 删除后该模型的成本将无法计算，相关的 Token 消耗统计与成本分析报表中，该模型对应的**成本**列会显示为 `-`（与 ￥0.00 不同，表示“未配置单价”而非“成本为零”）。

## 注意事项

- **未配置单价的影响**：只要某条请求所对应的“模型供应商 + 模型名称”未配置单价，该请求的成本字段为 null，在 Token 消耗统计与成本分析报表中显示为 `-`。
- **自定义模型**：私有化模型或非主流供应商，请先在“模型管理”中创建对应的模型服务，并保证模型服务返回的 provider 与 model 与单价配置完全一致。
- **货币与计量单位**：当前货币固定为 CNY，计量单位固定为 1M Tokens，暂不支持 USD 或 1K Tokens。
- **定价类型**：当前仅支持固定价，阶梯价等其他定价类型在后续版本规划中。

## 常见问题

### Q1：为什么我的成本分析报表里某些行成本显示为 `-`？

A：说明该“模型供应商 + 模型名称”未配置单价。请前往“模型单价配置”页面新增对应单价后，新发生的请求会立即开始计算成本。`-` 表示未配置单价，与 ￥0.00（实际成本为 0）含义不同。

### Q2：缓存输入单价不填会怎样？

A：成本计算时缓存命中部分会自动按 **输入单价** 计算。不影响计费的整体准确性，但您将无法享受到模型供应商提供的缓存优惠价（如 DeepSeek、Anthropic 等供应商对缓存命中部分有显著降价）。建议查阅供应商官方定价后准确配置。

### Q3：同一模型可以配置多个单价吗？

A：不可以。同一网关实例下“模型供应商 + 模型名称”组合具有唯一性，只能存在一条单价配置。如需调整价格，请使用**编辑**操作。

### Q4：修改单价后，历史调用的成本会重算吗？

A：不会。单价修改仅对新发生的请求生效。如果对历史成本数据有重算需求，请在备注中记录调价时点，在导出 CSV 后离线计算。

### Q5：模型供应商和模型名称下拉框为什么没有我要的选项？

A：下拉选项基于当前网关实例下**模型管理** > **模型服务**中已创建的模型服务的 provider/model 自动收敛而来。请先在模型管理中创建对应模型服务，单价配置页面即可看到该模型选项。

# 成本分析与用量统计

最近更新时间：2026-07-23 17:07:36

成本分析与用量统计是 AI 网关成本管理的核心功能模块，包含 **Token 消耗统计** 和 **成本分析报表** 两大子功能，支持多维度查看 AI 网关的调用用量和成本明细。

## Token 消耗统计

Token 消耗统计提供 AI 网关所有 AI 请求的 Token 消耗数据，支持按不同维度查看和筛选，并支持导出为 CSV 文件。

### 功能概述

- **6 个指标卡**：输入 Token 数、输出 Token 数、缓存命中 Token 数、请求总数、最高消费者、总费用
- **数据导出**：支持 CSV 格式导出

### 操作步骤

#### 进入 Token 消耗统计页面

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏选择**成本管理**。
4. 在成本管理页面，单击 **Token 消耗统计**标签页。

#### 查看指标卡

页面顶部展示统计指标卡：

| 指标 | 说明 |
| --- | --- |
| **输入 Token** | 输入 Token 的总量 |
| **输出 Token** | 输出 Token 的总量 |
| **缓存命中 Token** | 缓存命中的输入 Token 总量 |
| **总请求次数** | AI 调用的总请求次数 |
| **最高消费者** | 消耗量最高的消费者 |
| **总费用** | 使用 Token 的总费用 |

#### 导出数据

1. 单击列表右上角的 **导出 CSV** 按钮。
2. 系统将生成当前筛选条件下的 Token 消耗数据 CSV 文件并自动下载。

# 成本分析报表

成本分析报表基于模型单价配置和 Token 消耗数据，自动计算各维度的调用成本，支持多维度成本明细查看、成本趋势分析和数据导出。

## 功能概述

- **5 个维度 Tab**：消费者、消费者组、模型服务、模型 API、路由
- **成本明细表**：展示各维度下的 Token 消耗和成本金额明细
- **成本趋势图**：按时间维度展示成本变化趋势
- **统计卡片**：展示总成本和各项汇总数据
- **数据导出**：支持 CSV 格式导出

## 前置条件

- 已配置模型单价（否则成本金额显示为 0）。
- 网关实例已有 AI 调用记录。

## 操作步骤

### 步骤1：进入成本分析报表页面

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏选择**成本管理**。
4. 在成本管理页面，单击**成本分析报表**标签页。

### 步骤2：选择分析时间

在页面顶部 Tab 栏，选择需要分析的时间维度

### 查看统计卡片

页面顶部展示统计汇总卡片：

- **总成本**：选定时间范围内的总调用成本
- **各维度成本占比**：各维度在总成本中的占比分布

### 查看成本趋势图

统计卡片下方展示成本趋势图，可按月粒度切换查看，直观了解成本变化趋势。

### 成本明细表

趋势图下方为成本明细表，展示各维度下的详细数据
支持按时间范围筛选和按关键词搜索。

### 导出数据

1. 单击列表右上角的**导出 CSV** 按钮。
2. 系统将生成当前维度下的成本分析数据 CSV 文件并自动下载。

> ⚠ **注意事项：**
> - 成本金额基于已配置的模型单价计算，未配置单价的模型成本显示为 0。
> - 导出数据范围与当前页面筛选条件一致。

## 常见问题

### Q1：为什么成本金额显示为 0？

请检查以下原因：

- 是否已为该模型配置了单价（在“模型单价配置”页面配置）。
- 网关实例是否已有 AI 调用记录。
- 成本管理功能开关是否已开启。

### Q2：Token 消耗统计和成本分析报表有什么区别？

- **Token 消耗统计**：侧重于展示 Token 用量的原始数据，不依赖单价配置。
- **成本分析报表**：基于 Token 消耗数据和模型单价，计算实际的调用成本。

### Q3：数据多久更新一次？

Token 消耗和成本数据需要约 30s～1min 同步，等待数据更新后即可在页面查看。

# 数据观测
# 查看请求监控

最近更新时间：2026-07-24 16:45:00

## 操作场景

AI 网关对运行的网关实例提供了多维度的监控指标，用以全面监测实例运行状况与 AI 调用质量。监控指标涵盖通用网关性能指标（如请求数、时延、错误码）以及大模型场景特有的 LLM 指标（如 Token 消耗、模型响应耗时）。
您可以根据这些指标实时了解网关实例及各个模型 API 的运行状况，洞察AI调用成本与性能，针对可能存在的风险及时处理，保障 AI 服务的稳定性与成本可控性。本文为您介绍通过 TSF 控制台查看网关请求监控的操作。

## 操作步骤

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**数据观测**。
4. 单击页面顶部**请求监控**标签。

## 支持监控指标及含义

### 实例/节点

此部分指标适用于所有经过网关的流量，用于评估网关及后端服务的通用性能与健康状况。

| 指标名 | 指标含义 |
| --- | --- |
| 总请求数 | 总请求数。按照所选择的时间粒度统计求和。 |
| 请求平均时延 | 请求平均时延。按照所选择的时间粒度统计求平均值。 |
| 请求最大时延 | 请求最大时延。按照所选择的时间粒度统计求最大值。 |
| 网关直接返回的请求数 | 网关未转发到后端，直接返回响应的请求量（如鉴权失败、触发限流时）。按照所选择的时间粒度统计求和。 |
| 网关平均时延 | 网关自身处理请求的平均耗时。 |
| 网关最大时延 | 网关自身处理请求的最大耗时。 |
| 2xx请求数 | 客户端发送到 AI 网关，请求成功的次数（如 200 OK），按照所选择的时间粒度统计求和。 |

| | |
|---|---|
| 3xx请求数 | 客户端发送到 AI 网关，请求重定向的次数，按照所选择的时间粒度统计求和。 |
| 4xx请求数 | 客户端发送到 AI 网关的是非法请求，如鉴权不通过或者超过限流值的错误个数，网关直接返回的客户端错误的个数（如 401 鉴权失败、403 权限不足、429 限流）。按照所选择的时间粒度统计求和。 |
| 5xx请求数 | AI 网关将消息转发到后端服务，后端服务返回的服务端错误的个数（如 500 后端异常、502 后端无效响应、504 后端不可达）。按照所选择的时间粒度统计求和。 |
| 403请求数 | 权限错误，后端有能力处理该请求，但是拒绝授权访问 |
| 404请求数 | 请求后端服务失败，请求所希望的资源未在后端服务器上发现，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 429请求数 | 请求后端服务失败，请求被限流，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 499请求数 | 请求后端服务失败，客户端在后端响应前主动断开连接，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 502请求数 | 网关尝试执行后端请求时，从后端服务器接收到无效的响应（通常连接服务失败），此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 504请求数 | 网关尝试执行后端请求时，后端机器不可达，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 网关转到后端的请求数 | 网关成功转发到后端服务的请求量。按照所选择的时间粒度统计求和。 |
| 后端平均时延 | 后端服务处理请求的平均耗时。按照所选择的时间粒度统计求平均值。 |
| 后端最大时延 | 后端服务处理请求的最大耗时。按照所选择的时间粒度统计求最大值。 |
| 后端2xx请求数 | 后端服务请求成功的次数（如 200 OK），按照所选择的时间粒度统计求和。 |
| 后端3xx请求数 | 后端服务请求重定向的次数，按照所选择的时间粒度统计求和。 |
| 后端4xx请求数 | 后端服务是非法请求的次数。按照所选择的时间粒度统计求和。 |
| 后端5xx请求数 | 后端服务返回的服务端错误的个数（如 500 后端异常、502 后端无效响应、504 后端不可达）。按照所选择的时间粒度统计求和。 |
| 后端404请求数 | 后端服务资源未在后端服务器上发现，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 后端429请求数 | 后端服务请求失败，请求被限流，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |

| | |
|---|---|
| 后端499请求数 | 后端服务请求失败，客户端在后端响应前主动断开连接，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 后端502请求数 | 后端服务请求失败，后端服务接收到无效的响应，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |
| 后端504请求数 | 后端服务请求失败，后端机器不可达，此类错误的个数的统计，按照所选择的时间粒度统计求和。 |

## 模型 API

**此部分指标专门用于监控大模型调用场景，帮助您分析 Token 消耗成本与模型提供商性能。**

| 指标名 | 指标含义 |
|---|---|
| LLM HTTP 请求次数 | 网关向大模型供应商发起的 HTTP 调用次数。此指标直接反映模型 API 的调用频次。 |
| LLM 总消耗 Token 数 | 网关消耗大模型供应商的总 Token 数，即输入（Prompt）与输出（Completion）的实际消耗 Token 数之和。用于评估消耗 Token 总数据吞吐量。 |
| LLM prompt 消耗 token 数 | 大模型在处理请求时，模型实际所消耗的输入（Prompt）部分的总 Token 数。 |
| LLM completion 消耗 token 数 | 大模型在生成回复时，模型实际所消耗的输出（Completion）部分的总 Token 数。此指标是评估模型调用成本的核心依据之一。 |
| LLM 模型每次请求平均耗时(ms) | 从网关发送请求到模型提供商，到收到其完整响应的平均耗时。此指标反映模型提供商的端到端响应性能。 |
| LLM 提供商平均每 token 耗时(ms) | 模型提供商平均消耗每个 Token 所花费的时间，此指标反映模型提供商 Token 的消耗速度。 |

## MCP

**此部分指标用于监控 MCP 服务的运行状况。通过该页面，您可以查看各 MCP 服务的调用情况，包括请求次数、平均耗时和请求成功率，以便快速发现异常并进行容量评估。**

| 指标名 | 指标含义 |
|---|---|
| MCP 请求次数 | MCP 服务在所选时间范围内接收到的请求总数。此指标直接反映 MCP Server 的调用频次。 |
| MCP 请求平均耗时(ms) | MCP 服务处理请求的平均耗时（毫秒）。此指标反映 MCP Server 的性能表现。 |

| | |
|---|---|
| MCP 请求成功率(%) | MCP 服务调用成功的请求数占总请求数的百分比。此指标反映 MCP Server 的稳定性。 |

# 查看系统监控

最近更新时间：2026-07-24 16:45:00

## 操作场景

AI 网关对运行的网关实例及节点、公网负载均衡提供全维度系统级监控指标，覆盖 CPU、内存、带宽、连接数等核心资源维度，助力您全面掌握实例的资源利用率、流量波动与运行健康度。
您可通过这些指标实时洞察实例的资源瓶颈、流量峰值与连接压力，及时发现过载、异常流量等风险，针对性优化资源配置（如扩容、限流）或排查故障，保障 AI 服务的稳定性与资源成本可控性。本文为您介绍通过 TSF 控制台查看网关系统监控的操作。

## 操作步骤

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击**数据观测**。
4. 单击页面顶部**系统监控**标签。

## 支持监控指标及含义

此部分指标适用于所有经过网关的流量，用于评估网关及后端服务的通用性能与健康状况。

### 实例/节点监控指标

| 指标名 | 指标含义 |
| --- | --- |
| CPU 使用率 | AI 网关的 CPU 使用率，按照所选择的时间粒度统计求平均值。 |
| 内存使用率 | AI 网关的内存使用率，按照所选择的时间粒度统计求平均值。 |
| 入口带宽流量 | AI 网关的入口带宽流量，按照所选择的时间粒度统计求平均值。 |
| 出口带宽流量 | AI 网关的出口带宽流量，按照所选择的时间粒度统计求平均值。 |
| TCP 入连接数 | AI 网关的 TCP 连接数，按照所选择的时间粒度统计求平均值。 |
| 最大内存使用率 | AI 网关在所选时间粒度内的内存使用率最大值。用于观测内存使用峰值，判断是否存在内存突增风险（如内存泄漏、突发流量压力）。 |
| 最大 CPU 使用率 | AI 网关在所选时间粒度内的 CPU 使用率最大值。用于发现 CPU 负载峰值波动，定位计算密集型操作（如复杂鉴权、协议转换）导致的性能突增。 |
| 运行节点数 | 所选时间粒度内，AI 网关中正常运行的节点数量。反映部署规模与可用节点状态，节点数异常减少可能代表故障或伸缩操作。 |

| | |
|---|---|
| 客户端到网关进程的新建连接数 | 所选时间粒度内，客户端与网关进程之间新建立的 TCP 连接数量。观测短时间内连接建立频率，判断客户端连接活跃度。 |
| 客户端到网关进程的活跃连接数 | 所选时间粒度内，客户端与网关进程之间处于活跃通信状态的 TCP 连接数量。反映网关当前承载的有效连接负载。 |
| 客户端到网关进程的非活跃连接数 | 所选时间粒度内，客户端与网关进程之间建立但无活跃通信的 TCP 连接数量。辅助判断连接资源闲置情况，过多可能意味着连接回收 / 管理机制需优化。 |
| 客户端到网关进程的并发连接数 | 所选时间粒度内，客户端与网关进程之间同时存在的 TCP 连接总数（含活跃、非活跃）。直接反映网关的连接并发压力，是评估网关连接容量的关键指标。 |
| 客户端到网关进程的入流量 | 所选时间粒度内，从客户端发送到网关进程的总数据量。 |
| 网关进程到客户端的出流量 | 所选时间粒度内，从网关进程发送到客户端的总数据量。 |
| 客户端到网关进程的入带宽 | 所选时间粒度内，客户端到网关进程的平均带宽占用量（单位时间内的流量传输速率）。评估客户端到网关的带宽压力，避免带宽瓶颈导致连接 / 传输延迟。 |
| 网关进程到客户端的出带宽 | 所选时间粒度内，网关进程到客户端的平均带宽占用量（单位时间内的流量传输速率）。与 “入带宽” 结合，分析网关对外带宽负载，防止带宽瓶颈影响响应传输。 |

## 公网负载均衡监控指标

### 1. 客户端到 LB 的监控

| 指标名 | 指标含义 |
|---|---|
| 入流量 | 在统计粒度内，客户端流入到负载均衡的流量。 |
| 出流量 | 在统计粒度内，负载均衡流出到客户端的流量。 |
| 入包量 | 在统计粒度内，客户端向负载均衡每秒发送的数据包数量。 |
| 出包量 | 在统计粒度内，负载均衡向客户端每秒发送的数据包数量。 |
| 入带宽 | 在统计粒度内，客户端流入到负载均衡所用的带宽。 |
| 出带宽 | 在统计粒度内，负载均衡流出到客户端所用的带宽。 |
| 活跃连接数 | 在统计粒度内，从客户端到负载均衡的活跃连接数。 |
| 非活跃连接数 | 在统计粒度内，从客户端到负载均衡的非活跃连接数。 |
| 并发连接数 | 在统计粒度内，从客户端到负载均衡的并发连接数。 |

| | |
|---|---|
| 新建连接数 | 在统计粒度内，从客户端到负载均衡的新建连接数。 |

### 2. 丢弃/利用率监控

| 指标名 | 指标含义 |
|---|---|
| 入带宽利用率 | 在统计粒度内，客户端通过外网访问负载均衡所用的带宽利用率。 |
| 出带宽利用率 | 在统计粒度内，负载均衡访问外网所用的带宽使用率。 |
| 并发连接数利用率 | 在统计粒度内的某一时刻，从客户端到负载均衡的并发连接数相比规格的并发连接数性能上限的利用率。 |
| 新建连接数利用率 | 在统计粒度内，从客户端到负载均衡的新建连接数相比负载均衡规格的新建连接数性能上限的利用率。 |
| 丢弃连接数 | 在统计粒度内，负载均衡丢弃的连接数。 |
| 丢弃入带宽 | 在统计粒度内，客户端通过外网访问负载均衡时丢弃的带宽。 |
| 丢弃出带宽 | 在统计粒度内，负载均衡访问外网时丢弃的带宽。 |
| 丢弃流入数据包 | 在统计粒度内，客户端通过外网访问负载均衡时丢弃的数据包。 |
| 丢弃流出数据包 | 在统计粒度内，负载均衡访问外网时丢弃的数据包。 |
| 丢弃 QPS | 在统计粒度内，负载均衡丢弃的请求数。 |
| QPS 利用率 | 在统计粒度内，负载均衡的 QPS 相比负载均衡规格的 QPS 性能上限的利用率。 |

### 3. LB 到后端的监控

| 指标名 | 指标含义 |
|---|---|
| 出流量 | 在统计粒度内，后端服务器流出到负载均衡的流量。 |
| 入带宽 | 在统计粒度内，负载均衡流入到后端服务器所用的带宽。 |
| 出带宽 | 在统计粒度内，后端服务器流出到负载均衡所用的带宽。 |

### 4. 七层协议监控

| 指标名 | 指标含义 |
|---|---|
| CLB 返回的 3xx 状态码 | 在统计粒度内，负载均衡返回 3xx 状态码的个数（负载均衡和网关节点返回码之和）。 |

| | |
|---|---|
| CLB 返回的 4xx 状态码 | 在统计粒度内，负载均衡返回 4xx 状态码的个数（负载均衡和网关节点返回码之和）。 |
| CLB 返回的 5xx 状态码 | 在统计粒度内，负载均衡返回 5xx 状态码的个数（负载均衡和网关节点返回码之和）。 |
| CLB 返回的 404 状态码 | 在统计粒度内，负载均衡返回 404 状态码的个数（负载均衡和网关节点返回码之和）。 |
| CLB 返回的 499 状态码 | 在统计粒度内，负载均衡返回 499 状态码的个数（负载均衡和网关节点返回码之和）。 |
| CLB 返回的 502 状态码 | 在统计粒度内，负载均衡返回 502 状态码的个数（负载均衡和网关节点返回码之和）。 |
| CLB 返回的 503 状态码 | 在统计粒度内，负载均衡返回 503 状态码的个数（负载均衡和网关节点返回码之和）。 |
| CLB 返回的 504 状态码 | 在统计粒度内，负载均衡返回 504 状态码的个数（负载均衡和网关节点返回码之和）。 |
| 2xx 状态码 | 在统计粒度内，后端服务返回 2xx 状态码的个数。 |
| 3xx 状态码 | 在统计粒度内，后端服务返回 3xx 状态码的个数。 |
| 4xx 状态码 | 在统计粒度内，后端服务返回 4xx 状态码的个数。 |
| 5xx 状态码 | 在统计粒度内，后端服务返回 5xx 状态码的个数。 |
| 404 状态码 | 在统计粒度内，后端服务返回 404 状态码的个数。 |
| 499 状态码 | 在统计粒度内，后端服务返回 499 状态码的个数。 |
| 502 状态码 | 在统计粒度内，后端服务返回 502 状态码的个数。 |
| 503 状态码 | 在统计粒度内，后端服务返回 503 状态码的个数。 |
| 504 状态码 | 在统计粒度内，后端服务返回 504 状态码的个数。 |
| 最大请求时间 | 在统计粒度内，负载均衡的最大请求时间。 |
| 平均响应时间 | 在统计粒度内，负载均衡的平均响应时间。 |
| 最大响应时间 | 在统计粒度内，负载均衡的最大响应时间。 |
| 响应超时个数 | 在统计粒度内，负载均衡响应超时的个数。 |
| 每分钟成功请求数 | 在统计粒度内，负载均衡每分钟的成功请求数。 |

| | |
|---|---|
| 每秒请求数 | 在统计粒度内，负载均衡每秒钟的请求数。 |

### 5. 健康检查监控

| 指标名 | 指标含义 |
|---|---|
| 健康检查异常数 | 在统计周期内，负载均衡的健康检查异常个数 |

# 查看默认日志

最近更新时间：2026-07-23 17:07:34

## 操作场景

AI 网关默认为您提供网关实时日志服务和简单搜索能力，免费使用。
默认日志主要分为用户访问日志和网关错误日志。您可以通过查看 AI 网关的访问日志了解用户的请求相关信息，便于进行数据分析、审计、业务排障等，也可以查看 AI 网关的错误日志，以便排查问题。

- 访问日志（accessLog）记录了用户的请求相关信息，可用于进行数据分析、审计、业务排障等。
- 错误日志（errorLog）是网关内部错误日志，用于网关排障。

本文为您介绍 AI 网关默认日志功能的使用说明。

## 前提条件

已创建 AI 网关实例，具体操作请参见 新建 AI 网关 。

## 查看默认日志

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **数据观测** > **默认日志**。
4. 设置好您要查看的日志，页面即可展示相关日志内容。 可以通过关键字查询相关日志。输入关键词查询，例如：“info”，注意日志检索区分大小写。

## 编辑默认日志规则

在默认日志页面，单击右上角的**编辑日志规则**，即可修改默认日志规则。您可以选择继续使用默认规则，也可以根据您的业务需求自定义日志规则。

> ⚠ **注意：**
> 修改默认日志规则后，投递到 CLS 的日志规则也会同步修改，请谨慎操作。

## 日志字段

下表列出了 AI 网关支持的访问日志字段，您可以根据需要进行配置：

## HTTP/HTTPS 日志字段

| 字段 | 说明 |
| --- | --- |
| $remote_addr | 客户端地址。 |

| | |
|---|---|
| $status | HTTP 状态码。 |
| $remote_user | Basic authentication 提供的用户名。 |
| $time_local | 请求时间。 |
| $request | 完整的请求行。 |
| $body_bytes_sent | 发送给客户端的文件主体内容的大小。 |
| $request_method | 请求方法。 |
| $host | 请求携带 Host 请求头时为“Host”字段的值，未携带时为主机虚拟域名。 |
| $upstream_addr | 后端服务的 IP 地址。 |
| $upstream_status | 上游服务返回响应中的 HTTP 响应码。 |
| $upstream_response_time | 上游服务响应耗时（毫秒精度），包括网关向后端服务开始建立连接、接收数据、关闭连接的时间。 |
| $scheme | HTTP 或 HTTPS 协议。 |
| $url | 请求 URL。 |
| $request_length | 请求数据大小 bytes，包含请求行、请求头、请求体。 |
| $bytes_sent | 响应字节数。 |
| $http_referer | 页面来源，header Referer 引用页面 URL。 |
| $http_user_agent | 客户端代理信息。 |
| $request_time | 请求耗时，从接收请求开始到发送完响应数据的时间，包含接收请求数据、处理请求、返回响应数据的时间。 |

## Nginx 变量

不支持的 Nginx 变量如下：

1. 如下变量：
    - $connection_time
    - $http3
    - $jwt_claim_
    - $jwt_header_
    - $jwt_payload

- $memcached_key
- $mqtt_preread_clientid
- $mqtt_preread_username
- $otel_parent_id
- $otel_parent_sampled
- $otel_span_id
- $otel_trace_id
- $proxy_protocol_tlv_
- $proxy_protocol_tlv_aws_vpce_id
- $proxy_protocol_tlv_azure_pel_id
- $proxy_protocol_tlv_gcp_conn_id
- $secure_link
- $secure_link_expires
- $session_log_binary_id
- $session_log_id
- $slice_range
- $ssl_alpn_protocol
- $ssl_curve
- $upstream_queue_time

2. geo 开头的变量。

# 日志投递至 CLS

最近更新时间：2026-07-23 17:07:35

## 操作场景

如您需要日志持久化存储，用于排障、审计等场景，建议开启 CLS 日志服务，将网关日志投递到 CLS。开启前请先确认您已开通 CLS 日志服务。日志服务由 CLS 日志服务 提供，会产生费用，具体计费项请参见 费用详情 。

## 前提条件

已创建 AI 网关实例，具体操作请参见 新建 AI 网关 。

## 开启日志投递

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 在左侧导航栏单击 **数据观测 > 日志投递**。
4. 在 CLS 日志投递页面，单击**立即开启**，在弹窗中开启**投递日志**按钮，勾选要投递的日志项，支持自动创建和选择已有的日志集和日志主题，单击**确定**后，即可开始投递日志。

   > **注意：**
   > 关闭日志投递不会删除日志集和日志主题，如需删除，请前往 CLS 控制台操作。

5. 配置完成后您可以单击**关联的日志主题**链接地址，自动跳转至网关投递的日志库查看日志。

## 编辑 CLS 日志规则

在日志投递页面，单击右上角的**编辑日志规则**，即可修改 CLS 日志规则。您可以选择继续使用默认规则，也可以根据您的业务需求自定义日志规则。

> **注意：**
> 投递到 CLS 的日志与默认日志使用相同的日志规则。修改 CLS 日志规则后，默认日志规则也会同步修改，请谨慎操作。

## 编辑 CLS 日志投递

在日志投递页面，单击右上角的**编辑日志投递**，在弹窗中即可修改日志投递规则。


CLS 日志投递
1. 日志服务 CLS 独立计费，具体费用按照您的使用情况收取。计费标准请参考 CLS计费概述 。
2. 关闭日志投递不会删除日志集和日志主题，如需删除，请前往 CLS控制台 进行操作。
投递日志
投递日志项
访问日志
错误日志
访问日志
自动创建
选择已有
访问日志配置
tem-
4314
tem-
6e20
错误日志
自动创建
选择已有
错误日志配置
tse_l
9f0eda
gateway-
3444
确定
取消

# 日志大盘

最近更新时间：2026-07-23 17:07:34

## 操作场景

AI 网关对接腾讯云日志服务（CLS），提供基于日志大盘的 LLM 和 MCP 监控可视化能力。通过 CLS 日志大盘，您可以使用如下能力：

- **全局概览**：查看网关层面的核心指标，包括总请求数、延迟分位、错误率、Token 消耗量、首 Token 延迟等
- **LLM 监控**：深入分析大模型调用的性能指标，包括 Token 消耗趋势、Provider 响应时间、Prompt 长度分布等
- **MCP 监控**：监控 MCP 工具调用的成功率、延迟和调用趋势

> **说明：**
> 日志大盘的数据来源于已投递到 CLS 的 AI 网关访问日志。使用前请确保已在 **数据观测 > 日志投递** 中开启了 LLM 访问日志和 MCP 访问日志投递。

## 前置条件

1. 已创建 AI 网关实例且处于运行中状态
2. 已在 **数据观测 > 日志投递** 中配置并开启了 CLS 日志投递（日志主题已创建且数据正常写入）
3. 网关已有 LLM 或 MCP 请求流量（新开启投递后约 1-2 分钟开始产生数据）

## 查看日志大盘

1. 登录 微服务平台控制台 ，在左侧导航栏选择 **AI 网关**进入实例列表。
2. 在实例列表中，单击目标实例名称进入实例详情页
3. 在左侧导航栏选择**数据观测**，单击**日志大盘** Tab

   日志大盘页面包含三个子 Tab：

| Tab | 说明 | 适用场景 |
| --- | --- | --- |
| **全局概览** | 展示网关层面的聚合指标和 Top 榜单 | 快速了解网关整体运行状态 |
| **LLM 监控** | 展示大模型调用的专项指标 | 分析 LLM 请求的性能和消耗 |
| **MCP 监控** | 展示 MCP 工具调用的专项指标 | 监控 MCP Server 和 Tool 的健康状态 |

## 使用场景

### 场景一：全局概览

全局概览提供网关层面的核心指标聚合，帮助您快速掌握整体运行状态。

> **说明：**
> 目前全局概览仅包括 LLM 调用相关指标，暂未统计 MCP 与 Agent 调用指标。

## 1. 核心指标卡片

| 指标卡片 | 说明 |
| --- | --- |
| **总请求数** | 选定时间范围内的网关请求总数 |
| **平均请求延迟 P95** | 端到端请求延迟的 95 分位值（ms） |
| **请求最大延迟 P99** | 端到端请求延迟的 99 分位值（ms） |
| **错误率（%）** | 失败请求占总请求的比例（HTTP ≥ 400） |
| **Token 消耗量** | 输入 + 输出 Token 的总消耗量 |
| **首 Token 延迟 TTFT P95** | 流式响应首 Token 时间的 95 分位值（ms） |

## 2. 查看异常快速发现

**异常快速发现（消费者风险 Top 10）面板展示高风险消费者的排名：**

- 基于错误率、延迟、Token 消耗等维度综合评估消费者风险计算的异常评分。
- 帮助您快速定位问题消费者，及时排查异常

## 3. 查看分布与趋势面板

| 面板名称 | 说明 |
| --- | --- |
| **请求量趋势（QPS）** | 展示每秒请求数的变化趋势 |
| **错误率趋势（%）** | 展示错误率随时间的变化趋势 |
| **端到端延迟分布 P50/P95/P99** | 展示请求延迟的分位数分布趋势（ms） |
| **首 Token 延迟 TTFT P50/P95/P99** | 展示流式响应首 Token 延迟的分位数趋势（ms） |
| **Token 消耗趋势** | 展示输入/输出 Token 的堆叠趋势图 |

## 4. 查看 Top 榜单

| 面板名称 | 说明 | 用途 |
| --- | --- | --- |

| | | |
|---|---|---|
| Top Token 消耗（按模型 API） | Token 消耗量最高的 Top 模型 API | 识别高消耗 API，优化成本 |
| Top 请求量（按消费者） | 请求量最高的 Top 消费者 | 了解流量分布，合理分配配额 |
| Top 平均请求延迟（按模型 API） | 平均延迟最高的 Top 模型 API | 定位延迟高的 API，优化性能 |
| Top 错误率（按模型 API） | 错误率最高的 Top 模型 API | 快速发现异常 API，排查问题 |
| Top 首 Token 延迟 TTFT（按模型 API） | TTFT 延迟最高的 Top 模型 API | 优化流式响应体验 |

## 场景二：查看 LLM 监控

LLM 监控提供大模型调用的专项指标，帮助您深入分析 LLM 请求的性能特征。

### 1. 核心指标卡片

| 指标卡片 | 说明 |
|---|---|
| 总请求数 | LLM 请求的总数 |
| P95 延迟 | 端到端延迟的 95 分位值（ms） |
| P99 延迟 | 端到端延迟的 99 分位值（ms） |
| 错误率（%） | LLM 请求的错误率 |
| 首 Token 延迟 P95 | 流式响应 TTFT 的 95 分位值（ms） |
| Token 消耗量 | LLM 请求的总 Token 消耗 |
| Input Token 总量 | 输入 Token 的总数 |
| Output Token 总量 | 输出 Token 的总数 |
| Token 生成速率（tok/s） | 平均每秒生成的 Token 数 |
| Input Cache 命中率（%） | Prompt Cache 命中的 Token 占比 |
| LLM HTTP 请求数 | LLM HTTP 请求的总数 |
| Chat 操作数 | Chat Completion 请求数 |

### 2. 查看 LLM 趋势与分布面板

| 面板名称 | 说明 |
| --- | --- |
| 延迟分位数趋势（P50/P95/P99） | 展示 LLM 请求延迟的分位数变化趋势 |
| Input/Output Token 趋势（堆叠） | 展示输入/输出 Token 的堆叠趋势 |
| Prompt 长度分布（Input Token 直方图） | 展示 Prompt 长度的分布情况 |
| TTFT 首 Token 延迟 P50/P95/P99 | 展示流式响应首 Token 延迟的分位数趋势 |
| TTFT 分布直方图 | 展示 TTFT 的分布直方图 |
| 按 Provider 分组 TTFT P95 | 按不同供应商分组展示 TTFT P95 |
| Provider 响应时间 P95 | 各供应商的后端响应时间 P95 |
| Provider 吞吐量（请求数） | 各供应商的请求吞吐量 |
| Top 调用流量 Top10 | 调用量最高的 Top10 消费者/API |

**说明：**

`Input Cache` 命中率 反映模型侧 Prompt Cache 的命中情况，命中率越高说明重复 Prompt 越多，可通过启用 AI 网关缓存进一步降低成本。

## 场景三：查看 MCP 监控

MCP 监控提供 MCP 工具调用的专项指标，帮助您监控 MCP Server 和 Tool 的健康状态。

### 1. 核心指标卡片

MCP 监控页顶部展示 3 个核心指标卡片：

| 指标卡片 | 说明 |
| --- | --- |
| 工具调用次数 | MCP Tool 的总调用次数 |
| 工具成功率 | Tool 调用成功的比例 |
| 工具 P99 延迟 | Tool 执行时长的 99 分位值（ms） |

### 2. 查看 MCP 趋势与排行面板

| 面板名称 | 说明 |
| --- | --- |
| **工具调用趋势（按工具分组）** | 各 Tool 的调用次数趋势 |
| **成功率/参数错误率趋势（%）** | Tool 调用成功率和参数错误率的趋势 |
| **按工具 P99 延迟趋势** | 各 Tool 的 P99 延迟趋势 |
| **平均延迟趋势** | Tool 调用的平均延迟趋势 |
| **工具调用明细排行（按工具）** | 各 Tool 的调用次数、成功率、延迟明细排名 |

## 场景四：在 CLS 控制台自定义分析

当日志大盘的预设面板无法满足分析需求时，您可以跳转至 CLS 控制台进行自定义分析。

### 操作步骤

1. 在日志大盘页面右上角，单击 **在日志服务中查看更多**
2. 系统将自动跳转至腾讯云 CLS 控制台对应的日志主题
3. 在 CLS 控制台，您可以：
   - 使用 SQL 语句自定义查询日志数据
   - 创建自定义仪表盘和图表
   - 配置日志告警规则
   - 导出日志数据用于离线分析

# 使用 Prometheus 监控

最近更新时间：2026-07-24 16:45:01

## 操作场景

本文介绍如何使用 Prometheus 获取 AI 网关 的监控数据，当前支持直接关联腾讯云 Prometheus。关联 Prometheus 实例后，网关 将自动上报监控数据到已关联的 Prometheus。

## 前置条件

- 已购买 AI 网关实例，详情请参见 操作文档 。
- 已拥有腾讯云 Prometheus 实例 。

## 关联腾讯云 Prometheus

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 单击左侧菜单成本管理模块中的**数据观测**。
4. 切换到 **Prometheus** 页签，单击**关联 Prometheus 实例**。


5. 选择您要关联的腾讯云 Prometheus 实例（仅支持关联与实例在同一 VPC 下的 Prometheus 实例）。
6. 单击**确定**，完成关联。

## 注意事项

- 开启 Prometheus 插件对于网关的数据流性能有影响，建议只为需要监控的特定的 API(Route) 开启 Prometheus 插件。
- 在腾讯云上购买云监控 Prometheus 将产生费用，需要您自行承担。

# 日志包体采集

最近更新时间：2026-07-24 16:45:01

## 操作场景

本文档介绍如何配置 AI 网关日志中包体的记录格式。通过该功能，您可以选择将 Prompt/Completion 以 OpenAI 协议格式（跨供应商归一化）或原始 body（转发给供应商的实际请求体）写入 LLM Log，以满足日志分析和问题调试两类不同场景的需求。

## 前提条件

已创建 AI 网关实例，具体操作请参见 新建 AI 网关 。

## 开启包体采集

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关** > **实例列表**。
2. 在实例列表页面，单击需要配置的网关实例的“ID”，进入该网关实例的基本信息页面。
3. 选择**模型管理** > **模型 API**
4. 进入目标模型 API 详情页，切换到**基本信息**页签，单击右上角**编辑**，在**高级设置**区域（也可在新建模型 API 时直接开启）：
   - 根据需要，开启**请求包体采集（Prompt）**开关，网关将根据所配置的内容截断长度以及包体记录格式，记录入参 messages。
   - 根据需要，开启**响应包体采集（Completion）**开关，网关将根据所配置的内容截断长度以及包体记录格式，记录模型返回内容。

   目前包体记录格式支持以下两种：
   - 原始 Body：客户端发来的原始请求，原样保真
   - 归一化格式：网关实际转发给上游模型的请求，格式跟随上游供应商（如上游是 Claude 则为 Anthropic 格式）

> ⚠ **注意：**
> 开启包体采集后，请求中的 messages 内容将写入 LLM Log。请确认内容不含用户隐私数据，且符合企业合规要求。如需进行脱敏，可结合日志脱敏规则与数据安全配置。

# 操作记录

最近更新时间：2026-07-23 17:07:34

## 操作场景

操作记录功能记录所有用户/管理员对 AI 网关资源的主动操作，提供合规审计和问题追溯能力。通过操作记录，您可以：

- 查看特定时间范围内的资源变更历史。
- 按资源类型、操作人、操作状态筛选操作记录。
- 追溯配置变更的详细内容，定位问题变更。
- 下载操作记录用于合规审计。

> **说明：**
> 仅支持保存最近 90 天的操作记录

## 前置条件

1. 已创建 AI 网关实例且处于运行中状态。
2. 具备操作记录的查看权限。

## 查看操作记录

1. 登录 微服务平台控制台 ，在左侧导航栏单击 **AI 网关 > 实例列表**。
2. 在实例列表中，单击目标实例名称进入实例详情页。
3. 在左侧导航栏选择**数据观测**，单击**操作记录** Tab。
4. 操作记录列表展示以下字段：

| 字段 | 说明 |
| --- | --- |
| **资源名称** | 被操作资源的名称 |
| **资源类型** | 被操作资源的类型，包括 |
| **状态** | 操作执行结果 |
| **操作** | 具体操作类型 |
| **操作人** | 执行操作的账号 UIN |
| **操作内容** | 变更的详细内容（JSON 格式） |

| | |
|---|---|
| **操作时间** | 操作执行的时间 |

## 筛选操作记录

当操作记录较多时，可通过筛选条件快速定位目标记录。

### 按时间范围筛选

| 快捷选项 | 说明 |
|---|---|
| **近 7 天** | 展示最近 7 天的操作记录 |
| **近 30 天** | 展示最近 30 天的操作记录 |
| **自定义** | 单击日历图标选择自定义日期范围 |

### 按资源类型筛选

单击 **资源类型** 下拉框，选择要查看的资源类型：

| 资源类型 | 说明 | 典型操作 |
|---|---|---|
| **全部** | 展示所有资源类型的操作记录 | – |
| **消费者** | 消费者（Consumer）的增删改 | 创建、编辑、删除 |
| **模型 API** | 模型 API 的增删改及策略变更 | 创建、编辑、删除、上线、下线 |
| **服务来源** | 服务来源（注册中心/Private DNS）的增删改 | 创建、编辑、删除 |
| **MCP 服务** | MCP Server 的增删改及上下线 | 创建、编辑、删除、上线、下线 |
| **MCP Tool** | MCP Tool 的启用/禁用 | 启用、禁用 |
| **Agent API** | Agent API 的增删改 | 创建、编辑、删除 |
| **Agent API 路由** | Agent API 路由规则的增删改 | 创建、编辑、删除 |
| **证书** | SSL 证书的增删改 | 创建、编辑、删除 |

## 下载操作记录

如需将操作记录导出用于合规审计或离线分析，可使用下载功能。

1. 在操作记录页面，配置好筛选条件（时间范围/资源类型/操作人）。
2. 单击页面右上角的**下载**按钮。
3. 系统生成并下载操作记录文件。