负载均衡

模型路由

腾讯云

文档目录

模型路由
产品简介
模型路由产品介绍

最近更新时间：2026-04-27 10:57:38

产品概述

模型路由是一个统一的大模型访问服务，支持您通过单一 API 接口访问多家模型提供商的模型服务。
该服务与 VPC 环境深度集成，支持接入 VPC 内及混合组网环境中的模型服务；对于公网部署的模型服务，可结合公网加速能力提供更稳定的网络连接。模型路由内置自动重试和智能路由能力，可降低调用失败率，提升服务可用性。

核心能力

- **标准接口兼容**：兼容 OpenAI API，便于集成，降低适配成本。
- **多模型统一接入**：通过单一 API 接口访问多家模型提供商的服务。
- **混合网络接入**：支持接入 VPC 内及自建 IDC 等混合组网环境中的模型服务。
- **公网模型接入**：支持接入公网部署的模型服务，并可结合公网加速能力提升连接质量。
- **自动重试**：请求失败时自动重试，降低调用失败次数和运维成本。
- **智能故障切换**：单模型故障时，自动切换至同模型的其他供应商，保障服务连续性。

开通说明

当前模型路由处于内测阶段。如需体验产品功能，可联系您的架构师或提交 工单 申请。

支持的模型提供商

最近更新时间：2026-04-27 10:58:16

概述

模型提供商是模型路由与用户实际调用的具体模型之间的桥梁。通过模型路由，用户可以灵活选择不同的调用渠道和模型提供商，满足多样化的 AI 应用需求。

支持的提供商类型

模型提供商	说明	资源消耗	状态
平台预置模型提供商	直接调用模型路由预置的模型	消耗模型路由提供的免费 Token 或用户付费的资源配额	即将推出
BYOK 原厂模型提供商	通过模型路由调用原厂 AI 平台提供商的模型	在原厂 AI 平台或算力平台上消耗用户的 Token	已上线
BYOK 第三方代理提供商	通过模型路由调用第三方 AI 平台上的模型	在第三方 AI 平台上消耗用户的 Token	已上线
BYOK 用户自建模型提供商	通过模型路由调用用户私有化部署的模型	在用户自建的 AI 平台上消耗用户自己的 Token	已上线

BYOK 模式下预置的模型提供商和模型

> **注意：**
> 预置的模型可能因供应商调整而未能及时更新。如果未找到您期望的模型名称，可通过 BYOK 第三方代理提供商自行配置 URL 及模型名称。

模型提供商	模型
DeepSeek	deepseek-chat
	deepseek-reasoner
阿里百炼	Qwen3.6-Plus
	Qwen3.5-Plus

	Qwen3.5-Flash
	Qwen3-Max
	Qwen3-VL-Plus
	Qwen3-VL-Flash
智谱 AI	GLM-5
	GLM-5-Turbo
	GLM-4.7
	GLM-4.7-FlashX
	GLM-4.6
	GLM-4.5
	GLM-4.5-Air
	GLM-4.7-Flash
月之暗面	kimi-k2.5
	moonshot-v1-8k-preview
	moonshot-v1-8k
	kimi-k2-thinking
	moonshot-v1-auto
	moonshot-v1-32k-preview
	kimi-k2-turbo-preview
	kimi-k2-0711-preview
	kimi-k2-0905-preview
	kimi-k2-thinking-turbo
	moonshot-v1-32k
	moonshot-v1-128k
	kimi-latest

	moonshot-v1-128k-preview
MiniMax	MiniMax-M2.7
	MiniMax-M2.7-highspeed
	MiniMax-M2.5
	MiniMax-M2.5-highspeed
	MiniMax-M2.1
	MiniMax-M2.1-highspeed
	MiniMax-M2
火山引擎	doubao-seed-2-0-pro-260215
	doubao-seed-2-0-lite-260215
	doubao-seed-2-0-mini-260215
	doubao-seed-2-0-code-preview-260215

最佳实践建议

- 对于稳定性和可靠性要求较高的场景，建议使用 BYOK 原厂模型提供商。
- 对于需要灵活性和定制化的场景，建议使用 BYOK 第三方代理提供商或用户自建模型。
- 定期检查模型提供商列表，确保使用最新版本的模型。

使用约束

最近更新时间：2026-04-27 10:58:36

本文介绍模型路由的使用限制如下：

通用限制

类型	默认配额（单位：个）
单账号单地域可创建的共享型模型路由实例	1（不支持提升）
单账号单地域可创建的企业型模型路由实例	5
单账号单地域可创建的预置 BYOK 模型数量	20
单账号单地域可创建的第三方 BYOK 模型数量	20
单账号单地域可创建的自定义 BYOK 模型数量	20
每个标准/公网模型实例可添加 Key 数量	50
每个内网第三方模型实例可添加 Key 数量	50
每个共享型实例可添加 Key 数量	10
每个共享型实例可关联模型实例数量	20
每个企业型实例可添加 Key 数量	100
每个企业型实例可关联模型实例数量	50

如果上述配额不满足您的需求，请提交 工单 联系我们。

产品计费
模型路由资源包

最近更新时间：2026-04-27 10:58:54

CLB 模型路由资源包是针对使用 CLB 模型路由产品推出的一种计费资源包。购买后立即生效并自动抵扣 CLB 模型路由实例所产生的处理费。本文将为您介绍 CLB 模型路由资源包的基本信息，并引导您完成购买。

资源包介绍

优势

- **使用便捷**：资源包购买后立即生效，系统自动完成费用抵扣，无需额外配置。
- **灵活配置**：支持按需选择购买容量，支持叠加购买多个资源包，支持设置续订，可根据不同业务场景灵活调整购买策略。

使用限制

CLB 模型路由资源包仅适用于抵扣 CLB 模型路由实例产生的处理费，不支持用于抵扣其他云服务产品的费用，请注意按需购买。

资源包服务区域

CLB 模型路由资源包为地域级资源包。每个地域的资源包相对独立，互不影响。

> **说明：**
> 每个资源包仅支持抵扣同地域 CLB 模型路由实例产生的处理费。

计费周期

资源包默认有效期为 3 年（自然年），购买后立即生效。假如您在2026年06月06日 10:00:00购买一个资源包，则到期时间为2029年06月06日 09:59:59。

计费价格

地域	单价（元/u）
所有地域	0.01

当前已支持的地域：北京、上海、广州、硅谷。其他地域陆续上线中，地域开通请留意官网公告。

资源包抵扣规则

当账号中存在多个资源包时，系统优先抵扣先到期的资源包。

> **注意：**
> - 若计费区域内所有资源包用尽或到期且未续订时，CLB 模型路由服务将在下一个计费周期停止。
> - 到期后未用完的资源包容量将自动清零，不支持余额转移。

资源包购买

1. 登录腾讯云 模型路由资源包 控制台。
2. 在左侧导航栏选择**资源包**。
3. 在资源包页面，单击**购买**。
4. 在购买资源包页面，按需配置**购买容量**、**是否续订**，并**确认**。

> **说明：**
> 建议您选择**自动续订**，避免资源包用尽或到期影响您的业务。

资源包续订规则

CLB 模型路由资源包支持设置自动续订。开启自动续订后，当您账户余额充足时，系统将根据自动续订规则为您自动购买容量等配置与原资源包一致的新资源包。

触发自动续订的条件

对于开启自动续订的资源包，触发自动续订的条件如下：

- 在资源包有效期内，当抵扣区域内的所有的资源包耗尽时，针对设置续订的资源包自动续订一个相同配置的新资源包。
- 如果抵扣区域内的资源包没有用尽，则会在抵扣区域内的所有的资源包到期后，针对设置续订的资源包自动续订一个相同配置的新资源包。

> **说明：**
> 所有资源包耗尽是指同计费区域所有的有效资源包均用完而非指设置的某个资源包用完。

自动续订的注意事项

- 同一个抵扣区域内仅支持对一个资源包做续订设置。如果重复设置，则新设置的资源包将会代替原有的设置。
- 当资源续订失败时，该资源自动续订状态将变更为关闭。

资源包退费规则

资源包到期前，未用完的容量可在资源包页面进行退费操作。

> **注意：**

- 免费资源或者代金券不予退还。
- 退还金额将全部原路退回到您的腾讯云账户。

退款流程

1. 登录腾讯云 模型路由资源包 控制台。
2. 在左侧导航栏选择**资源包**。
3. 在资源包页面，找到还有剩余容量的资源包。在指定资源包的**操作列**，单击**退费**。
4. 在退款页面确定**容量**和**金额**，并单击**确认**，完成退款。

常见问题

1. **CLB 模型路由资源包可以抵扣哪些费用?**

   CLB 模型路由资源包可用于抵扣 CLB 模型路由实例所产生的处理费。
2. **CLB 模型路由资源包购买后需要设置吗?**

   购买后不需要任何配置，系统会自动进行抵扣。
3. **购买的 CLB 模型路由资源包什么时候生效?**

   模型路由资源包在购买后立即生效。
4. **CLB 模型路由资源包可以购买多个，叠加使用吗?**

   支持叠加购买多个资源包。当账号中存在多个资源包时，系统优先抵扣先到期的资源包。
5. **先购买的 CLB 模型路由资源包会优先抵扣吗?**

   当账号中存在多个资源包时，系统优先抵扣先到期的资源包。资源包到期时间是根据购买时选择的生效时间计算的。

模型路由处理费

最近更新时间：2026-04-27 10:59:13

概述

模型路由处理费用覆盖平台为您提供的模型路由、请求/响应中转、流量优化及系统运维等服务成本。

> **注意：**
> 为助力您更流畅地开始使用，在当前的推广期内，我们免费提供以下两项高价值网络加速与互联能力：
> - 公网链路加速：当前所有转发至主流大模型服务的网络链路，均已集成 Agent 公网加速产品能力，有效降低延迟、提升稳定性。推广期(到2026年12月31日)内免费。
> - 混合云原生互联：我们为您免费提供了混合云组网的原生 VPC 能力。您可以将部署在自有私有云或本地数据中心的服务，与我们网关的 VPC 网络进行安全、高速、低延迟的内网级互联，避免公网传输带来的安全与性能风险。推广期内免费。
>
> 请注意：模型路由处理费与上述免费权益独立。推广期结束后，网络加速与混合云组网服务将可能按相应规则独立计费，届时会有正式公告。

模型路由处理费与业界常见模式的差异如下：

特性维度	模型路由网关	业界常见代理模式
计费基础	处理的 Token 总量。无论上游模型单价如何，我们均对通过网关的 Token 总数进行计费。	按比例传递上游成本。通常基于用户实际上游开销（即模型 API 调用成本）按一定比例加成收费。
定价模式	固定单价。按照每百万 Token（M tokens）收取固定的处理费，价格不随上游模型定价变化。	浮动价格。收费随上游模型官方定价的波动而波动，用户最终成本随上游定价变化而变化。
覆盖范围	完整处理流量。计费依据是经过网关路由的所有输入与输出 Token 总和，包括上游模型可能产生的、但未直接返回给客户端的部分（例如工具调用的额外 Token、内部思考/推理过程的 Token）。	通常基于返回的 usage。计费主要依据上游模型返回的标准 usage 对象，覆盖范围取决于上游模型返回的 usage 字段定义，可能不包含部分模型内部消耗的 Token。

模型路由处理费不随上游模型单价变化，仅与您通过网关处理的总 Token 数量正相关，为您提供了更稳定的成本控制预期。

计费介绍

计算公式

总模型路由处理费 = Token 处理费
Token 处理费 = 总处理 Token 数 * 处理费单价

- 总处理 Token 数 ≈ 上游模型返回 usage 中的 total_tokens。总处理 Token 数是指单次请求中，经过网关路由的所有输入 Token 与输出 Token 的总和。

> **说明：**
> total_tokens 由输入和输出两部分构成，以下都会计入处理费用中：
> - 输入 Token (prompt_tokens)：来源于 usage.prompt_tokens_details 的细项总和，包括：
> - 非缓存的输入 Token
> - 输入 Token 读缓存
> - 输入 Token 写缓存
> - 输出 Token (completion_tokens)：来源于 usage.completion_tokens_details 的细项总和，包括：
>   - 普通回复 Token (completion token)。
>   - 模型的推理/思考 Token (reasoning_tokens，如有)。
>   - 为完成工具调用所产生的额外 Token。

- 处理费单价：当前定价为 0.49 元 / 百万 Token。

计费示例

假设您通过我们的网关调用了一次上游模型（以 Kimi 为例）。
用户请求与上游模型响应示例：


```
{
  "id": "cmpl-a1b2c3d4e5f6a7b8c9d0e1f2",
  "object": "chat.completion",
  "created": 1774675200,
  "model": "kimi-k2-0905-preview",
  "choices": [
    {
      "index": 0,
      "finish_reason": "stop",
      "message": {
        "role": "assistant",
```

```
        "content": "你好！"
      },
      "logprobs": null
    }
  ],
  "usage": {
    "prompt_tokens": 98,
    "completion_tokens": 12,
    "total_tokens": 110,
    "prompt_tokens_details": {
      "cached_tokens": 64          // ← 缓存命中的 token 数，自动填充，是
prompt_tokens 的子集
    },
    "completion_tokens_details": {
      "reasoning_tokens": 0        // ← thinking 模型时此处非零，单价与普通
output 相同
    }
  }
}
```

模型路由处理费用计算过程如下：

1. 确定总处理 Token 数：

   网关从响应中提取 usage.total_tokens 字段，数值为 110 Tokens。
2. 计算 Token 处理费：
   - 总处理 Token 数 = 110 Tokens
   - 处理费单价 = 0.49 元 / 百万 Token = 0.49 元 / 1,000,000 Tokens
   - Token 处理费 = 110 / 1,000,000 * 0.49 ≈ 0.0000539 元
3. 对于此次调用，您将产生的模型路由处理费约为0.0000539元。无论本次调用的上游模型是 Kimi、GLM 还是 DeepSeek，只要处理的 Token 总数是110，此费用固定不变。

快速入门

最近更新时间：2026-04-24 18:11:29

本教程介绍了如何快速开始使用 CLB 模型路由。

> **说明：**
> 使用本教程前，建议您先阅读 使用约束 、支持的模型提供商 ，了解相关信息。

前提条件

您已经获得 CLB 模型路由使用资格。如需获得 CLB 模型路由的使用资格，请提交 工单申请 。

操作步骤

步骤1：创建模型路由实例

1. 登录 CLB 模型路由控制台 。
2. 在左侧导航栏中，单击**入口管理**。
3. 在实例列表页中，单击**新建**，参数说明如下。

参数	说明
实例类型	可选共享型、企业型。共享型实例适用于开发测试与功能验证环节；企业型实例适用于生产环境，保障业务安全可控。
网络类型	仅企业型实例支持，可选公网、内网。
监听协议	网络类型选择内网，监听协议可选 HTTP（80）、HTTPS（443）。网络类型选择公网，监听协议仅可选 HTTPS（443）。
证书	共享型实例本身自携带证书，仅企业型实例需要绑定证书。
所属网络	企业型实例需要选择所属网络。
实例名称	最多支持 255 个字符。
标签	选择标签键和标签值，也可选择添加标签，详情请参见 创建标签 。
TPM	每分钟允许处理的最大 Token 数（Tokens Per Minute），单位：千/分钟。
RPM	每分钟允许的最大请求次数（Requests Per Minute），单位：次/分钟。

4. 完成以上参数配置后，单击**确定**创建实例。
5. 在实例列表中，即可查看您创建的实例。

步骤2：生成 API Key

1. 在左侧导航栏中，单击**入口管理**。单击您创建的实例，进入**实例管理**页面，切换至 **API Key** 页签。
2. 单击**新建 Key**，参数说明如下。

参数	说明
Key 名称	最多支持 255 个字符。
标签	选择标签键和标签值，也可选择添加标签，详情请参见 创建标签 。
限制类型	可选择 API Key 或积分预算。
积分预算	若限制类型为积分预算则需要填写具体的积分预算内容。
TPM	若限制类型为 API Key 则需要填写 TPM。每分钟允许处理的最大 Token 数（Tokens Per Minute），单位：千/分钟。
RPM	若限制类型为 API Key 则需要填写 RPM。每分钟允许的最大请求次数（Requests Per Minute），单位：次/分钟。

3. 完成以上参数配置后，单击**确定**完成新建 Key。请妥善保存以下 API Key，关闭弹窗后将无法再次查看完整 Key。

步骤3： 新增 BYOK 模型

1. 在左侧导航栏单击 **BYOK** 进入 BYOK 列表页。
2. 单击**新建**创建 BYOK，参数说明如下。

参数	说明
模型来源	可选原厂模型、第三方代理、自建模型。原厂模型：自带官方 API Key，平台自动补全 APIBase 并提供公网加速，最易接入。第三方代理：接入 OpenRouter 等代理商 API 自定义 APIBase，统一管理 Key，灵活切换模型厂商。自建模型：通过 VPC 内网直连企业自建 GPU 集群，支持云联网/专线打通 IDC 机房，数据零出网。
所属厂商	请参考 支持的模型提供商 。
API 地址	模型来源选择第三方代理、自建模型时需要填写 API 地址。API 地址仅支持 VIP，不支持域名。
域名	模型来源选择自建模型时需要填写域名。域名为往上游模型发送请求时携带的 http header。
选择模型	支持手动输入自定义模型名称，最多选择 20个。
所属网络	模型来源选择自建模型时需要填写所属网络。

API Key	需要填写您在上游大模型上使用的 API Key
实例名称	最多支持255个字符。
标签	选择标签键和标签值，也可选择添加标签，详情请参见 创建标签 。

3. 完成以上参数配置后，单击**确定**完成新建 BYOK。

步骤4：关联模型

1. 在左侧导航栏中，单击**入口管理**。
2. 单击您创建的实例，进入**实例管理**页面，切换至**模型路由**页签。
3. 在关联模型列表右侧点击**批量关联**，并选择关联模型，确认后进行关联。
4. 配置路由策略。路由策略分为模型间策略和模型内策略，具体介绍如下：
   - 模型间策略：当请求未指定具体模型时，系统将根据当前实时状态或语义复杂度，智能选择最合适的模型进行处理。模型间策略分为简单随机路由、最低系数路由、语义复杂度路由（暂未开放）。
     - 简单随机路由：在可用模型中随机选择。
     - 最低系数路由：优先分发到积分较低的模型。
     - 语义复杂度路由（暂未开放）：开放后将支持按语义复杂度分级，每级可选多个模型并复用调度策略。
   - 模型内策略：当模型确定后，系统将根据实时性能指标，从该模型下不同的服务所属厂商中，动态选择最优的访问节点。模型内策略分为简单随机路由、最低繁忙路由、最低延迟路由、用量均衡路由。
     - 简单随机路由：在可用模型中随机选择。
     - 最低繁忙路由：将请求分配给当前最空闲的模型。
     - 最低延迟路由：自动选择当前延迟最低的模型。
     - 用量均衡路由：按用量均衡分配请求到各模型。
5. 配置 Fallback 策略，当关联模型路由中的模型服务失败时会使用 Fallback 中的模型。在 Fallback 策略列表右侧点击**编辑**。选择对应模型并点击**确定**。

步骤5：调用模型路由 API

在左侧导航栏中，单击**入口管理**。单击您创建的实例，进入**实例管理**页面，您可以根据调用示例中的举例并使用 OpenAI 请求方式编写请求即可访问各种配置的模型。

后续操作

在左侧导航栏中，单击**入口管理**。单击您创建的实例，进入**实例管理**页面，切换至**用量详情**页签。关注资源消耗，随时监控模型网关的使用情况（比如 token 和模型资源包使用情况）,避免额度不足造成调用失败。

用户指南
创建模型路由实例

最近更新时间：2026-05-06 16:21:24

模型路由实例是模型路由能力的承载单元，通过创建模型路由实例，用户可统一管理模型接入、流量分发、限流控制、网络访问等策略。本文介绍如何创建模型路由实例。

实例类型说明

模型路由实例包括以下两种类型：共享型、企业型。

- 共享型适用于开发调试、功能验证等轻量场景，无需复杂网络与安全配置，开箱即用。
- 企业型面向生产级业务，提供证书配置、VPC 选择能力，满足企业级稳定性与合规要求。

下表展示了共享型和企业型模型路由实例的功能对比。

对比项目	共享型	企业型
公网访问	支持	支持
内网访问（VPC）	不支持	创建的企业型内网实例，系统会从您的 VPC 中获取一个内网 VIP 进行关联。
HTTPS	支持。固定 HTTPS，不支持用户配置。	支持。出于安全性考虑，企业型公网模型路由实例仅支持通过 HTTPS 进行访问。
证书配置	不支持	支持。您可以选择 SSL 证书平台 中已有的证书，或新建上传证书。

前提条件

您已经获得 CLB 模型路由使用资格。如需获得 CLB 模型路由的使用资格，请提交 工单申请 。

操作步骤

1. 登录 CLB 模型路由控制台 。
2. 在左侧导航栏中，单击**实例管理**进入实例列表页。
3. 在实例列表页中，单击**新建**，参数说明如下。

参数	说明
实例类型	可选共享型、企业型。共享型实例适用于开发测试与功能验证环节；企业型实例适用于生产环境，保障业务安全可控。

网络类型	仅企业型实例支持，可选公网、内网。
监听协议	网络类型选择内网，监听协议可选 HTTP（80）、HTTPS（443）。网络类型选择公网，监听协议仅可选 HTTPS（443）。
证书	仅企业型实例需要绑定证书，您可以选择 SSL 证书平台 中已有的证书，或新建上传证书。
所属网络	企业型实例需要选择所属网络。
实例名称	最多支持 255 个字符。
标签	选择标签键和标签值，也可选择添加标签，详情请参见 创建标签 。
限制类型	仅企业型实例支持，可选速率限制、积分预算。若限制类型为积分预算则需选择具体的积分预算模板。
TPM	每分钟允许处理的最大 Token 数（Tokens Per Minute），单位：千/分钟。默认值 10 千/分钟，共享型实例 TPM 取值范围 1-10 千/分钟，企业型实例 TPM 取值范围 1-100000 千/分钟。如需调整，请提交 工单申请 。 仅限制类型为速率限制时支持调整，限制类型为积分预算则沿用积分预算模板设置。
RPM	每分钟允许的最大请求次数（Requests Per Minute），单位：次/分钟。默认值 10 次/分钟，共享型实例 RPM 取值范围 1-10 次/分钟，企业型实例 RPM 取值范围 1-10000 次/分钟。如需调整，请提交 工单申请 。 仅限制类型为速率限制时支持调整，限制类型为积分预算则沿用积分预算模板设置。

4. 完成以上参数配置后，单击**确定**创建实例。在实例列表中，即可查看您创建的实例。

创建模型路由访问密钥（API Key）

最近更新时间：2026-05-08 18:08:08

本文介绍如何创建和使用**模型路由访问密钥**（API Key）。

简介

模型路由产品形态包括两种类型的密钥：**模型路由访问密钥**和 **BYOK 密钥**。

- **模型路由访问密钥（API Key）**：用户访问模型路由实例时携带的密钥。模型路由根据该密钥进行鉴权和调用模型控制。

> **说明：**
> - 模型路由访问密钥（API Key）是实例级别资源，一个实例下创建的访问密钥仅支持在当前实例下使用，不允许跨实例使用。
> - 当 API Key 设置速率限速或者绑定积分预算后，除了 API Key 本身的限制外，还受到模型路由实例上所配置的速率限速或积分预算的限制。

- **BYOK 密钥**：用户在第三方大模型服务提供商处持有的自有密钥。支持将 BYOK 密钥配置到 CLB 模型路由实例上，作为连接其他大模型服务提供商的密钥。

> **说明：**
> 用户具体的消费产生在其他大模型服务提供商上，模型路由不会重复计费，仅收取模型路由处理费，详见 计费说明 。

前提条件

1. 已获得 CLB 模型路由的使用资格。如未获得，请提交 工单申请 。建议提前阅读 使用约束 与 支持的模型提供商 了解相关信息。
2. 已经创建模型路由实例，详细操作指导请参见 创建模型路由实例 。

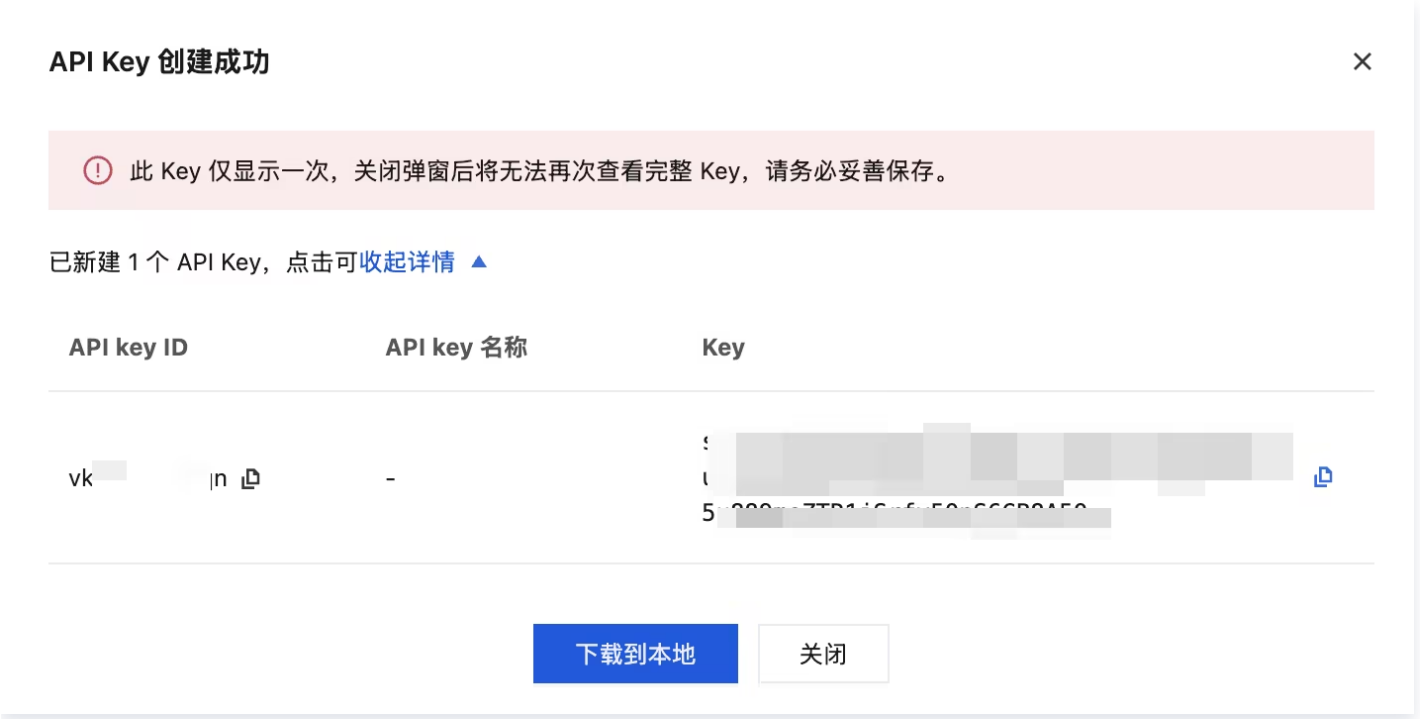
创建密钥

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例名称**。

2. 在实例详情页，选择 **API Key** 页签。
3. 在 **API Key** 页签详情中，单击**新建 Key**。

4. 在弹窗中配置 **Key 名称** 等参数，并单击**确定**。
5. **API Key** 创建成功后，请您妥善保管 Key 信息。您也可以单击**下载到本地**，进行保存。

> **说明：**
> 关闭弹窗后将无法再次查看完整的 Key，请谨慎操作并妥善保管。



后续操作

1. 新增 BYOK 模型，详细操作指导请参见 BYOK 模型介绍 。
2. 关联模型并配置模型路由策略，详细操作指导请参见 配置模型调度管理 。

配置模型调度管理

最近更新时间：2026-05-08 11:28:08

本文介绍如何关联模型，以及配置模型间和模型内的路由策略。配置完成后，用户请求经由 CLB 模型路由统一接入，平台完成计费、限流与日志记录后，根据您关联的模型，以及所配置的路由策略进行匹配与决策，最终分发至对应的后端模型中。

前提条件

- 已获得 CLB 模型路由的使用资格。如未获得，请提交 工单申请 。建议提前阅读 使用约束 与 支持的模型提供商 了解相关信息。
- 已创建模型路由实例，详细操作指导请参见 创建模型路由实例 。
- 已完成新增 BYOK 模型，详细操作指导请参见 BYOK 模型配置管理 。
- 已创建模型路由访问密钥（API Key），详细操作指导请参见 创建模型路由访问密钥 。

操作指导

步骤一 关联模型

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例名称**，进入目标实例的**实例管理**页面。
2. 切换至**模型调度管理**页签，单击**批量关联**，选择关联模型，确认关联信息后单击**确认**。

步骤二 配置模型间路由策略

当请求未指定具体模型时，系统将根据当前实时状态（如负载、延迟）或用户意图，智能选择最合适的模型进行处理。

策略	说明
简单随机路由	在可用模型中随机选择。
最低系数路由	优先分发到积分较低的模型。
意图路由	叠加增强路由，根据用户意图智能分级，每级可选多个模型并复用模型间路由策略。

意图路由配置指导（可选）

1. 在意图路由中单击**新建规则**。

2. 在新建意图路由规则中，添加**意图路由规则名称**，配置复杂度，单击**确定**。
3. 配置完成后，可查看规则或对已有规则进行编辑。

步骤三 配置模型内路由策略

当模型确定后，系统将根据实时性能指标（如可用性、响应速度），从该模型下不同的服务所属厂商（或 API 密钥）中，动态选择最优的访问节点。

策略	说明
简单随机路由	在可用模型中随机选择。
最低繁忙路由	将请求分配给当前最空闲的模型。
最低延迟路由	自动选择当前延迟最低的模型。
用量均衡路由	按用量均衡分配请求到各模型。

步骤四 配置 Fallback 策略

当主模型服务失败时，系统将自动切换至 Fallback 中的模型，保障业务连续性。系统采用两层故障退避机制：

- 第一层（模型内退避）：优先在同一模型下的不同服务供应商（或 API 密钥）之间进行切换尝试。
- 第二层（模型间退避）：若当前模型无可用供应商，则根据您预设的模型优先级，自动切换到备选模型继续提供服务。

1. 在 Fallback 策略中，单击**去设置**。

2. 在编辑 Fallback 策略中，选择模型，并单击**确定**，完成配置。

后续操作

- 创建完成后，您可以在实例管理页面切换至**聊天测试**页签，对配置进行验证， 详细操作请参见 聊天测试 。
- 创建完成后，您可在实例管理页面切换至**用量详情**页签，关注 Token 消耗和模型资源包使用情况，避免因额度不足导致业务调用失败。

聊天测试

最近更新时间：2026-05-07 17:32:21

本文将为您介绍如何通过模型路由的聊天测试功能，验证模型配置是否正确、API Key 是否有效，以及模型响应质量是否符合预期。

对系统影响

- 聊天测试会实际调用模型 API，产生费用。
- 聊天测试记录仅保存在当前浏览器会话中，刷新页面后记录将丢失。
- 测试不会影响线上业务，仅用于验证和调试。

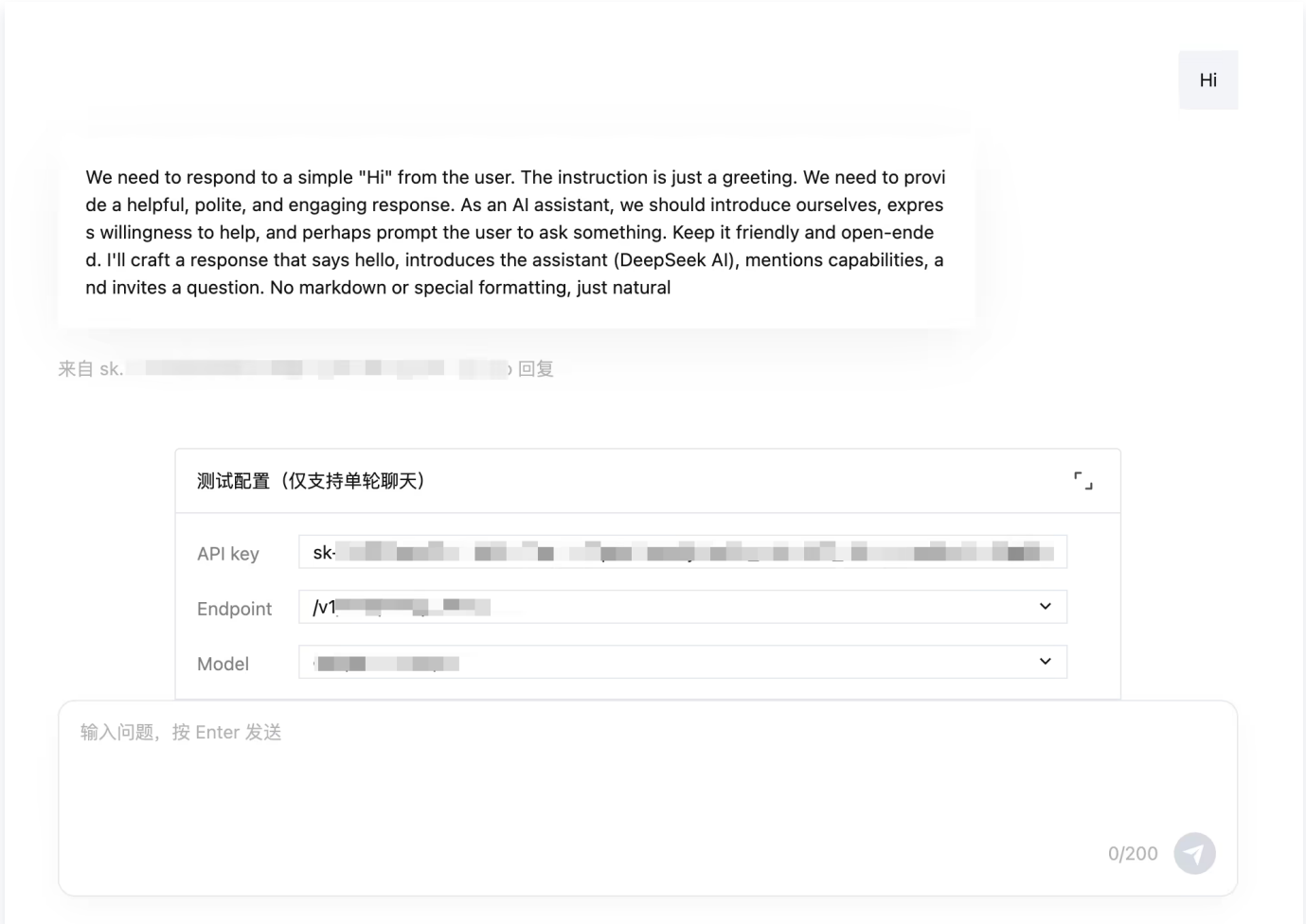
前提条件

1. 已获得 CLB 模型路由的使用资格。如未获得，请提交 工单申请 。建议提前阅读 使用约束 与 支持的模型提供商 了解相关信息。
2. 已创建模型路由实例，详细操作指导请参见 创建模型路由实例 。
3. 已完成新增 BYOK 模型，详细操作指导请参见 BYOK 模型介绍 。
4. 已创建模型路由访问密钥（API Key），详细操作指导请参见 创建模型路由访问密钥 。
5. 已完成模型关联和模型路由策略配置，详细操作指导请参见 配置模型调度管理 。

操作步骤

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例名称**，进入目标实例的**实例管理**页面。
2. 切换至**聊天测试**页签，输入 **API Key**，并选择指定模型。

3. 在底部的输入框中，输入测试消息（例如"Hi"），单击**发送**。
4. 等待模型响应（通常在 30 秒内返回）。



在聊天窗口中查看模型响应，重点关注以下信息：

信息项	说明
响应内容	模型返回的文本内容，用于判断模型是否按预期工作。
响应时间	模型响应的耗时（毫秒），用于评估模型性能。
Token 消耗	本次测试消耗的 Token 数量（输入 + 输出），用于估算费用。
错误信息	若测试失败，会显示错误码和错误描述，可根据错误信息进行排查。

后续操作

在实例管理页面中，切换至**用量详情**页签，关注 Token 消耗和模型资源包使用情况，避免因额度不足导致业务调用失败。

常见错误及解决方法

错误信息	可能原因	解决方法
API Key 无效 或 Unauthorized	API Key 错误或未授权	在 BYOK 页面检查 API Key 是否正确，必要时重新录入。
余额不足 或 Insufficient Balance	API Key 额度已用完	更换 API Key，或联系供应商充值。
模型不存在 或 Model Not Found	模型 ID 配置错误	检查模型配置中的模型 ID 是否与供应商提供的模型 ID 一致。
请求超时 或 Request Timeout	网络问题或模型响应慢	检查网络连接，或增加超时时间（如有配置项）。
速率限制 或 Rate Limit Exceeded	触发供应商的速率限制	降低请求频率，或升级供应商套餐。

用量详情

最近更新时间：2026-05-08 18:06:44

腾讯云可观测平台为 CLB 模型路由产品提供数据收集与展示功能。腾讯云默认为所有用户开通，无需手动配置。只要您使用了 CLB 模型路由，腾讯云可观测平台即可自动收集相关监控数据。

监控指标说明

CLB 模型路由提供如下监控指标。这些指标反映路由模型实例、API Key 及模型的使用状态和用量信息等。支持1分钟、5分钟、1小时和1天四种时间粒度。

类型	指标	说明	单位
核心用量	Token 总数(Count)	调用对话类模型时，输入和输出的 Token 数量总和。	个
	输入 Token 数(Count)	调用对话类模型时，输入的 Token 数量总和。	个
	输出 Token 数(Count)	调用对话类模型时，输出的 Token 数量总和。	个
	请求积分消耗(Count)	根据配置的积分计算系数以及本次请求消耗的输入 Token 数、输出 Token 数计算	个
请求信息	CMR 成功请求次数(Count)	成功的 CMR 请求总数。	个
	CMR 失败请求次数(Count)	失败的 CMR 请求总数。	个
	CMR 调用上游模型失败的请求次数(Count)	CMR 调用上游模型失败的请求总数。	个
	CMR 请求总数(Count)	CMR 请求总数。	个
	请求返回的400状态码个数(Count)	CMR 返回400状态码含义为请求参数错误，常见原因包括：请求参数错误、上下文窗口超限等。	个
	请求返回的401状态码个数(Count)	CMR 返回401状态码含义为鉴权失败，常见原因包括：访问 BYOK 模型时用户提供的 API Key 无效或过期、请求未携带 API Key 等。	个

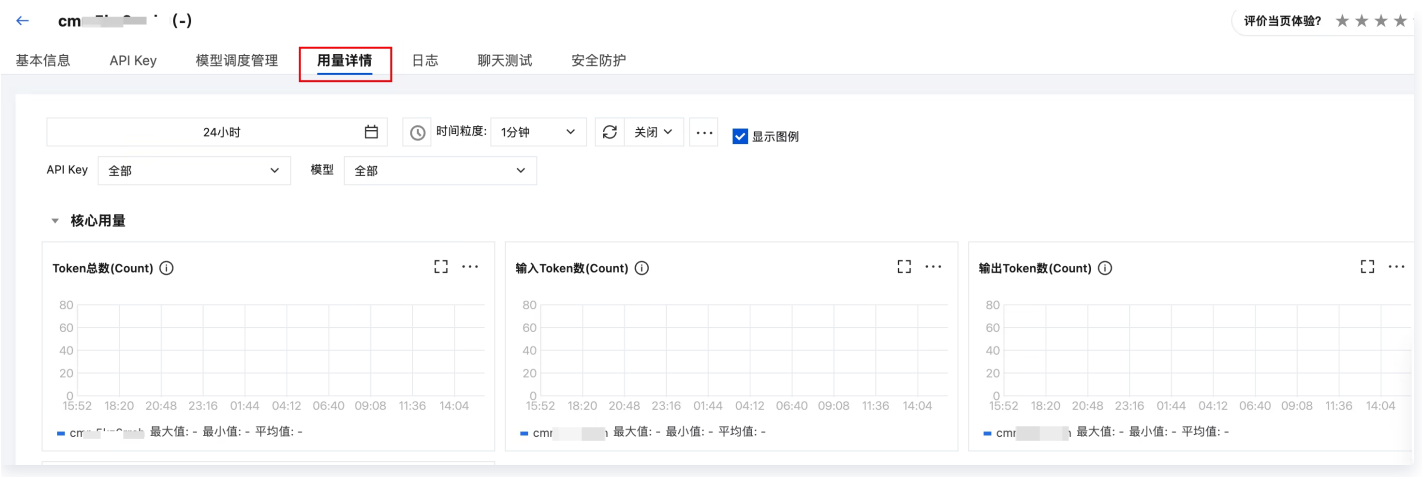
	请求返回的403状态码个数(Count)	CMR 返回403状态码含义为权限不足，常见原因包括：访问 BYOK 模型时 API Key 没有访问请求中模型的权限、厂商侧帐户被暂停或受限等。	个
	请求返回的404状态码个数(Count)	CMR 返回404状态码含义为资源不存在，常见原因包括：请求的模型名称在厂商侧不存在、BYOK 模型自定义 API Base 的路径错误等。	个
	请求返回的408状态码个数(Count)	CMR 返回408状态码含义为请求超时，常见原因包括：上游模型响应超时、与上游模型建立连接超时等。	个
	请求返回的422状态码个数(Count)	CMR 返回422状态码含义为请求不可处理，常见原因为请求体语义错误。	个
	请求返回的429状态码个数(Count)	CMR 返回429状态码含义为请求被限流，常见原因包括：每分钟请求数、消耗 Token 数超过上游模型厂商配额、并发请求数超过上游模型厂商限制等。	个
	请求返回的500状态码个数(Count)	CMR 返回500状态码含义为上游模型服务端内部错误，常见原因包括：上游模型服务端内部异常、上游模型 API 连接失败等。	个
	请求返回的502状态码个数(Count)	CMR 返回502状态码含义为上游模型厂商网关错误，常见原因包括：上游模型服务不可达、上游模型网关层异常等。	个
	请求返回的503状态码个数(Count)	CMR 返回503状态码含义为上游模型服务不可用，常见原因包括：上游模型服务暂不可用、特定模型负载过高暂不可用、流式响应过程中连接中断等。	个
用量明细	读缓存 Token 数(Count)	部分模型支持 Input 方向命中缓存的 Token 计数。	个
	写缓存 Token 数(Count)	部分模型支持 Input 方向写缓存的 Token 计数。	个
	常规非缓存 Token 数(Count)	Input 方向没有命中任何缓存的 Token 计数。	个
	上游模型内置工具使用次数(Count)	部分模型支持工具调用的次数计数。	个
	推理思考 Token 数(Count)	部分模型支持把推理思考部分的 Token 进行计数。	个
	常规文本输出 Token 数(Count)	模型 Output 方向输出的 Token 计数。	个

时延	上游模型调用时延(ms)	模型路由访问上游模型的调用时延。	ms
	流式请求的首 Token 时延(ms)	模型路由从输入到输出首个 Token 间隔的时间。	ms
	CMR 自身处理开销时延(ms)	模型路由自身进行处理逻辑的耗时。	ms
	CMR 请求时延(ms)	CMR 请求时延，为上游模型调用时延与 CMR 自身处理开销时延之和。	ms
上游模型	上游模型失败响应次数(Count)	上游模型失败响应次数	个
	上游模型请求总数(Count)	上游模型的总请求次数	个
	上游模型成功响应次数(Count)	上游模型成功响应次数	个
	上游模型 fallback 调用成功次数(Count)	上游模型 fallback 调用成功次数	个
	上游模型 fallback 调用失败次数(Count)	上游模型 fallback 调用失败次数	个

查看监控指标

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例名称**，进入目标实例的**实例管理**页面。

2. 切换至**用量详情**页签，查看相关指标。也可以指定 API Key 或者模型进行筛选查看。


(-)
评价当页体验?
基本信息
API Key
模型调度管理
用量详情
日志
聊天测试
安全防护
24小时
时间粒度:
1分钟
关闭
显示图例
API Key
全部
模型
全部
核心用量
Token总数(Count)
输入Token数(Count)
输出Token数(Count)
80
60
40
20
0
15:52
18:20
20:48
23:16
01:44
04:12
06:40
09:08
11:36
14:04
最大值: - 最小值: - 平均值: -

日志

最近更新时间：2026-05-07 17:32:38

CLB 模型路由提供基本的日志功能，您可以在实例详情中查看最近 24 小时内的 API 请求记录，包括请求状态、Token 消耗、响应耗时等信息，帮助您快速排查问题与分析调用情况。

使用限制

- 仅支持查询最近 24 小时内的请求日志。
- 后续将支持接入 日志服务 CLS，实现更长时间的日志存储与高级检索能力。

查看日志

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例名称**，进入目标实例的**实例管理**页面。

2. 切换至**日志**页签，查看请求信息。

3. 通过页面顶部的时间按钮筛选日志范围，支持以下选项：

选项	说明
近 30 分	查看最近 30 分钟内的请求。
近 1 时	查看最近 1 小时内的请求。
近 3 时	查看最近 3 小时内的请求。
近 12 时	查看最近 12 小时内的请求。

近 24 时	查看最近 24 小时内的请求。
自定义	手动选择起止时间范围。

日志字段说明

请求日志列表包含以下字段：

字段	说明
Key ID	本次请求使用的 API Key 标识。
模型名称	本次请求实际调用的模型名称。
所属厂商	模型所属的提供商名称。
状态	请求的响应状态，可据此判断调用是否成功。
总 Token	本次请求消耗的 Token 总量（含输入与输出）。
请求耗时	从发起请求到收到完整响应的时间。
请求 IP	发起本次请求的客户端 IP 地址。
开始时间	请求发起的时间。

BYOK 模型配置管理
BYOK 模型介绍

最近更新时间：2026-05-07 17:34:31

CLB 模型路由全面支持自带密钥（Bring Your Own Key，BYOK）模式。您可以灵活选择以下三种方式接入模型：

- **原厂模型**：直接使用 DeepSeek 等模型厂商官方提供的 API 密钥，享受原厂性能与最新能力。平台自动补全 API Base，用户仅需提供官方 API Key 即可接入，支持公网加速，适合快速上手。
- **第三方代理**：接入由代理商、云市场等提供的合规模型服务密钥，满足特定的采购或区域要求。用户可自定义 API Base，统一管理 Key 与模型配置，灵活切换厂商，适合有降本需求或多厂商聚合场景。
- **自建模型**：对接您私有化部署或本地化模型服务的密钥，数据不出域，满足您的合规严要求。通过 VPC 内网或专线直连企业自建 GPU 集群与 IDC 机房，数据全程不出网，适合对数据主权和安全合规有严格要求的场景。

原厂模型

最近更新时间：2026-05-08 17:59:21

本文介绍如何创建和使用**原厂模型**类型 BYOK 实例。

操作指导

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK**。
2. 在 BYOK 页面，单击**新建**，并选择**原厂模型**。

3. 在**新建原厂模型弹窗**中，按需选择或填写相关参数，并单击**确定**。

参数	说明
所属厂商	可通过下拉选项选择您的模型提供商（如 DeepSeek）。若无合适选项，可手动输入并单击 Enter 键确认。 同时需要选择对应模型提供商的兼容协议（如 DeepSeek、OpenAI）。
API Key	按需填写您自有的 API Key，如果您有多个 API Key 可依次填入。模型路由会随机的使用其中的 Key，如果有 Key 不可用，模型路由会自动切换至其他的可用 Key。 添加后即可单击**检查**，探测 API Key 的健康情况。也可在所有的 API Key 添加完毕后，在下方的健康检查中，单击**开始检查**，统一探测。

选择模型	所属厂商和 API Key 填入后，在此处下拉框中选择即可。若没有展示完全，可单击输入框后面的**探测其他模型**，进行探测和刷新。
健康检查	单击**开始检查**，可统一探测 API Key 的健康情况，方便您及时剔除不健康的 API Key。
实例名称	选填，最多支持 255 个字符。
标签	选择标签键和标签值，也可选择添加标签，详情请参见 创建标签 。

第三方代理

最近更新时间：2026-05-07 17:34:31

本文介绍如何创建和使用**第三方代理**类型 BYOK 实例。

操作指导

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK**。
2. 在 BYOK 页面，单击**新建**，并选择**第三方代理**。

3. 在**新建第三方代理弹窗**中，按需选择或者填写相关参数，并单击**确定**。

参数	说明
所属厂商	可通过下拉选项选择您的模型提供商（如 DeepSeek）。若无合适选项，可手动输入并单击 Enter 键确认。 同时需要选择对应模型提供商的兼容协议（如 DeepSeek、OpenAI）。
API 地址	业务请求使用的就是用户输入的 API 地址。不同厂商叫法可能不同，（如 API Base 或 Base URL）。以 DeepSeek 为例，其 API 地址为：https://api.deepseek.com/v1 。
API Key	按需填写您自有的 API Key，如果您有多个 API Key 可依次填入。模型路由会随机的使用其中的 Key，如果有 Key 不可用，模型路由会自动切换至其他的可用 Key。

	添加后即可单击**检查**，探测 API Key 的健康情况。也可在所有的 API Key 添加完毕后，在下方的健康检查中，单击**开始检查**，统一探测。
选择模型	所属厂商、API 地址和 API Key 填入后，在此处下拉框中选择即可。若没有展示完全，可单击输入框后面的**探测其他模型**，进行探测和刷新。
健康检查	单击**开始检查**，可统一探测 API Key 的健康情况，方便您及时剔除不健康的 API Key。
实例名称	选填，最多支持 255 个字符。
标签	选择标签键和标签值，也可选择添加标签，详情请参见 创建标签 。

自建模型

最近更新时间：2026-05-08 17:59:01

本文介绍如何创建和使用**自建模型**类型 BYOK 实例。

操作指导

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK**。
2. 在 BYOK 页面，单击**新建**，并选择**自建模型**。

3. 在**新建自建模型弹窗**中，完成网络配置并单击**下一步：模型& Key 配置**。

参数	说明
所属网络	选择所属网络与子网 IP。
实例名称	选填，最多支持 255 个字符。
标签	选择标签键和标签值，也可选择添加标签，详情请参见 创建标签 。

4. 按需选择或者填写**模型& Key 配置**相关参数，并单击**完成**。

参数	说明

所属厂商	可通过下拉选项选择您的模型提供商（如 DeepSeek）。若无合适选项，可手动输入并单击 **Enter** 键确认。 同时需要选择对应模型提供商的兼容协议（如 DeepSeek、OpenAI）。
API 地址	业务请求使用的就是用户输入的 API 地址。不同厂商叫法可能不同，（如 API Base 或 Base URL）。以 DeepSeek 为例，其 API 地址为：https://api.deepseek.com/v1 。
域名	往上游模型发送请求时携带的 http header。
API Key	按需填写您自有的 API Key，如果您有多个 API Key 可依次填入。模型路由会随机的使用其中的 Key，如果有 Key 不可用，模型路由会自动切换至其他的可用 Key。 添加后即可单击**检查**，探测 API Key 的健康情况。也可在所有的 API Key 添加完毕后，在下方的健康检查中，单击**开始检查**，统一探测。
选择模型	所属厂商、API 地址和 API Key 填入后，在此处下拉框中选择即可。若没有展示完全，可单击输入框后面的**探测其他模型**，进行探测和刷新。
健康检查	单击**开始检查**，可统一探测 API Key 的健康情况，方便您及时剔除不健康的 API Key。

管理 BYOK 模型

最近更新时间：2026-05-09 11:53:20

本文介绍如何管理 BYOK 模型，包括 Key 管理、模型管理与查看用量详情。

Key 管理

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK**，在列表页中可见已创建的 BYOK 模型。
2. 点击具体的 BYOK 模型进入**基本信息页**，支持操作添加/删除 Key。或在列表页操作列中点击**管理 Key** 可直接跳转至 **基本信息页**，支持操作添加/删除 Key。添加 Key 需要填写参数说明如下。

参数	说明
API Key	按需填写您自有的 API Key，如果您有多个 API Key 可依次填入。模型路由会随机的使用其中的 Key，如果有 Key 不可用，模型路由会自动切换至其他的可用 Key。 添加后即可单击**检查**，探测 API Key 的健康情况。也可在所有的 API Key 添加完毕后，在下方的健康检查中，单击**开始检查**，统一探测。

> **说明：**
> BYOK 模型需要至少保留一个 Key，模型仅有一个 Key 的情况下不允许删除该 Key。

模型管理

点击具体的 BYOK 模型进入**基本信息页**，支持操作添加/删除模型。添加模型需要填写参数说明如下。

参数	说明
API Key	按需填写您自有的 API Key，如果您有多个 API Key 可依次填入。模型路由会随机的使用其中的 Key，如果有 Key 不可用，模型路由会自动切换至其他的可用 Key。 添加后即可单击**检查**，探测 API Key 的健康情况。也可在所有的 API Key 添加完毕后，在下方的健康检查中，单击**开始检查**，统一探测。
选择模型	填入 API Key 后，在此处下拉框中选择即可。若没有展示完全，可单击输入框后面的**探测其他模型**，进行探测和刷新。

> **说明：**
> 仅完成网络配置未完成模型 & Key 配置的自建模型，可以在列表页操作列中点击**配置模型**继续完成模型 & Key 配置。

查看用量详情

点击具体的 BYOK 模型进入**基本信息页**，切换至**用量详情页**，即可查看用量详情，详情可参见 用量详情 。

积分管理
积分管理介绍

最近更新时间：2026-05-09 11:53:33

积分管理旨在为用户提供灵活、可控的模型使用与资源消耗管理能力。通过配置模型系数和积分预算，实现资源分配的精细化控制，帮助团队或项目有效规划、监控与优化 token 消耗。
资源可控：预算与资源绑定，实现消耗的精细化管控。
成本可视：实时查看积分消耗情况，支持成本分析与预测。
灵活适配：模型系数可调，适应不同业务场景与成本策略。
无缝集成：关联后即可自动启用，无需调整业务代码。

核心功能说明

模型系数

支持为不同模型设置不同的系数，包含输入系数和输出系数。您可根据模型能力、性能、成本等因素，自定义积分消耗比例。从而实现多模型使用场景下的成本差异化控制。比如直接将系数与模型单价同比例配置，假如模型单价为2.5元/百万 token，系数则配置为25。

积分预算

您可根据需要设置不同的积分预算，积分预算包含积分最大消耗、重置周期和速率限制。支持为不同的 CLB 模型路由实例或 API Key 绑定相同或者不同的积分预算，即授权不同的额度，从而实现对 token 消耗的主动控制与预警。

> **说明：**
> - 一个积分预算可同时关联多个实例或者多个 API Key，关联后对应对象的调用将受该预算配置的约束。
> - API Key 绑定积分预算后，除了受到 API Key 本身的积分预算限制外，还受到模型路由实例上所配置的速率限制或积分预算的限制。

比如积分预算 test-A，最大积分额度是1百万，积分预算 test-B，最大积分额度是1千。模型路由实例 A 下面有两个 API Key，分别为 Key_01 和 Key_02。
模型路由实例 A 的积分消耗=Key_01的积分消耗 + Key_02 的积分消耗
场景一：模型路由实例和其中某个 API Key绑定相同的积分预算
模型路由实例 A 额度耗尽后，Key_01 和 Key_02 均无法继续使用，即使 Key_01 自身仍有余额

关联对象	关联关系	最大积分额度
模型路由实例 A	积分预算 test-A	1百万

Key_01	积分预算 test-A	1百万
Key_02	不绑定积分预算	无积分预算限制（实际会有速率限制，在这里不继续展开）

场景二：模型路由实例下的 API Key 绑定不同积分预算

Key_02 的积分预算额度较小，一旦耗尽将会停止服务，即使此时模型路由实例 A 还有额度。

关联对象	关联关系	最大积分额度
模型路由实例 A	积分预算 test-A	1百万
Key_01	积分预算 test-A	1百万
Key_02	积分预算 test-B	1千

典型使用场景

- 团队项目管理为不同项目组分配独立 API Key 及积分预算，实现成本分拆与管控。
- 多模型成本优化通过调整模型系数，引导业务方在满足需求的前提下选用更具性价比的模型。
- 资源预警与防控为生产环境 CLB 模型路由实例或 API Key 设置预算预警，避免突发流量导致积分超额消耗。

常见问题

1: 积分预算用完后会发生什么?

当积分预算耗尽时，关联的 CMR 实例或 API Key 将停止服务，直到预算重置或增加。

2: 模型系数调整后何时生效?

系数调整将立即生效，无需重启服务。

3: 如何查看历史积分消耗记录?

可以在用量详情中通过请求积分消耗(Count)指标查看。

相关文档

- 配置模型系数
- 配置积分预算

配置模型系数

最近更新时间：2026-05-09 11:54:01

本文档介绍在腾讯云 CLB 模型路由控制台中配置积分管理模型系数的操作方法。

前提条件

- 已获得 CLB 模型路由的使用资格。如未获得，请提交 工单申请 。建议提前阅读 使用约束 与 支持的模型提供商 了解相关信息。
- 已完成新增 BYOK 模型，详细操作指导请参见 BYOK 模型介绍 。BYOK 模型新增后，在积分管理的模型系数页面自动出现，无需手动操作。

操作步骤

1. 登录 模型路由 控制台，在左侧导航栏选择**积分管理**。
2. 在**模型系数**页面，找到目标模型，在操作列单击**编辑**。
3. 在**编辑系数**弹窗，配置**输入系数**和**输出系数**，并单击**确定**完成配置。
4. （可选）也可在目标模型的**输入系数**列和**输出系数**列，通过单击**编辑图标**，单独配置。

其他操作

单击目标模型用量详情列的**图标**，可查看模型的相关用量指标。

注意事项

- 模型系数修改后，立即生效。请根据实际业务需求合理配置系数。
- 模型系数的修改可能影响积分的消耗速度，建议提前评估影响。
- 当模型删除时，与模型有关的模型系数将会一并删除。

配置积分预算

最近更新时间：2026-05-09 11:54:01

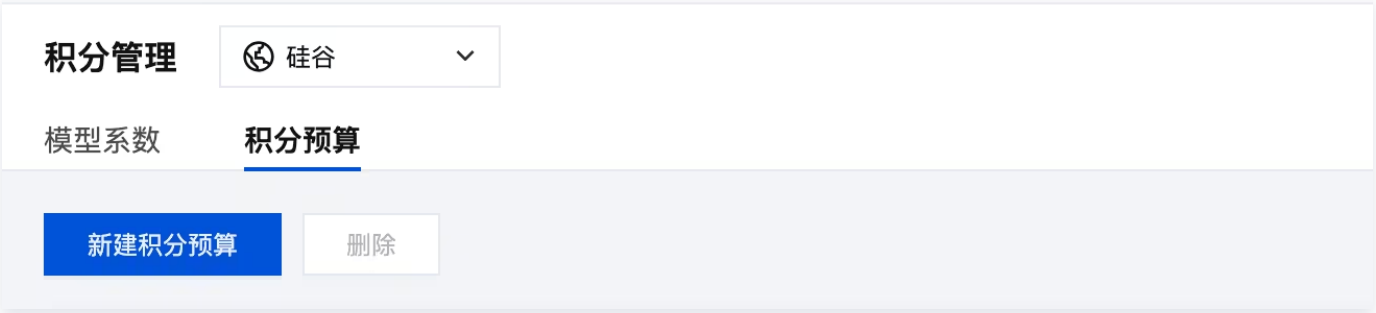
本文档介绍在腾讯云 CLB 模型路由控制台中配置积分预算的操作方法。通过配置积分预算，控制模型调用的消费额度，并将积分预算关联到指定实例或 API Key，实现精细化的成本管控。

前提条件

- 已获得 CLB 模型路由的使用资格。如未获得，请提交 工单申请 。建议提前阅读 使用约束 与 支持的模型提供商 了解相关信息。
- 已完成新增 BYOK 模型，详细操作指导请参见 BYOK 模型介绍 。BYOK 模型新增后，在积分管理的模型系数页面自动出现，无需手动操作。

新建积分预算

1. 登录 模型路由 控制台，在左侧导航栏选择**积分管理**。
2. 在**积分预算**页面，单击**新建积分预算**。



3. 在**新建积分预算**弹窗，按需选择或者填写相关参数，并单击**下一步关联对象**。

参数	说明
积分预算名称	自定义预算名称，用于标识不同预算用途。最多支持 255 个字符。
预算配置	包含最大积分额度和重置周期。 • 最大积分额度：预算周期内允许消耗的最大积分数，关联对象消耗达到此额度时将被自动拦截。单位可选千、万、百万、亿。积分消耗 = 输入 Token（每百万） × 输入系数 + 输出 Token（每百万） × 输出系数。实际消耗取决于每次调用的 Token 数量和使用的模型（主要取决于 模型配置的系数 ）。 • 重置周期：积分额度自动重置的时间间隔，可选1天、7天、30天。
限速配置	可针对 TPM 和 RPM 进行配置。 TPM：Token per minute，每分钟允许处理的最大 Token 数，单位：千/分钟。 取值范围：1-100000。

	RPM：Requests per minute，每分钟允许处理的最大请求次数，单位：次/分钟。 取值范围：1–10000。

4. 可选择关联实例或者 API Key，也可以暂时不选择关联对象，直接单击**完成**配置。

编辑积分预算

1. 在**积分预算**页签的列表中，找到目标预算条目。
2. 在操作列单击**编辑**。
3. 按需修改积分预算名称、积分预算配置或限速配置等参数。
4. 单击**确定**保存修改。

关联对象

1. 在**积分预算**页签的列表中，找到目标预算条目。
2. 在操作列单击**关联对象**。
3. 选择需要关联的实例或 API Key，单击**确定**。

> **说明：**
> - 一个积分预算可同时关联多个实例或者多个 API Key，关联后对应对象的调用将受该预算配置的约束。
> - API Key 绑定积分预算后，除了受到 API Key 本身的积分预算限制外，还受到模型路由实例上所配置的速率限速或积分预算的限制。

解除关联对象

1. 在**积分预算**页签的列表中，找到目标预算条目。
2. 在关联对象列单击**关联对象**。
3. 在弹窗中勾选后单击**批量解关联**或者在指定目标的操作列，单击**解关联**。
4. 确认解关联对象信息，并单击**确认**，完成解除关联对象操作。

删除积分预算

1. 在**积分预算**页签的列表中，勾选目标预算条目。
2. 单击列表上方的**删除**，或在操作列单击**删除**。
3. 在确认弹窗中单击**确定**。

> **说明：**
> 无法删除已有关联对象的积分预算，请先解除绑定对象，再执行删除操作。