模型路由

操作指南

腾讯云

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

操作指南
创建模型路由实例

最近更新时间：2026-09-15 15:34:31

本文介绍如何创建模型路由实例。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。

简介

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。
- 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。

前提条件

建议提前阅读 使用约束 与 支持模型 了解相关信息。

操作步骤

1. 登录 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
地域	选择实例地域。
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	监听协议可选 HTTP（80）、HTTPS（443）。网络类型选择公网时，监听协议建议选择 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。
计费模式	支持按规格计费和按用量计费两种模式。 • 按规格计费：不同规格的 CMR 实例对应的费用不同，超过规格限制会自动限流。仅企业型实例支持。 • 按用量计费：零调用零费用，若用量较大可能产生高额账单。

资源包抵扣	• 按规格计费：可选择是否使用资源包抵扣。若选择资源包抵扣，资源包用尽后实例将转为后付费按小时扣费。 • 按用量计费：仅支持选择资源包抵扣。资源包用尽后实例将停止服务，请及时续费以保障业务连续性。
规格	选择按规格计费时，可按需选择不同规格，不同规格的 TPM 限速值和费用，详见 **计费概述** 。
限制类型	默认可设置 TPM 和 RPM，也可选择合适的 **积分预算** 。其中，按规格计费的实例，TPM 限速值默认为所选规格的上限，不支持修改。
TPM	每分钟允许处理的最大 Token 数（Tokens Per Minute），单位：千/分钟。 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。 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。
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** 。

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。

后续操作

1. 创建模型路由访问密钥，详细操作指导请参见 **创建模型路由访问密钥** 。
2. 新增 BYOK 模型，详细操作指导请参见 **创建 BYOK 实例** 。
3. 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** 。

创建模型路由访问密钥

最近更新时间：2026-09-15 15:34:31

本文介绍如何创建和使用**模型路由访问密钥**（API Key）。模型路由根据 API Key 获取身份凭证，以完成模型全链路管控与调度。

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）**：调用方访问模型路由实例时使用的身份凭证。模型路由根据 API Key 完成请求鉴权，并执行模型访问、用量统计、限流和预算控制。
- **BYOK 密钥**：用户在第三方模型服务提供商处持有的自有密钥，用于模型路由连接原厂模型、第三方代理模型或用户自建模型。

模型路由访问密钥（API Key）属于实例级资源。在一个模型路由实例中创建或导入的 API Key，只能用于访问当前实例，不支持跨实例使用。
一个模型路由实例可以创建多个 API Key。新建或导入的 API Key 默认归属系统内置的**未分组 API Key** 用户组。后续可根据管理需求，将 API Key 分配至其他用户组。关于用户组的详细介绍，请参见 用户组介绍 。

> **说明：**
> - 模型路由实例、用户组和 API Key 均可配置积分预算及速率限制。当多个层级同时配置限制时，以最先达到的额度或速率限制为准。
> - BYOK 模型产生的模型调用费用由相应模型服务提供商收取。模型路由不会重复收取模型调用费用，仅收取模型路由处理费，详见 计费概述 。

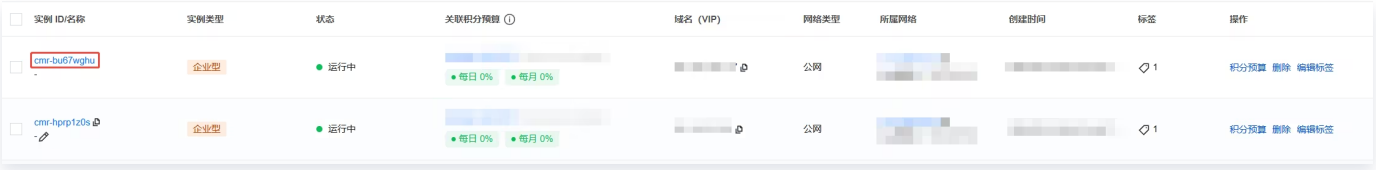
前提条件

1. 建议提前阅读 使用约束 与 支持模型 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 。

操作步骤

步骤1：创建 API Key

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例 ID**。



2. 在实例详情页，选择**用户组**页签。

3. 在**用户组**页签中，单击**新建 Key**。

4. 在**新建 Key** 界面中配置 **Key 名称**等参数，并单击**确定**。
5. **API Key** 创建成功后，会在创建成功的弹窗中展示 API Key 信息，请您妥善保管。您也可以单击**下载到本地**，进行保存。

> **说明：**
> 1. 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。
> 2. 新建 API Key 默认归属至**未分组 API Key** 用户组，后续可根据业务需要分配至其他用户组。

步骤2：批量创建或导入 API Key

1. 如果需要批量创建密钥，在**用户组**页签中，单击**批量新建 Key**。

2. 在**批量新建 Key** 界面中，可以选择**自动生成 Key** 或**导入已有 Key**。

按照弹窗中的格式要求输入配置信息，单击**确定**，即可批量创建或导入 Key。

- **自动生成 Key**：输入要批量生成的 **Key 名称**，可用中英文逗号、换行分隔多值，支持批量粘贴。
- **导入已有 Key**：每行输入一个要导入的 Key，格式为：**Key 名称，Key 内容**（中英文逗号隔开）。

> **说明：**
> **导入已有 Key** 用于导入访问模型路由实例的凭证，不用于导入模型服务提供商的 API Key。

3. 批量创建或导入 API Key 完成后，系统将在创建成功的弹窗中展示 API Key 信息，请您妥善保管。您也可以单击**下载到本地**，进行保存。

> **说明：**
> 1. 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。
> 2. 批量创建或导入的 API Key 默认归属至**未分组 API Key** 用户组。

后续操作

1. 将 API Key 分配至相应用户组。详细操作指导请参见 分配路由访问密钥 。
2. 新增 BYOK 模型，详细操作指导请参见 创建 BYOK 实例 。
3. 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：2026-09-15 15:34:31

本文介绍如何关联模型，以及对模型调度策略进行管理。配置完成后，用户请求经由模型路由统一接入，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。

简介

模型路由调度管理支持以下路由策略：

- 意图路由：按输入意图匹配模型。
- 模型间路由：按路由策略选择模型。
- 模型内路由：按路由策略选择供应商（或 API 密钥）。
- 粘性路由：多轮会话场景支持粘性路由。
- Fallback：支持模型内和模型间两层 Fallback。

前提条件

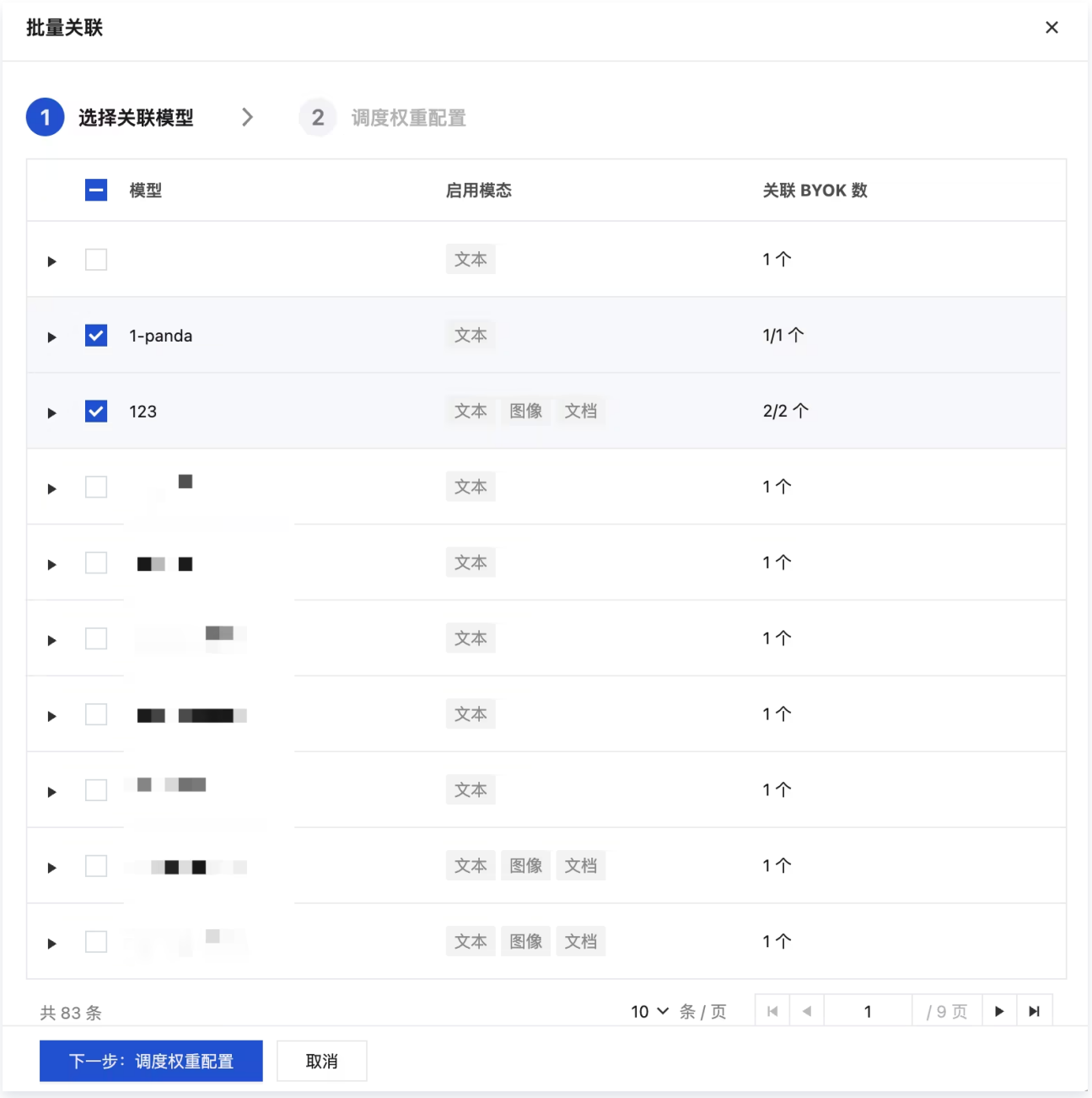
- 建议提前阅读 使用约束 与 支持模型 了解相关信息。
- 已创建模型路由实例，详细操作指导请参见 创建模型路由实例 。
- 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。
- 已创建模型路由访问密钥（API Key），详细操作指导请参见 创建模型路由访问密钥 。

操作步骤

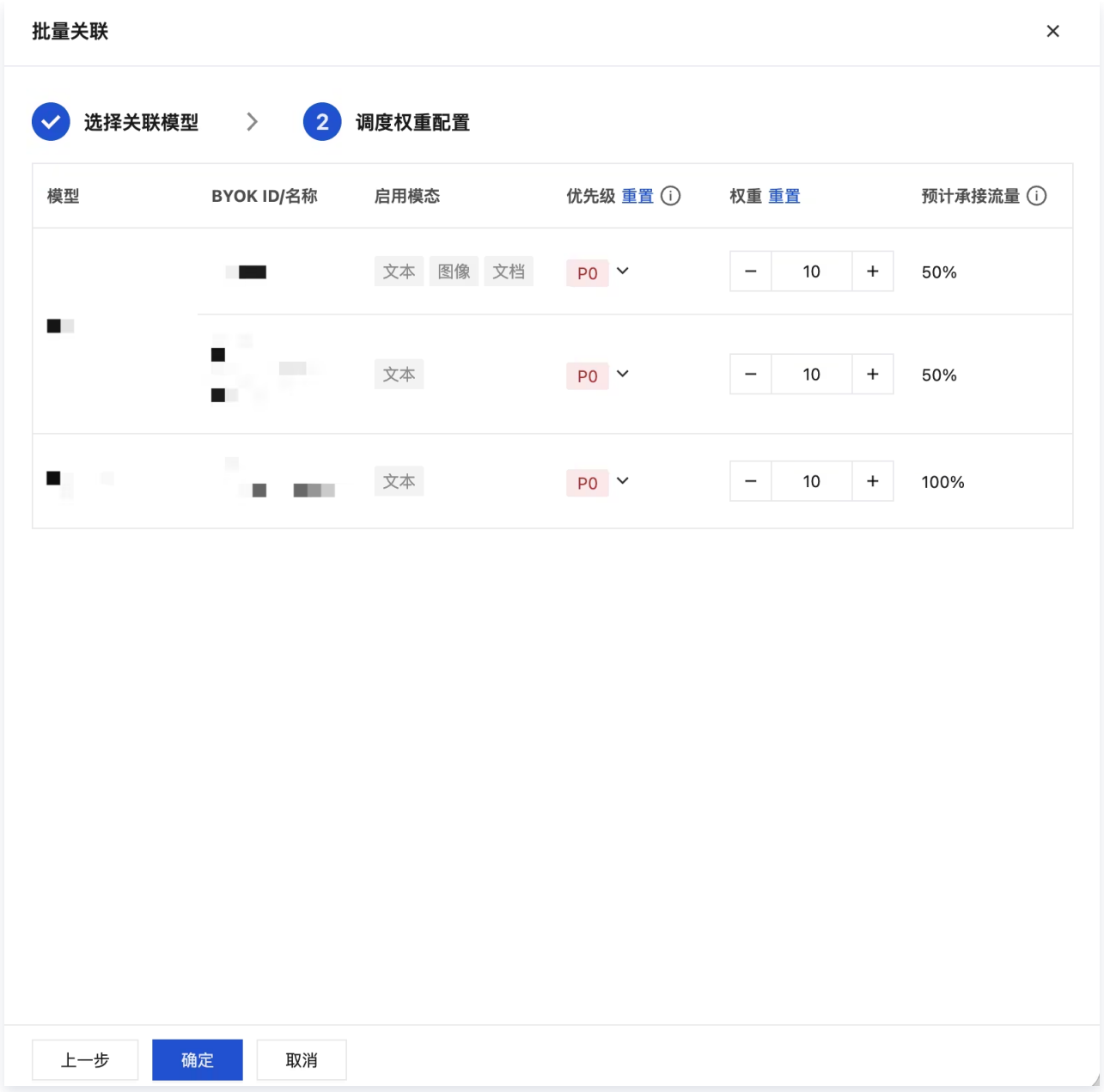
步骤1：关联模型

1. 登录 模型路由 控制台，在**实例管理**页面，单击目标**实例 ID**，进入目标实例的基本信息页面。
2. 切换至**模型调度管理**页签，在**路由策略**示意图中的**关联模型**节点，单击**关联**。或者在页面**关联模型**模块单击**批量关联**。

3. 确认模型名称、启用模态等信息，选择要关联的模型，并确认信息无误后单击**下一步：调度权重配置**。



4. 为所选模型配置优先级与权重。网关将按 P0 → P1 → P2 顺序选择模型，高优先级不可用时自动降级。且同一优先级按权重排序选择模型，并按同一优先级下的权重占比，计算出预计承接流量。确认信息无误后单击**确定**。



5. 关联完成后，可在列表查看已关联模型的模型名称、来源、启用模态、模型连接状态、BYOK 数量、优先级与权重等信息，并可以执行**重定向模型**、**解关联**和**调度配置**。
6. 已关联模型支持配置**重定向**。当用户请求中传入已关联的模型 ID 时，可通过重定向将该模型的实际调用模型指定为其他模型或意图路由规则；当 model=auto 或未传模型 ID 时，系统默认按照 L1 模型路由策略指定模型，您也可以通过重定向指定模型或意图路由规则。
7. 如需配置重定向，在目标模型所在行的**重定向模型**列单击**去配置**，选择**重定向类型**以及**目标模型**后完成配置。

重定向目标支持以下两种类型：

重定向类型	说明
重定向模型	当请求命中当前模型时，系统不再直接调用当前模型，而是调用重定向后指定的模型。
重定向意图路由规则	当请求命中当前模型时，系统不再直接调用当前模型，而是进入指定的意图路由规则，并由意图分类对应的处理模型继续完成请求分发。

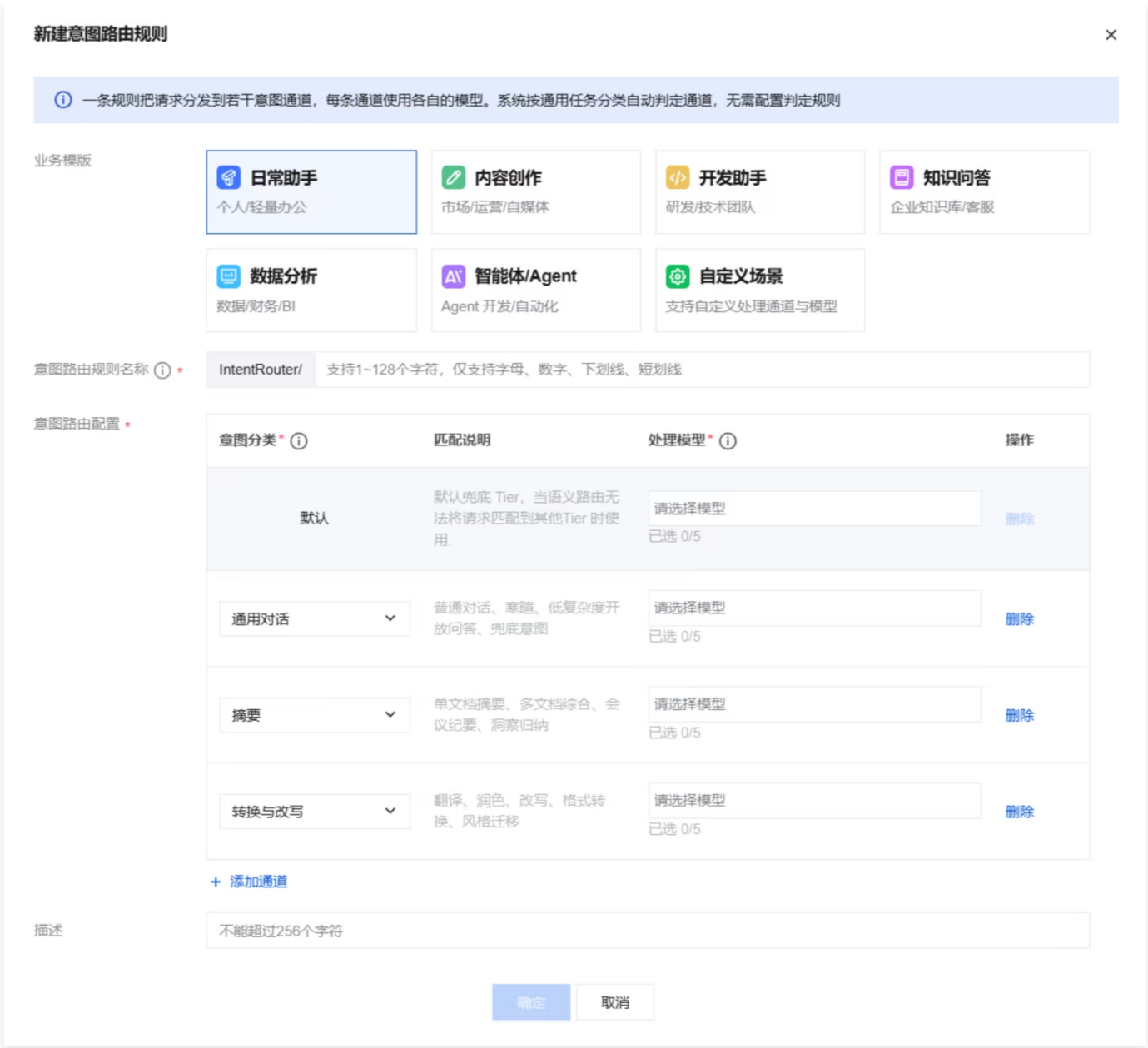
8. 如需解除模型关联，在目标模型所在行单击**解关联**；如需批量解除关联，可单击**批量解关联**。解除关联后，该模型将从当前实例移除，不再作为模型路由策略的可选模型。
9. 如需修改优先级权重配置，在目标模型所在行单击**调度配置**，重新为关联模型配置调度优先级与权重。

步骤2：配置意图路由（推荐）

> **说明：**
> 若不配置意图路由规则，则跳过该策略。

1. 在**模型调度管理**页签，**路由策略**示意图中的**意图路由**节点，单击**新建**。或者在页面**意图路由**模块，单击**新建规则**。

2. 在**新建意图路由规则**弹窗，配置**业务模板**、**意图路由规则名称**、**意图路由配置**和**描述信息**。一条意图路由规则可将请求分发到多个意图分类，每个意图分类可配置各自的处理模型。系统按通用任务分类自动判定用户请求所属的意图分类，无需配置判定规则。



字段	说明
业务模板	用于快速生成不同业务场景下的意图路由配置。支持日常助手、内容创作、开发助手、知识问答、数据分析、智能体/Agent 等模板，并支持自定义场景。选择业务模板后，系统会显示对应的意图分类和匹配说明。
意图路由规则名称	意图路由规则名称，可按需输入。
意图路由配置	用于配置不同意图分类对应的处理模型。系统会根据用户请求内容自动匹配意图分类，并调用该意图分类下配置的处理模型。
意图分类	表示用户请求所属的任务类型，系统会根据用户消息自动判断意图分类。
处理模型	命中该意图分类后实际调用的模型。

	每个意图分类最多可选择 5 个处理模型；当配置多个处理模型时，系统将依据模型间路由策略进行调度。
默认	默认兜底分类。当意图路由无法将请求匹配到其他意图分类时，使用默认分类下配置的处理模型。 建议为默认分类配置可用模型，避免请求无法处理。
添加通道	单击添加通道，可新增一条意图分类配置，并为该意图分类选择处理模型。
删除	删除当前行的意图分类配置。
描述	可按需输入相关描述，帮助您识别和定位不同的意图路由规则。

3. 单击**确定**完成配置。

配置完成后，可执行如下操作：

- 单击**详情**：查看意图路由规则的配置信息。
- 单击**编辑**：对意图路由规则进行配置。
- 单击**删除**：删除指定意图路由规则。
- 单击**复制**：快速复制当前意图路由的配置。

步骤3：配置模型间路由策略

在**路由策略**示意图中的**模型间路由**节点，单击**切换**来配置模型间路由策略。当模型关联后，系统将根据所选路由算法，动态选择所关联模型下的不同模型。

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。

步骤4：配置模型内路由策略

在**路由策略**示意图中的**模型内路由**节点，单击**切换**来配置模型内路由策略。当模型确定后，系统将根据所选路由算法，动态选择该模型下的不同服务供应商（或 API 密钥）。

策略	说明
简单随机路由	在可用供应商中随机选择。
最低繁忙路由	将请求分配给当前最空闲的供应商。
最低延迟路由	自动选择当前延迟最低的供应商。
用量均衡路由	按用量均衡分配请求到各供应商。
最低系数路由	优先分发到积分较低的供应商。

步骤5：配置粘性路由策略

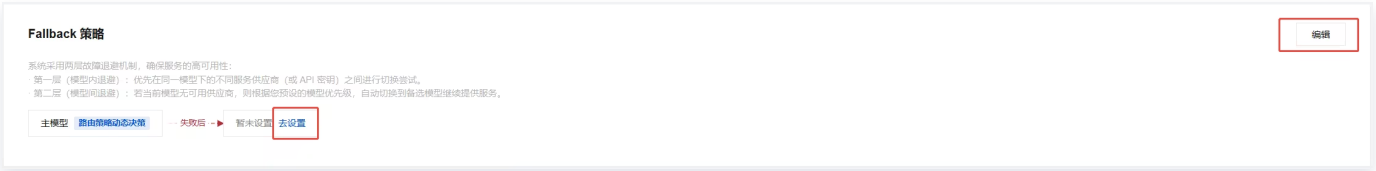
在**路由策略**示意图中的**粘性路由**节点，可配置粘性路由策略。粘性路由默认开启，可选择关闭。开启后可在多轮会话场景支持粘性路由，有效提升上游缓存命中率；会话空闲超过 30 分钟后，才会重新路由模型。

步骤6：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. 在路由策略示意图中 Fallback 节点，单击**设置**来配置 Fallback 策略。或点击 Fallback 策略区域的**编辑**按钮和底栏未配置 Fallback 策略时的**去设置**按钮。



2. 在**编辑 Fallback 策略**弹窗中，设置模型内递进次数，选择模型间 Fallback 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 配置聊天测试 。
- 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 消耗和模型资源使用情况，详细操作请参见 查看用量详情 。

配置提示词管理

最近更新时间：2026-09-15 15:34:31

本文介绍如何配置提示词，并对其进行版本管理与用户组关联配置。配置完成后，用户请求经由模型路由统一接入，平台完成计费、限流与日志记录后，根据您配置的提示词及改写策略，在请求转发至后端模型前注入对应的提示词，实现对企业级对话行为的统一约束与管控。

简介

提示词管理具有以下核心能力：

- 集中维护：所有提示词在统一页面创建维护，包括提示词内容、改写策略、最大 Token 数。
- 版本管理：每次改动生成版本，支持查看历史、一键变更。
- 按组下发：不同业务线关联不同提示词，规矩相互隔离。
- 零代码：注入和改写都在网关侧完成，业务方无需发版。

前提条件

1. 建议提前阅读 使用约束 与 支持模型 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. 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。
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. 已创建用户组，详细操作指导请参见 用户组介绍 。

操作步骤

步骤1：创建提示词

1. 登录 模型路由 控制台，在**实例管理**页面，单击目标**实例 ID**，进入目标实例的基本信息页面。
2. 切换至**提示词管理**页签，单击**新建**。

3. 在**新建提示词**弹窗中，配置提示词名称、提示词内容、改写策略、最大输出 Token 数及版本说明。

参数	说明
提示词名称	提示词的名称，为必填项，用于区分不同提示词的管控目的。
提示词内容	平台注入到请求中的系统提示词正文，为必填项，用于约束模型行为。
改写策略	当请求中已包含客户端提示词时，平台配置与客户端内容之间的合并方式。支持三种策略，详见下方说明。
最大输出 Token 数	限制该提示词占用的 Token 上限，避免过长提示词挤占上下文窗口。
版本说明	可按需输入本次创建或变更的备注，便于版本追溯。

改写策略支持以下三种方式：

改写策略	说明
完全覆盖	无论客户端请求是否自带提示词，都会用平台配置的提示词替换，管控力度强。
头部追加	把平台配置的提示词追加到客户端请求提示词的头部，两者共同生效。
尾部追加	把平台配置的提示词追加到客户端请求提示词的尾部，两者共同生效。

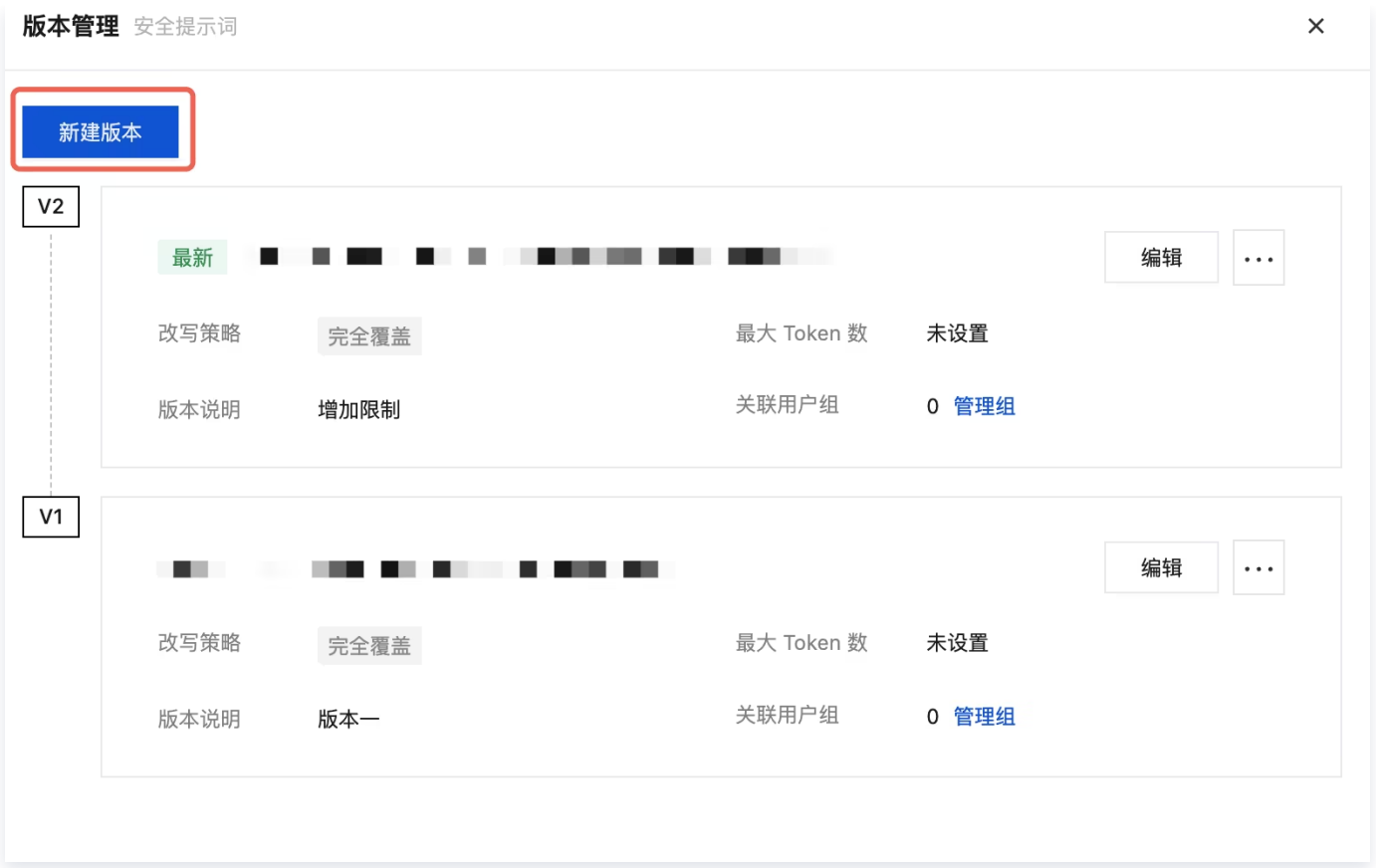
4. 单击**确定**，完成提示词的创建，系统自动生成首个版本。

步骤2：管理提示词版本

1. 在**提示词管理**列表，可查看各提示词的名称、版本数量、关联用户组、版本覆盖度以及修改时间。单击目标提示词操作中的**版本管理**，展开版本列表页面。

2. 在**版本**列表中，可查看各版本的版本号、提示词内容、关联用户组及操作。
3. 单击目标版本操作列的**编辑**，可修改该版本提示词的内容、改写策略、最大输出 Token 数及版本说明；修改版本说明外的其他内容将直接生成新版本。

4. 单击**新建版本**，可编辑内容并生成新版本，用于在不影响线上配置的前提下迭代提示词。



说明：

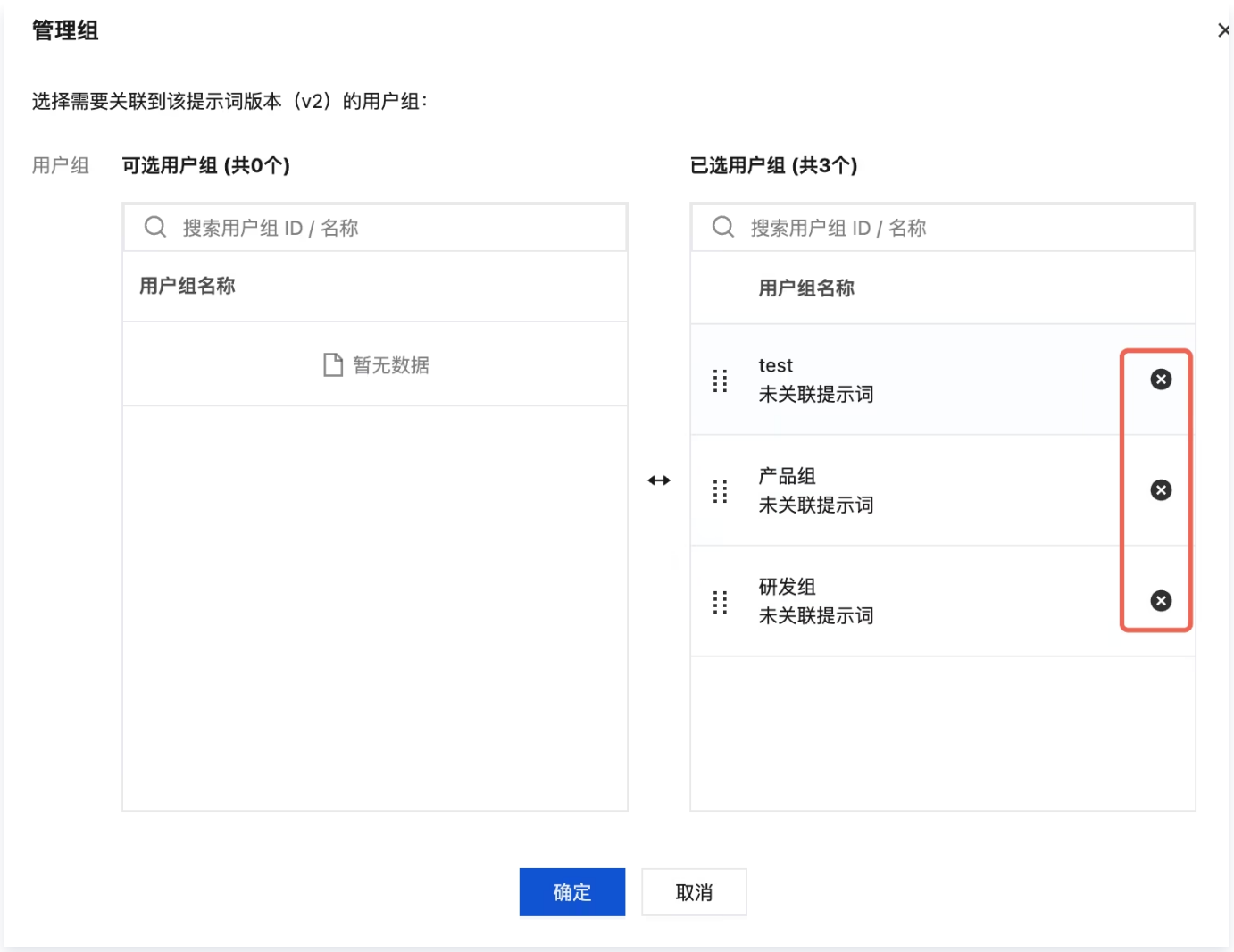
- 灰度发布提示词：您可以通过将少量用户组先分配至新版本进行验证，确认无误后再逐步扩大关联范围。
- 一键全量发布：可单击版本操作中的**全量发布**，将该提示词下的所有用户组关联至此版本。
- 历史版本清理：若需要删除历史提示词版本，可单击版本操作中的**删除**，移除指定版本。

步骤3：配置关联用户组

1. 在提示词版本管理页面，单击目标版本**关联用户组**列的**管理组**，打开用户组关联弹窗。

2. 在左侧**可选用户组**区域，选择需要关联的用户组，单击**移动图标**将其添加至右侧**已选用户组**区域。

3. 若需解关联用户组，可在右侧已选用户组中，选择需要解关联的用户组，单击**删除图标**将其解关联。



4. 单击**确定**完成配置。

> **说明：**
> 将用户组加入另一版本时，系统会自动完成跨版本的唯一归属迁移，确保一个用户组在同一份提示词下仅归属一个版本。

步骤4：查看用户组关联的提示词

1. 在**实例基本信息**页面，切换至**用户组**页签，单击左侧目标**用户组名称**，进入用户组详情页面。
2. 在**关联提示词**区域， 可直接查看本组关联的提示词名称及对应版本。
3. 单击关联提示词区域的提示词名称，可跳转至对应的提示词页面，进行版本与关联管理。

步骤5：删除提示词

1. 在实例管理页面，单击目标实例 ID，进入目标**实例基本信息**页面。
2. 切换至**提示词管理**页签，在提示词列表中找到需要删除的提示词。
3. 确认该提示词关联用户组数为0。若仍有用户组关联，请先在**版本管理页**通过管理用户组解除所有关联，再执行删除。

> **说明：**
> 若某提示词存在已关联的用户组，则该提示词的删除按钮将置灰不可用，需先解除所有关联用户组后才可删除。

4. 单击目标提示词操作列的**删除**，在弹出的确认对话框中单击确定，完成删除。

> **注意：**
> 删除操作不可撤销。提示词删除后，其所有历史版本及关联记录将被清除。请确认已无业务依赖后再执行删除。

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 配置聊天测试 。

- 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 消耗和模型资源使用情况，详细操作请参见 查看用量详情 。

查看用量详情

最近更新时间：2026-09-15 15:34:31

本文介绍如何查看用量详情。在用量详情可进行用量分析与性能观测。

简介

腾讯云可观测平台为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。只要您使用了模型路由，腾讯云可观测平台即可自动收集相关监控数据。

指标说明

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	个

		过期、请求未携带 API Key 等。	
	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
	输入图片折算 Token 数(Count)	Input 方向图像内容折算后的 Token 数量。	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

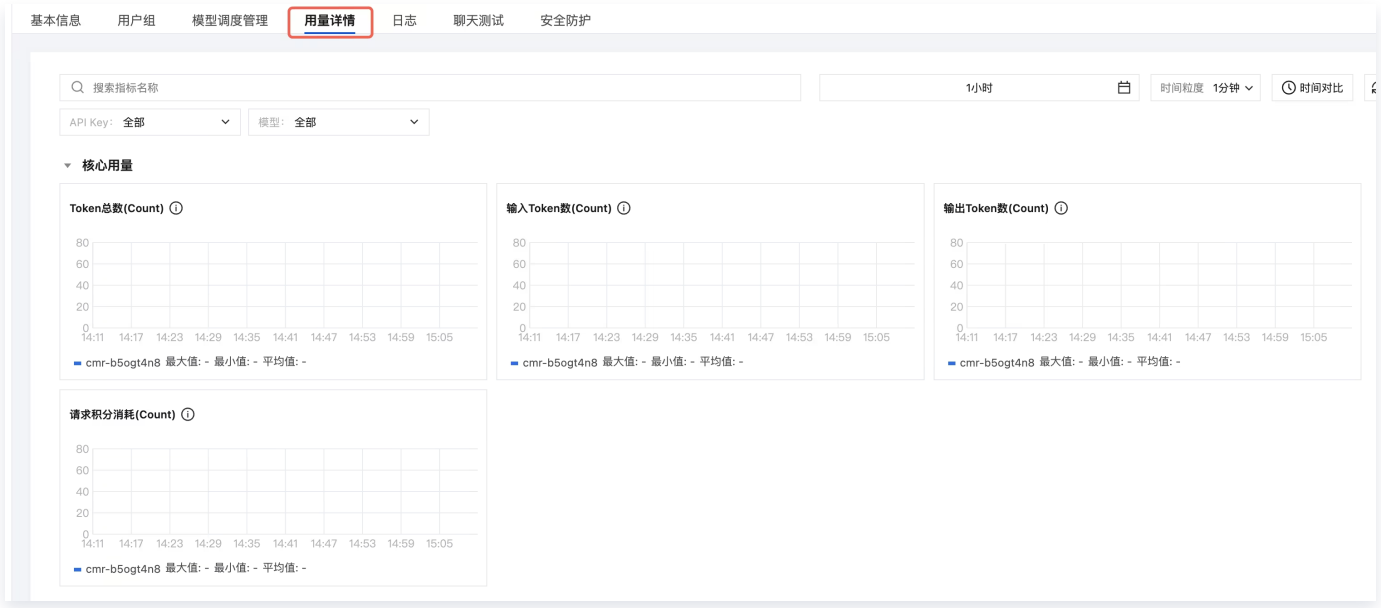
前提条件

- 建议提前阅读 使用约束 与 支持模型 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 。
- 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。
- 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 ID**，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。



查看日志

最近更新时间：2026-09-15 15:34:30

本文介绍如何查看日志。接入日志后支持进行检索分析与监控告警等。

简介

模型路由提供以下两种日志能力：

- **CLS 日志：推荐使用，**依托 CLS 日志服务，提供从日志采集、日志存储到日志检索分析、监控告警等多项服务。

> ⚠ **注意：**
> CLS 日志服务为腾讯云独立计费产品，使用后 CLS 侧将按照资费标准收费，计费标准请参见 CLS 计费概述。

- **基本日志：**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。**后期计划下线。**

模型路由接入 CLS 后，您将获得以下能力：

- **故障排查：**通过检索访问日志，快速定位请求失败、超时等问题。
- **安全审计：**记录所有模型调用请求，满足审计和合规要求。
- **用量分析：**统计模型调用量、Token 消耗、响应延迟等指标，辅助容量规划和成本优化。
- **监控告警：**基于 CLS 的告警能力，对异常访问模式、错误率上升等指标设置告警规则。

字段说明

模型路由投递至 CLS 的访问日志包含以下字段：

字段名称	类型	说明
timestamp	long	请求时间戳（毫秒级）
request_id	text	请求唯一标识
client_ip	text	客户端 IP 地址
model	text	请求的模型名称
provider	text	模型提供商
request_path	text	请求路径
method	text	请求方法（如 POST）

status_code	long	响应状态码
latency_ms	long	请求延迟（毫秒）
request_tokens	long	请求消耗的输入 Token 数
response_tokens	long	响应消耗的输出 Token 数
total_tokens	long	请求消耗的总 Token 数
upstream_status	long	后端模型服务返回的状态码
upstream_latency_ms	long	后端模型服务响应延迟（毫秒）
error_code	text	错误码（请求失败时返回）
error_message	text	错误信息（请求失败时返回）
user_agent	text	客户端 User-Agent

> ⚠ **注意：**
> 以上日志字段为模型路由默认投递的字段。实际投递字段可能随产品版本更新有所调整，请以 CLS 控制台中实际收到的日志内容为准。如需了解更详细的日志字段信息，请联系您的架构师或 提交工单 咨询。

前提条件

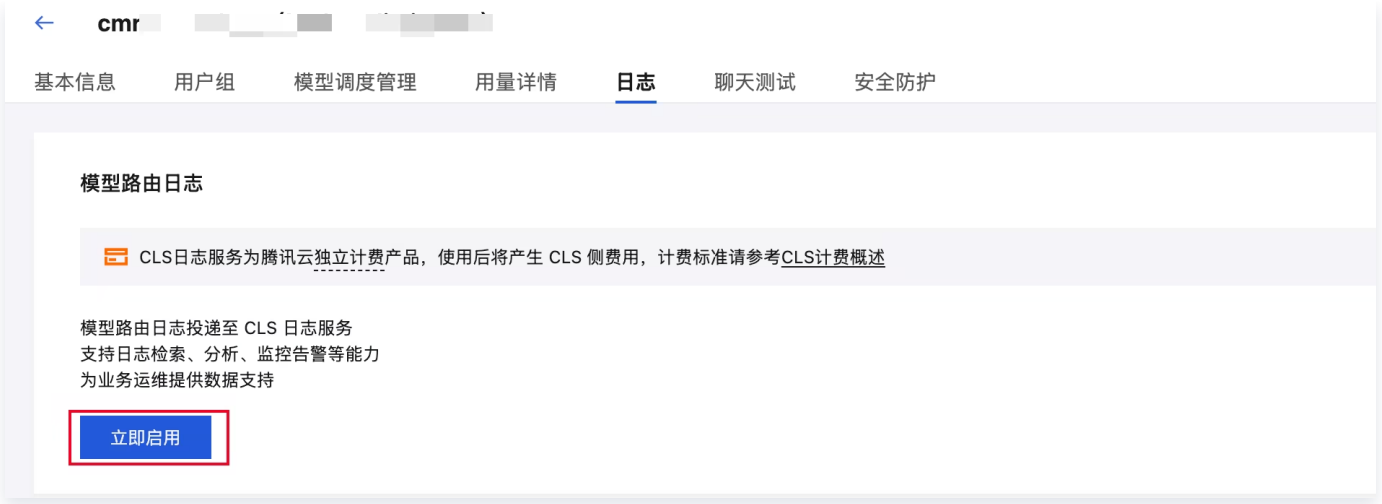
1. 已开通 日志服务 CLS，创建日志集和日志主题。日志主题（Topic）是日志数据进行采集、存储、检索和分析的基本单元，日志集则是对日志主题的分类，方便用户管理日志主题。
2. 已创建模型路由实例，详细操作指导请参见 创建模型路由实例 。

操作步骤

步骤1：开启日志投递

1. 登录 模型路由 控制台，在**实例管理**页面，单击目标**实例 ID**，进入目标实例的**实例基本信息**页面。

2. 切换至**日志**页签，单击**立即启用**。



3. 在**开启日志投递**弹窗中，选择已有**日志集**和**日志主题**，然后单击**立即开启**完成配置。推荐您**记录输入输出内容**。

> **说明：**
> 日志从投递到可检索存在约1 – 3分钟的延迟。日志投递开启后，您可以在 CLS 控制台的**检索分析**页面实时查看日志内容。如果检索不到日志，请检查索引配置是否正确。

步骤2：修改日志投递策略

1. 在模型路由实例的**日志**页签，单击**修改投递**。

步骤3：关闭日志投递策略

1. 在模型路由实例的**日志**页签，单击关闭投递。

2. 在**关闭访问日志**弹窗，勾选**确认关闭**后，单击**确认**。

> ⚠ **注意：**
> 关闭后，将无法继续记录模型路由的访问情况，请谨慎操作。

常见问题

1. 日志投递至 CLS 后检索不到数据?

日志投递至 CLS 后存在约1 – 3分钟的延迟。如果超过5分钟仍检索不到日志，请检查以下内容：

- 确认日志主题的索引配置是否正确，是否已开启全文索引或对应字段的键值索引。
- 确认检索的时间范围是否覆盖实际日志产生的时间。
- 确认日志投递是否处于开启状态。可以在模型路由控制台的日志页面查看投递状态。

2. CLS 日志投递是否会产生额外费用?

日志投递至 CLS 后，CLS 将按照日志存储量、日志读写流量等维度收取费用。费用详情请参见 日志服务计费概述 。低频存储的成本低于标准存储，适用于查询频率较低的日志。

3. 能否将日志同时投递至多个日志主题?

每个模型路由实例仅限添加至一个日志主题。如需将日志分发至多个目的地，可以在 CLS 控制台使用日志投递功能将日志投递至 COS 或 CKafka。

相关文档

- 配置日志索引
- 管理日志集
- 管理日志主题
- 查看与分析日志

配置聊天测试

最近更新时间：2026-09-15 15:34:30

本文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。

简介

- 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。

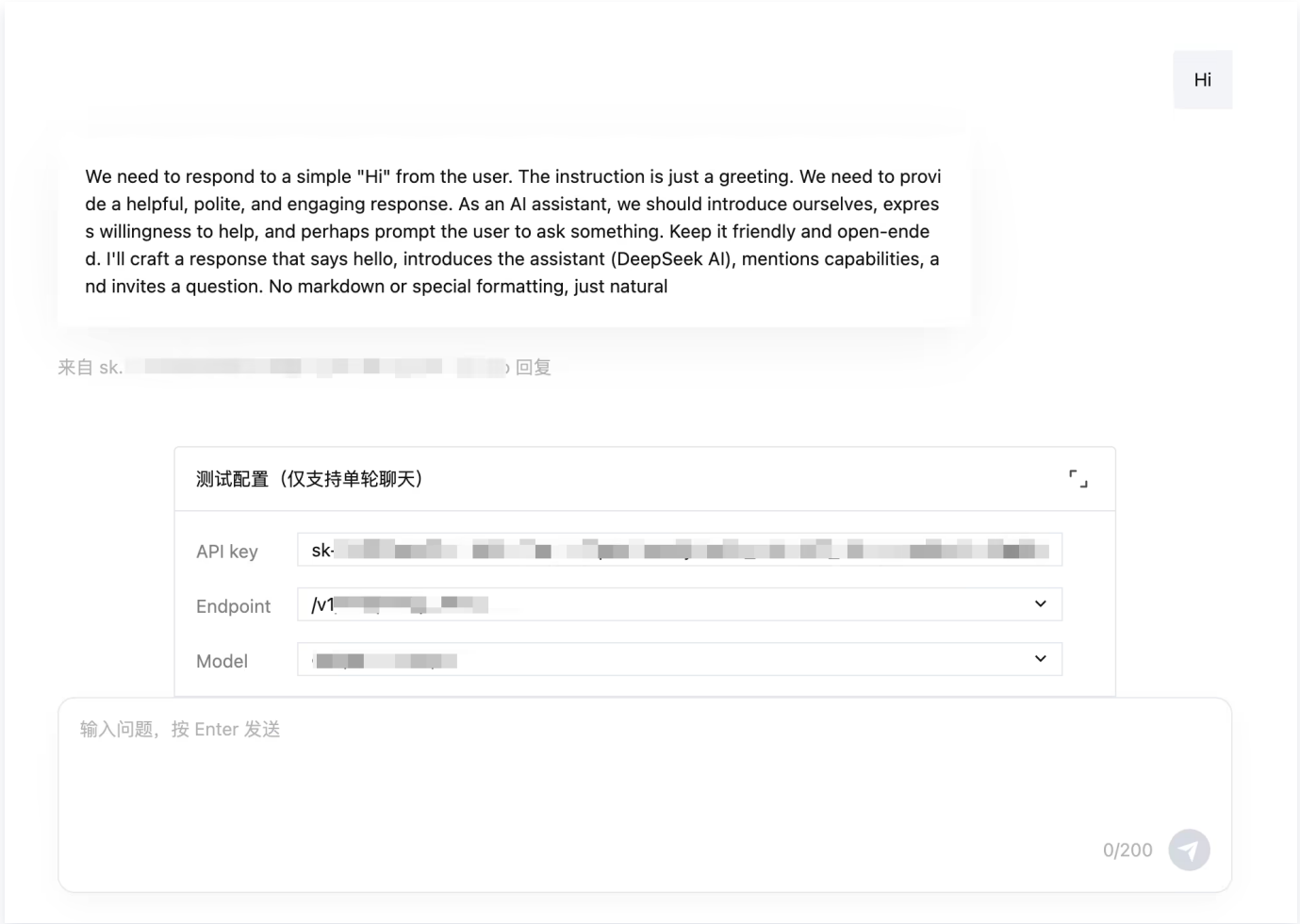
前提条件

1. 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. 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。
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 ID**，进入目标实例的**实例基本信息**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"），按 Enter 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

- 配置完成后，您可在实例管理页面切换至**用量详情**页签，关注 Token 消耗和模型资源使用情况，详细操作请参见 查看用量详情 。

常见问题

错误信息	可能原因	解决方法
BYOK API Key 无效 或 Unauthorized	BYOK 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	BYOK API Key 额度已用完	更换 BYOK 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。

配置安全防护

最近更新时间：2026-09-15 15:34:31

本文介绍如何配置安全防护，帮助您从 AI 内容安全和网络访问控制两个维度保障模型服务的安全运行。

简介

WAF 大模型安全

由 Web 应用防火墙（WAF）提供的 AI 内容安全防护能力，专门针对大模型应用场景，对用户与大模型之间的交互内容进行实时安全检测，防范 Prompt 注入攻击、敏感信息泄露、有害内容生成等风险，保障大模型业务的安全合规运行。详细介绍请参见 业务接入 。

核心功能

- **Prompt 注入攻击检测**：识别并拦截尝试绕过模型安全限制的恶意 Prompt 注入行为。
- **敏感信息泄露防护**：检测模型响应中是否包含身份证号、手机号、银行卡号等敏感信息。
- **有害内容生成检测**：识别模型生成内容中的涉黄、涉暴、涉政等违规内容。
- **连续对话内容检测**：支持对多轮对话的上下文进行关联分析，检测隐藏在多轮对话中的攻击行为。

安全组

一种有状态的虚拟防火墙，用于设置模型路由实例的网络访问控制。您可以通过配置安全组规则，允许或拒绝指定协议和端口的入站和出站流量。

核心功能

- **逻辑分组**：您可以将同一地域内具有相同网络安全隔离需求的模型路由实例加入同一个安全组。
- **默认拒绝**：安全组未添加任何规则时，默认拒绝所有入站和出站流量，您需要按需添加允许规则。
- **有状态**：对于已允许的入站流量，系统将自动允许其流出。
- **即时生效**：修改安全组规则后，新规则立即生效，无需重启实例。
- **规则优先级**：安全组内的规则按照列表顺序从上至下依次匹配，先匹配成功的规则直接执行，不再继续匹配后续规则。

规则组成

每条安全组规则包含以下组成部分：

组成部分	说明
来源或目标	流量的源地址（入站规则）或目标地址（出站规则）。支持 CIDR 段、IP 地址、IP 地址组、安全组与当前登录 IP。

协议	协议类型（如 TCP、UDP）。
端口	端口号。
策略	允许或拒绝。
备注	规则的描述信息。

多安全组

一个模型路由实例可以绑定一个或多个安全组。当实例绑定多个安全组时，系统按照安全组优先级从高到低依次匹配规则。您可以在安全组列表中调整安全组的优先级。

> **注意：**
> 如果较高优先级的安全组内没有匹配的规则，系统会继续匹配下一个安全组。建议您将最严格的规则放在优先级较高的安全组中。

前提条件

1. 建议提前阅读 使用约束 与 支持模型 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. 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。
4. 已购买 WAF 实例 以及 大模型安全 增值功能。

操作步骤

WAF 大模型安全操作步骤

步骤1：开启安全防护

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例 ID**，进入目标实例的**实例基本信息**页面。
2. 切换至**安全防护**页签。
3. 在安全防护页面，**WAF 大模型安全**区域，单击**开启防护**。
4. 在弹出的配置窗口中，配置以下参数：

参数	说明
WAF 实例	选择已购买的 WAF 实例。若列表为空，请先 购买 WAF 实例 。
Service 信息	选择需要防护的 Service。系统将自动拉取该实例下已配置的 Service 列表。

最大检测对话轮数	设置连续对话内容检测的最大轮数，取值范围：1 – 20，默认值为5。仅在连续对话内容检测开启后生效。

5. 单击**确定**，完成 WAF 安全防护配置。操作完成后，实例的安全防护状态将变更为**已开启**，表示该实例已成功接入 WAF 安全防护。

> **注意：**
> - 开启 WAF 安全防护后，WAF 实例将开始计费，请确保账户余额充足。
> - 首次开启防护建议选择观察模式，确认业务流量正常后再切换为拦截模式，避免因误拦截影响业务。
> - 若 WAF 实例到期未续费，防护功能将自动失效，请及时续费。

步骤2：查看防护效果

WAF 安全防护开启后，您可以在 WAF 控制台 的防护概览页面查看攻击统计和防护日志，查看防护效果：

1. 登录 WAF 控制台 。
2. 在**安全概览**页面，查看攻击统计信息，包括攻击次数、攻击类型分布、攻击趋势等。
3. 在左侧导航栏选择**攻击日志**，可以查看详细的攻击日志，包括攻击时间、攻击类型、攻击源 IP、拦截动作等。
4. 在左侧导航栏选择**大模型安全**，可以查看大模型安全专项的防护统计和详细日志。

> **说明：**
> 当您不再需要 WAF 防护时，可以在模型路由实例的安全防护页面关闭 WAF 大模型安全功能。关闭后，WAF 将停止对模型路由实例的防护，但不会影响 WAF 实例的其他防护配置。

安全组操作步骤

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例 ID**，进入目标实例的**实例基本信息**页面。
2. 切换至**安全防护**页签。
3. 在安全防护页面，安全组配置区域，单击**配置**。
4. 在弹出的配置窗口中，选择安全组，并单击**确定**。

> **说明：**
> 如果您还未创建安全组，可以在 安全组控制台 中创建新的安全组。创建安全组时，您可以选择腾讯云提供的安全组模板快速配置，也可以选择自定义模板手动添加规则。安全组创建完成后，返回模型路由控制台完成绑定操作。

最佳实践

1. 多层防护组合

建议同时启用 WAF 大模型安全和安全组，构建多层安全防护体系：

防护层级	能力	推荐配置
内容安全	WAF 大模型安全	开启 Prompt 注入检测和有害内容检测，根据业务需求配置连续对话检测轮数
网络安全	安全组	仅放通业务必需的端口和来源 IP，遵循最小权限原则

2. 安全组最小权限原则

- 仅放通业务必需的端口，例如模型服务的监听端口。
- 限制来源 IP 范围为业务可信 IP，避免使用 `0.0.0.0/0` 放通所有来源。
- 定期审计安全组规则，删除不再使用的规则。

3. 防护模式切换建议

- **首次接入 WAF**：建议先使用观察模式运行 1 – 2 周，观察攻击日志和业务流量情况。
- **切换拦截模式**：确认观察期间无误拦截后，在业务低峰期切换为拦截模式。
- **定期巡检**：每隔 1 – 2 个月检查 WAF 防护日志和安全组规则，确保防护策略与业务需求匹配。

常见问题

1. 安全组规则配置后未生效?

安全组规则配置后立即生效，无需重启实例。如果规则未生效，请检查以下内容：

- 确认规则匹配的协议和端口是否正确。
- 确认规则的优先级顺序，高优先级规则可能先匹配并执行，导致后续规则被跳过。
- 确认实例是否绑定了正确的安全组。
- 确认规则的方向（入站/出站）是否正确。

2. WAF 误拦截了正常请求该如何处理?

如果发现 WAF 误拦截了正常请求，可以采取以下措施：

- 在 WAF 控制台的攻击日志中查找被拦截的请求，确认拦截规则和原因。
- 针对该规则配置白名单，放行特定的 IP、URL 或参数。
- 如果是规则本身存在问题，可以在 WAF 控制台调整规则的动作模式为观察模式，确认后向腾讯云提交规则优化反馈。

3. 如何关闭模型路由实例的安全防护?

- **关闭 WAF 防护**：在模型路由实例的安全防护页面，找到 WAF 大模型安全区域，单击**关闭防护**。
- **解绑安全组**：在模型路由实例的安全防护页面，在安全组配置区域，单击**配置**，取消已绑定安全组的勾选。

BYOK
BYOK 介绍

最近更新时间：2026-09-15 15:34:32

模型路由 CMR 全面支持自带密钥（Bring Your Own Key，BYOK）模式。无论是原厂模型、第三方代理还是自建模型，均可支持。

- 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。

创建 BYOK 实例

最近更新时间：2026-09-15 15:34:32

本文介绍如何创建 BYOK 实例。

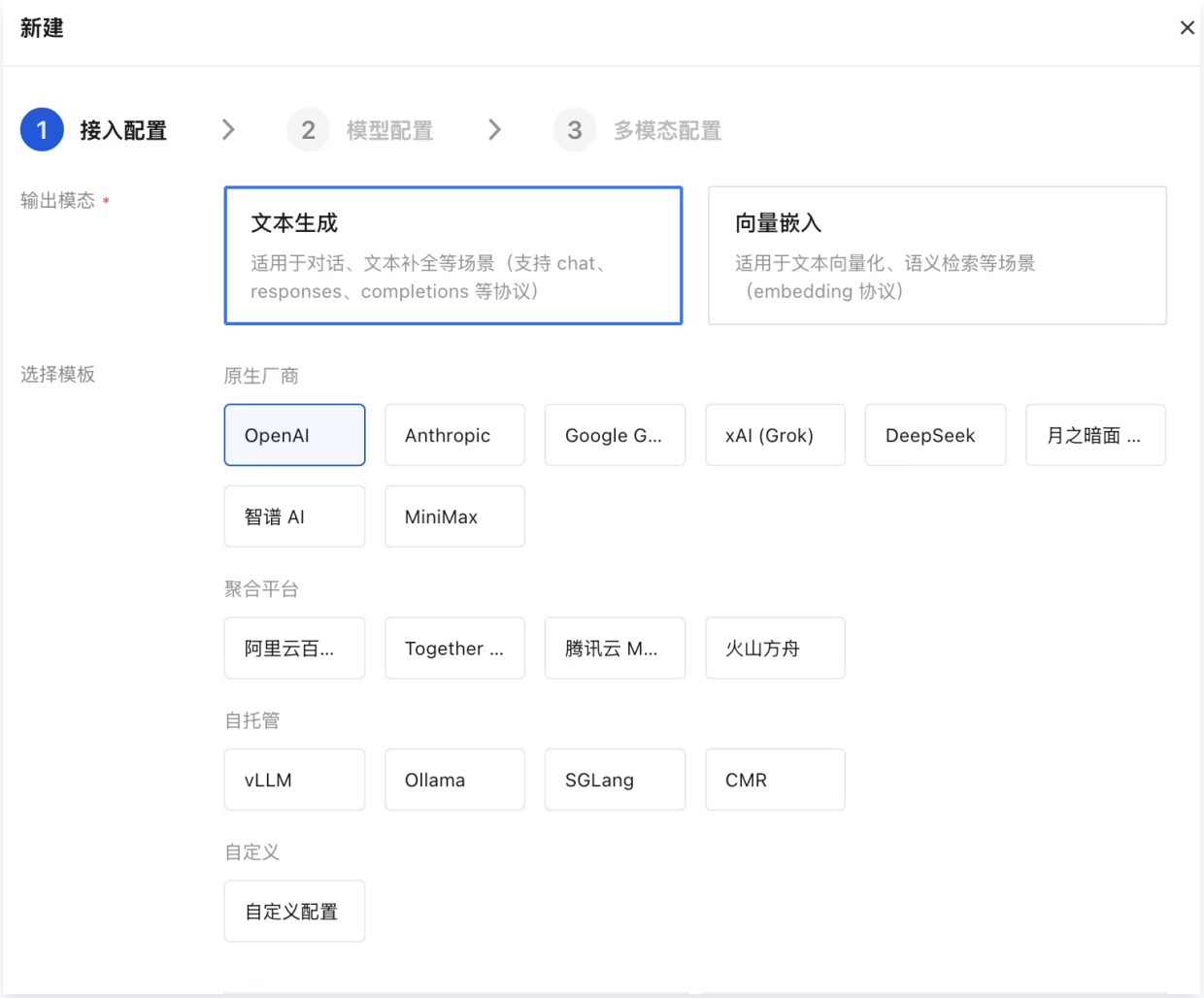
前提条件

- 您已获得上游模型的 API key，支持原厂模型、第三方代理或自建模型的 API Key。
- 如果您的上游模型为自建模型，请先完成 CMR 私网管道的创建 。

操作指导

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

3. 在**新建**页面中，按需选择或者填写相关参数。



参数	说明
输出模态	当前已支持文本生成和向量嵌入两种模态。 • 文本生成：适用于对话、文本补全等场景（支持 chat、responses、completions 等协议）。 • 向量嵌入：适用于文本向量化、语义检索等场景（支持 embedding 协议）。
选择模板	支持选择预置模板快速配置或自定义配置两种方式。 • 预置模板快速配置：为方便用户配置，产品已预置部分厂商的配置模板，如腾讯云 MaaS TokenHub，可按需选择。 • 自定义配置：当预置模板尚未覆盖您的使用场景，比如，使用自建模型时，可选择。

所属厂商与协议	• 所属厂商：
	  ○ 当选择预置模板快速配置时，此项已经预填，无需修改。
	  ○ 当选择自定义配置时，可通过下拉选择您的模型提供商（如 DeepSeek），或者通过自定义的方式输入厂商名称。
	• 协议：
	  ○ 当选择预置模板快速配置时，此项已经预填，无需修改。

	○ 当选择自定义配置时，可通过下拉选择对应的协议。
所属站点	在**输出模态**为**文本生成**时，部分模型如智谱，需要选择所属站点，您可按需选择。
Endpoint 协议 & BaseURL	在**输出模态**为**文本生成**时需要配置，即 OpenAI 兼容 Base URL 或者 Anthropic Base URL。为方便用户配置，产品已预置部分厂商的相关参数，包括多协议，您可按需配置。
BaseURL & EndpointPath	在**输出模态**为**向量嵌入**时需要配置，最终请求地址 = BaseURL + EndpointPath。为方便用户配置，产品已预置部分厂商的相关参数，您可按需配置。
实例名称	选填，最多支持 255 个字符。
标签	您可选择已有的标签键和标签值；若所需标签不存在，也可单击添加标签。详情请参见 创建标签 。

4. **接入配置完成后，进入模型配置步骤。按需选择或者填写相关参数，并单击确定。**

参数	说明
网络配置	支持通过以公网或者内网的方式配置模型。 • 公网：通过公共互联网配置模型，如原厂模型。 • 内网：通过私有网络配置模型，如自建模型。
CMR 私网管道	仅在**网络配置**选择**内网**时涉及。通过下拉选择已存在的 CMR 私网管道，也可以 创建 CMR 私网管道 。
域名	选填，仅在**网络配置**选择**内网**时涉及。往上游模型发送请求时携带的 HTTP Header。 通常无需填写，默认使用「BaseURL」中的域名；当「BaseURL」填写为 IP 或 VIP 时，且上游需按域名路由的情况下，在此指定对应域名（如 api.deepseek.com ）。
API 填写方式	支持手动填写和批量导入两种方式。 • 手动填写：手动将 API key 和密钥依次粘贴到对应的输入框中。 • 批量导入：通过下载模板文件，填写 API key 和密钥后，再以导入文件的方式实现批量快速导入。
选择模型	支持手动输入模型名称，也支持通过单击**获取模型列表**来拉取支持模型。如需更改模型对外名称，可以在**对外展示名称**输入框中，输入符合您标准的名称。
连通性探测	添加 API Key 并选择正确的模型后即可单击**开始探测**，开始探测 API Key 的健康情况，方便您及时剔除不健康的 API Key。

5. 模型配置完成并通过模型连通性探测后，单击**下一步：多模态配置**。在多模态配置步骤中，您可以根据模态可用性探测结果按需选择各模型需要启用的模态，完成配置。仅**输出模态**为**文本生成**时，涉及该配置。

参数	说明
模型名称	展示已选择的模型名称。不同模型支持的模态能力可能不同，请以页面展示结果为准。
启用模态	选择当前模型需要启用的模态。启用后，该模型可在对应模态场景下被调用。
模态可用性探测	用于探测当前 API Key 下各模型模态的可用情况。单击**开始探测**，探测完成后，可在上方模型列表的模态可用性中查看探测结果。
模态可用性	展示当前模型各模态的可用状态，可根据提示调整启用模态或重新探测。
按探测结果启用	单击后，系统将根据当前探测结果自动启用可用模态。
重新探测	单击后，系统将重新探测当前 API Key 下各模型模态的可用情况，并刷新模态可用性结果。

6. 完成以上参数配置后，单击**确定**完成 BYOK 实例的创建。

相关文档

- CMR 私网管道介绍
- 创建 BYOK 实例

管理 BYOK 实例

最近更新时间：2026-09-15 15:34:32

本文介绍如何管理 BYOK 实例，包括基本信息、Key 管理、模型管理与查看用量详情等。

基本信息

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK > BYOK 管理**，在列表页中可查看已创建的 BYOK 实例。
2. 单击具体的 BYOK 实例，进入 **基本信息页**。在基本信息页，可对如下参数进行配置或修改。

参数	说明
BYOK ID	展示已创建 BYOK 实例的 ID 信息，可复制。
BYOK 名称	展示已创建 BYOK 实例的名称信息，单击**编辑图标**，可对名称进行修改。
Endpoint 协议 & BaseURL	**输出模态为文本生成**时涉及，即 OpenAI 兼容 Base URL 或者 Anthropic Base URL，在支持多协议时，可编辑、新增或者删除。
BaseURL & EndpointPath	在**输出模态为向量嵌入**时需要配置，最终请求地址 = BaseURL + EndpointPath，支持复制和编辑。
标签	可对标签键和标签值进行编辑，也可选择添加标签，详情请参见 创建标签 。

健康检查

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK > BYOK 管理**，在列表页中可查看已创建的 BYOK 实例。
2. 单击具体的 BYOK 实例，进入 **基本信息页**。在基本信息页，可对健康检查配置进行编辑。

Key 管理

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK > BYOK 管理**，在列表页中可查看已创建的 BYOK 实例。
2. 单击具体的 BYOK 实例或在列表页操作列中单击**管理 Key**，可直接跳转至 **基本信息页**，支持添加/删除 Key。**添加 Key** 时，支持手动填写，或批量导入，要填写参数说明如下。

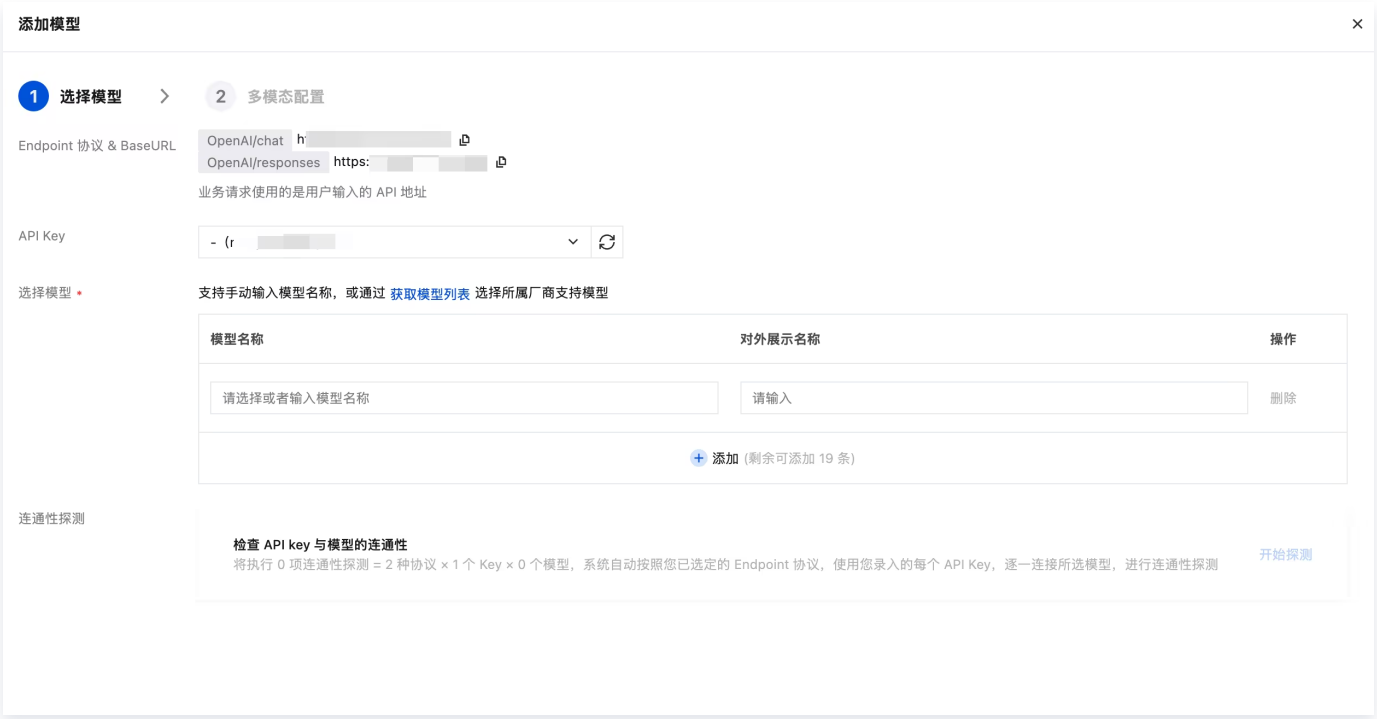
参数	说明
API 填写方式	支持手动填写和批量导入两种方式。 • 手动填写：手动将 API key 和密钥依次粘贴到对应的输入框中。 • 批量导入：通过下载模板，填写 API key 和密钥，再以导入文件的方式实现批量快速导入。

连通性探测	系统自动按照您已选定的 Endpoint 协议，使用您录入的每个 API Key，逐一连接所选模型，进行连通性探测。探测 API Key 的健康情况，方便您及时剔除不健康的 API Key。

> **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

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK > BYOK 管理**，在列表页中可见已创建的 BYOK 实例。
2. 单击具体的 BYOK 实例，进入 **基本信息页**。在基本信息页，支持添加/删除模型及批量探测多模态功能（多模态仅在**输出模态**为**文本生成**时涉及）。**添加模型**需要填写的参数说明如下。
3. 选择模型



参数	说明
Endpoint 协议 & BaseURL	**输出模态**为**文本生成**时涉及，即 OpenAI 兼容 Base URL 或者 Anthropic Base URL，在支持多协议时，可编辑、新增或者删除。
BaseURL & EndpointPath	在**输出模态**为**向量嵌入**时需要配置，最终请求地址 = BaseURL + EndpointPath，支持复制和编辑。
API Key	选择您在当前 BYOK 实例中添加的 API Key。
选择模型	支持手动输入模型名称，也支持通过单击**获取模型列表**来拉取支持模型。如需更改模型对外名称，可以在**对外展示名称**输入框中，输入符合您标准的名称。

连通性探测	添加 API Key 并选择正确的模型后即可单击**开始探测**，开始探测 API Key 的健康情况，方便您及时剔除不健康的 API Key。

填写完成后，单击**下一步：多模态配置**。

4. 多模态配置，仅输出模态为文本生成的 BYOK 涉及。

参数	说明

模型名称	展示已选择的模型名称。不同模型支持的模态能力可能不同，请以页面展示结果为准。
启用模态	选择当前模型需要启用的模态。启用后，该模型可在对应模态场景下被调用。
模态可用性探测	用于探测当前 API Key 下各模型模态的可用情况。单击**开始探测**，探测完成后，可在上方模型列表的模态可用性中查看探测结果。
模态可用性	展示当前模型各模态的可用状态，可根据提示调整启用模态或重新探测。
按探测结果启用	单击后，系统将根据当前探测结果自动启用可用模态。
重新探测	单击后，系统将重新探测当前 API Key 下各模型模态的可用情况，并刷新模态可用性结果。

5. 确认无误后单击**完成**。

> ⓘ **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单击**配置模型**继续完成模型 & Key 配置。

查看用量详情

单击具体的 BYOK 实例 ID 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 。

相关文档

- CMR 私网管道介绍
- 创建 BYOK 实例

CMR 私网管道介绍

最近更新时间：2026-09-15 15:34:32

本文介绍如何创建和管理 CMR 私网管道。

创建 CMR 私网管道

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK > CMR 私网管道**。
2. 在 **CMR 私网管道** 页面，单击**新建**。
3. 在**新建**弹窗中，按需选择或填写相关参数。

参数	说明
CMR 私网管道名称	选填，留空则自动生成，可支持输入 255 个字符。
私有网络	选择您的模型所在的私有网络。
子网	选择空闲的子网，一个 CMR 私网管道将会占用该子网中的4个 IP 地址。
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. 完成以上参数配置后，单击**确定**完成创建。

管理 CMR 私网管道

1. 登录 模型路由 控制台，在左侧导航栏选择 **BYOK > CMR 私网管道**。
2. 在 **CMR 私网管道** 页面，单击 **CMR 私网管道 ID** 或者操作列的**编辑**。
3. 在基本信息页查看 CMR 私网管道的基本信息，并支持对 CMR 私网管道 ID 进行复制，对 CMR 私网管道名称和标签进行编辑。

相关文档

创建 BYOK 实例

积分管理
积分管理介绍

最近更新时间：2026-09-15 15:34:32

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。

功能概述

- **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

1. 配置模型系数

支持为不同模型设置不同的系数，包含输入系数、输出系数、缓存读取系数和缓存写入系数。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，建议配置十倍比例将系数配置为25。

2. 管理积分预算

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，实现成本拆分与管控。
- 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：2026-09-15 15:34:31

本文档介绍在腾讯云模型路由控制台中配置积分管理模型系数的操作方法。

简介

在实际使用场景中，同一个模型可能来自不同的 BYOK。产品支持按照模型（不考虑 BYOK 来源）设置模型系数，同时也支持按照不同的 BYOK 对模型设置更细粒度的系数。

前提条件

1. 建议提前阅读 使用约束 与 支持模型 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. 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。

操作步骤

步骤1： 配置模型系数

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**编辑模型系数**。
3. 在**编辑系数**弹窗，配置**输入系数**、**输出系数**、**缓存读取系数**和**缓存写入系数**，并单击**确定**完成配置。
4. 也可在目标模型的**输入系数**列、**输出系数**列、**缓存读取系数**列和**缓存写入系数**列，通过单击**编辑图标**，单独配置。（可选）

步骤2： 配置 BYOK 系数

1. 可在目标 BYOK 行的操作列，单击**编辑 BYOK 系数**，配置**输入系数**、**输出系数**、**缓存读取系数**和**缓存写入系数**，并单击**确定**完成配置。
2. 也可在目标 BYOK 行的**输入系数**列、**缓存输入系数**列和**输出系数**列，通过单击**编辑图标**，单独配置系数。（可选）

步骤3：查看用量详情

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。

模型	用量详情	BYOK 数量	输入系数	输出系数	缓存读取系数	缓存写入系数	操作
▸		1	25	200	3	0	编辑模型系数 批量设置 BYOK 系数

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：2026-09-15 15:34:32

本文档介绍在腾讯云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。

简介

模型路由积分预算支持配置最大积分额度和重置周期。可关联实例、用户组或 API Key 进行预算限制。

前提条件

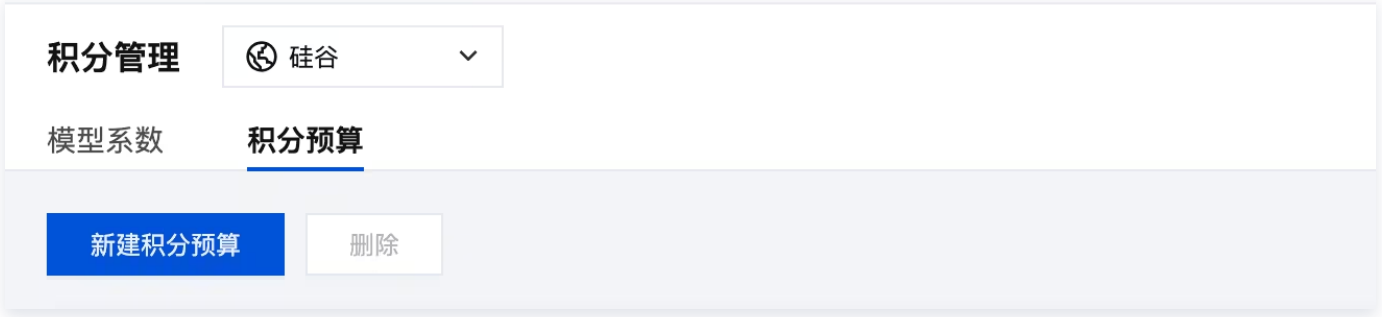
1. 建议提前阅读 使用约束 与 支持模型 了解相关信息。
2. 已创建模型路由实例，详细操作指导请参见 创建模型路由实例 。
3. 已创建模型路由访问密钥（API Key），详细操作指导请参见 创建模型路由访问密钥 。
4. 已创建用户组，详细操作指导请参见 用户组介绍 。

操作步骤

关联积分预算操作步骤

步骤1：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。单位可选个、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，可选每日、每周、每月。
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、用户组或者 API Key，也可以暂时不选择关联对象，直接单击**完成**配置。

步骤2：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。

步骤3：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、用户组或 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。

解除积分预算操作步骤

步骤1：解关联对象

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**关联对象的数量**。
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。

步骤2：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。

用户组
用户组介绍

最近更新时间：2026-09-15 15:34:32

用户组用于对模型路由访问密钥（API Key）进行分组管理。您可以按照团队、业务线、项目或环境创建用户组，并统一配置组内 API Key 可访问的模型、意图路由规则和资源限制。

> **说明：**
> 用户组功能仅适用于模型路由企业版实例，共享版实例不支持该功能。

功能概述

- **分组管理**：按照组织或业务边界集中管理 API Key。
- **访问可控**：限制组内 API Key 可以访问的模型和意图路由规则。
- **资源可控**：通过积分预算和速率限制控制用户组的整体资源消耗。
- **灵活配置**：支持通过标签对用户组进行分类，满足不同业务场景的管理需求。

核心功能

1. API Key 分组管理

一个模型路由实例可以创建多个用户组，一个用户组可以关联多个 API Key。您可以根据 API Key 所属的团队、业务线、项目、应用或环境，将其分配至对应的用户组。
每个 API Key 只能归属一个用户组。将 API Key 分配至其他用户组后，该 API Key 将从原用户组中移出，并归属至新的用户组。
新建或导入的 API Key 默认归属系统内置的“未分组 API Key”用户组。后续可根据业务需要，将其分配至自定义用户组。

2. 模型访问控制

您可以为用户组指定允许访问的模型或意图路由规则。组内 API Key 只能使用该用户组允许访问的模型和意图路由规则。
如果创建用户组时未指定模型或意图路由规则，组内 API Key 默认可以使用当前模型路由实例下的全部模型和意图路由规则。
通过用户组配置模型访问范围，可以为不同团队或业务开放不同的模型能力。例如，为开发测试团队开放低成本模型，为生产业务开放指定的高可用模型或意图路由规则。

3. 用户组限制

用户组可以关联积分预算，用于限制组内 API Key 的总体积分消耗和调用速率。

积分预算可以包含以下限制：

- 最大积分额度。
- 积分重置周期。
- TPM（Token per minute）限制。
- RPM（Request per minute）限制。

模型路由实例、用户组和 API Key 可以分别配置积分预算和速率限制。API Key 发起调用时，需要同时满足实例、所属用户组和 API Key 三个层级的限制条件。任一层级达到额度或速率上限后，相关调用都会受到限制。

4. 标签管理

您可以为用户组配置标签，根据部门、项目、环境或成本中心等维度对用户组进行分类，便于检索和管理相关资源。

典型场景

- 为不同团队或业务线创建独立用户组，将相应 API Key 分配至对应用户组，并分别配置可访问的模型和积分预算，实现不同团队之间的资源隔离。
- 按照业务环境创建用户组，为不同环境配置独立的模型访问范围、积分预算和速率限制，避免开发测试流量影响生产业务。
- 将高成本模型或特定意图路由规则仅授权给指定用户组，其他用户组使用通用模型或更具性价比的模型，实现模型能力的按需开放。
- 在实例层控制整体资源消耗，在用户组层控制团队或业务线的资源消耗，在 API Key 层控制具体应用、员工或脚本的调用行为，实现分层资源管控。

常见问题

1. 新建 API Key 会归属哪个用户组?

新建或导入 API Key 时，如果未指定用户组，该 API Key 默认归属系统内置的“未分组 API Key”用户组；如果在目标用户组下新建或导入，该 API Key 将直接归属该用户组。创建后，您也可以根据业务需要调整其所属用户组。

2. 一个 API Key 可以同时归属多个用户组吗?

不可以。每个 API Key 只能归属一个用户组。将 API Key 分配至新的用户组后，该 API Key 将从原用户组中移出。

3. 用户组未指定模型或意图路由规则时，组内 API Key 可以访问哪些模型?

如果用户组未指定模型或意图路由规则，组内 API Key 默认可以使用当前模型路由实例下的全部模型和意图路由规则。

4. 用户组的积分预算耗尽后会发生什么?

当用户组关联的积分预算耗尽时，组内 API Key 将无法继续调用模型，直到预算额度重置或调整。

相关文档

- 创建用户组
- 分配 API Key 至用户组
- 创建模型路由访问密钥（API Key）
- 配置积分预算
- 配置模型调度管理

管理用户组

最近更新时间：2026-09-15 15:34:31

本文介绍如何在模型路由实例中创建用户组，并配置用户组限制、关联 API Key 及模型访问范围。

简介

用户组用于对模型路由访问密钥（API Key）进行分组管理。您可以按照团队、业务线、项目或环境创建用户组，并统一配置组内 API Key 可访问的模型、意图路由规则和资源限制。

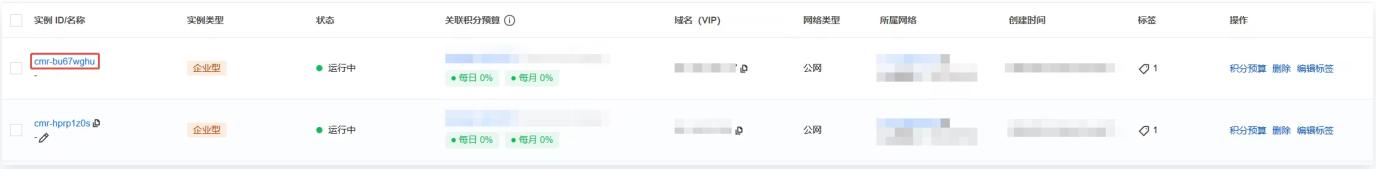
前提条件

1. 建议提前阅读 使用约束 与 支持模型 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. 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。
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. 如需为用户组配置积分预算，请先创建积分预算。详细操作指导请参见 配置积分预算 。

操作步骤

创建用户组操作步骤

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例 ID**。

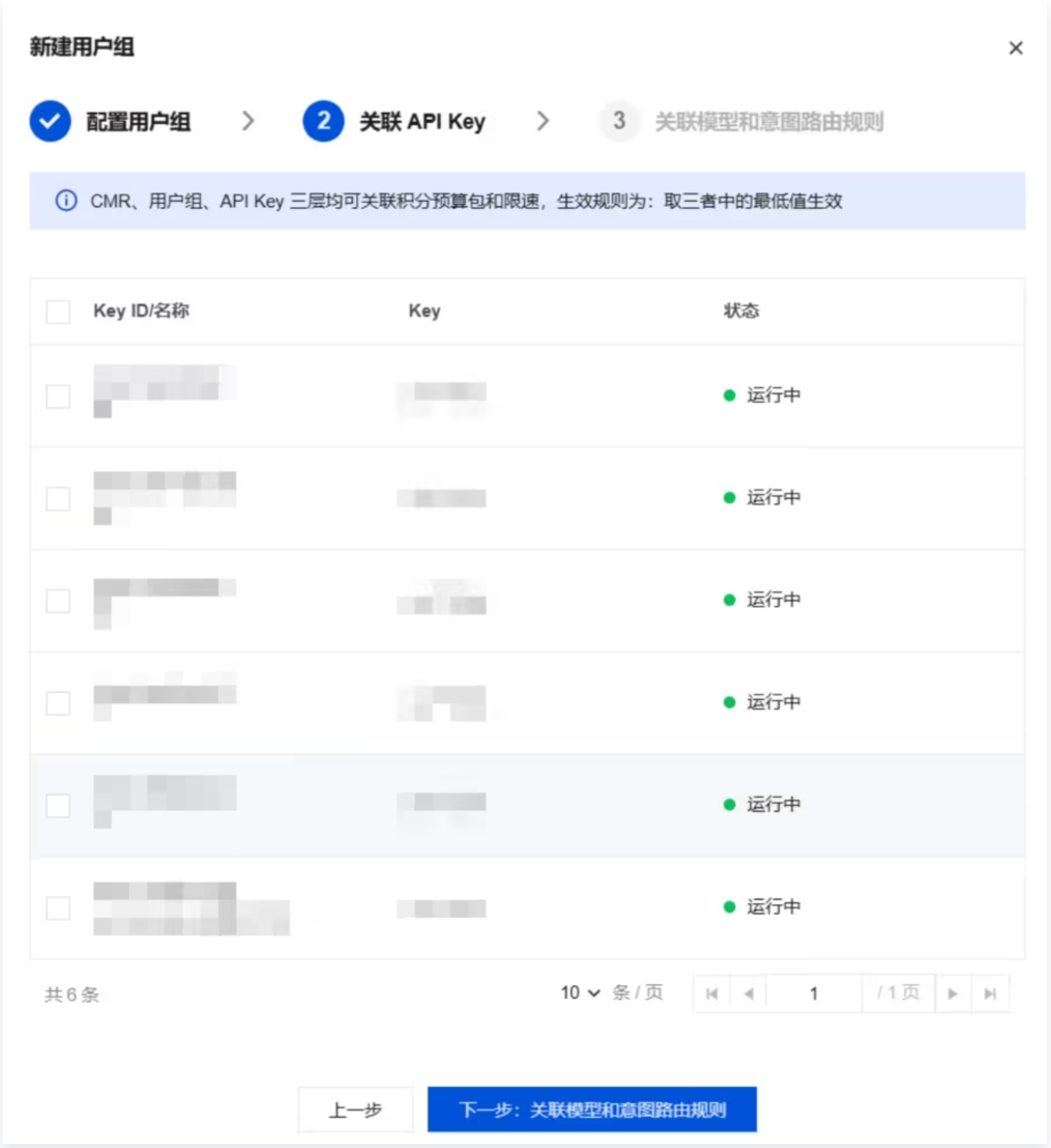


2. 在实例详情页，选择**用户组**页签。
3. 在**用户组**页签中，单击**新建用户组**，在**新建用户组**弹窗中，按照页面提示完成用户组配置。

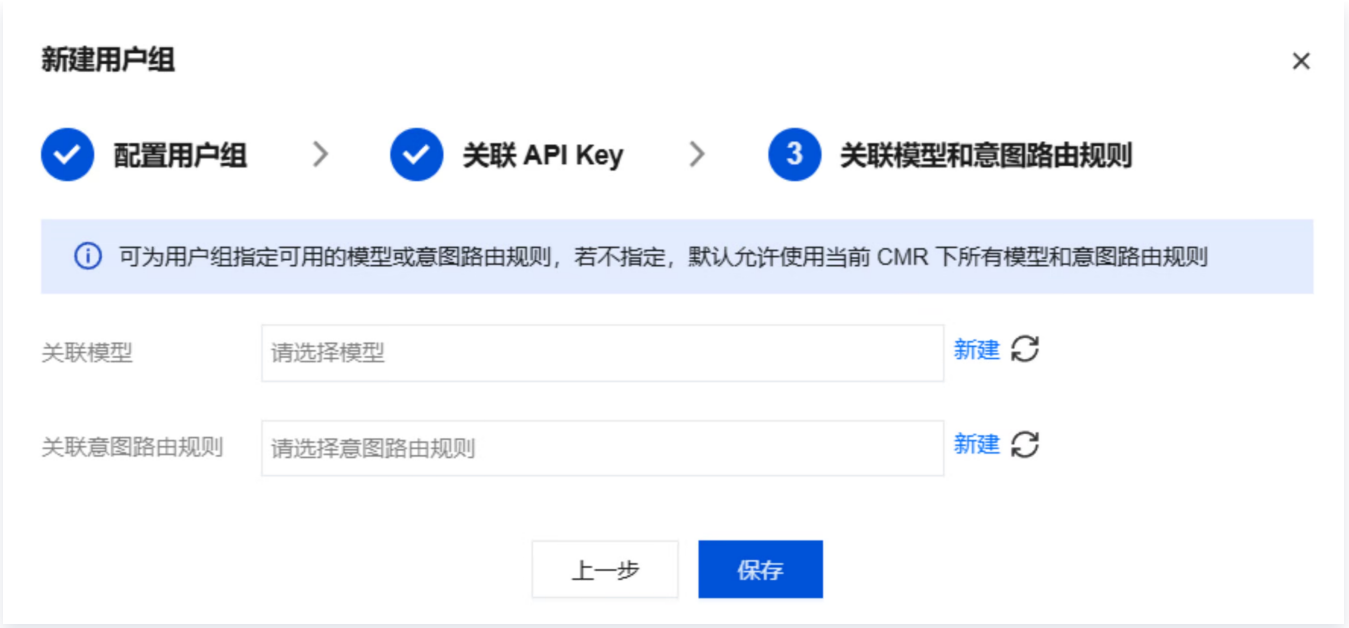
4. 配置用户组，填写用户组的基本信息。

参数	说明
用户组名称	自定义用户组名称，用于标识团队、业务线、项目或环境。最多支持 255 个字符。
用户组限制	可选。为用户组关联积分预算，用于限制组内 API Key 的积分消耗及调用速率。
标签	可选。配置标签键和标签值，用于对用户组进行分类管理。

5. 关联 API Key，选择需要分配至该用户组的 API Key。如果暂时不需要关联 API Key，可跳过该步骤，待用户组创建完成后再进行分配。详细操作指导请参见 分配 API Key 至用户组 。



6. **关联模型和意图路由规则：**为用户组指定可用的模型或意图路由规则。若不指定，默认允许使用当前 CMR 下所有模型和意图路由规则。



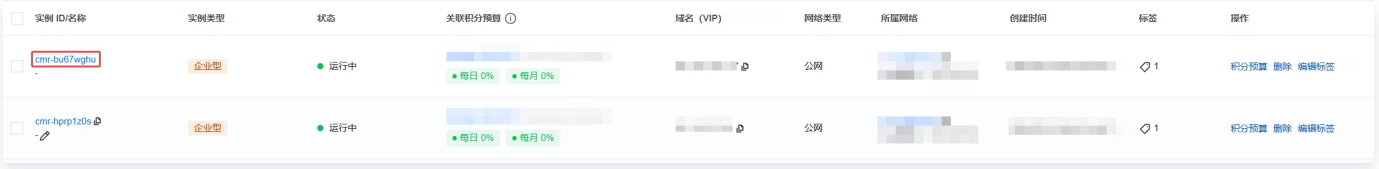
参数	说明
关联模型	选择用户组内 API Key 可以直接访问的模型。
关联意图路由规则	选择用户组内 API Key 可以使用的意图路由规则。

7. 配置完成后，单击**保存**。

删除用户组操作步骤

当业务不再需要按用户组维度管理 API Key 的可用模型、意图路由、积分预算、限速或标签策略时，您可以删除对应的自定义用户组，以简化资源管理。

1. 登录 模型路由 控制台，在实例管理页面，单击目标**实例 ID**。



2. 在实例详情页，选择**用户组**页签。
3. 在**用户组**页签中，在页面左侧的用户组列表中选择需要删除的用户组。

4. 清空组内 API Key。
   - 若用户组内仍存在 API Key，请在下方 Key 列表中选择对应 Key，单击**切换用户组**，并将其迁移至其他用户组。

   - 您也可以在**更多操作**中删除选中的 API Key。

> **说明：**
> 当用户组内仍有关联 API Key 时，用户组详情区域的**删除**按钮将置灰，无法执行删除操作。

5. 发起删除。在用户组详情区域，单击**删除**。

6. 确认删除。在弹出的确认窗口中核对用户组信息，确认无误后单击**确定**。

其他操作

- 向用户组添加或转移 API Key，详细操作指导请参见 分配 API Key 至用户组 。
- 调整用户组的积分预算，详细操作指导请参见 配置积分预算 。

- 配置模型和意图路由规则，详细操作指导请参见 配置模型调度管理 。

分配访问密钥

最近更新时间：2026-09-15 15:34:31

本文介绍如何将模型路由访问密钥（API Key）分配至用户组，包括单个分配和批量分配两种方式。

简介

模型路由访问密钥（API Key）是调用方访问模型路由实例时使用的身份凭证。将其分配至用户组，可进行用户组维度的统一管理。

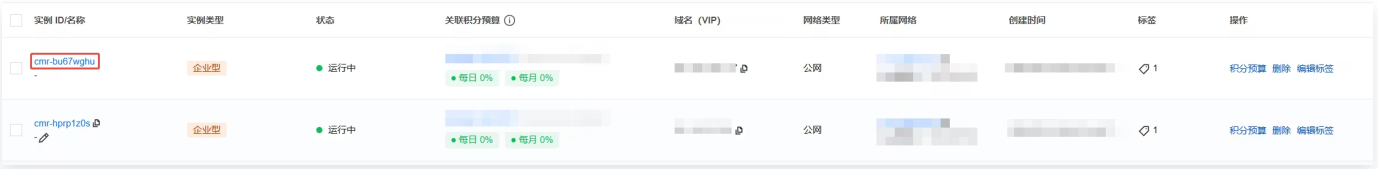
前提条件

1. 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 。
2. 已创建或导入模型路由访问密钥。详细操作指导请参见 创建模型路由访问密钥 。
3. 已创建目标用户组。详细操作指导请参见 配置用户组 。

操作步骤

步骤1：分配单个 API Key

1. 登录 模型路由 控制台，在**实例管理**页面，单击目标**实例 ID**。



2. 在实例详情页，选择**用户组**页签。
3. 在页面左侧的用户组列表中，选择目标 API Key 当前所属的用户组。新建或导入 API Key 时，如果未指定用户组，该 API Key 默认归属**“未分组 API Key”**用户组。

4. 在当前 API Key 列表中，找到目标 API Key，单击**分配用户组**。

5. 在弹出的对话框中，选择当前实例下的目标用户组，然后单击**确定**。

步骤2：批量分配 API Key

1. 在**用户组**页签左侧的用户组列表中，选择目标 API Key 当前所属的用户组。
2. 在 API Key 列表中，选择需要分配的多个 API Key。
3. 单击列表上方的**分配用户组**。

4. 在弹出的对话框中，选择当前实例下的目标用户组，然后单击**确定**。

> **说明：**
> - 每个 API Key 只能分配至当前实例下的一个用户组。
> - 将 API Key 分配至其他用户组后，该 API Key 将从原用户组中移出。

- API Key 分配完成后，将受目标用户组配置的模型访问范围、意图路由规则、积分预算和速率限制约束。

AI 工具接入 WorkBuddy

最近更新时间：2026-09-15 15:34:32

WorkBuddy 是腾讯云推出的全场景桌面 AI 智能体，适配 QQ / 企微生态，支持本地操作电脑、多模型切换。本文介绍如何在 WorkBuddy 中配置使用腾讯云模型路由 CMR。

安装 WorkBuddy

访问 WorkBuddy 官网 下载和安装 WorkBuddy：

- 如果您使用 Mac 系统，请参考 Mac 系统安装指南 进行安装。
- 如果您使用 Windows 系统，请参考 Windows 系统安装指南 进行安装。

获取 API Key

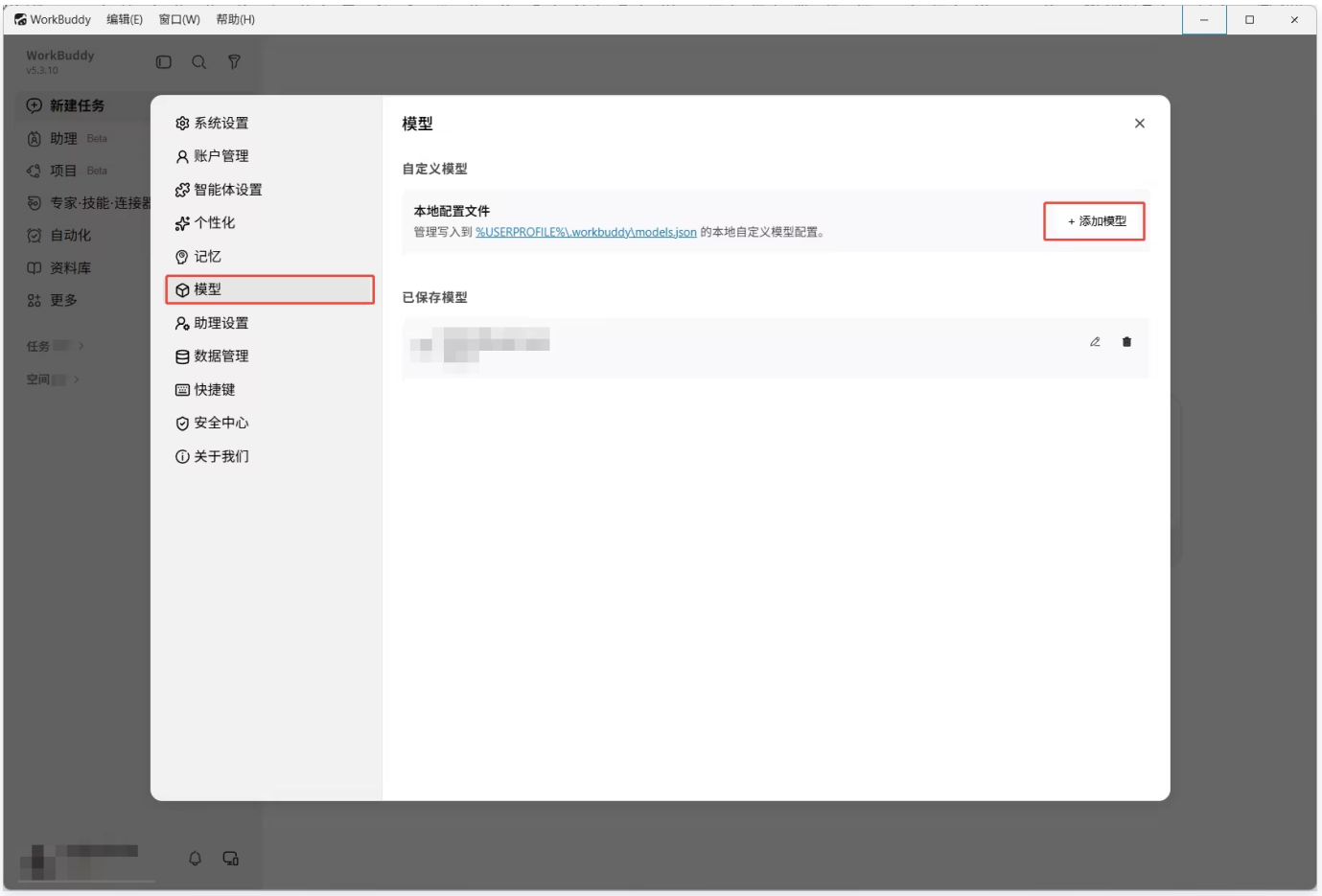
登录 模型路由 控制台，创建模型路由 API Key。操作详情请参见 模型路由访问密钥 。

> ⚠ **注意：**
> 创建完成后，请您务必复制并妥善保管 API Key，关闭创建成功弹窗后，将无法再次查看完整的 API Key。

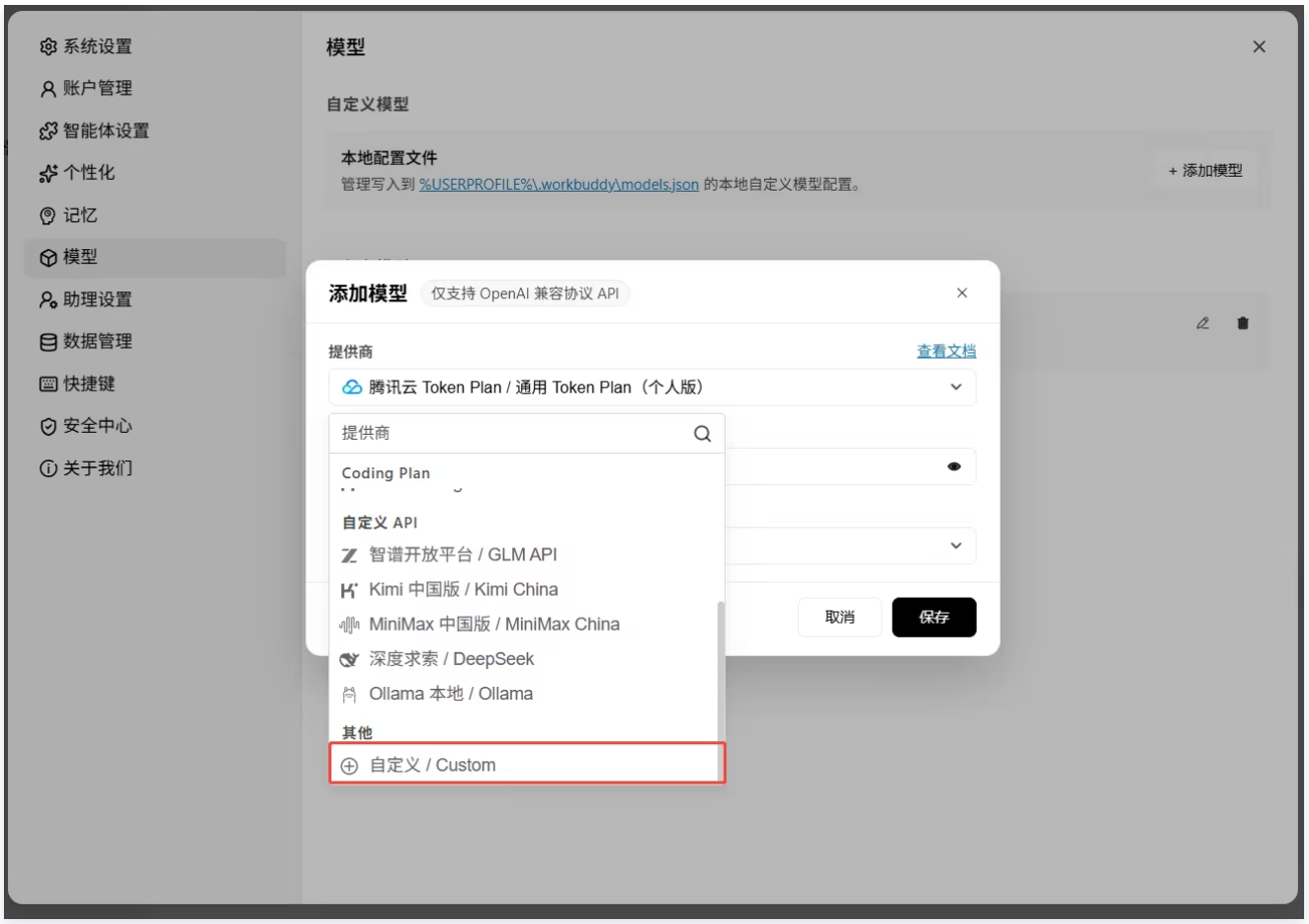
配置 WorkBuddy

1. 启动 **WorkBuddy**，单击左下角账户，选择**设置**。

2. 在左侧导航中选择**模型**，在**自定义模型**中，单击**添加模型**。



3. 配置以下信息，单击**保存**：
   - 在**提供商**选择框中，选择**自定义 / Custom**。如下图所示：



- 填写以下字段信息：
  - 接口地址：https://<用户域名>/v1/chat/completions
  - API Key：填入您创建的模型路由 API Key。
  - 模型名称：填写当前 API Key 支持的模型名称。
  - 高级工具：非必填，按需使用。

说明：

输入模型名称时需确认当前 API Key 的可访问范围，您需要填写当前 API Key 可用的意图路由或模型名称，您也可以设置 model=ModelRouter/auto，将按照 L1 模型路由策略指定可用模型。

4. 配置完成后，在 WorkBuddy 模型选择框中选择已添加的模型，即可开始对话。

CodeBuddy

最近更新时间：2026-09-15 15:34:32

CodeBuddy 是腾讯云推出的 AI 辅助编程工具，支持 CodeBuddy IDE 和 CodeBuddy Code 两种形态。本文分别介绍如何通过 CodeBuddy IDE 图形界面和 CodeBuddy Code 命令行配置使用腾讯云模型路由 CMR。

安装 CodeBuddy

CodeBuddy IDE

访问 CodeBuddy IDE 官网，下载并安装与本机操作系统和处理器匹配的版本。CodeBuddy IDE 支持 Windows 和 macOS。

CodeBuddy Code

根据本地环境安装 CodeBuddy Code。操作详情请参见 CodeBuddy Code 安装 。如果已安装 CodeBuddy Code，可以跳过此步骤。

获取 API Key

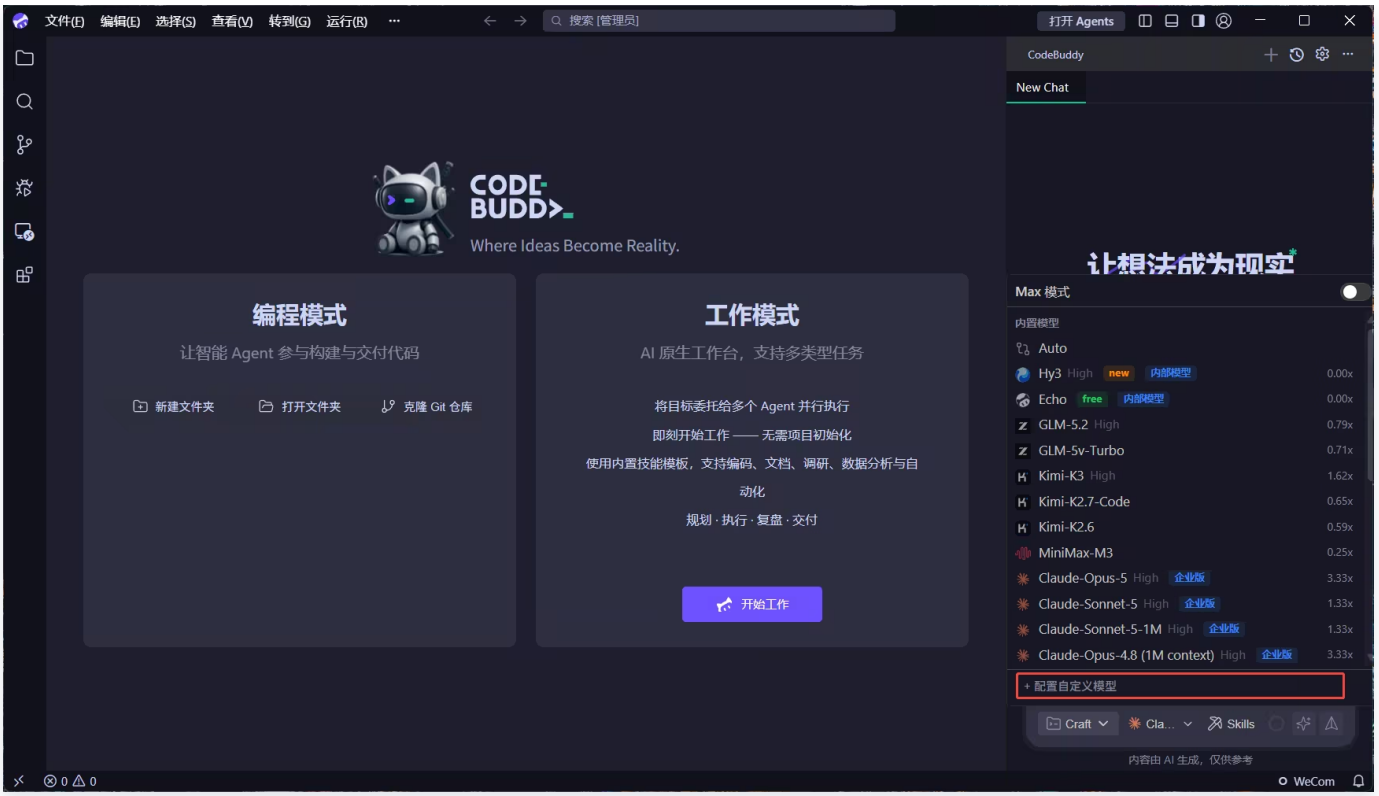
登录 模型路由 控制台，创建模型路由 API Key。操作详情请参见 创建模型路由访问密钥 。

> ⚠ **注意：**
> 创建完成后，请您务必复制并妥善保管 API Key，关闭创建成功弹窗后，将无法再次查看完整的 API Key。

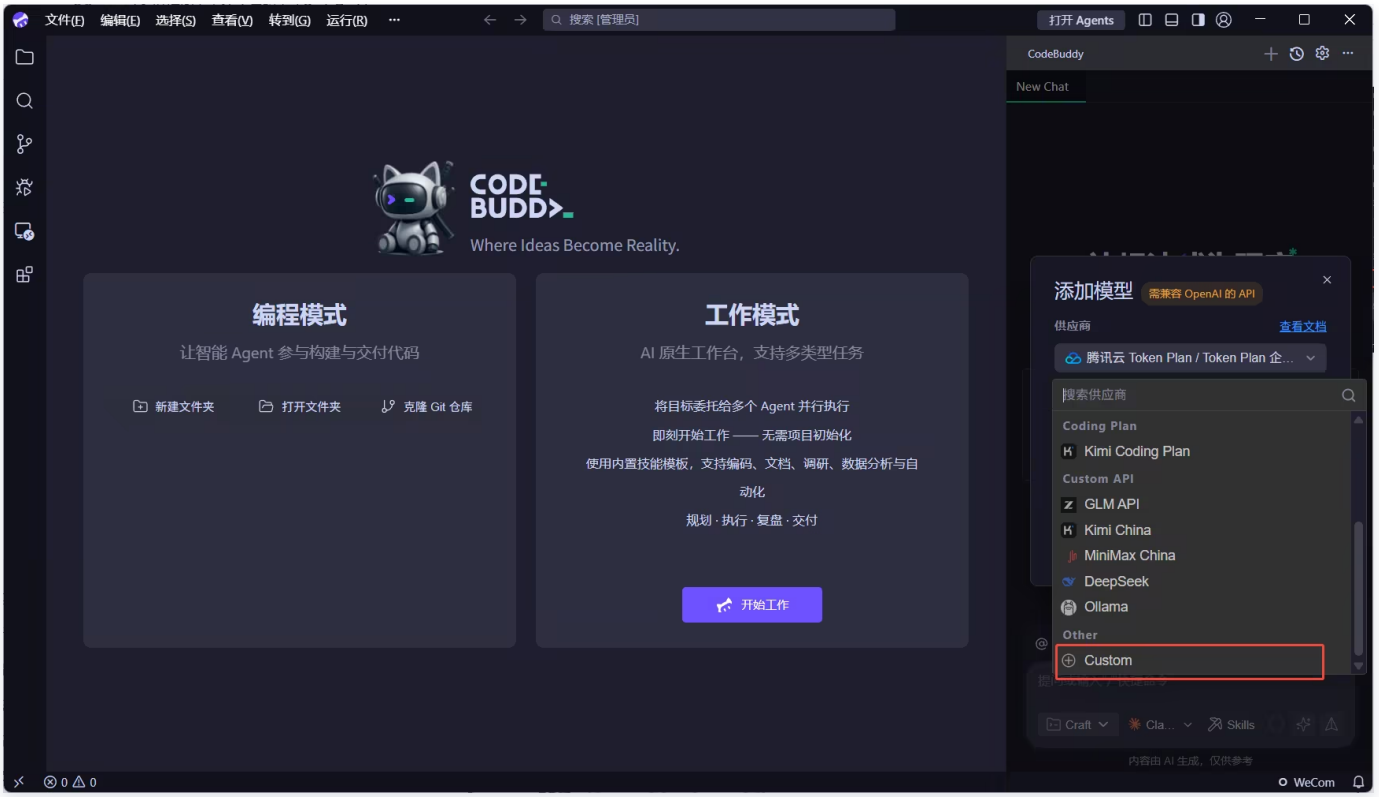
配置 CodeBuddy IDE

CodeBuddy IDE 支持通过图形界面添加 OpenAI 兼容模型，无需手动编辑配置文件。

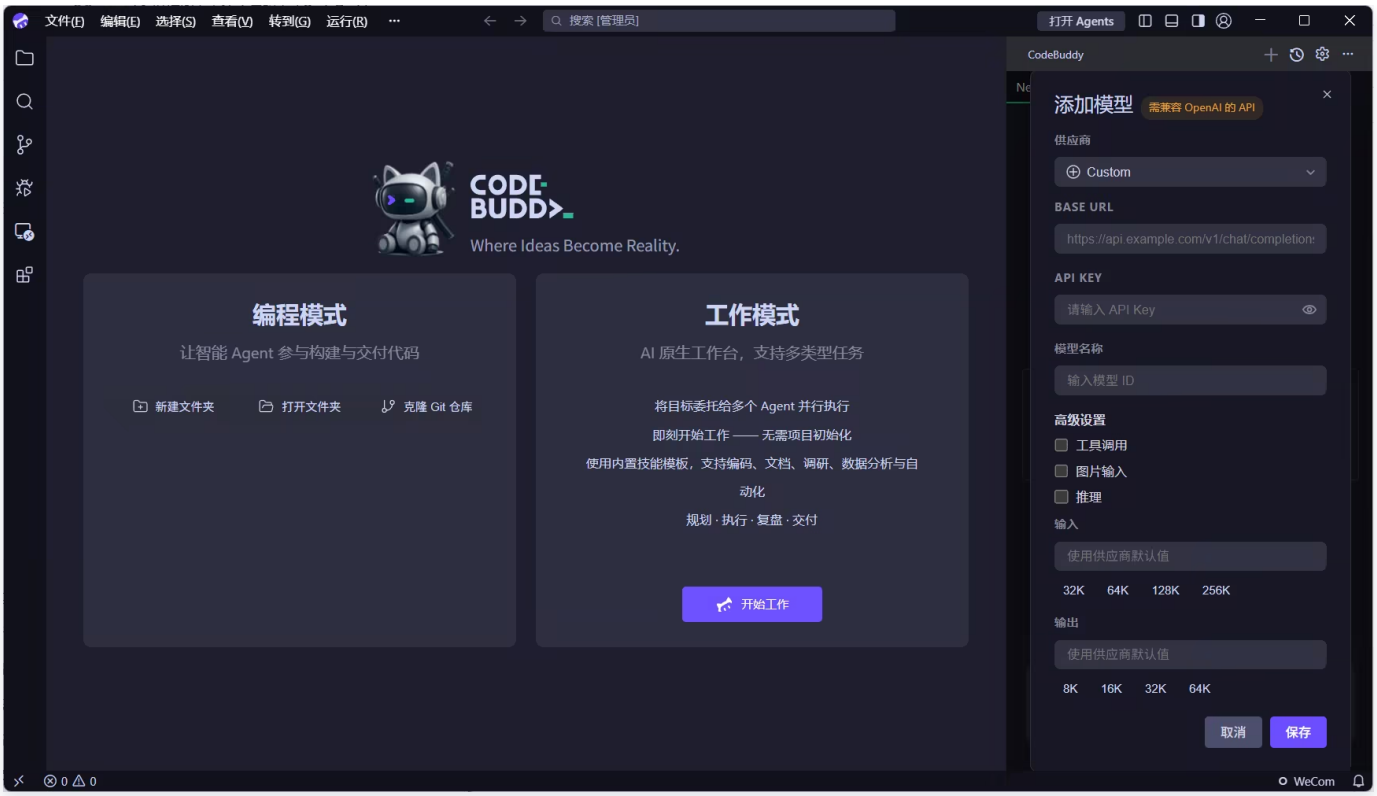
1. 启动 CodeBuddy IDE。在对话面板下方单击模型选择框，然后单击**配置自定义模型**。



2. 在添加模型页面，将**供应商**设置为 **Custom**。



3. 填写以下配置，并单击**保存**。



配置项	配置值
Base URL	`https://<用户域名>/v1/chat/completions`
API Key	模型路由 API Key
模型名称	填写当前 API Key 可调用的模型名称或意图路由名称，默认 `ModelRouter/auto`
高级设置	按需勾选

4. 返回对话面板，打开模型选择框，在 **Custom Models** 中选择已添加的模型，即可开始使用。

配置 CodeBuddy Code

1. CodeBuddy Code 通过 `models.json` 配置自定义模型。配置文件路径如下，配置文件不存在时新建即可。
   - Windows：`C:\Users\<用户名>\.codebuddy\models.json`
   - macOS/Linux：`~/.codebuddy/models.json`
2. 将以下配置复制到 `models.json`，并将 `<USER_API_KEY>` 和 `<用户域名>` 替换为当前模型路由实例的实际配置信息。

```
{
  "models": [
    {
```

```
      "id": "ModelRouter/auto",
      "name": "Tencent Cloud CMR",
      "vendor": "Tencent Cloud",
      "apiKey": "<USER_API_KEY>",
      "url": "https://<用户域名>/v1/chat/completions"
    }
  ]
}
```

说明：

- `ModelRouter/auto` 表示按照 L1 模型路由策略选择可用模型。您也可以将 `id` 中的值替换为当前实例已关联的模型名称或意图路由名称，`name` 是模型显示在 CodeBuddy 中展示的名称。
- 如果 `models.json` 已存在，请将上述模型追加到 `models` 中。更多相关配置，请参见 CodeBuddy Code 模型配置 。

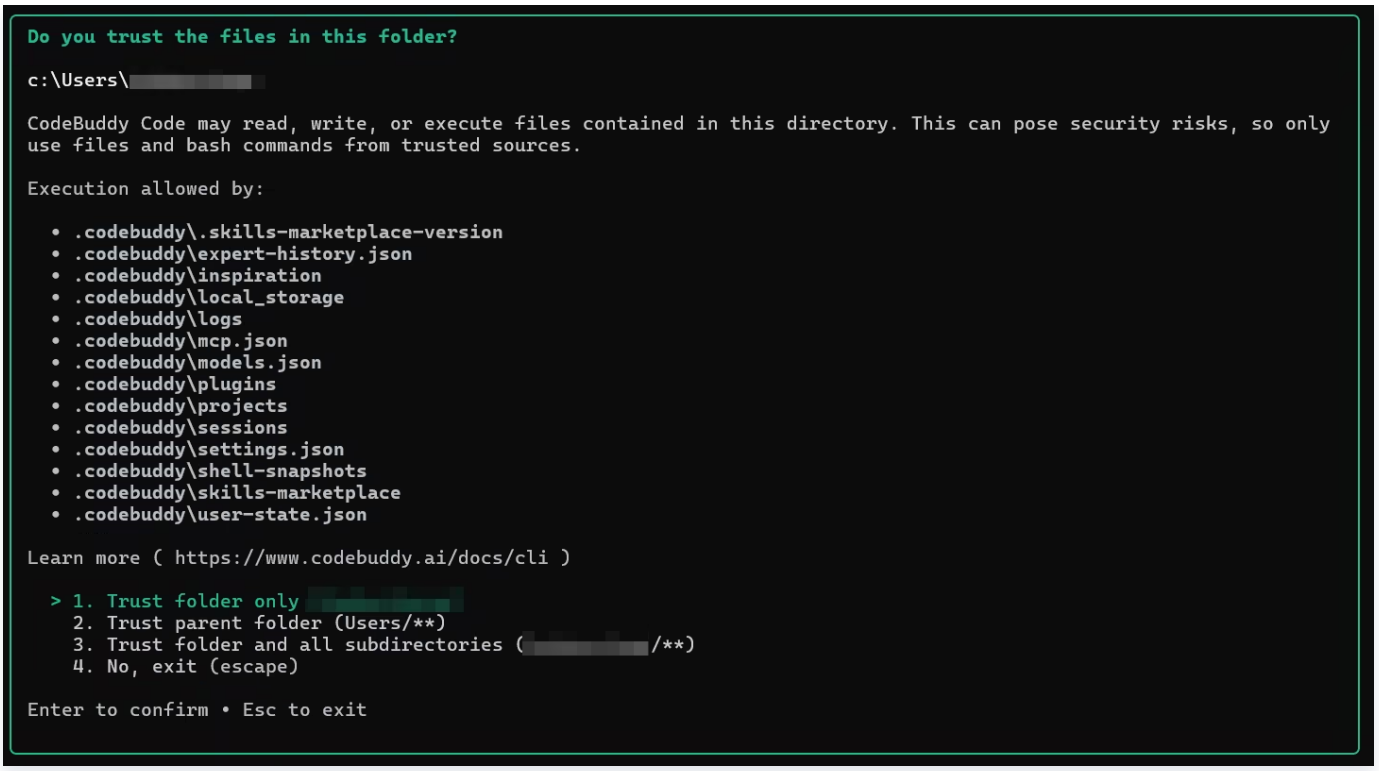
开始使用

1. 执行以下命令，启动 CodeBuddy Code。

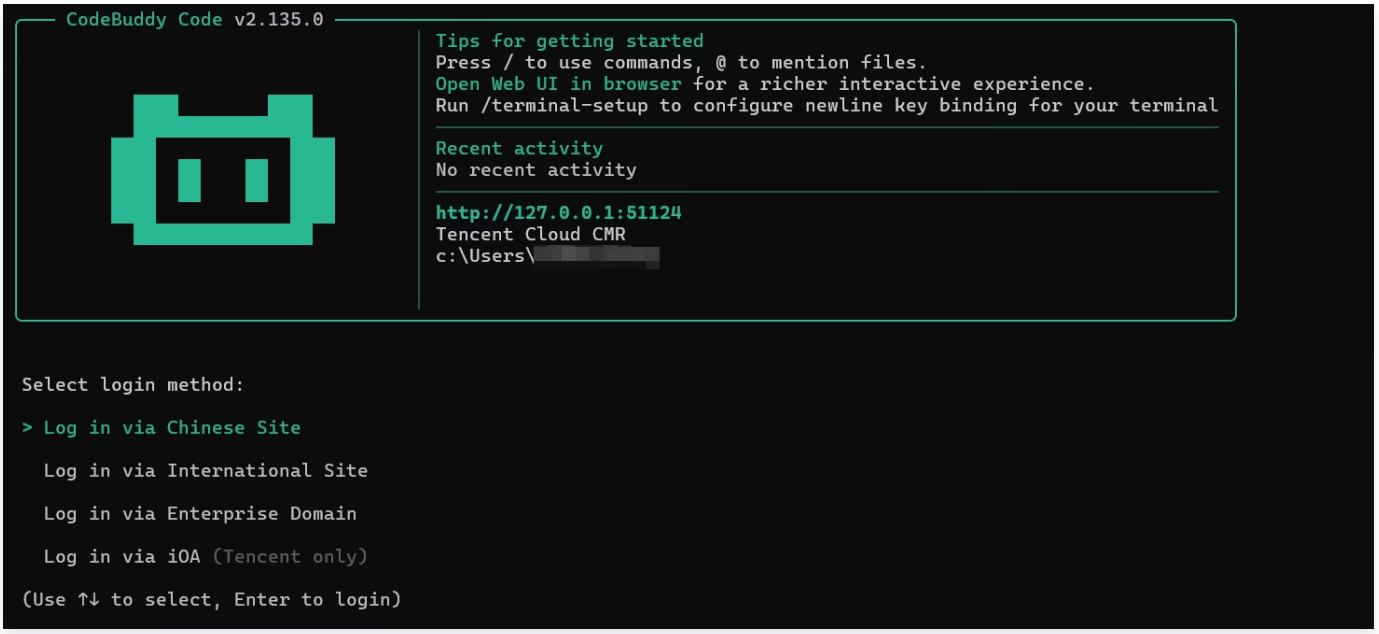
```
codebuddy
```

```
Running Environment

Operating System: Ubuntu 24.04.3 LTS / x86_64

Runtime Version: GNU bash, version 5.2.21(1)-release (x86_64-pc-linux-gnu)
```

2. 首次使用 CodeBuddy Code 时，根据页面提示确认是否信任当前工作目录。仅当您信任目录中的所有文件时，才选择信任并继续。



3. 根据实际场景选择登录方式，并按 **Enter** 键完成认证。



登录方式说明如下：

登录方式	适用场景	说明
Chinese Site	中国站用户	通过腾讯云中国站 (copilot.tencent.com) 进行认证，支持境内主流模型。
International Site	国际站用户	通过腾讯云国际站 (codebuddy.ai) 进行认证，支持境外主流模型。

Enterprise Domain	专享版/私有化部署	连接企业专享版或自建的 CodeBuddy 服务，需要输入企业提供的服务地址。
iOA	腾讯内部员工	通过腾讯 iOA 零信任系统进行认证，仅限腾讯内部员工使用。

4. 返回 `CodeBuddy Code` 命令行交互界面，输入 `/model` 并按 Enter 键。
5. 在模型列表中选择 **Tencent Cloud CMR**，即可开始使用。

> **警告：**
> 请勿泄露 API Key，或将包含实际 API Key 的配置文件和 Shell 配置提交到代码仓库。

Cursor

最近更新时间：2026-09-15 15:34:32

Cursor 是一款基于 VS Code 打造的 AI 代码编辑器，支持通过自然语言进行代码生成、代码理解、跨文件重构和智能调试。本文介绍如何在 Cursor 中配置使用腾讯云模型路由 CMR。

安装 Cursor

访问 Cursor 官网，下载并安装 Cursor。

> ⚠ **注意：**
> 由于 Cursor 的限制，只有订阅 Cursor Pro 及以上版本套餐的用户才支持自定义配置模型。

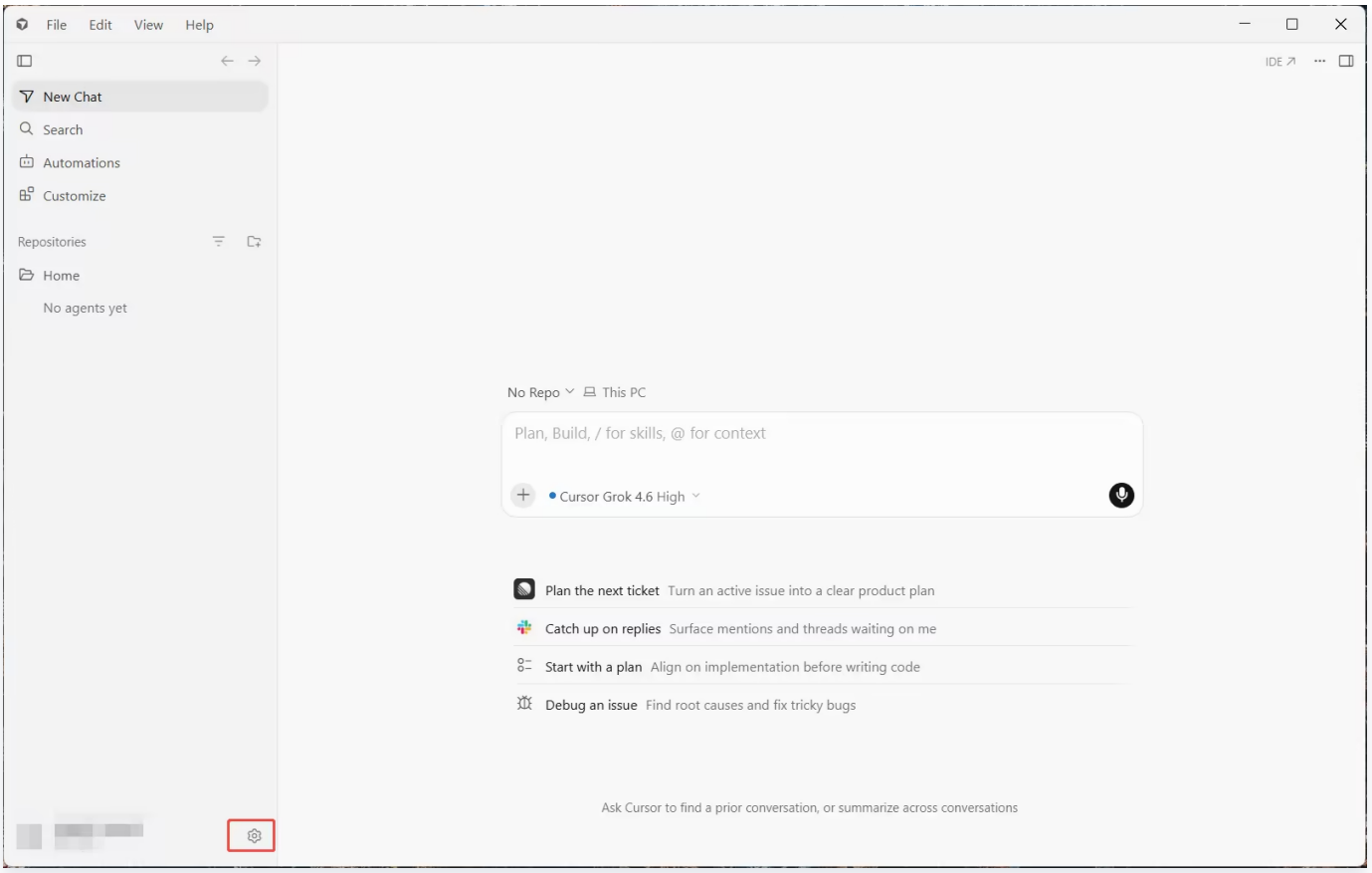
获取 API Key

登录 模型路由 控制台，创建模型路由 API Key。操作详情请参见 模型路由访问密钥 。

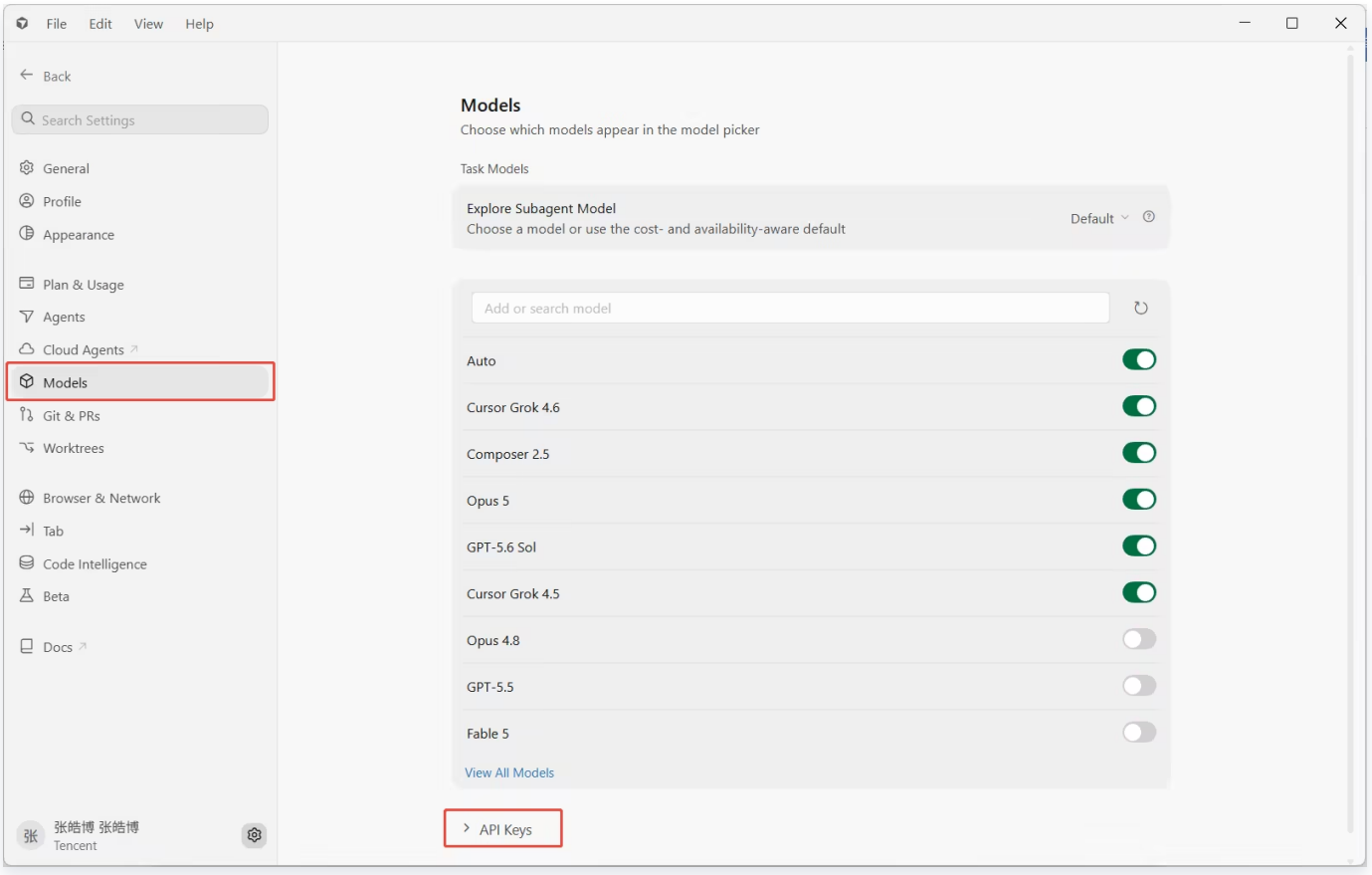
> ⚠ **注意：**
> 创建完成后，请您务必复制并妥善保管 API Key，关闭创建成功弹窗后，将无法再次查看完整的 API Key。

配置 Cursor

1. 启动 Cursor，单击左下角的 **Settings**。您也可以按 **Cmd/Ctrl+Shift+P**，输入 `Cursor Settings` 并按 **Enter** 键。

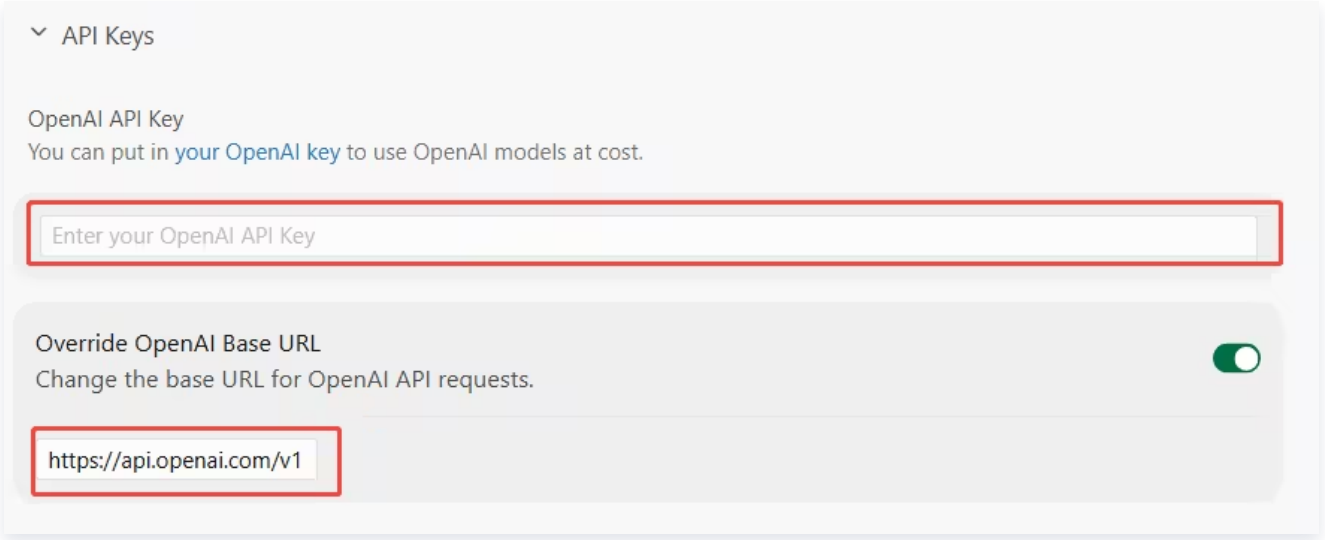


2. 在左侧菜单栏中选择 Models，展开 API Keys。

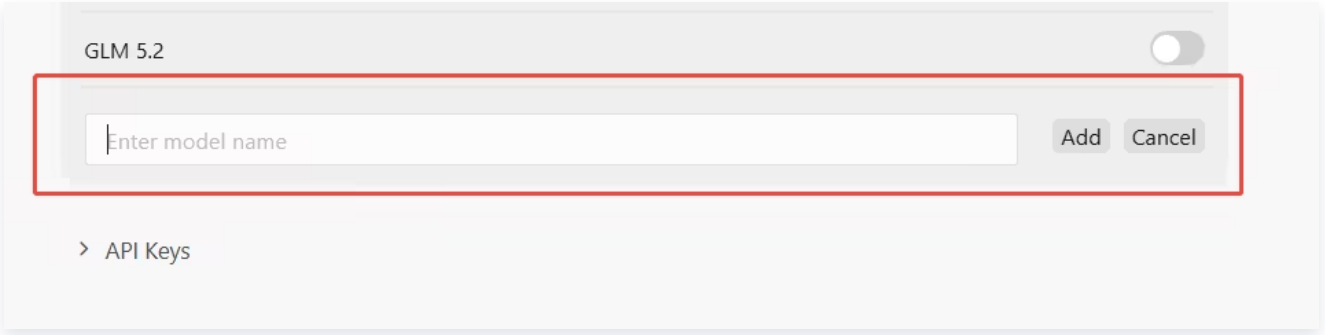


3. 配置以下信息：

- **OpenAI API Key**：填入模型路由 API Key。
- **Override OpenAI Base URL**：开启该配置，并填写 `https://<用户域名>/v1` 。



- **Add Custom Model**：单击 **View All Models** > **Add Custom Model**，填写当前模型路由实例支持的模型名称或意图路由名称。



> **说明：**
> 填写模型名称时，请确认模型已经关联到当前模型路由实例。您也可以填写 `ModelRouter/auto` ，由 L1 模型路由策略选择可用模型。

4. 配置完成后，在 Cursor 的对话窗口中选择已添加的模型，即可开始使用。

Claude code

最近更新时间：2026-09-15 15:34:32

Claude Code 是一款运行在终端中的 AI 编程工具，可以通过自然语言辅助开发者编码、调试和管理代码。本文介绍如何在 Claude Code 中配置使用腾讯云模型路由 CMR。

安装 Claude Code

根据本地环境安装 Claude Code。操作详情请参见 Claude Code 文档 。如果已安装 Claude Code，可以跳过此步骤。
安装完成后，执行以下命令查看版本。终端显示版本号，表示安装成功。

```
claude --version
```

获取 API Key

登录 模型路由 控制台，创建模型路由 API Key。操作详情请参见 模型路由访问密钥 。

> ⚠ **注意：**
> 创建完成后，请您务必复制并妥善保管 API Key，关闭创建成功弹窗后，将无法再次查看完整的 API Key。

配置 Claude Code

跳过登录验证

编辑或新建 `~/.claude.json`（Windows 路径：`C:\Users\<用户名>\.claude.json`），将 `hasCompletedOnboarding` 设为 `true`，跳过 Anthropic 官方登录验证。

```
{
  "hasCompletedOnboarding": true
}
```

创建配置文件

1. 编辑或新增 `settings.json` 文件，不同系统的配置文件路径不同，具体可参考：
   - MacOS/Linux 对应路径为：`~/.claude/settings.json`
   - Windows 对应路径为：`用户目录/.claude/settings.json`

   若 `settings.json` 文件不存在，可使用文本编辑器（如 VSCode、记事本等）手动创建，或在终端中执行以下命令创建：

macOS 或 Linux

执行以下命令，创建用户级配置文件：

```
mkdir -p ~/.claude && touch ~/.claude/settings.json
```

Running Environment

Operating System: Ubuntu 24.04.3 LTS / x86_64

Runtime Version: GNU bash, version 5.2.21(1)-release (x86_64-pc-linux-gnu)

配置文件路径为 `~/.claude/settings.json` **。**

Windows

在 PowerShell 中执行以下命令，创建用户级配置文件：

```
New-Item -ItemType Directory -Force -Path "$HOME\.claude" | Out-
Null; New-Item -ItemType File -Force -Path
"$HOME\.claude\settings.json" | Out-Null
```

配置文件路径为 `$HOME\.claude\settings.json` **。**

2. **在** `settings.json` **中写入以下配置，并将** `<USER_API_KEY>` **和** `<用户域名>` **替换为当前模型路由实例的实际信息。**

```
{
  "env": {
    "ANTHROPIC_BASE_URL": "https://<用户域名>",
    "ANTHROPIC_AUTH_TOKEN": "<USER_API_KEY>",
    "ANTHROPIC_MODEL": "ModelRouter/auto",
    "ANTHROPIC_DEFAULT_OPUS_MODEL": "ModelRouter/auto",
    "ANTHROPIC_DEFAULT_SONNET_MODEL": "ModelRouter/auto",
    "ANTHROPIC_DEFAULT_HAIKU_MODEL": "ModelRouter/auto",
    "CLAUDE_CODE_SUBAGENT_MODEL": "ModelRouter/auto",
    "ENABLE_TOOL_SEARCH": "false"
  }
```

```
}
```

环境变量	是否必填	说明
`ANTHROPIC_BASE_URL`	是	Claude Code 请求的 API 网关地址。
`ANTHROPIC_AUTH_TOKEN`	是	用于鉴权的 API Key，需替换为您在 CMR 控制台创建的真实 API Key。请妥善保管，避免泄露。
`ANTHROPIC_MODEL`	是	Claude Code 默认调用的模型名，本文以 `ModelRouter/auto` 为例。
`ANTHROPIC_DEFAULT_OPUS_MODEL`	否	Claude Code 中 Opus 档位（高复杂度任务）映射到的模型。
`ANTHROPIC_DEFAULT_SONNET_MODEL`	否	Claude Code 中 Sonnet 档位（默认日常任务）映射到的模型。
`ANTHROPIC_DEFAULT_HAIKU_MODEL`	否	Claude Code 中 Haiku 档位（轻量/快速任务）映射到的模型。
`CLAUDE_CODE_SUBAGENT_MODEL`	否	Claude Code 创建子任务（subagent）时使用的模型。
`ENABLE_TOOL_SEARCH`	否	是否启用 Claude Code 内置的原生 web search 工具。

说明：

- `ANTHROPIC_BASE_URL` 填写模型路由实例域名，不需要追加 `/v1/messages`。Claude Code 会自动拼接 Anthropic Messages 请求路径。
- `ModelRouter/auto` 表示按照 L1 模型路由策略选择可用模型。您也可以将各模型变量替换为当前实例已关联的模型名称或意图路由名称。
- 将 Opus、Sonnet、Haiku 和子代理模型设置为同一值，可以避免 Claude Code 在不同任务中切换到未配置的模型。
- Claude Code 禁止第三方模型使用内置原生搜索工具，因此将 `ENABLE_TOOL_SEARCH` 设置为 `false`。如需联网搜索，可以按需配置搜索 MCP 服务。

3. 保存配置后，关闭当前终端并打开新的终端窗口，使配置生效。

验证配置

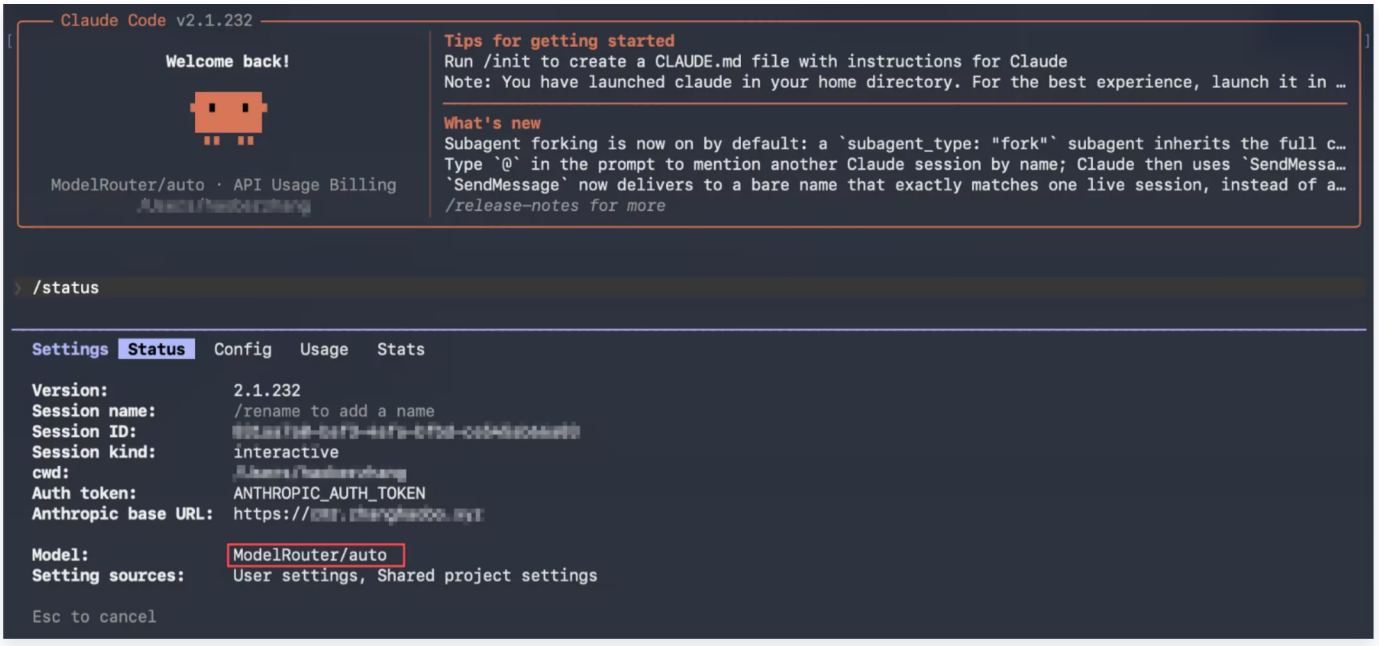
1. 在终端中进入需要使用 Claude Code 的项目目录。

```
cd <项目目录>
```

2. 执行以下命令，启动 Claude Code。

```
claude
```

3. 启动成功后，输入 `/status` 并按 Enter 键，检查当前配置。



当 `Anthropic base URL` 显示模型路由实例域名，且 `Model` 显示 `ModelRouter/auto` 或您配置的模型名称时，表示配置已生效。

4. 确认配置生效后，即可开始对话。


Claude Code v2.1.232
Welcome back!
ModelRouter/auto · API Usage Billing
Tips for getting started
Run /init to create a CLAUDE.md file with instructions for Claude
Note: You have launched claude in your home directory. For the best experience, launch it in …
What's new
Subagent forking is now on by default: a `subagent_type: "fork"` subagent inherits the full c…
Type `@` in the prompt to mention another Claude session by name; Claude then uses `SendMessa…
`SendMessage` now delivers to a bare name that exactly matches one live session, instead of a…
/release-notes for more
/status
Settings dialog dismissed
你好
Thought for 8s (ctrl+o to expand)
● 你好！很高兴见到你。我是 Claude Code，有什么我可以帮你的吗？
我可以帮你完成各种软件工程任务，比如：
- 写代码 — 实现新功能、重构、修复 bug
- 浏览代码库 — 查找文件、理解代码结构
- 运行命令 — 执行 shell 命令、安装依赖、运行测试
- 搜索和研究 — 查找相关信息、API 文档
- 调试 — 分析问题、提出解决方案
- CLI 使用 — 回答关于 Claude Code 本身的问题
你今天想做什么？
* Brewed for 9s
⏸ manual mode on · ? for shortcuts · ← for agents

Codex

最近更新时间：2026-09-15 15:34:32

Codex 是 OpenAI 推出的 AI 编程工具，可以通过自然语言完成代码库理解、文件编辑和编程任务。本文介绍如何在 Codex 中配置使用腾讯云模型路由 CMR。

安装 Codex

根据本地环境安装 Codex。操作详情请参见 Codex 快速入门。如果已安装 Codex，可以跳过此步骤。
安装完成后，执行以下命令查看版本。终端显示版本号，表示安装成功。

```
codex --version
```

> ⚠ **注意：**
> 新版 Codex 通过 OpenAI Responses API 调用模型。配置前，请确认模型路由实例及目标模型支持 Responses API。

获取 API Key

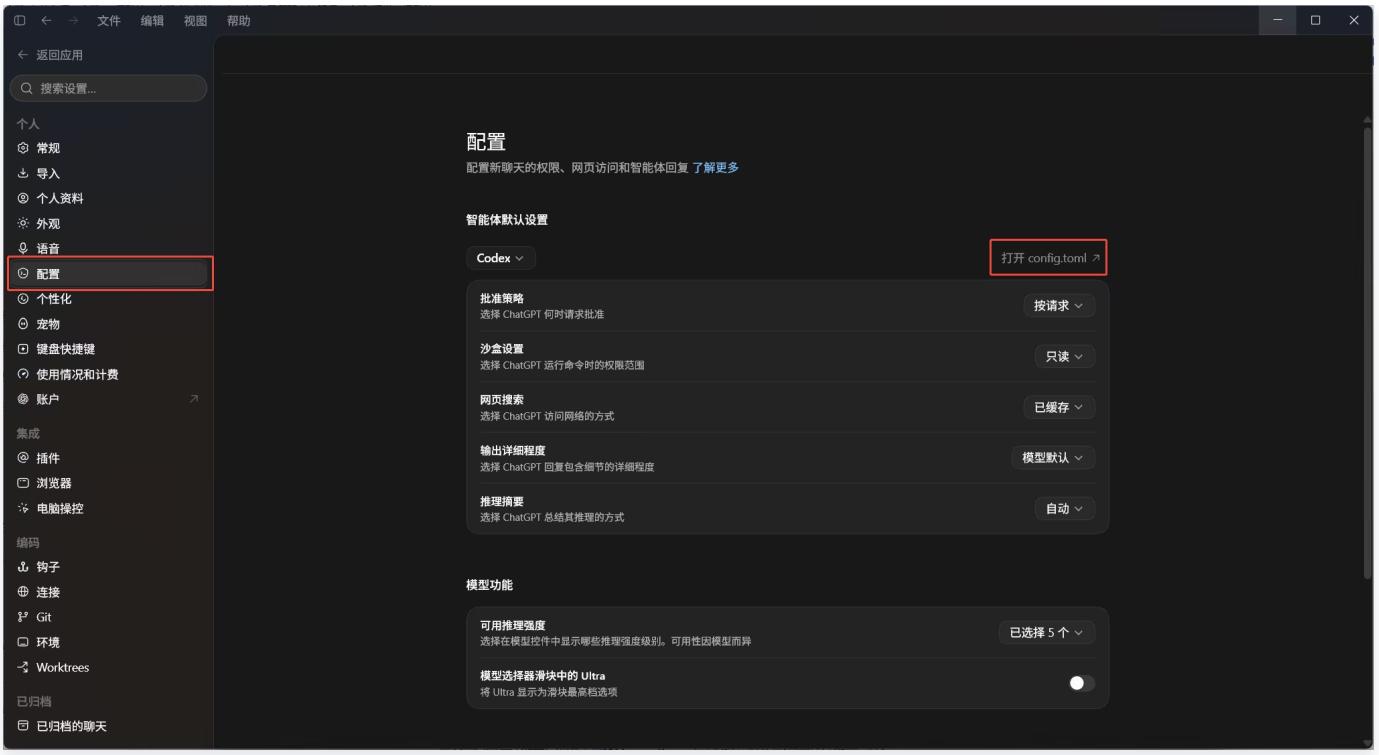
登录 模型路由 控制台，创建模型路由 API Key。操作详情请参见 模型路由访问密钥 。

> ⚠ **注意：**
> 创建完成后，请您务必复制并妥善保管 API Key，关闭创建成功弹窗后，将无法再次查看完整的 API Key。

配置 Codex

配置模型

1. 打开 Codex，单击左下角的**设置**。
2. 在设置页面选择**配置**，然后单击**打开 config.toml**，使用系统默认编辑器打开配置文件。



> **说明：**
> 您也可以直接打开用户级配置文件：macOS/Linux 为 `~/.codex/config.toml`，Windows 为 `$HOME\.codex\config.toml`。如果文件不存在，请先创建 `.codex` 目录和 `config.toml` 文件。

3. 在 `config.toml` 文件顶部写入以下内容。文件中已有其他配置时，将以下配置添加至文件开头。

```
model_provider = "tencent-cloud-cmr"
model = "ModelRouter/auto"
disable_response_storage = true

[model_providers.tencent-cloud-cmr]
name = "Tencent Cloud CMR"
base_url = "https://<用户域名>/v1"
env_key = "CMR_API_KEY"
wire_api = "responses"
```

配置项说明如下：

配置项	说明
`model_provider`	指定 Codex 使用的自定义模型提供商，值需与 `[model_providers.tencent-cloud-cmr]` 中的名称保持一致。

`model`	指定 Codex 调用的模型。`ModelRouter/auto` 表示按照 L1 模型路由策略选择可用模型，也可填写实例已关联的模型名称或意图路由名称。
`disable_response_storage`	禁止上游服务存储响应数据。
`name`	指定客户端中对外展示的模型名称。
`base_url`	填写模型路由实例域名，并在末尾添加 `/v1` 。Codex 会自动请求 `/v1/responses` 。
`env_key`	指定保存模型路由 API Key 的环境变量名称。
`wire_api`	Codex 使用的接口协议，固定填写 `responses` 。

4. 保存并关闭 `config.toml` 。

配置 API Key

请根据操作系统配置 `CMR_API_KEY` 环境变量。

macOS

在终端中执行以下命令：

```
launchctl setenv CMR_API_KEY "<USER_API_KEY>"
```

Windows

CMD

1. 在 CMD 中运行以下命令，设置环境变量。

```
setx CMR_API_KEY "<USER_API_KEY>"
```

2. 打开一个新的 CMD 窗口，运行以下命令检查环境变量是否生效。

```
echo %CMR_API_KEY%
```

PowerShell

1. 在 PowerShell 中运行以下命令，设置环境变量。

```
[Environment]::SetEnvironmentVariable("CMR_API_KEY",
"USER_API_KEY", [EnvironmentVariableTarget]::User)
```

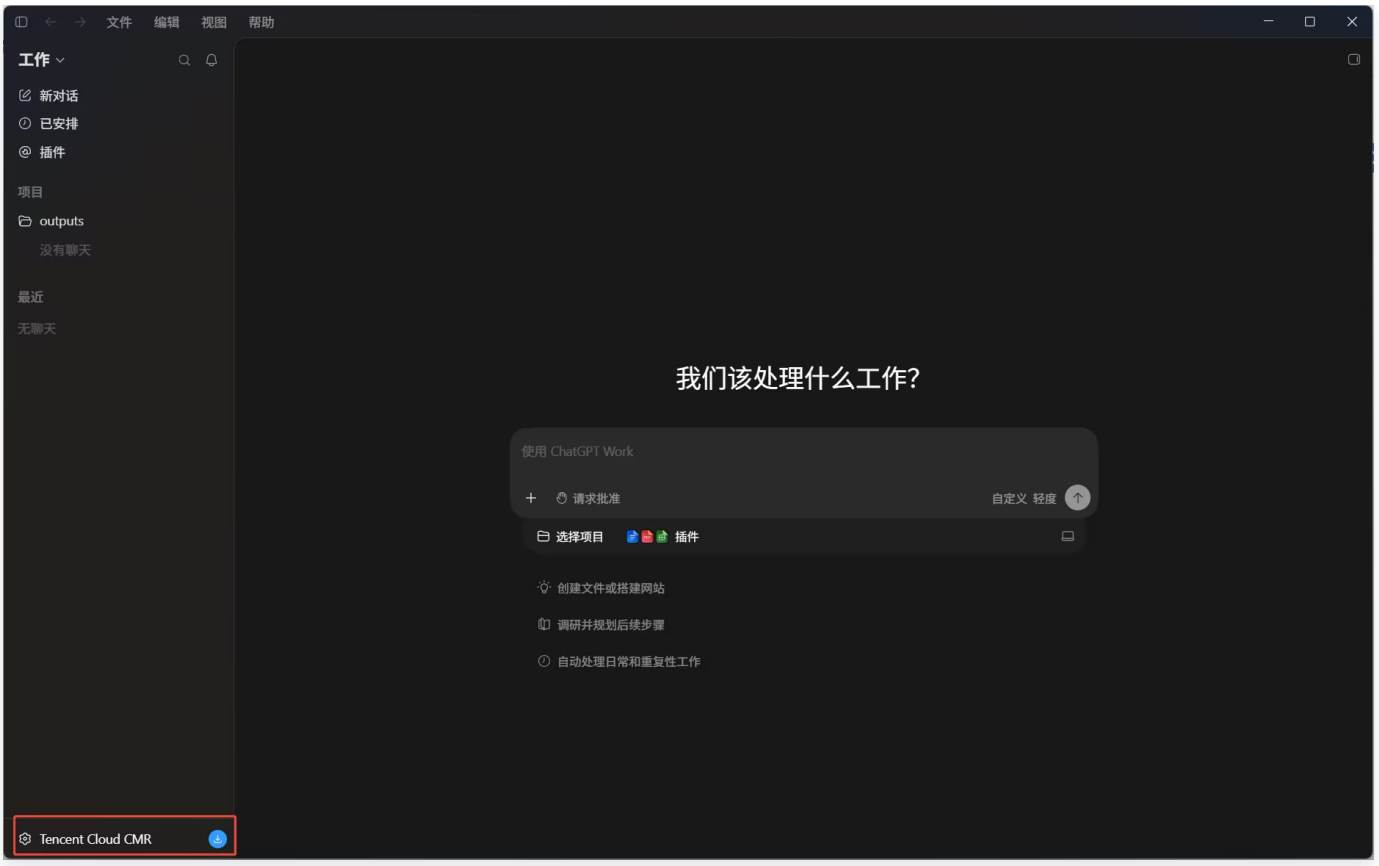
2. 打开一个新的 PowerShell 窗口，运行以下命令检查环境变量是否生效。

```
echo $env:CMR_API_KEY
```

警告：
请勿泄露 API Key，或将包含真实 API Key 的命令、Shell 配置和日志提交到代码仓库。

验证配置

1. 完全退出 Codex。仅关闭窗口不会结束 Codex 进程。
2. 重新启动 Codex，使 `config.toml` 和环境变量生效。客户端右下角显示 Tencent Cloud CMR，与配置的模型名称一致。



3. 在 Codex 输入框中输入测试消息，例如“请用一句话说明当前使用的模型”，然后发送。
4. Codex 正常返回响应，表示 CMR 配置已生效。
5. 如果调用失败，请按以下顺序检查：
   - `base_url` 是否为模型路由实例域名，并以 `/v1` 结尾。
   - `CMR_API_KEY` 是否已在 Codex 进程环境中生效。
   - `model` 是否为当前实例可用的模型名称、意图路由名称。
   - 目标模型是否支持 Responses API。