模型路由

快速入门

腾讯云

快速入门

最近更新时间：2026-09-15 15:34:30

本教程介绍了如何快速开始使用模型路由。

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

步骤1：创建模型路由实例

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

参数	说明
地域	可按需选择模型路由 CMR 实例所在的地域。
实例类型	支持共享型和企业型两种类型， 其中共享型单地域仅支持创建1个实例。
网络类型	支持公网和内网两种网络类型。 • 公网实例，直接面向互联网提供服务。 • 内网实例，仅在 VPC 内部提供服务，适合内部系统互联。仅企业型实例支持选择。
监听协议	支持 HTTP 和 HTTPS 两种协议。
证书	仅选择 HTTPS 协议时需要选择。
所属网络	选择私有网络。私有网络是用户在腾讯云上建立的一块逻辑隔离的网络空间，在私有网络内，您可自由定义网段划分、IP 地址和路由策略。
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。
实例名称	可输入255个字符，允许字母、数字、中文、句号、下划线、短划线。不填写时默认自动生成。
标签	选择标签键和标签值，也可选择新建标签，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。
5. 在实例列表中，即可查看您创建的实例。

步骤2：创建模型路由访问密钥

1. 在左侧导航栏中，单击**模型路由 CMR** 标签下的**实例管理**。单击您创建实例的**实例 ID**，进入**实例基本信息**页面，切换至**用户组**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	可选择速率限制或积分预算。
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**速率限制**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**速率限制**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。
4. **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：**
> 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。

步骤3： 创建 BYOK 实例

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**。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 。

3. 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。

4. 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。

5. 完成以上参数配置后，单击**确定**完成 BYOK 实例的创建。

步骤4：配置模型调度管理

1. 在左侧导航栏中，单击**实例管理**。
2. 单击您创建实例的**实例 ID**，进入**实例基本信息**页面，切换至**模型调度管理**页签。
3. 在**关联模型**列表右侧单击**批量关联**，并选择需要关联的模型，确认后进行权重配置与模型关联。
4. 配置意图路由。在**路由策略示意图**中，在**意图路由**模块单击**新建**。或者在页面**意图路由**模块，单击**新建规则**。
   - 可以选择内置的业务模板，也可以自定义意图路由业务。
   - 通过配置意图路由的意图分类和对应的模型，系统会自动识别用户消息的意图分类，将请求路由到对应场景的模型处理。
5. 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。
     - 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。
   - 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、最低系数路由。
     - 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。
     - 最低系数路由：优先分发到积分较低的模型。
6. 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。
7. 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 策略列表右侧单击**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**实例管理**。单击您创建的**实例 ID**，进入**实例基本信息**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

1. 监控资源消耗：切换至**用量详情**页签。在此可查看 Token 消耗和模型资源包使用情况，避免额度不足导致调用失败。
2. 查看访问日志：切换至**日志**页签，将访问日志投递至日志服务（CLS），用于故障排查与安全审计。
3. 测试配置效果：切换至**聊天测试**页签，验证路由策略、Fallback 等配置是否按预期生效。

4. 开启安全防护：切换至**安全防护**页签，根据需要开启 WAF 大模型防护或配置安全组防护，保障模型服务的内容与访问安全。