

模型路由 常见问题



腾讯云

【 版权声明 】

©2013–2026 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【 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。

常见问题

最近更新时间：2026-09-15 15:34:30

基础概念问题

- 共享型企业型实例有什么区别？如何选择？
- 速率限制和积分预算两种限制类型有什么区别？
- TPM 和 RPM 分别代表什么？应该如何设置？
- 什么是 BYOK（Bring Your Own Key）？
- 模型路由支持哪些模型提供商？
- 模型路由的意图路由是什么？如何配置？

计费配额问题

- 模型路由处理费是如何计算的？
- 配额不足怎么办？如何提升配额？

共享型企业型实例有什么区别？

模型路由提供两种实例类型，适用场景不同：

对比维度	共享型实例	企业型实例
适用场景	开发测试、功能验证、POC 验证	生产环境、业务正式上线
网络类型	仅公网	可选公网或内网
证书管理	实例自带证书	需自行绑定 SSL 证书
所属网络	无需选择	需要选择 VPC 网络

选型建议：开发测试阶段使用共享型实例快速验证；生产环境务必使用企业型实例以获得更高的安全隔离和资源配额。

速率限制和积分预算两种限制类型有什么区别？

在创建 API Key 时，需要选择限制类型。两者对比如下：

对比维度	速率限制	积分预算
控制目标	控制请求频率和 Token 吞吐速度	控制总消耗积分上限
配置参数	TPM + RPM	积分预算总额

适用场景	防止某个 Key 调用过频导致服务压力过大	防止某个 Key 消耗过多积分超出预算
生效方式	实时限流（滑动窗口）	累计扣减，耗尽后停止服务
典型用法	给不同团队/应用分配不同的速率配额	给不同项目/客户分配固定的积分额度

选型建议：如果您关注的是“不要调用太频繁”，选择速率限制；如果您关注的是“不要花太多积分”，选择积分预算。两者也可以配合使用——在实例层面设速率限制，在 Key 层面设积分预算。

TPM 和 RPM 分别代表什么？应该如何设置？

TPM（Tokens Per Minute）：每分钟允许处理的最大 Token 数量，单位为千/分钟。例如：设置为100表示每分钟最多处理100,000个 Token。

RPM（Requests Per Minute）：每分钟允许的最大请求次数，单位为次/分钟。例如：设置为60表示每分钟最多接受60次请求。

这两个参数出现在两处配置位置，含义有所不同：

配置位置	TPM 含义	RPM 含义
创建实例时	整个实例每分钟的最大 Token 吞吐上限	整个实例每分钟的最大请求频率上限
创建 API Key 时（速率限制模式）	单个 Key 每分钟的最大 Token 吞吐上限	单个 Key 每分钟的最大请求频率上限

设置建议：

- 实例级别的 TPM/RPM 应根据您的业务峰值和预算合理设置，避免因限制过低导致请求被拒绝。
- Key 级别的 TPM/RPM 用于细粒度控制不同用户/应用的配额，通常小于或等于实例级别配额。
- RPM 限制主要用于防止突发流量冲击；TPM 限制主要用于控制 Token 成本。

什么是 BYOK（Bring Your Own Key）？

BYOK（Bring Your Own Key，自带密钥）是模型路由的核心接入模式，允许您使用自己在各大模型平台上申请的 API Key 通过模型路由网关进行统一管理和调度。

模型路由支持哪些模型提供商？

BYOK（Bring Your Own Key，自带密钥）模式允许用户携带自己的模型密钥，灵活接入原厂模型、第三方代理或自建模型。

提供商类型	说明	资源消耗	状态
-------	----	------	----

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 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

模型路由的意图路由是什么？如何配置？

意图路由是模型路由的智能分发能力。系统自动识别用户消息的意图分类，将请求路由到对应场景的最优模型进行处理。

工作原理：

1. 客户端发送请求到模型路由网关。
2. 网关对请求内容进行意图识别（例如：通用对话、内容生成、编码与技术等）。
3. 根据预先配置的意图-模型映射规则，将请求分发到对应的模型。

配置方式：

1. 登录 [模型路由](#) 控制台，在实例管理页面单击目标实例 ID 进入实例详情页，切换至模型调度管理页面。
2. 在路由策略示意图中的意图路由模块单击新建，或在页面意图路由模块单击新建规则。
3. 可选择内置的业务模板，也可以自定义意图路由业务。
4. 配置意图分类与对应模型的映射关系。

模型路由处理费是如何计算的？

模型路由计费项参考 [计费概述](#) 。其中模型路由的处理费采用按 Token 量计费的模式，与业界常见的“按比例加成”模式有本质区别：

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（固定单价）	按比例传递上游成本（浮动价格）
定价模式	固定单价，不随上游定价变化	收费随上游模型定价波动
覆盖范围	经过网关的所有输入+输出 Token 总和	主要依据上游返回的 usage 字段

计算公式：模型路由处理费 = 模型路由处理费单价（元/百万 Token）× Token 总量。

配额不足怎么办？如何提升配额？

模型路由的各项默认配额如下：

实例维度

类型	企业型		共享型	
	默认配额（单位：个）	是否支持提升	默认配额（单位：个）	是否支持提升
单账号单地域可创建的实例数量	5	可提升，请 提交工单 咨询。	1	不支持提升。
每个实例可添加的 Key 数量	200	可提升，请 提交工单 咨询。	10	不支持提升。
每个实例可关联的模型数量	50	可提升，请 提交工单 咨询。	20	不支持提升。
每个实例可创建意图路由的数量	5	不支持提升。	—	
每个实例的意图路由每层可添加的模型数量	5	不支持提升。	—	
每个实例可关联的用户组数量	20	可提升，请 提交工单 咨询。	—	
每个实例可添加提示词的数量	20	可提升，请 提交工单 咨询。	—	

BYOK 维度

类型	默认配额（单位：个）	是否支持提升
单账号单地域可创建的原厂 BYOK 数量	20	可提升，请 提交工单 咨询。
单账号单地域可创建的第三方代理 BYOK 数量	20	可提升，请 提交工单 咨询。
单账号单地域可创建的自建模型 BYOK 数量	20	可提升，请 提交工单 咨询。
每个原厂/第三方代理 BYOK 实例可添加的 Key 数量	50	可提升，请 提交工单 咨询。

每个自建模型 BYOK 实例可添加的 Key 数量	50	可提升，请 提交工单 咨询。
每个 BYOK 实例可添加的模型数量	20	可提升，请 提交工单 咨询。

其他维度

类型	默认配额（单位：个）	是否支持提升
单账号可创建的积分预算对象数量	50	可提升，请 提交工单 咨询。
每个积分预算对象可关联的资源数量	200	可提升，请 提交工单 咨询。
每个提示词可添加的版本数量	20	可提升，请 提交工单 咨询。
每个提示词可关联的用户组数量	20	可提升，请 提交工单 咨询。

提升配额方式：如果上述配额不满足您的需求，请 [提交工单](#) 联系我们，说明需要的配额类型、目标数量及业务场景。