

模型路由 购买指南



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。

文档目录

购买指南

计费概述

计费项

处理费

实例费

TCU费

资源包

资源包介绍

资源包购买

资源包退订

资源包续订

购买方式

欠费说明

购买指南

计费概述

最近更新时间：2026-09-15 15:34:30

本文介绍模型路由 CMR 的计费项、计费模式、计费规则及账单查询方法，便于您快速了解模型路由 CMR 的费用构成。

- ① 说明：
- 模型路由（CMR）当前处于公测阶段，暂未全面开放。使用前需申请开通资格。请提交 [工单申请](#)。
 - 产品公测期间，提供价值1000元首购免费资源包，用尽后将按照计费规则正常收费，若不使用请注意删除，避免产生计划外费用。

计费模式

付费模式	实例类型	网络类型	计费项	说明
按用量计费	共享型	公网	- 模型路由处理费 - 公网网络费（限时免费）	此计费方式仅支持通过资源包抵扣，详见 资源包介绍 。
	企业型	公网	- 模型路由处理费 - 公网网络费（限时免费）	
		内网	模型路由处理费	
按规格计费	企业型	公网	- 模型路由实例费 - 模型路由 TCU 费 - 公网网络费（限时免费）	此计费方式仅企业型实例支持。
		内网	- 模型路由实例费 - 模型路由 TCU 费	

- ① 说明：
- 模型路由收取的模型路由处理费、模型路由实例费和模型路由 TCU 费，均按照模型路由处理能力计算，与上游模型定价无关。
 - 按规格计费的实例，不同规格对应的模型路由实例费的系数不同，所以价格也不相同，建议按需选择。

- 推广期结束后将对限时免费的计费项进行收费，收费时将提前通过官网公告、站内信等方式通知。

计费规则

模型路由处理费

模型路由处理费基于 CMR 实例处理的 Token 总量计算，与上游模型的定价无关。

计费公式

模型路由处理费 = 模型路由处理费单价（元/百万 Token）× Token 总量

模型路由实例费

不同规格的 CMR 实例对应不同的规格系数，实例费按照所选实例规格及使用时长计费。

计费公式

实例费 = 实例单价（元/小时）× 规格系数 × 实际使用时长（小时）

模型路由 TCU 费

CMR 性能容量单元（Cloud Model Routing Token Capacity Unit，TCU）是衡量 CMR 处理业务时所消耗性能指标的单位。

计算公式

TCU 费用 = TCU 单价（元/个）× TCU 个数

公网网络费（限时免费）

仅公网 CMR 实例产生公网网络费。内网 CMR 实例，不产生公网网络费用。

账单查询

您可以通过以下方式查询 CMR 的用量及费用明细：

登录 [费用中心](#)，进入[费用账单](#) > [账单查看](#)，查看 CMR 产品的计费账单。

计费价格

- [处理费](#)
- [实例费](#)
- [TCU 费](#)
- [资源包](#)

计费项 处理费

最近更新时间：2026-09-15 15:34:32

按用量计费模式下的模型路由 CMR 实例收取处理费。实例创建成功，有业务处理以后会产生对应费用。

计费周期

- 按照模型路由 CMR 实例实际处理的 Token 量计费，每小时结算一次。
- 若用户创建实例后，完全不使用，则不会产生费用。
- 模型路由 CMR 实例收取的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。
- 模型路由 CMR 实例产生的处理费，仅支持通过资源包抵扣，详见 [资源包介绍](#)。

注意：
需要先购买资源包，才能创建按用量计费的 CMR 实例。

计费价格

地域	单价（元/百万 token）
广州/上海/北京/新加坡/硅谷	0.49

计费公式

按用量计费实例费用计算公式如下：

CMR 处理费 = 处理费单价（元/百万 token）* 处理总 token 数（百万 token）

处理总 token 数 $\approx$ 上游模型返回 usage 中的 total_tokens。处理总 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。

计费示例

场景：在广州地域创建一个按用量计费的 CMR 实例，2026年7月1日 00:00:00 创建成功，2026年7月1日 02:30:00 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
    }
  }
}
```

CMR 实例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
- CMR 处理费 = $110/1,000,000 * 0.49 \approx 0.0000539$ 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。

相关文档

[计费概述](#)

实例费

最近更新时间：2026-09-15 15:34:32

按规格计费的 CMR 实例收取实例费和 TCU 费，其中，实例费按照所选规格和使用时长计费。按规格计费模式下，实例创建成功即开始收取实例费。本文将介绍 CMR 实例费。

计费周期

- 按小时计费，每小时结算一次。
- 计费的起点以实例创建成功的时间点为准，终点以您发起销毁操作的时间点为准。
- 使用不足一小时将按照存续时间收费。
- 创建实例时会按照所选实例规格，预先扣除一个小时的规格费（即规格系数*实例费）。资源释放后（含主动释放或因欠费释放），对应冻结金额将自动解冻。冻结金额未产生实际消耗，解冻后即可恢复正常使用。有关冻结的详细信息可前往费用中心 [账号冻结说明](#) 中查看。

注意：
若您的账户余额不足，预扣费将失败并导致实例创建失败，请确保账户余额充足后再创建实例。

计费价格

地域	实例费单价（元/个/小时）
广州/上海/北京/新加坡/硅谷	1

不同实例规格对应规格系数不同，收取的实例费不同，规格系数及费用如下：

规格	TPM 限速值	规格系数（个）	总实例费（元/小时）
极简版	100 万	0.3	0.3
微型版	200 万	0.5	0.5
轻量版	500 万	1	1
轻量增强版	1,000 万	1.8	1.8
基础版	2,000 万	3.5	3.5
标准版（推荐）	5,000 万	8	8
进阶版	2 亿	30	30

专业版	8 亿	110	110
旗舰版（最高）	60 亿	640	640

计费公式

按规格计费下 CMR 实例的实例费用计算公式如下：

实例费 = 实例费单价（元/小时）× 规格系数 × 实际使用时长（小时）

① 说明：

如果在同一个计费周期内多次调整规格，则按各规格生效时长分段计费。

计费示例

场景：在广州地域创建一个规格为标准版的 CMR 实例，2026年7月1日 00:00:00 创建成功，2026年7月1日 02:30:00 发起销毁。

计算过程：标准版的 CMR 实例的规格系数是8，使用时长为 2.5 小时。

费用计算：实例费 = 1 元/小时 × 8 × 2.5 小时 = **20 元**

相关文档

- [计费概述](#)
- [TCU 费](#)

TCU费

最近更新时间：2026-09-15 15:34:32

性能容量单位 TCU（Token Capacity Unit）用来衡量模型路由 CMR 处理业务的性能指标。本文将介绍 TCU 费用。

TCU 衡量指标

TCU 衡量指标包括处理总 token 指标。单个 TCU 对应的性能指标如下：

指标名称	描述	TCU 系数	每小时 TCU 个数计算
处理总 token 数	处理总 token 数 $\approx$ 上游模型返回 usage 中的 total_tokens。处理总 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。	1百万	TCU 个数 = 处理总 token 数 $\div$ TCU 系数

TCU 的计费周期为 1 小时，在 1 小时内 CMR 消耗的 TCU 数量计算逻辑如下：首先计算计费周期内处理的总 token 数，再除以 TCU 系数，最终的结果作为该小时内的 TCU 计费个数。

计算公式：

每小时 TCU 费用 = 计费 TCU 个数 $\times$ TCU 单价

计费价格

地域	TCU 单价（元/个）
广州/上海/北京/新加坡/硅谷	0.001

计费示例

假设一个 CMR 实例平均每分钟处理100个 token，则 TCU 费的计费规则如下：

指标名称	使用示例	TCU 系数	TCU 个数计算
处理总 token 数	CMR 实例平均每分钟处理100个 token，则一个小时累计处理6000个 token。	1百万	TCU 个数 = 6000 $\div$ 1000000 = 0.006

在本示例中，TCU 计费个数为0.006个，因此，对应的 TCU 费用如下：

TCU 本小时的费用 = 0.001 元/个 $\times$ 0.006 个 = **0.000006 元**

相关文档

- [计费概述](#)
- [实例费](#)

资源包

资源包介绍

最近更新时间：2026-09-15 15:34:32

模型路由资源包是针对使用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按用量计费的 CMR 实例所产生的处理费。如果购买时选择按规格计费的 CMR 实例，您可按需选择是否通过资源包抵扣。本文将为您介绍模型路由资源包的基本信息。

优势

- **使用便捷**：资源包购买后立即生效。
- **自动续订**：支持开启续订功能，保障业务连续性。

使用限制

模型路由资源包仅适用于抵扣按用量计费的 CMR 实例所产生的处理费和按规格计费的 CMR 实例所产生的实例费和 TCU 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。

资源包服务地域

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 CMR 实例产生的费用。

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/容量单位）
广州/上海/北京/新加坡/硅谷	0.01

容量单位为模型路由资源包的计量单位，资源包售价为0.01元/容量单位。

支持地域

当前已支持的地域：北京、上海、广州、硅谷、新加坡。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：

- 若计费地域内所有资源包用尽或到期且未续订时，按用量计费的 CMR 实例将在下一个计费周期停止提供服务，按规格计费的 CMR 实例将在下一个计费周期转为后付费，按小时扣费。
- 到期后未用完的资源包容量将自动清零，不支持余额转移。

常见问题

1. 模型路由资源包可以抵扣哪些费用？

模型路由资源包可用于抵扣按用量计费的 CMR 实例所产生的处理费和按规格计费的 CMR 实例所产生的实例费和 TCU 费。

2. 模型路由资源包购买后需要设置吗？

- 针对按用量计费的 CMR 实例所产生的费用，购买后不需要任何配置，系统会自动进行抵扣。
- 针对按规格计费的 CMR 实例所产生的费用，可选择是否使用资源包抵扣。

3. 购买的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. 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。

资源包购买

最近更新时间：2026-09-15 15:34:32

本文将为您介绍模型路由资源包的购买流程。

使用须知

- 模型路由资源包仅支持抵扣 CMR 实例所产生的费用，包含按用量计费实例所产生的处理费和按规格计费实例所产生的规格费（实例费*规格系数）和 TCU 费。
- 模型路由资源包购买时，可以选择是否设置续订，多次设置续订，则仅最后一次设置的生效。
- 资源包最大可购买10,000,000容量单位，按当前售价计算，最大购买金额为100,000元。
- 当账号中有多个资源包时，按照资源包到期时间进行抵扣，优先抵扣先到期的资源包。
- 不支持抵扣资源包生效前产生的消耗。

⚠ 注意：

- 若计费区域内所有资源包用尽或到期且未续订时，按用量计费的 CMR 实例将在下一个计费周期停止提供服务，按规格计费的 CMR 实例将在下一个计费周期转为后付费，按小时扣费。
- 到期后未用完的资源包容量将自动清零，不支持余额转移。

购买资源包

1. 登录腾讯云 [模型路由](#) 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**、**是否自动续订**，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。
- 开启自动续订后，资源包到期将自动按当前价格续费，同一时间仅支持一个资源包开启续订。
- 账户余额足够时，资源包到期后会触发自动续订，自动续订详情请参见《[续订规则说明](#)》。

资源包退订

最近更新时间：2026-09-15 15:34:32

退款规则

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。
- 退款金额将退回您的腾讯云账户余额。
- 资源包中存在抵扣中的 CMR 实例时，建议先删除实例或者退出资源包抵扣。

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**。

资源包 ID	状态	剩余可用/总额度	创建/生效时间 ↓	到期时间 ①	抵扣截止时间	自动续订状态	操作
cr-1234567890	有效	1,000 / 1,000	创建: 2026-09-13 10:00:00 生效: 2026-09-13 10:00:00	-	2026-09-13 10:00:00	不续订	设置续订 退费 查看消耗

4. 在退款页面确定**剩余用量**和**金额**，并单击**确认**，完成退款。

资源包续订

最近更新时间：2026-09-15 15:34:32

自动续订规则

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，当前地域所有的资源包到期或用尽时，系统将根据自动续订规则为您自动购买一个新的资源包。续订的资源包是新购一个与设置续订资源包规格相同，设置相同，有效期为3年的新资源包，而非延长原资源包的有效期。

触发条件

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，当抵扣地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。
- 如果抵扣地域内的资源包没有用尽，则会在抵扣地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。

① 说明：

所有资源包耗尽是指同计费地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。

注意事项

- 同一个抵扣地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。

购买方式

最近更新时间：2026-09-15 17:20:00

腾讯云提供了两种购买模型路由 CMR 的方式：官网购买和 API 购买。本文将详细介绍两种购买方式。

官网购买

所有用户均可通过 腾讯云官方网站 购买 CMR 实例。

1. 登录腾讯云 [模型路由 CMR 实例列表页](#)，单击**新建**，进入 [模型路由 CMR 购买页](#)。



说明：
推荐您优先购买资源包，使用模型路由资源包抵扣用量更优惠，详见 [资源包介绍](#)。

2. 按需选择以下配置。

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 。

3. 完成以上参数配置后，勾选**服务协议**，单击**确定**创建实例。

后续操作

购买成功后，您需要为 CMR 实例配置 API key 和关联模型，才能正常使用：

- 若需创建 API key，请参见 [创建模型路由访问密钥](#)。
- 若需关联模型，请参见 [配置模型调度管理](#)。

相关文档

[计费概述](#)

欠费说明

最近更新时间：2026-09-15 15:34:31

- i* 说明：
- 您创建的 CMR 实例资源不再使用时请及时删除，以免继续扣费。
 - 您的实际资源消耗可能不断变化，因此余额预警可能存在一定的误差。

余额预警

系统每天会根据您名下按量付费资源在过去24小时的消费情况以及账户余额情况，预估余额可支撑的时间。若可支撑时间小于5天，系统将会向您推送余额预警消息。预警消息将通过已设置的开启方式（例如邮件、短信、站内信等），通知到腾讯云账户的创建者及已订阅消息的协作者。

您可以通过以下路径自助订阅接收方式，同时支持增加消息接收人：

消息中心控制台 > 消息中心 > [订阅管理](#)，配置方式可参见 [账户欠费预警指引](#)。

欠费预警

按量计费资源每个整点进行结算。在您的账户被扣为负值时，预警消息将通过已设置的开启方式（例如邮件、短信、站内信等），通知到腾讯云账户的创建者及已订阅消息的协作者。

欠费处理

从您的账户余额被扣为负值时刻起：

阶段	时间节点	状态	行为
宽限期	余额扣为负值起2小时内	继续使用	继续扣费。
隔离状态	2小时后	停服处理	- 服务不可用，仅保留配置数据，停止扣费。 - 隔离后续费：账户充值并补足欠费后，按规格计费的模型路由 CMR 实例将自动恢复服务。账户充值并补足欠费后，资源包购买后，按用量计费的模型路由 CMR 实例自动恢复服务。
实例释放	隔离满15天	永久删除	释放 CMR 实例，相关配置等数据被永久删除，不可恢复。

示例：

假如客户账号在2026年7月1日20:00:00欠费，则 CMR 实例在2026年7月1日20:00:00进入宽限期，继续使用，继续扣费。在2026年7月1日22:00:00开始进入隔离期，此时停止提供服务，也停止扣费。在2026年7月16日22:00:00隔离期满，启动释放。释放后不支持恢复数据或找回实例。

相关文档

[账单查询](#)