模型路由

产品简介

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

产品简介
产品概述

最近更新时间：2026-09-15 15:34:31

简介

模型路由（Cloud Model Routing，CMR）是腾讯云提供的面向企业级 AI 应用的统一大模型网关服务。CMR 支持以标准兼容接口统一客户端接入，纳管公有云、第三方及私有化部署的各类大模型服务。企业无需为每家模型厂商单独适配协议和配置密钥，配置接入 CMR 即可调用多方模型资源，并获得统一的鉴权、配额、审计与可观测能力。
模型路由与私有网络 VPC 深度集成，支持接入 VPC 内、自建 IDC 及公网部署的模型服务，并可结合公网加速优化跨域访问。内置自动重试、健康检查与多级 Fallback，在屏蔽各厂商 API 差异的同时降低调用失败率、提升服务可用性。产品提供覆盖整个调用链路的监控与分析能力，协助您掌握业务的运行状态。

核心功能

多模型统一接入

减少各厂商 API 适配的操作，将主流模型统一纳管，业务应用无需逐一适配。

功能描述	解决问题
OpenAI / Anthropic 兼容接口：面向客户端侧，提供统一接口协议，配置接入 CMR 即可透明切换后端供应商。	各厂商 API 格式、鉴权、参数不统一，应用侧需适配多套 SDK 与调用逻辑；切换或新增供应商需重复改造。
多厂商纳管：模型侧支持接入腾讯混元、DeepSeek 等主流模型，并支持自研、微调及私有化部署模型的 endpoint 的纳管。	自研、微调、私有化部署模型与公有云模型分散在不同入口，缺少统一纳管与统一调用方式。

智能路由调度

基于意图路由等高级路由引擎的路由转发策略，将请求自动分发至符合预设路由策略的模型，无需在应用侧编写复杂判定逻辑。

功能描述	解决问题
意图路由：按业务意图（客服问答、代码生成、文档总结等）自动匹配模型，支持基于 Prompt 内容的意图分类与动态路由。	模型选择依赖人工经验，应用侧 if-else 规则复杂且难以持续优化。
负载感知调度：依据用量、繁忙度、延迟等实时指标，动态调度至负载最低的节点，规避热点过载。	请求集中到少数节点造成热点过载，延迟升高、限流增多。
会话亲和：多轮对话中同一会话始终路由至相同实例，保持上下文连续。	多轮对话被分散到不同实例，上下文断裂影响体验和缓存命中率。
灰度发布与流量染色：支持配置优先级与权重，将部分请求导向新模型，实现安全灰度。	新模型上线缺少可控的小流量验证手段，全量切换风险高。

网络加速与混合组网

为不同部署形态的企业提供灵活的网络接入方案，并优化跨域模型调用的访问速度[1]。

功能描述	解决问题
混合组网：支持公网、VPC、IDC、专线 / 云联网等多种网络环境的统一接入。	VPC / IDC 内模型、公网大模型分散在不同网络与权限体系内，网络难以可控互通。
公网加速：提供公网加速通道，优化跨域调用的延迟与稳定性。	模型跨域链路波动，延迟高且不稳定，影响业务体验。

高可用保障

模型路由提供健康检查、自动重试和故障切换等能力，可辅助提升模型调用链路的可用性。

功能描述	解决问题
健康检查与熔断：持续探测后端模型健康状态，异常节点自动隔离，恢复后自动加回。	模型不可用时请求仍被打到故障节点，故障扩散且需人工介入摘除。
自动重试：调用失败或超时时自动重试，支持可配置的重试策略（次数、间隔、退避算法）。	上游限流、超时或偶发错误直接透传到业务侧，调用失败率偏高。

多级故障切换：主模型不可用时，自动降级至备用模型或备用供应商，支持多级 Fallback 链配置。	单一模型或单一供应商故障即导致业务中断，应用侧缺少自动容灾能力。

企业级安全管控

安全能力深度集成到模型调用链路中，提供网络隔离、访问控制、内容风险检测和审计日志等可配置能力，可辅助客户开展安全管理与审计留存。

功能描述	解决问题
网络层安全：路由实例原生部署于用户 VPC 内部，支持安全组和网络 ACL 等访问控制。	面向客户端侧的模型访问控制需要达到企业网络安全等级要求。
内容安全 WAF：内置可配置的 LLM-WAF 审核管道，可按配置规则对输入 Prompt 和返回内容进行实时内容风险检测，支持敏感信息脱敏、规则命中后的拦截及自定义规则[2]。	输入输出内容缺少合规检测，存在敏感信息泄露与违规内容风险。
细粒度访问控制：结合 CAM 与用户组能力，实现模型级、API Key 级权限管理。	模型 API Key 分散在业务代码中，权限无法按团队/应用隔离。
全链路审计（CLS）：全部调用记录写入日志服务 CLS，支持检索、报表与合规审计。	调用链路缺少审计日志，安全合规与事后追溯要求无法满足。

成本治理

按 CMR 实例、API Key、用户组、BYOK 、标签等维度拆分 Token 消耗与费用明细，并设定预算与配额，超限自动告警或阻断。

功能描述	解决问题
多维成本拆分：按业务维度精细核算 Token 消耗与费用，支撑内部成本归因。	不同团队、应用、模型的 Token 用量难以拆分，成本归因与优化缺少依据。
预算与配额管控：为实例、API Key、用户组设定调用上限，超限自动告警或阻断，防止费用失控。	Token 用量不可控，异常调用或突增导致费用失控。
语义缓存：对语义相似的重复请求进行缓存匹配，提升上游缓存命中率，减少无效 Token 消耗。	重复或高度相似的请求反复调用模型，产生大量无效 Token 消耗。
智能成本路由：依据用户配置、模型质量、延迟与价格等策略进行调度。	缺少质量、延迟与价格之间的权衡机制，高价模型被无差别使用。

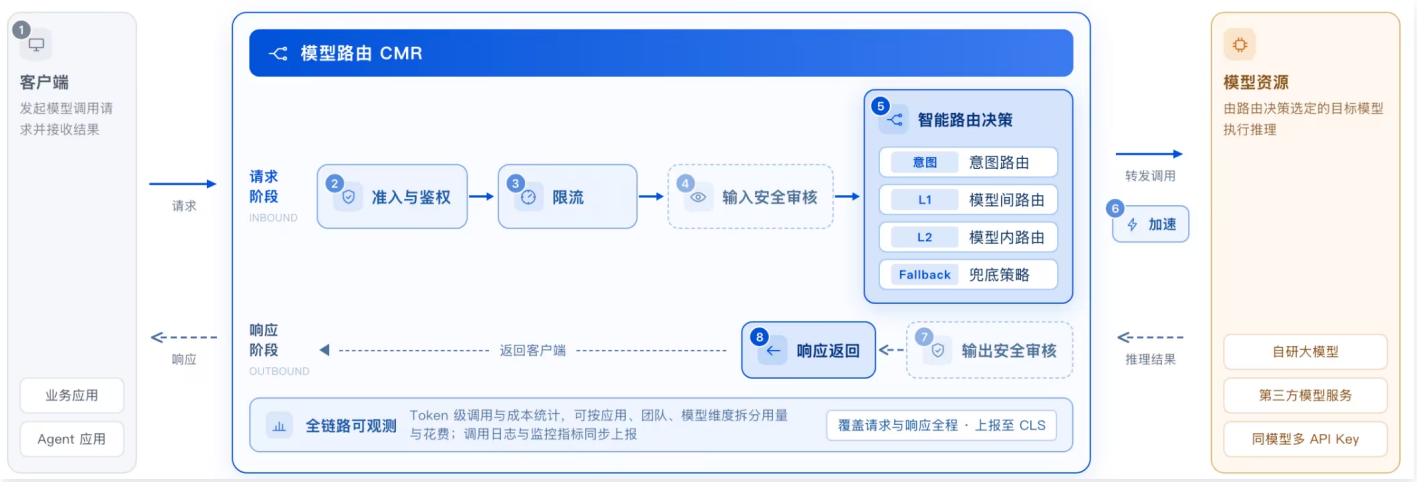
全链路可观测

覆盖模型调用全链路的监控与分析能力，实时掌握 AI 应用运行状态。

功能描述	解决问题
统一调用监控：支持展示已接入且已开启相应监控采集的模型调用指标，如模型调用的请求量、成功率、延迟分布与 Token 消耗，并在支持的调用链路范围内提供追踪能力。	多模型调用状态分散，缺少统一视图掌握请求量、成功率与消耗情况。
LLM 专属指标：除传统 QPS 与延迟外，提供首 Token 延迟（TTFT）、Token 产出速率等指标，支持 P95 / P99 延迟分位统计。	传统 QPS / 延迟指标无法反映大模型体验，首包与产出速率无处衡量。
全链路追踪：基于 OpenTelemetry 协议，从网关入口到模型推理完成，追踪单次调用的完整链路。	单次调用耗时或失败原因难以定位，问题排查依赖人工比对日志。
智能告警：基于异常检测，Token 消耗突增、成功率骤降或延迟异常时自动通知。	异常与用量突增发现滞后，往往事后才被察觉。

转发路径

模型路由 CMR 提供 Token 级的调用与成本统计，可按应用、团队、模型等维度拆分用量与花费，帮助企业清晰掌握大模型调用的成本结构[3]。在转发过程中，提供全链路可观测能力。



1.客户端发起请求

客户端向模型路由 CMR 发起请求，请求体包含模型标识、Prompt 消息及 thinking 开流式开关等字段。

2.准入与鉴权

在客户端发起请求后，根据客户配置的安全组或者用户组中的策略，进行准入与鉴权审核。只有符合策略要求的请求才允许通过。

3.限流

模型路由 CMR 支持在 CMR 实例、用户组和 API Key 维度设置 TPM/RPM 限速值，或者绑定积分预算。可提供有效的模型使用与资源消耗管理能力，帮助团队或项目有效规划、监控与优化 Token 消耗。

4.入方向安全审核

腾讯云 WAF 大模型安全针对大模型应用场景提供专属安全防护，包括提示词攻击防御、输入内容安全审核、敏感数据泄露防护、算力消耗防护、运行时防护等核心能力[2]。

5.智能路由决策

模型路由 CMR 提供意图路由、L1 模型间路由和 L2 模型内路由。CMR 根据请求客户关联的模型和路由策略，逐级匹配，最终选定模型。

6.加速

依托腾讯云支持的网络节点和线路，模型路由 CMR 基于节点就近调度，可为跨域调用场景提供网络加速能力，有助于降低跨域网络延迟、提升连接稳定性。

7.出方向安全审核

腾讯云 WAF 大模型安全针对大模型应用场景提供专属安全防护，包括输出内容安全审核、敏感数据泄露防护、算力消耗防护、运行时防护等核心能力[2]。

8.响应返回

模型推理结果以流式（SSE）或一次性方式返回客户端，调用日志及监控指标同步上报至 CLS，供后续审计与分析。

开通说明

当前模型路由 CMR 处于公测阶段。如需体验产品功能，可联系您的客户经理或 提交工单 申请。

⚠ **注意：**

- 官网文档里的产品概述、优势、应用场景等文档中的数据描述，均来源于2026年腾讯内部实验测试结果，测试基于特定环境、条件及时间范围，仅反映相应测试场景下的情况，实际效果可能因业务场景、配置及使用情况不同而存在差异。
- 文中涉及的第三方名称、商标及软件归其各自权利人所有，仅用于说明兼容性或适配范围，不构成认可、担保或背书。

ⓘ **说明：**

1. 使用跨域模型或跨域网络链路前，客户应结合实际数据类型、处理目的和适用地域，依法完成必要的数据与个人信息保护义务。具体支持的模型、地域、网络线路和数据处理范围以产品文档及控制台展示为

准。

2. 客户使用内容检测及处理相关数据时仍应依法履行数据和个人信息保护义务。
3. 实际命中率和 Token 消耗以调用场景及配置为准，模型、场景等不同可能导致命中率不同。

产品优势

最近更新时间：2026-09-15 15:34:31

核心优势

云原生部署，降低对接成本

模型路由 CMR 实例部署在用户 VPC 内，并与腾讯云 VPC、CLS、CAM、云监控等平台原生集成。您在腾讯云上已有的网络规划与 CAM 权限体系可按需复用。后端模型可按实际网络配置通过 VPC、IDC 或公网链路接入[1]。

全球加速网络，优化模型访问体验

依托腾讯云支持的网络节点和线路，模型路由 CMR 可以基于节点就近调度，为跨域调用场景提供网络加速能力，有助于降低跨域网络延迟、提升连接稳定性[2]。

智能意图路由，兼顾成本与效果

基于语义的智能路由引擎，支持根据用户配置的策略，将路由决策的依据从请求的调用特征扩展到请求的内容语义。复杂推理、长文分析与关键业务请求可交由高阶模型承接，保障输出质量；高频出现的简单问答、分类与抽取类请求由轻量模型处理，有助于降低单次调用成本。在控制调用成本的同时，兼顾响应质量，帮助企业在规模化调用下取得成本与效果的平衡[3]。

成本可视与优化，规模化更可控

模型路由 CMR 提供 Token 级的调用与成本统计，可按应用、团队、模型等维度拆分用量与花费，帮助企业清晰掌握大模型调用的成本结构。结合意图路由的分级调度与重复请求的缓存复用，有助于在保障效果的前提下减少不必要的高阶模型调用与冗余请求，让规模化使用下的成本更可控、更透明。

安全合规能力，一站式集成

模型路由 CMR 内置从网络层到应用层的多层安全防护，帮助企业更省心地满足安全与合规需求。网络层提供安全组与访问控制策略，帮助限制非法访问、隔离异常流量；应用层集成腾讯云 CLS 日志审计、WAF 安全防护，并支持按配置对手机号、证件号码、银行卡号等指定类型信息进行识别与脱敏处理，同时通过 CAM 实现细粒度的权限管控。以上能力作为产品内置的安全合规模块统一提供，开通可快速接入，有助于降低部署门槛[4]。

全托管服务，简化运维

模型路由 CMR 以托管方式提供服务，平台负责相关基础设施的运维、扩缩容与版本升级。底层依托腾讯云高可用架构，支持调用量的弹性伸缩，并提供模型间、模型内不同 key 与 Fallback 三级容错机制：在用户开启并完成相关配置的前提下，支持按策略进行重试等切换动作。当某个模型出现异常时自动切换，同模型内，某个 key 不可用时切换至其他 key，仍无法满足时回退到预设的兜底模型。异常发生时自动尝试切换，减少人工介入，保障业务连续性[5]。

与自建方案的综合对比

企业可在自有云主机或 IDC 服务器上部署大模型网关。与自建方案相比，模型路由 CMR 在以下方面可减少企业的建设与长期维护投入。

对比项	客户自建	模型路由 CMR
部署与接入	自行选型、部署与配置，能力以插件或自研模块组合实现，与企业既有云上资源需逐项对接。	开通后快速接入，实例可部署于用户 VPC 内，并与 VPC、CLS、CAM、云监控等平台原生集成，已有网络与权限规划可按需复用。
跨域访问质量	网关软件本身不改变链路质量，需自行采购加速线路或自建海外中转节点，链路波动自行排查。	依托腾讯云支持的网络节点与线路，基于节点就近调度为跨域调用提供加速能力。
路由智能化	语义级意图选路与负载感知调度需自行开发实现，并结合业务效果持续调优。	内置基于语义的智能路由引擎，可按内容语义在高阶模型与轻量模型间分级调度。
成本拆分维度	按应用、团队、模型等业务维度归集用量与费用，需自行设计归属标识与聚合逻辑并持续维护。	通过用户组与标签的组合，便于按应用、团队、模型等维度拆分用量与花费。
安全与审计	需自行接入审核服务或开发过滤逻辑，并自建日志存储与检索体系，规则效果、留存周期与容量扩容自行设计实现。	网络层至应用层多层防护内置提供，集成 CLS 日志审计、WAF 与指定类型信息脱敏，通过 CAM 实现细粒度权限管控。
可用性保障	网关自身即为调用链上新增的单点，多副本、多可用区及依赖组件（数据库 / 缓存）的高可用需自行建设。	托管服务形态，底层依托腾讯云高可用架构，支持自助扩缩容。
维护投入与服务保障	版本升级、安全补丁、上游接口变更与新模型适配需自行跟进；除服务器等资源开销外，还需专业人员持续投入与值守。	平台负责基础设施运维与版本升级，无需自建基础设施与专职运维，并提供支持渠道与服务等级承诺。

说明：

1. 不同调用场景下的数据传输路径以模型部署位置、网络配置和控制台展示为准，兼顾数据安全与模型选择的自由度。
2. 依托腾讯云支持的网络节点和线路，为部分跨域调用场景提供网络加速能力，实际效果受模型服务地域、网络配置和业务负载等因素影响，以控制台和产品文档为准。跨域调用可按产品支持的网络线路进行转发，具体网络路径和可用范围以控制台及产品文档为准。客户应结合实际处理的数据类型、业务场景和适用要求，完成必要的法律评估与管理。
3. 支持根据用户配置模型的质量、时延和价格等策略，将请求分发至相应模型。实际输出质量、时延和费用受模型、请求内容及配置影响，以实际调用和账单为准。

4. 支持集成安全组、CAM、CLS、WAF 及数据脱敏等可配置能力，具体开通方式、适用范围、依赖服务和计费规则以产品文档及控制台展示为准。相关能力可辅助客户开展安全管理，不替代客户基于自身业务履行的数据和内容管理责任。支持按配置对手机号、证件号码、银行卡号等指定类型信息进行识别与脱敏处理，实际效果受内容格式、语言类型和规则配置等因素影响。
5. 容错机制下的切换可能导致模型输出、时延、Token 消耗和费用发生变化，具体以实际调用结果为准。用户仍需结合业务需求关注配额、成本、版本兼容和变更通知。公测阶段的服务范围、可用性和支持策略以届时有效的公测规则及服务协议为准。

应用场景

最近更新时间：2026-09-15 15:34:31

模型路由 CMR 面向企业级大模型统一接入与智能调度场景，在提供多模型兼容、混合网络打通、智能路由和高可用保障的同时，支持 Token 级成本治理与全链路安全审计，广泛适用于多供应商纳管、跨域访问加速、AI Agent 开发及企业合规管控等多元化业务场景。

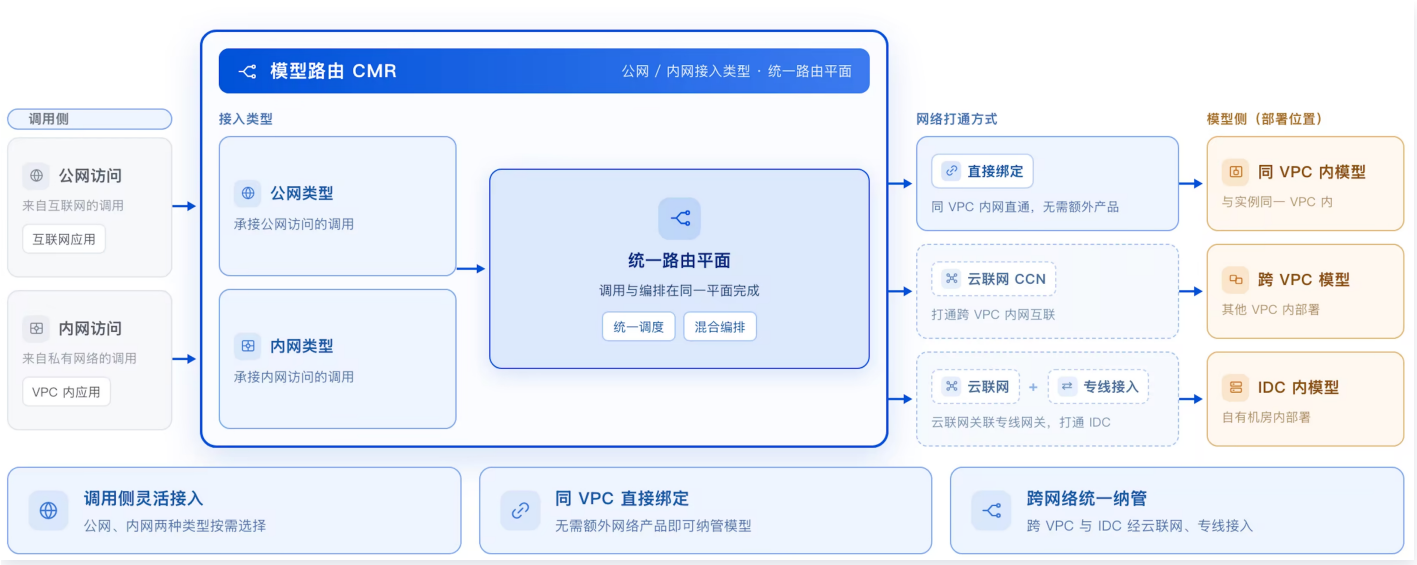
多模型、多供应商的统一接入与智能分发

在企业 AI 应用快速迭代的过程中，业务团队往往需要同时接入多家模型厂商的服务，既包括开源与自建模型（如 DeepSeek、Llama、Qwen 等），也包括各类商业大模型。不同厂商的 API 格式、鉴权方式、参数定义与错误码各不相同，即便部分厂商提供了兼容接口，应用侧仍需分别适配、维护多套调用逻辑，开发与维护成本较高。

模型路由 CMR 提供兼容 OpenAI 与 Anthropic 的统一 API 接口，将国内外主流模型服务纳管到同一个路由平面。业务应用无需为每个厂商单独编写适配代码，配置接入模型路由 CMR 即可透明调用多个后端模型。同时，模型路由 CMR 的意图路由引擎可按业务意图（如客服问答、代码生成、文档总结）将请求自动分发至合适的模型，减少在应用侧维护复杂 if-else 判定规则的工作量。

此外，模型路由 CMR 支持为不同业务线、团队或应用分配独立的 API Key，并可按 Key 配置调用额度、权限范围与成本归属。多团队共享同一路由平面时，便于实现用量隔离、成本拆分与权限管控，也便于在 Key 泄露或异常调用时快速定位与吊销，降低统一管理的风险。

VPC、IDC 与公网的混合模型统一治理



企业在落地大模型时，往往会同时存在多种部署形态：出于数据安全与自主可控的考虑，将开源模型（如 DeepSeek、Llama、Qwen 等）私有化部署在自有的 VPC 内网或 IDC 机房；同时，部分场景又需要调用部署在公网的外采模型服务，以补齐特定能力。这些分散在不同网络与权限体系中的模型，难以实现统一的接入与管控。模型路由 CMR 原生支持混合组网部署，路由实例可按需部署到企业的 VPC 内部，通过专线或云联网打通 IDC 环境，并支持以安全代理方式访问公网模型。自有部署的开源模型与外采的公网模型可在同一路由平面下混合编排、统一调度，模型调用流量经由统一的 模型路由 CMR 实例入口管控，实现多网络环境下的一体纳管。结合 私有网络 VPC 与 云联网，企业无需改变现有网络架构，即可将「自建开源模型 + 外采公网模型」纳入同一套调用治理平面。

模型的跨域访问加速

企业在国内调用跨域的主流模型时，跨域链路容易受到网络波动影响，出现延迟增大、连接不稳定等问题，影响业务体验。模型路由 CMR 依托底层已集成的全球加速能力，通过合规的加速链路对跨域调用进行网络路径优化，并结合智能 DNS 解析就近接入节点，有助于降低传输延迟、提升连接稳定性。企业无需自行搭建或额外配置跨域加速方案，通过 模型路由 CMR 统一入口即可获得开箱即用的跨域访问优化体验。

多级容错与业务连续性保障

大模型服务在生产环境中难免遇到上游限流、超时或临时不可用等情况，若缺少自动容灾能力，业务链路可能直接中断。

模型路由 CMR 提供三级容错机制，逐层兜底以保障请求得到响应：

- 模型间切换：当所选模型异常时，自动切换至其他可用模型
- 模型内切换：同一模型的多 API Key（供应商）切换，同一模型可配置多个 API Key（对应不同供应商或渠道），当某个 Key 限流或对应供应商不可用时，自动切换至该模型的其他可用 Key。
- Fallback 兜底：当上述切换仍无法满足时，回退到预先设定的兜底模型，避免请求得不到响应。

针对性配置健康检查策略（检查间隔、超时阈值、不健康阈值），异常节点自动熔断并隔离，恢复后自动重新加入可用列表。整个过程对调用方透明，帮助业务在模型异常场景下维持稳定运行。面向对可用性要求较高的核心业务，建议结合 腾讯云可观测平台 配置监控与告警策略，在模型调用出现异常时第一时间感知并介入，进一步保障业务连续性。

Token 消耗的精细化管理与成本治理

在多团队、多应用共享模型资源的场景下，不同业务线、不同 API Key 的 Token 消耗难以拆分，成本因此缺少依据，容易出现预算超支或资源浪费。

模型路由 CMR 提供按 CMR 实例、API Key、用户组、BYOK ID 等维度的 Token 消耗拆分与费用明细，支持输入 / 输出 Token 分别统计，让每一笔调用开销都可归因、可核算。您可以为不同用户组或 API Key 设定调用配

额与预算上限，超限时自动触发告警或阻断，把成本约束在可控范围内。
在看清成本构成的基础上，模型路由 CMR 进一步提供两项主动降本能力，帮助从“成本可见”走向“成本更优”：一是语义缓存，对语义相近的重复请求可直接复用已有结果，提升缓存命中率，从而减少重复计算带来的无效 Token 消耗；二是智能成本路由，可结合模型质量、延迟与价格画像，将请求自动调度至性价比更优的模型，在不牺牲服务质量的前提下降低模型调用开支。两者与前述的用量拆分、配额管控相互配合，形成「看清成本 → 约束成本 → 优化成本」的闭环。

企业级安全与全链路审计

当企业内多个团队、多个应用同时接入大模型时，模型 API Key 往往散落在各业务代码中，既增加泄露与滥用风险，又使调用行为缺乏统一记录，难以满足内部安全管理与审计要求。
模型路由 CMR 支持将 BYOK 密钥统一托管至路由实例，业务代码中不再直接持有密钥。结合 访问管理 CAM 的权限体系和模型路由的用户组能力，您可按模型、供应商维度配置细粒度访问策略：不同团队与业务线只能调用被授权的模型，各自的配额与用量独立核算，权限边界与资源用量互不影响。
内置 LLM-WAF 对输入 Prompt 与模型返回内容进行实时检测，支持敏感信息脱敏与违规内容拦截。所有调用记录实时写入 日志服务 CLS，涵盖请求来源、调用参数、返回结果等字段，支持检索与报表分析，为企业的内部审计与合规举证提供完整依据。

基本概念

最近更新时间：2026-09-15 15:34:31

一个提供服务的模型路由 CMR 实例由以下核心组件构成：

CMR 实例

CMR 实例是模型路由对外提供服务的最小载体，也是您接入大模型能力的统一入口。

- **实例类型**：分为企业型和共享型两类，满足不同业务规模和隔离要求。
- **网络类型**：CMR 实例支持公网或内网接入，可按需选择。
- **协议与端口**：兼容 HTTP（80）、HTTPS（443）等标准协议。

API Key

API Key 是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。

模型调度管理

模型调度配置是模型路由 CMR 的“调度大脑”和“核心引擎”，负责根据用户配置的模型策略与用户请求特征，将请求分发到合适的模型通道。支持意图智能路由， L1模型间路由和 L2模型内路由两层调度策略，并提供配置默认 Fallback 兜底策略。

BYOK

BYOK（Bring Your Own Key）让您将自有的大模型厂商 API Key 托管到模型路由，由路由统一代您调用，避免在业务代码中硬编码密钥。产品采用分层隔离和加密存储的安全设计，保护您的上游模型 API Key。客户的 API Key 仅在数据面模型代理组件中通过业界标准的认证加密算法（AEAD）加密后落库，全链路 TLS 传输，日志与查询接口中的 Key 字段自动脱敏（如 sk-****1234），确保即使数据库泄露也无法还原原始密钥。

产品对比

最近更新时间：2026-09-15 15:34:30

模型路由 CMR 实例分为企业型和共享型两种类型，本文将为您介绍两种类型实例的差异。

对比项	企业型	共享型
定位	适用于生产环境，各项配置更加灵活，性能更强。	可用于快速体验，不适用生产环境。
配额	单地域5个，支持调整配额。	单地域1个，不支持调整。
计费	支持按用量计费和按规格计费2种计费模式，支持按用量计费的实例切换为按规格计费。	仅支持按用量计费，不支持切换计费模式。
网络类型	公网和内网均支持	仅支持公网
监听协议	支持 HTTP 和 HTTPS	仅支持 HTTPS，默认配置一本证书，不支持扩展或修改。
证书	支持选择自有证书	不支持选择证书

限速（TPM/RPM）	• 按用量计费：
	  ○ TPM 上限：10万千次/分钟，如需提高配额，可 提交工单 申请。
	  ○ RPM 上限：1万次/分钟，如需提高配额，可 提交工单 申请。
	• 按规格计费：
	  ○ TPM 上限：与所选规格上限一致，最高可支持60次/分钟。
	  ○ RPM 上限：1万次/分钟，如需提高配额，可 提交工单 申请。
		TPM 上限：10千次/分钟，不支持继续提高。
		RPM 上限：10次/分钟，不支持继续提高。

积分预算	支持	不支持
非标参数转换	支持	不支持
向量嵌入	支持	不支持
意图路由	支持	不支持
用户组	支持	不支持

是否支持自建模型	支持	不支持
CLS 日志	支持	支持
WAF	支持	支持

其他使用限制详见 使用约束 。

使用约束

最近更新时间：2026-09-15 15:34:31

本文为您介绍模型路由（CMR）的使用限制。

通用配额限制

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。

相关服务

最近更新时间：2026-09-15 15:34:31

模型路由 CMR 作为腾讯云网络生态的核心组件，可与以下产品协同使用：

网络与安全类：

- 私有网络 VPC：CMR 可部署在 VPC 内，并可结合安全组、ACL 和访问控制策略进行网络隔离与访问管理。
- 安全组和 ACL：安全组 是一种虚拟防火墙，能够通过自定义的访问规则，控制业务流量。ACL 规则 是一种子网级别的可选安全层，用于控制进出子网的数据流，可以精确到协议和端口粒度，实现子网粒度流量的精细化控制。
- Web 应用防火墙 WAF：集成到 CMR，接入用户与大语言模型（Large Language Model，LLM）之间的交互链路，对用户输入的提示词（Prompt）和模型输出的生成内容进行检测，可识别提示词注入、内容违规、敏感数据泄露、算力异常消耗等大模型应用场景下的常见风险。
- SSL 证书：支持为 HTTPS 监听器配置符合条件的 SSL 证书，证书申请、托管和续期规则以 SSL 证书服务及控制台展示为准。

运维与监控类：

- 腾讯云可观测平台：支持对已接入且已开启相应监控采集的 CMR 调用查看 Token、请求次数、时延等指标，具体指标范围、统计周期和数据延迟以控制台展示为准。
- 访问管理 CAM：帮助您安全、便捷地管理对腾讯云服务和资源的访问。
- 日志服务 CLS：收集和分析 CMR 的访问日志。

支持模型

最近更新时间：2026-09-15 15:34:30

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。其中 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。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

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

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
腾讯云 TokenHub	hy3
	hy3-preview
	hy-mt2-pro
	hy-mt2-plus
	hy-mt2-lite
	hunyuan-role-latest
	hy-role
	deepseek-v4-flash-202605
	deepseek-v4-pro-202606
	deepseek-v4-flash
	deepseek-v4-pro
	deepseek-v3.2
	glm-5.2
	glm-5.1
	glm-5v-turbo
	glm-5-turbo

glm-5
kimi-k2.7-code-highspeed
kimi-k3
kimi-k2.7-code
kimi-k2.6
kimi-k2.5
minimax-m3
minimax-m2.7
minimax-m2.5
qwen3.5-flash
qwen3.5-plus
hy-image-v3.0
hy-image-lite
hy-video-1.5
yt-video-2.0
yt-video-humanactor
yt-video-fx
kl-video-v3
kl-video-v2-6
kl-video-v2-5-turbo
kl-video-v2-1-master
kl-video-v2-1
kl-video-v2-master
kl-video-v1-6
kl-video-v1-5

	kl-video-v1
	vd-video-q3-pro
	vd-video-q3-turbo
	vd-video-q2-pro
	vd-video-q2-pro-fast
	vd-video-q2-turbo
	vd-video-q2
	hy-3d-3.0
	hy-3d-3.1
	hy-3d-express
	youtu-vita
	hy-vision-2.0-instruct
	hunyuan-t1-vision-20250916
	hunyuan-turbos-vision-video-20250728
	kinfra-text-embedding-0.6b
	kinfra-text-embedding-4b
	kinfra-vl-embedding-2b
	kinfra-vl-embedding-8b

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 BYOK 自建模型提供商。