模型路由

最佳实践

腾讯云

版权所有：腾讯云计算（北京）有限责任公司

【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】

腾讯云

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

版权所有：腾讯云计算（北京）有限责任公司

最佳实践
企业模型精细化管理

最近更新时间：2026-09-15 15:34:30

概述

当大模型能力由单一团队的试点扩展为全公司共享的基础设施时，管理复杂度随之非线性上升。多个业务线、多个团队、大量员工与自动化任务同时消费模型，将带来以下问题：

- 密钥管理困难，以单一 Key 覆盖全部调用，一旦泄露则全面失守。
- 用量难以拆分，账单为一个总数，无法归集到具体业务线。
- 成本难以控制，缺乏预算约束，超支往往到月底才被发现。
- 厂商难以更换，业务代码写死模型名，更换底层厂商需改动大量业务代码。

不少企业为此自建网关以集中处理鉴权、路由与统计，但自建方案意味着持续的研发与运维投入，且功能往往不完整，尤其在费用管控与用量归集等治理能力上难以做深。模型路由 CMR 的定位是以开箱即用的托管网关替代此类自建方案，将上述四类问题收敛为统一治理模型：以分 Key 管控可用用户，以调度管控请求模型，以用量统计明确成本观测，以积分预算约束超额使用。

方案价值

治理能力的价值在于将一系列事后问题转化为事前治理，并将大模型由一项难以核算的浮动开销，转变为一项可预算、可分摊、可审计的企业资源。财务可按业务线获得清晰的消耗数据，管理者可为各业务线设定额度上限，安全团队可确保凭证不外泄，架构团队则保留随时更换底层模型的自主权。对内，其替换了成本较高且功能不完整的自建网关；对外，其使企业在与模型厂商的合作中保持主动权。

维度	客户收益	实现机制
替代自建网关	开箱即用地获得鉴权、路由、限流、统计、费控，替换自研网关，降低研发与运维成本	托管式统一网关
密钥安全	API Key 与模型 Key 分离，人员流动与 Key 泄露不波及真实模型凭证	双层 Key 隔离
成本管控	按业务线/员工分配积分预算，超额限流，成本由月底对账转为实时约束	模型系数与积分预算

架构说明

架构以 CMR 网关为中枢，网关统一治理业务侧的 API Key 与网关侧托管的模型 Key。业务侧的 API Key，由业务线、员工、Agent 等分别持有。网关侧托管的模型 Key，包含原厂模型、第三方代理、自建模型等真实凭证，统一托管于网关侧，对业务不可见。在费控机制上，积分是连接“Token 消耗”与“预算约束”的中间计量单位。网关为各模型配置系数，将不同单价的模型折算至统一积分口径，再通过积分预算为实例或指定 API Key 设定额度，额度耗尽即自动限流。由此，“企业额度 → 业务线额度 → 员工 Key 额度”的层级化费用管控得以成立，且各层口径一致。

前提条件

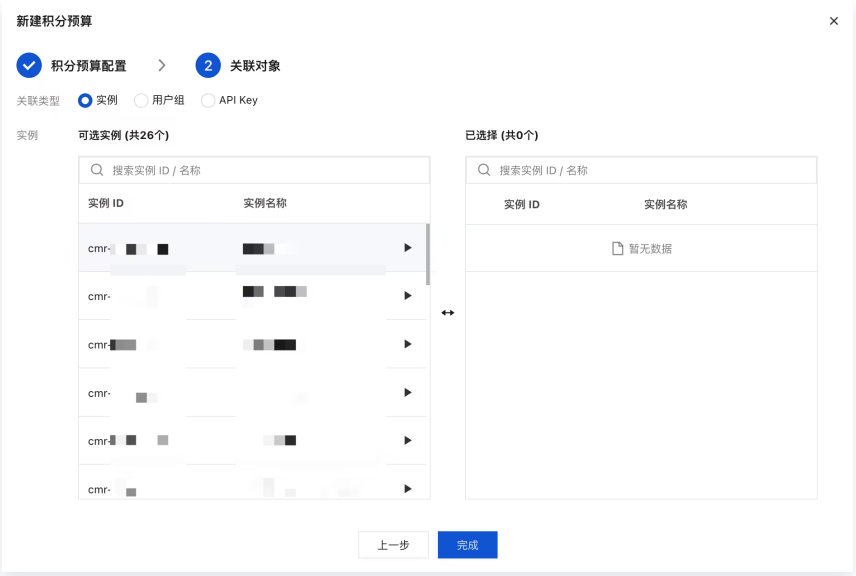
1. 已创建模型路由实例，详细操作指导请参见 创建模型路由实例 。
2. 已创建模型路由访问密钥，详细操作指导请参见 创建模型路由访问密钥 。
3. 已创建 BYOK 实例，详细操作指导请参见 管理 BYOK 实例 。
4. 如需控制积分消耗或调用速率，请提前创建对应的积分预算，详细操作指导请参见 配置积分预算 。

操作步骤

步骤1：创建积分预算

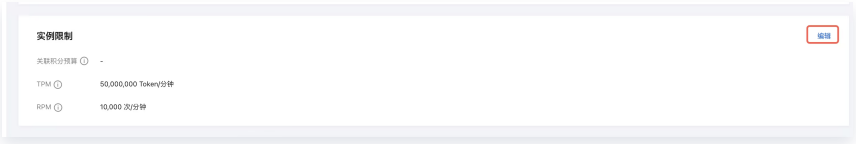
1. 登录 模型路由 控制台，在左侧导航栏选择**积分管理**。在**积分预算**页面，单击**新建积分预算**。

2. 在**新建积分预算**页面输入积分预算名称，并配置最大积分额度、重置周期，以及 TPM、RPM 速率限制。单击**下一步：关联对象**。
3. 在**关联对象**页面，选择关联类型与实例，并单击**完成**。

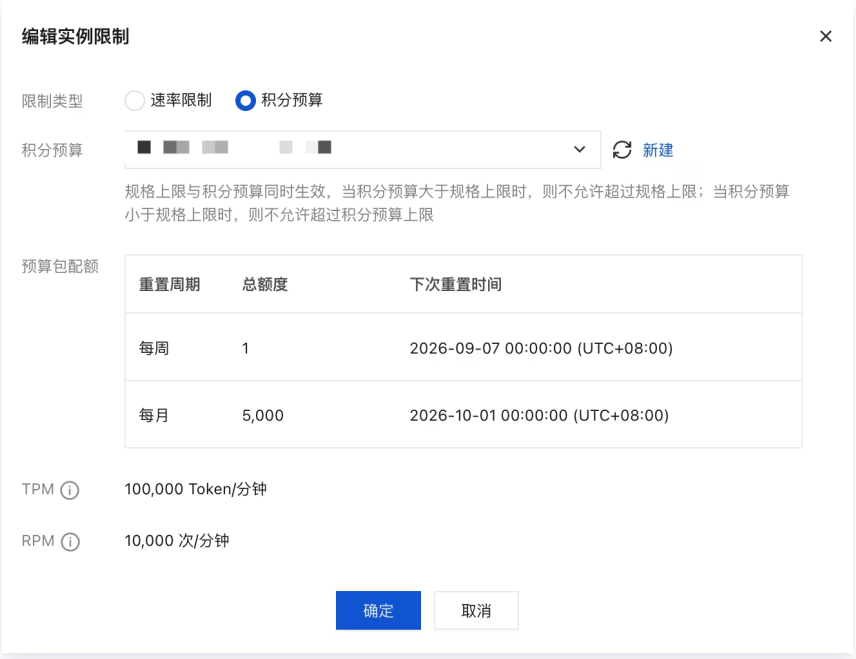


步骤2：为 CMR 实例关联积分预算

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. 在**实例限制**模块下，单击**编辑**，选择**限制类型**为积分预算，为当前 CMR 实例关联已经配置好的积分预算。



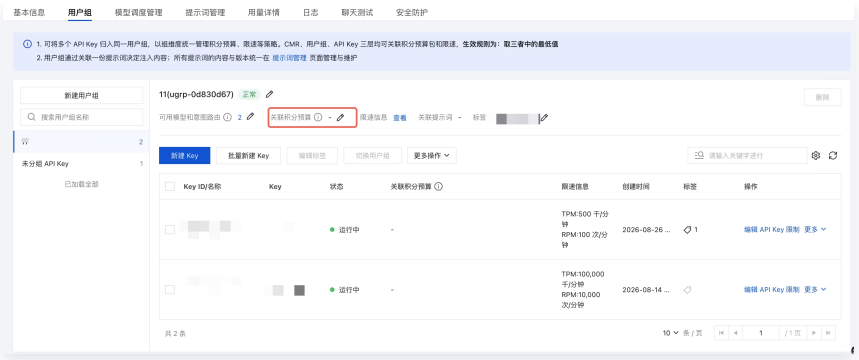
3. 确认积分预算的预算包配额、TPM 和 RPM 配置。



4. 单击**确定**完成配置。

步骤3：为用户组关联积分预算

1. 如已创建用户组，可在对应的用户组页面单击**关联积分预算**进行关联。



2. 如未创建用户组，选择**用户组**页签，在用户组页签中，单击**新建用户组**。在**新建用户组**页面，勾选**积分预算限制**，并选择已配置好的积分预算。

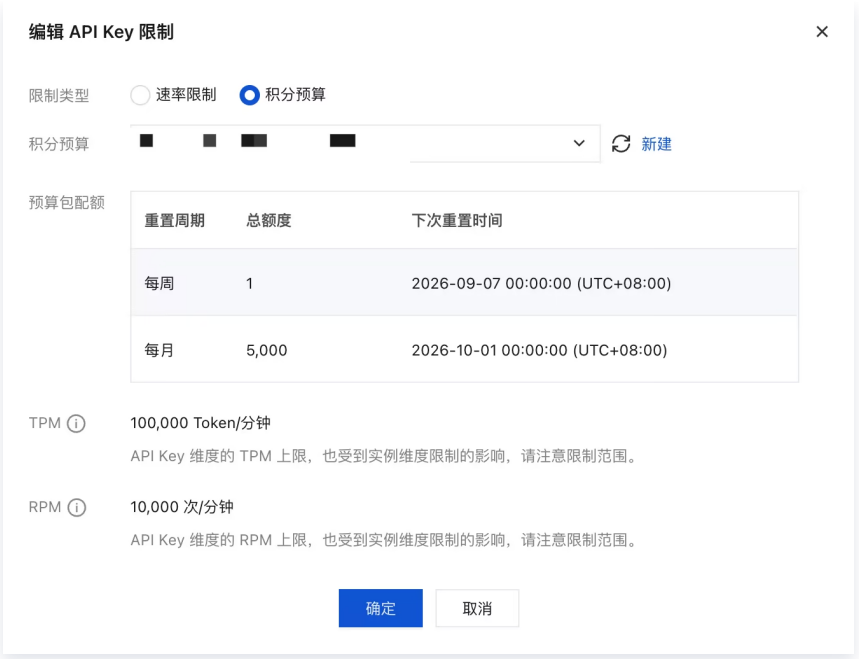
步骤4：为 API Key 关联积分预算

如果需要进一步控制单个应用、员工或脚本的资源消耗，可以为 API Key 配置独立的积分预算。

1. 在**用户组 API Key** 列表或**未分组 API Key** 列表中找到要进行积分限制的 API Key，单击**编辑 API Key 限制**。

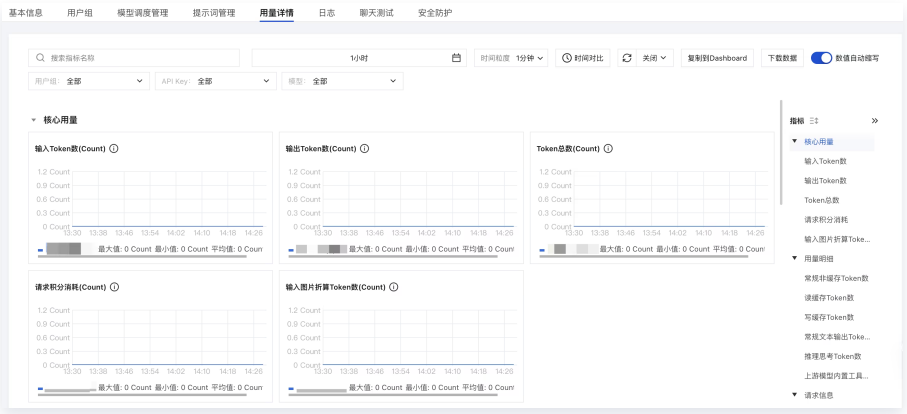
2. **限制类型**选择**积分预算**，在积分预算下拉框中选择已配置好的用于 API Key 的积分预算。

3. 确认积分预算的预算包配额、TPM 和 RPM 配置。



步骤5：验证结果

1. 不同 API Key 通过业务应用或测试工具连续发送多次模型请求。
2. 在用量详情中按实例、用户组、API Key 等维度查看消耗。



相关文档

- 创建模型路由实例
- 创建模型路由访问密钥
- 管理用户组
- 分配访问密钥
- 配置积分预算
- 查看用量详情

Buddy 私有模型统一接入

最近更新时间：2026-09-15 15:34:31

概述

以 WorkBuddy、CodeBuddy IDE 为代表的智能助手，已经成为研发和办公场景中的重要生产力工具。此类助手通常直接接入云端大模型，能够快速投入使用。但在金融、政企和大型制造等对数据安全与合规要求较高的行业，代码仓库、内部文档和经营数据经由公网发送至第三方模型，可能带来数据出域、商业机密泄露和合规审计等风险。对于已经在私有网络中部署大模型的企业，关键是如何在不显著改变 Buddy 使用方式的前提下，让 Buddy 安全、稳定地调用企业私有模型。

模型路由 CMR 可以作为 Buddy 与私有模型之间的统一接入层：对外提供兼容 OpenAI 协议的访问地址，对内完成鉴权、协议转换、模型路由、限流和用量统计，并将请求转发至企业部署的私有模型。本最佳实践聚焦以下同一 VPC 内的纯内网接入场景：

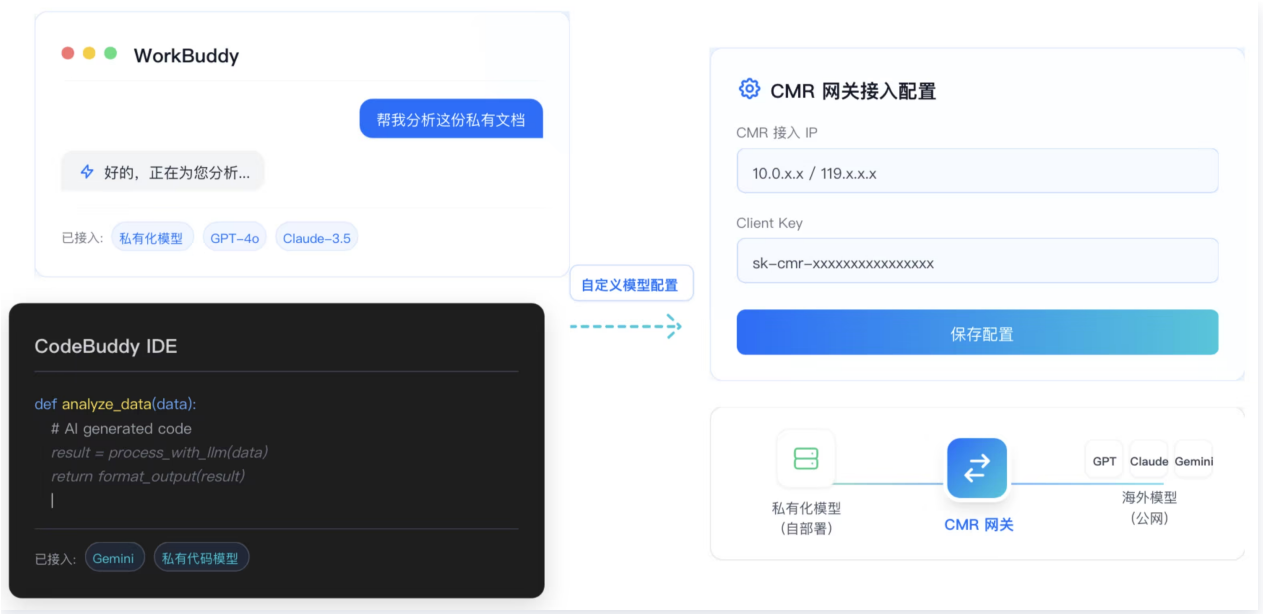
- 私有模型部署在客户 VPC 内。
- 运行 CodeBuddy IDE 或 WorkBuddy 的服务器部署在同一 VPC。
- CMR 创建在同一 VPC，并通过内网访问私有模型。
- Buddy 通过 CMR 内网接入地址调用私有模型。

方案价值

CMR 在 Buddy 与私有模型之间提供统一接入和治理能力，主要具有以下价值。CMR 的作用不只是转发请求，而是将模型接入、凭证管理和调用治理集中到统一网关。Buddy 无需直接感知私有模型的真实地址和凭证，企业也可以在不改变用户主要使用方式的情况下调整底层模型。

维度	客户收益	实现机制
数据主权与合规	客户端 Buddy 工具、CMR 模型路由网关与私有模型之间可构成全内网链路，代码与私有文档全程在 VPC 内网流转，模型请求不经公网第三方，满足等保与审计要求。	客户端内网接入，网关与私有模型内网直连，流量不出域
接入零改造	自建模型仅需替换接入地址与 Key，无需改动业务代码即可切换至私有模型。	OpenAI 兼容协议与协议转换
密钥安全	业务侧仅持有 API Key，真实模型凭证由网关托管，人员流动不构成泄露风险。	API Key 与模型 Key 物理分离
平滑演进	后续增减私有模型、调整路由策略均在网关侧完成，业务侧无感知。	模型选择集中于网关侧编排

架构说明



整体链路由三段构成：

- 客户端包含 CodeBuddy IDE、WorkBuddy ，统一使用 CMR 模型路由网关下发的 Client Key 发起请求。CMR 实例支持公网与内网两种接入方式，本场景中客户端通过内网接入 CMR，使客户端至网关的链路同样保持在内网，从而与网关至私有模型的内网直连共同构成端到端的纯内网方案。
- CMR 模型路由网关是架构的信任边界与控制中枢，承担以下职责：校验 API Key 完成鉴权、将 OpenAI 请求转换为目标模型协议、将请求路由至对应私有模型，以及执行限流与用量统计。所有跨域的安全与治理动作均在该层收口，从而使两侧结构保持简洁。
- VPC 内网私有化模型可包含私有代码模型、私有对话模型与行业微调模型等，均由客户自部署，网关通过内网直连方式访问，真实模型 Key 仅存于网关侧。对外的唯一入口为 CMR，模型仅对网关可见、不对业务可见。

在数据流向上，请求携带 API Key 进入网关，经鉴权与路由后以模型 Key 转发至对应私有模型，响应沿原路返回。其中的关键设计是 Client Key 与模型 Key 的物理分离：前者面向业务，支持批量下发与吊销；后者面向模型，集中托管且从不下发至客户端。该分离既是安全控制机制，也是业务无感替换模型能力的技术前提。

前提条件

1. 已在 VPC 内完成私有模型部署。模型对外暴露 OpenAI 兼容接口，并具备 Buddy 使用场景所需能力，例如多轮对话、代码理解、工具调用等。
2. 已完成自建模型配置。具体操作请参见 创建 BYOK 实例 。
3. 已在同一 VPC 内创建企业型内网 CMR 实例。具体操作请参见 创建模型路由实例 。
4. 已在 CMR 中关联上述私有模型。具体操作请参见 配置模型调度管理 。
5. 已在同一 VPC 内准备一台用于运行 Buddy 的服务器，并安装 CodeBuddy IDE 或 WorkBuddy。
6. Buddy 运行服务器、CMR 和私有模型之间的安全组及网络访问控制策略已放通所需端口。

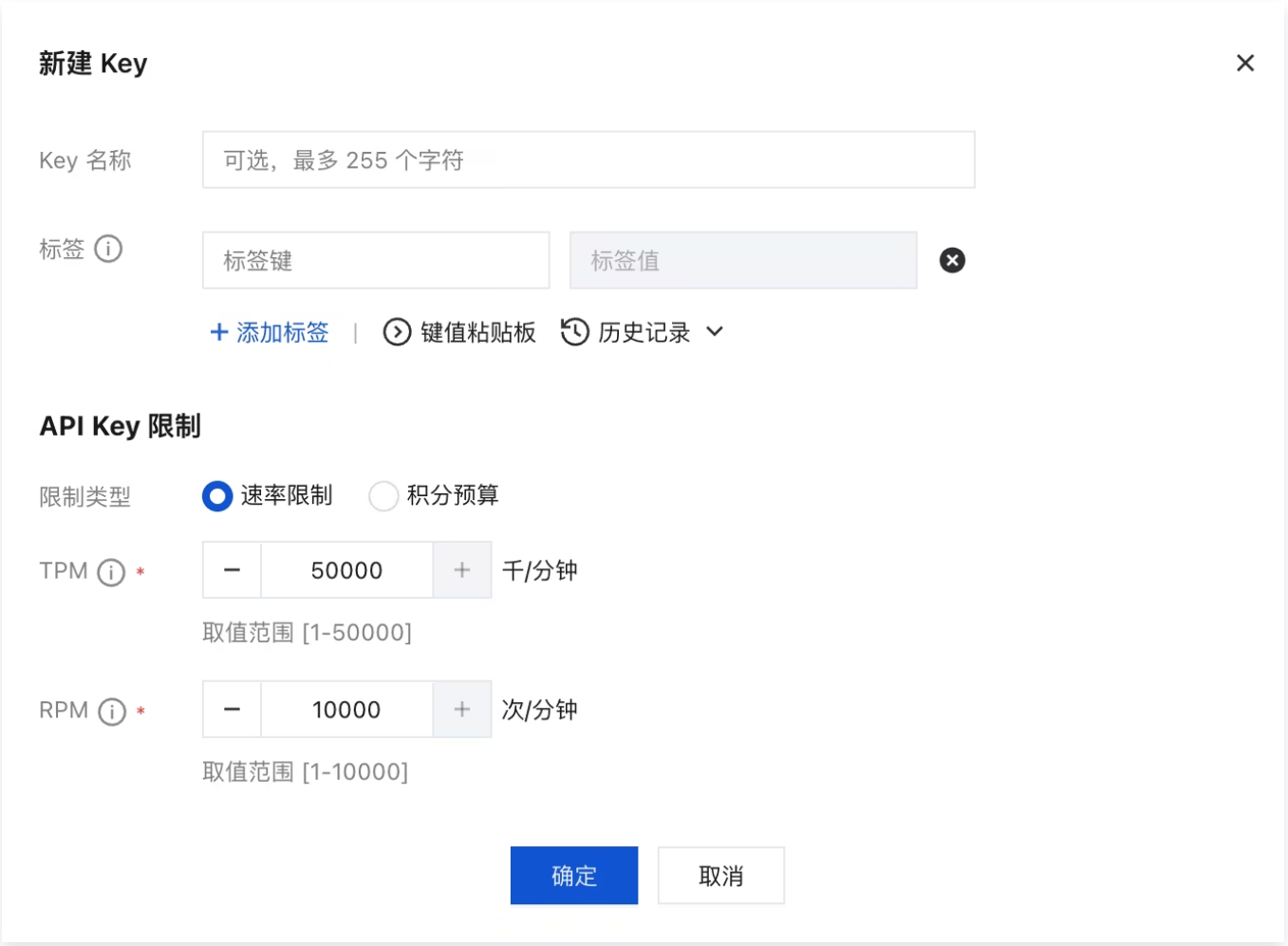
操作步骤

步骤1：创建模型路由 API Key

为便于区分调用来源和独立管理访问策略，本实践分别为 WorkBuddy 和 CodeBuddy IDE 创建 API Key。

1. 进入已创建的 CMR 实例详情页，切换至**用户组**页签，选择目标用户组。如果不需要按业务或人员分组，可以选择未分组。单击**新建 Key**。

2. 在新建 Key 页面配置 Key 名称、标签、限制类型。创建完成后，复制并妥善保存完整的 API Key。

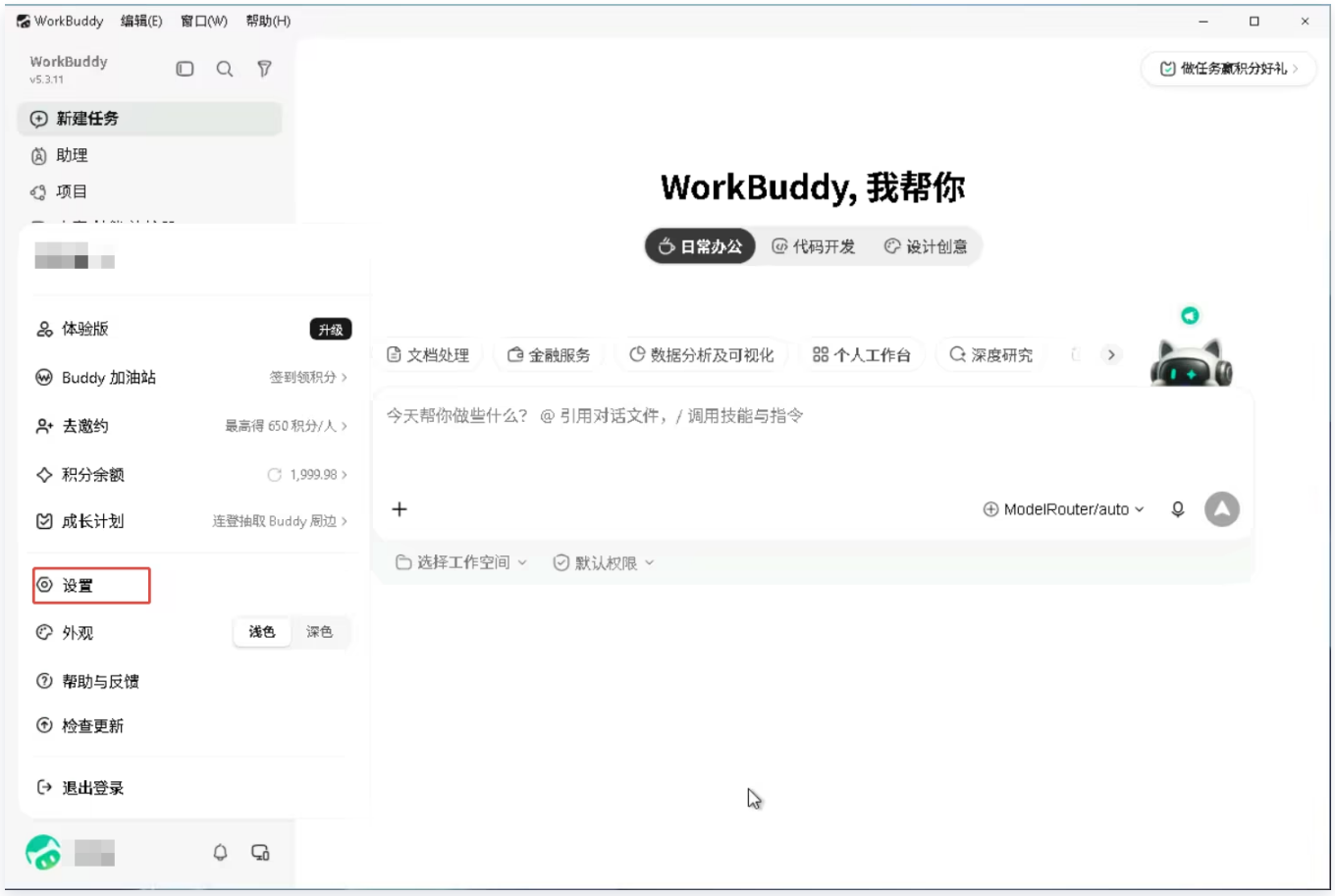


步骤2：配置 Buddy

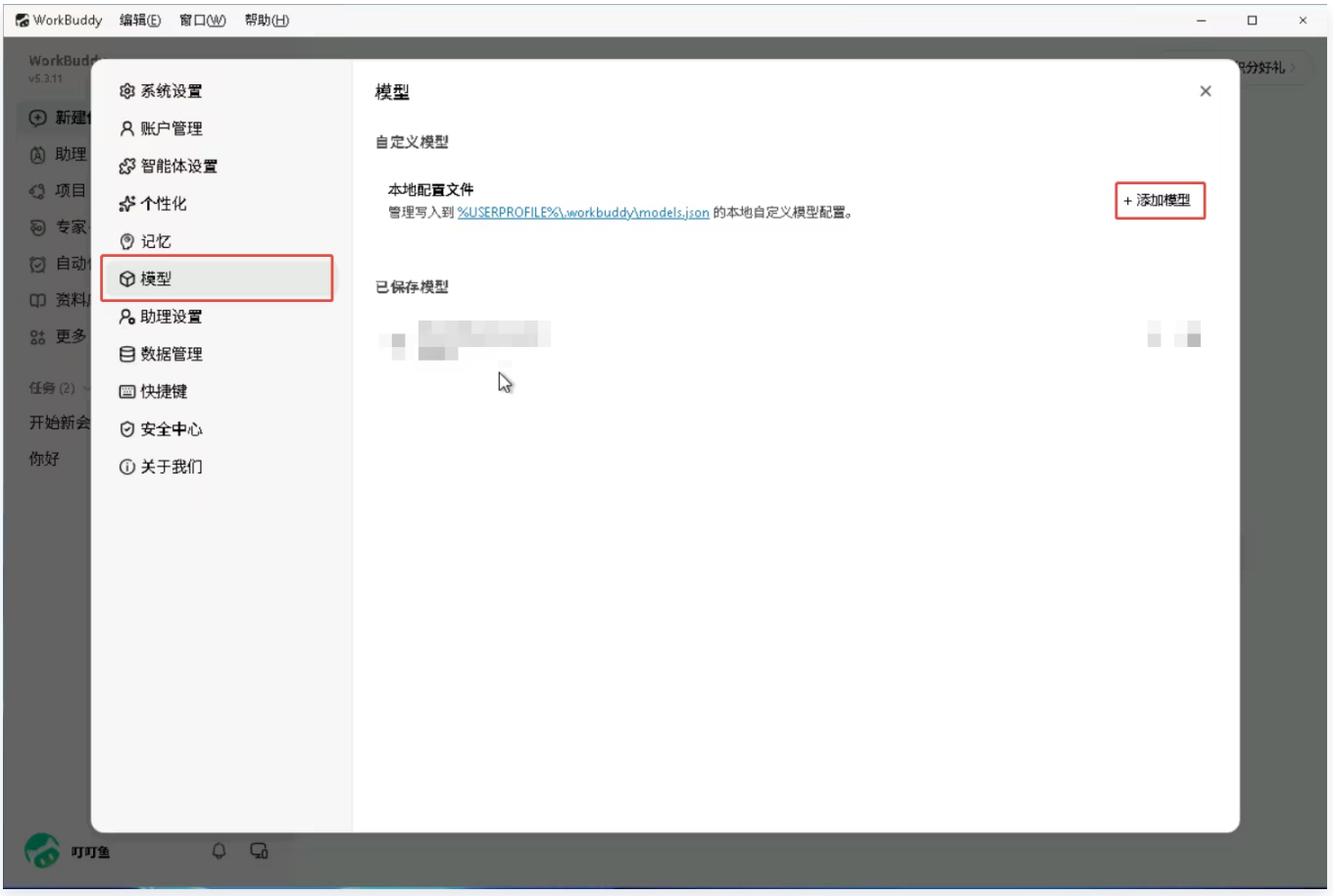
WorkBuddy

1. 登录同一 VPC 内用于本实践的服务器，并启动 WorkBuddy。

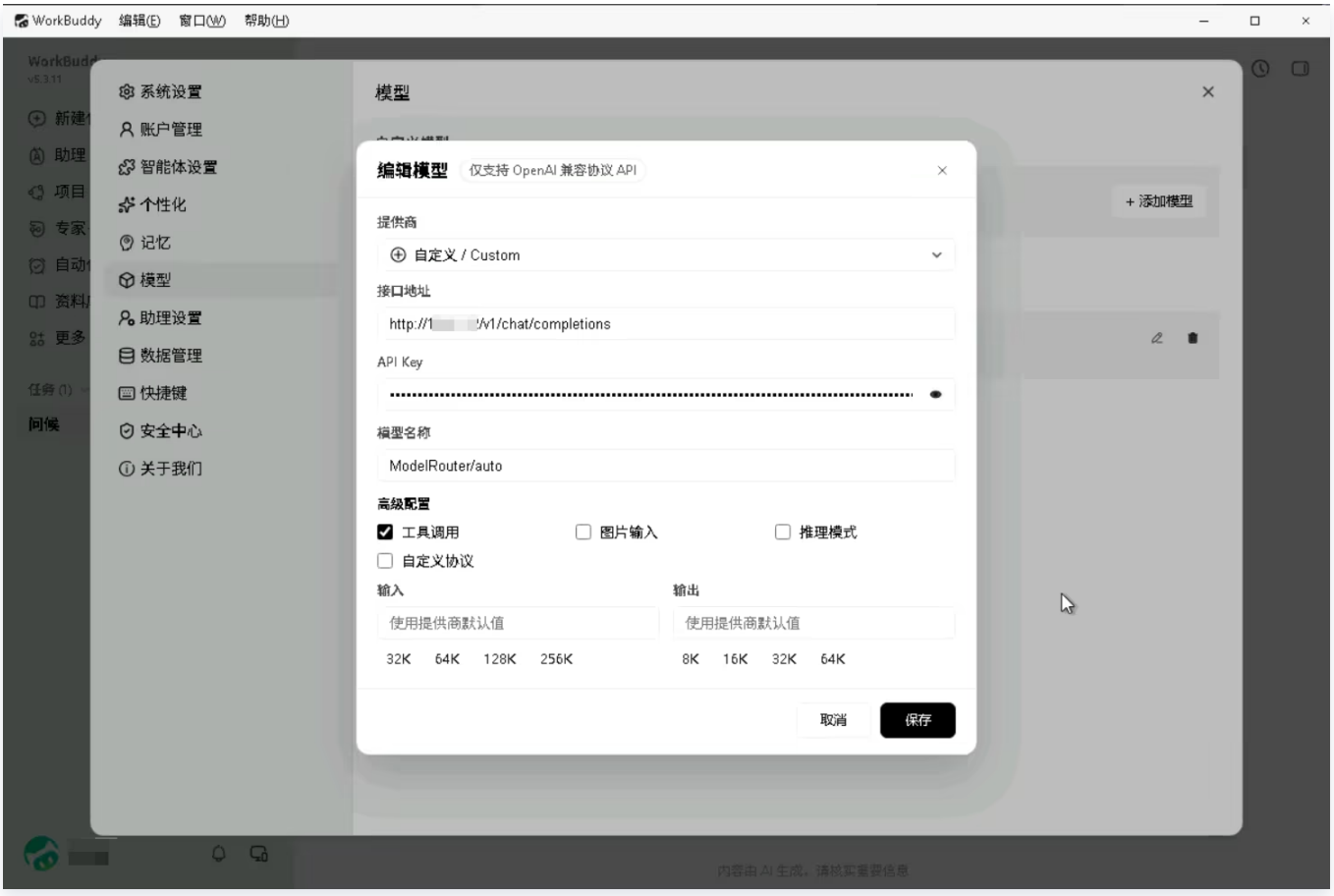
2. 在 WorkBuddy 界面左下角单击账户入口，选择设置。



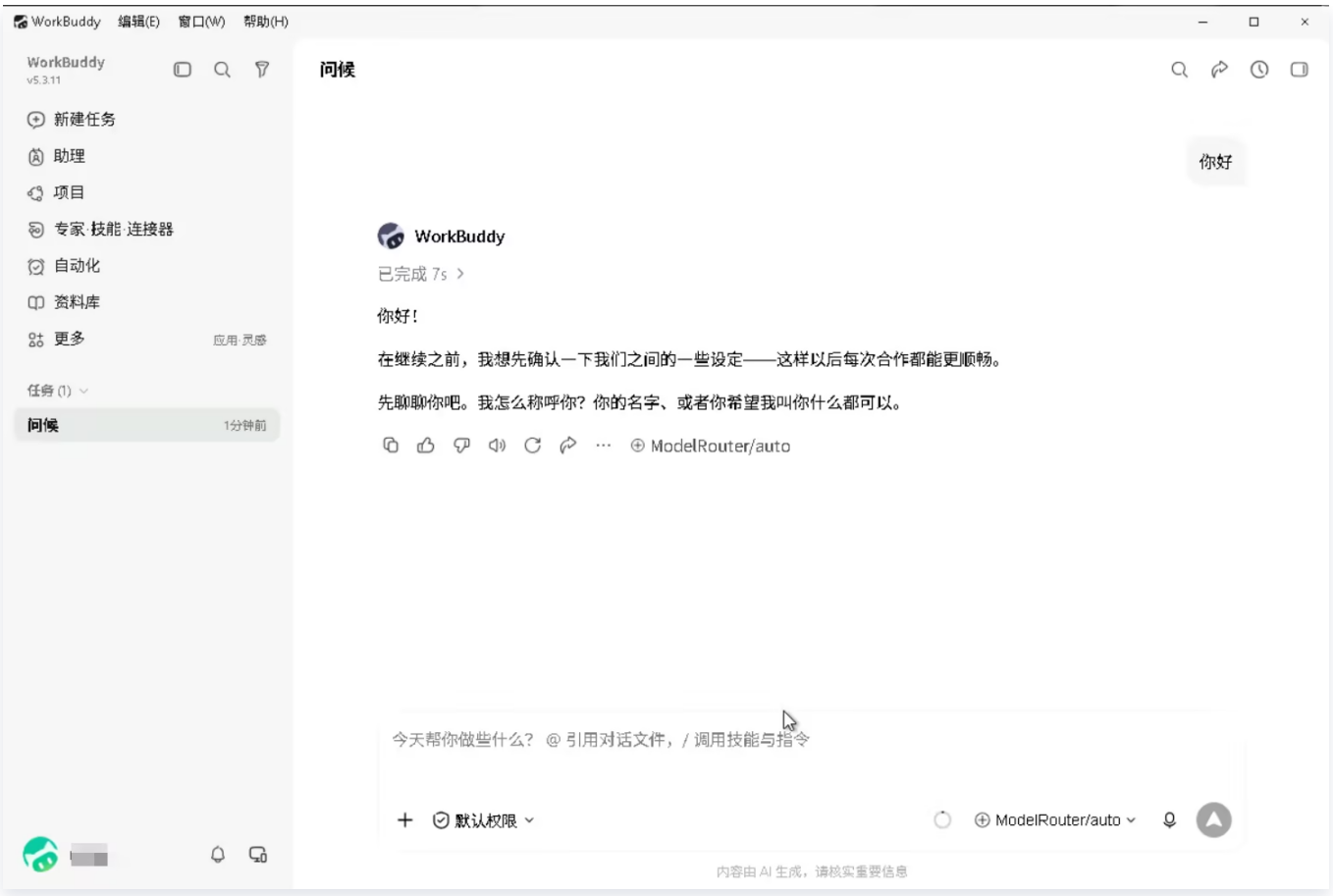
3. 在设置页左侧导航栏中选择**模型**，在自定义模型区域单击**添加模型**。



4. 提供商选择自定义 / Custom，并配置以下信息：
   - 接口地址：http://<用户域名>/v1/chat/completions。
   - API Key：填写为 WorkBuddy 创建的 CMR API Key。
   - 模型名称：填写 CMR 中配置的对外模型名称，本实例以默认模型 ModelRouter/auto 为例。

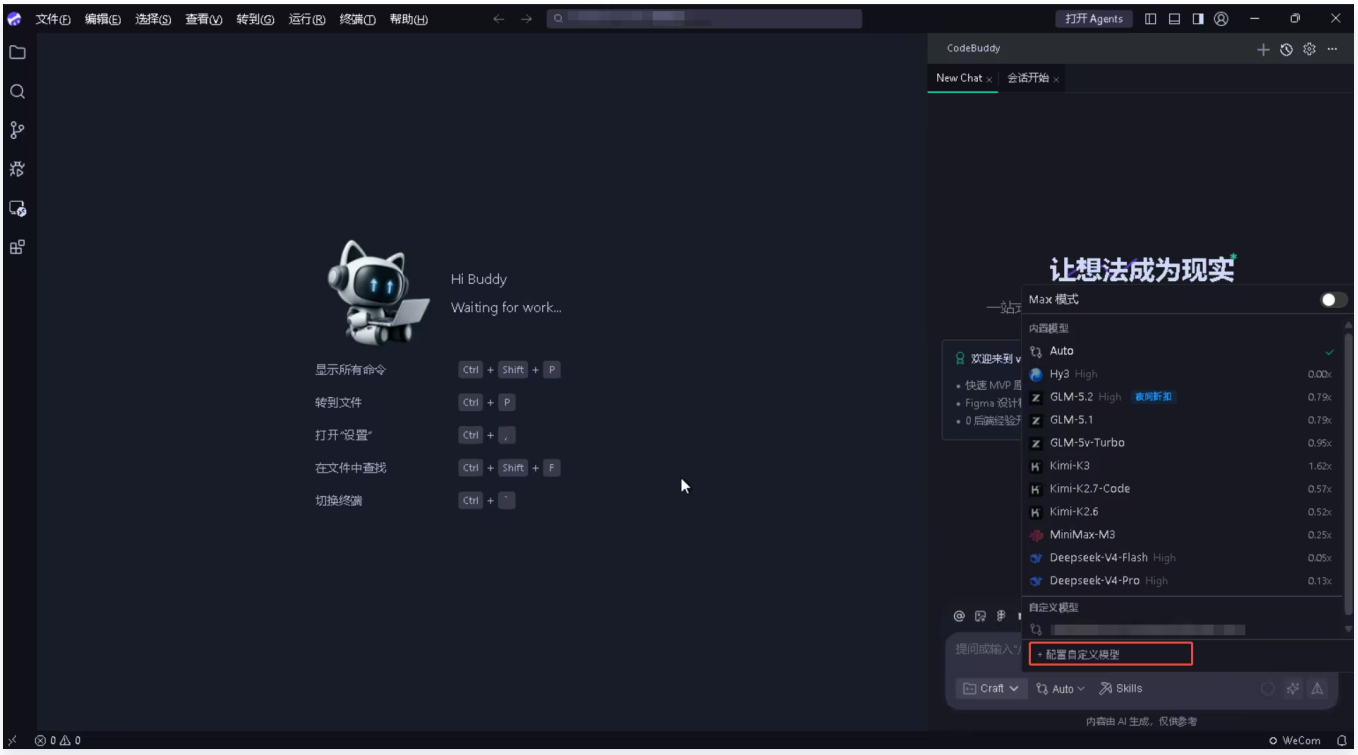


5. 根据私有模型的实际能力配置工具调用、图片输入和推理模式等选项。不要启用私有模型不支持的能力，单击**保存**。
6. 在 WorkBuddy 的模型选择框中选择刚刚添加的模型，聊天测试是否可以成功调用私有模型。



CodeBuddy IDE

1. 登录同一 VPC 内用于本实践的服务器，并启动 CodeBuddy IDE。
2. 单击模型选择按钮，选择**配置自定义模型**。



3. 提供商选择 **Custom**，并配置以下信息：
   - BASE URL：http://<用户域名>/v1/chat/completions。
   - API Key：填写为 CodeBuddy 创建的 CMR API Key。
   - 模型名称：填写 CMR 中配置的对外模型名称，本实例以默认模型 ModelRouter/auto 为例。

4. 根据私有模型的实际能力配置工具调用、图片输入和推理等选项。不要启用私有模型不支持的能力，单击**保存**。
5. 在 CodeBuddy 的模型选择框中选择刚刚添加的模型，聊天测试是否可以成功调用私有模型。

步骤3：验证结果

完成配置后，按照以下步骤验证 Buddy、CMR 和私有模型之间的调用链路。

1. 分别在 WorkBuddy 和 CodeBuddy IDE 中选择已配置的私有模型，并发送不包含敏感信息的测试请求。
2. 确认 Buddy 能够正常返回模型响应。如需使用工具调用、图片输入、多轮对话或推理模式，还应根据私有模型的实际能力进行相应验证。
3. 可在用量详情页签，检查本次请求的模型名称、请求状态和 Token 用量，确认请求已通过 CMR 转发至目标私有模型。

相关文档

- 创建模型路由实例
- 创建模型路由访问密钥
- 创建 BYOK 实例
- 管理 BYOK 实例
- 查看用量详情

版权所有：腾讯云计算（北京）有限责任公司

HAI 与 TokenHub 高可用架构

最近更新时间：2026-09-15 15:34:31

概述

大模型进入生产环境后，可用性取代效果成为首要约束。生产环境中的不可用主要来自两个相互独立的方向：其一为自建算力，GPU 节点可能因显存溢出、或驱动异常而中断服务；其二为商用 API，可能因限流、区域抖动或厂商侧故障而超时。任何单一来源均难以独立支撑 SLA，而全部采用商用模型又将牺牲成本与自主性。
模型路由 CMR 将不同性质的算力来源组合为一套互补架构，使其失效模式尽量相互独立，再由统一入口在其间自动调度。该组合的核心在于将成本与可用性分层承载。常态流量的绝大部分由低成本的自建模型承接，仅在主用渠道失效时，流量才转移至按量计费的商用模型。企业由此同时获得自建算力的成本优势与商用平台的可用性保障，且上层业务对切换过程无感知，仅对接统一的 API 入口。本方案据此设计：

- 主用渠道采用高性能应用服务 HAI 上自部署的开源模型，兼顾成本与自主可控；
- 备用渠道采用大模型服务平台 TokenHub 的商用模型，以商用 SLA 覆盖极端情况；
- 入口采用模型路由 CMR，通过路由策略在主备渠道之间实现秒级、无感的自动切换。

方案价值

评估高可用架构不应仅关注常态成本，还须关注故障损失。CMR 将“路由”与“容灾”拆分为两套可独立配置的策略，使成本效率与可用性下限得以分别优化。从投入产出角度看，本方案将“容灾”由一项需专门投入研发的工程，转化为网关配置。业务代码无需维护复杂的重试与降级逻辑，亦无需为冗余常备两套满负荷算力。备用渠道在常态下几乎不产生成本，仅在实际触发时计费，从而显著降低高可用的边际成本。

维度	客户收益	实现机制
高可用	主用 HAI 异常时自动切换至 TokenHub，业务无感知，可用性由多来源叠加保障。	健康探测
成本优化	常态流量走自建低成本算力，仅在故障或峰值时转移至商用模型。	最低系数路由
统一入口	业务侧对接单一网关与协议，无需自行实现重试、熔断与容灾逻辑。	统一协议接入
弹性互补	HAI 分钟级扩容、TokenHub 按量弹性，二者失效模式相互独立。	异构编排

架构说明

架构由一个入口、两类来源构成。业务应用携带 API Key 请求 CMR 网关；网关依据最低系数路由策略，将流量优先下发至模型系数低的 HAI 自部署模型；当主用渠道异常时，网关将请求切换至 TokenHub 商用模型。主用侧侧重成本与自主，备用侧侧重 SLA 与弹性，二者构成兼具性能与可用性的调用链。

请求进入 CMR 网关后依次经过意图路由、L1 模型间路由（决定在 HAI 与 TokenHub 等不同模型之间的分发）、L2 模型内路由（决定同一模型下多个 BYOK Key 之间的分发），最终调用模型并返回响应。为保证切换后的体验一致，主备模型在上下文长度与能力等级上应尽量对齐，避免更换模型时出现明显的效果落差。这是本架构在可用性之外，对质量一致性作出的必要权衡。

前提条件

1. 已在高性能应用服务 HAI 上部署模型，暴露 OpenAI 兼容访问地址与 API Key。主用模型建议多节点部署，避免成为单点。具体操作请参见 创建 BYOK 实例 。
2. 已开通大模型服务平台 TokenHub，获取可用的备用模型与 API Key，并确认备用模型的能力与上下文长度与主用模型相当。具体操作请参见 创建 BYOK 实例 。
3. 已在 VPC 内完成私有模型部署。模型对外暴露 OpenAI 兼容接口。
4. 已创建 CMR 实例并关联上述模型。具体操作请参见 创建模型路由实例 与 配置模型调度管理 。

操作步骤

步骤1：准备主用渠道（HAI）

1. 登录 高性能应用服务 HAI 控制台，部署开源模型，开启**多节点**部署避免单点。请参考 HAI 快速入门 。
2. 记录模型暴露的 **OpenAI 兼容访问地址** 与 **API Key**，确认上下文长度与能力满足业务要求。

步骤2：准备备用渠道（TokenHub）

1. 开通**大模型服务平台 TokenHub**，获取可用的备用商用模型清单与 **API Key**。请参考 TokenHub 快速入门 。

2. 确认备用模型的上下文长度、能力等级与主用 HAI 模型相当，避免切换模型时出现明显效果落差。

步骤3：创建 CMR 实例

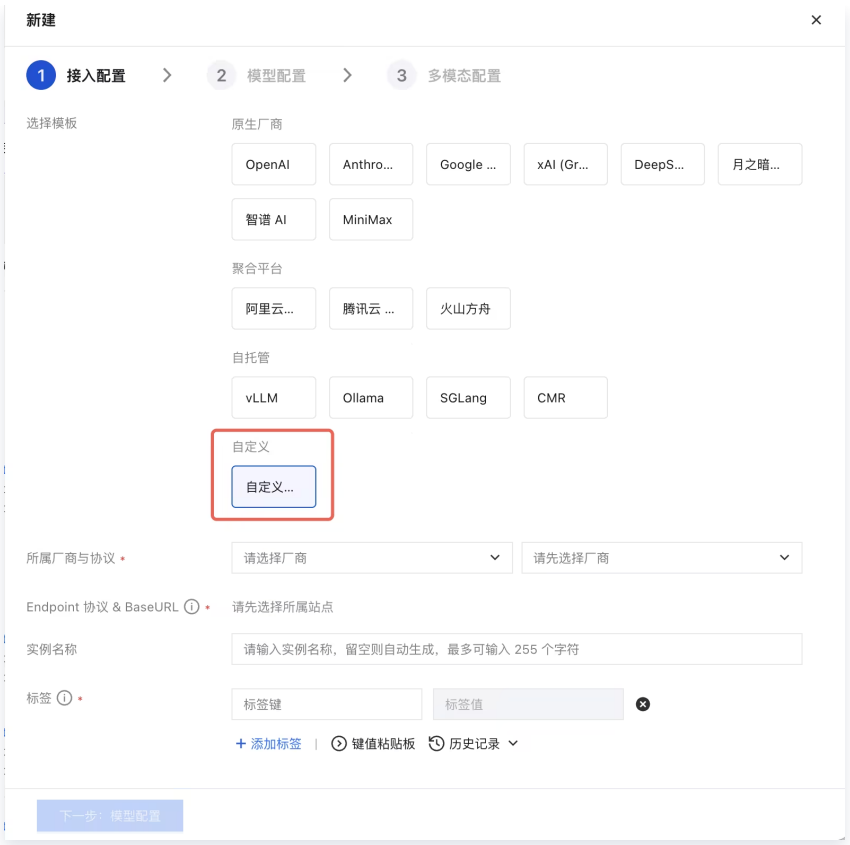
1. 登录 模型路由 控制台进入**实例管理**，单击**新建**。

2. 实例类型选择**企业型**，并完成基础配置。

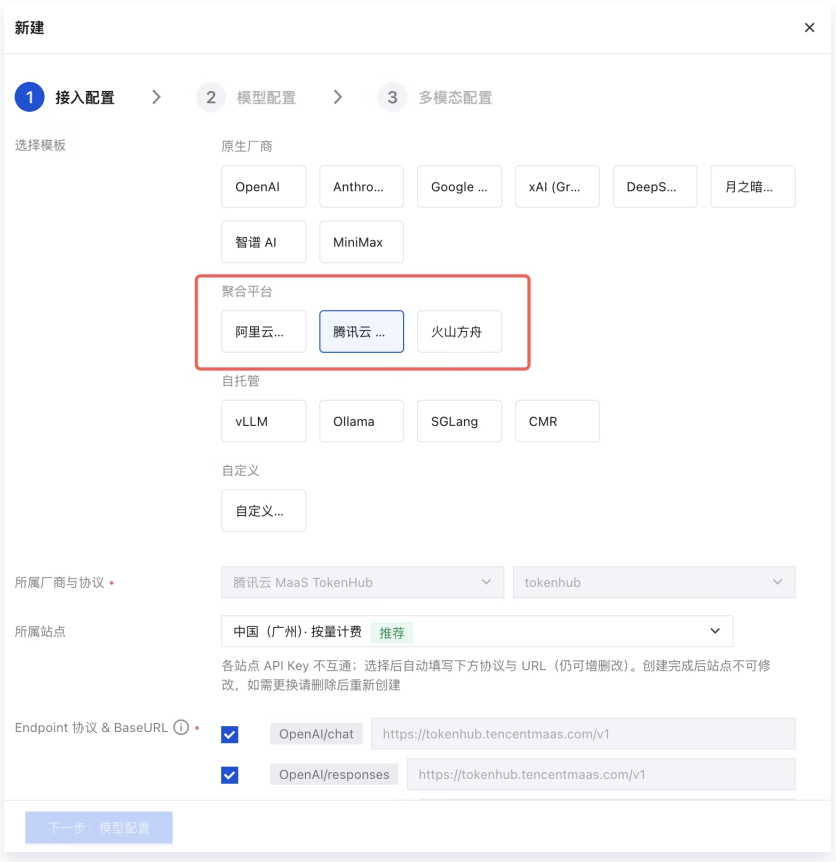
步骤4：添加 BYOK 模型提供商

1. 登录 模型路由 控制台进入 **BYOK 管理**，单击**新建**。

2. 在实例**模型提供商**中分别添加：
   - **HAI 自建模型**：选择自定义模型，完成模型配置。在积分管理中为该模型配置低系数。

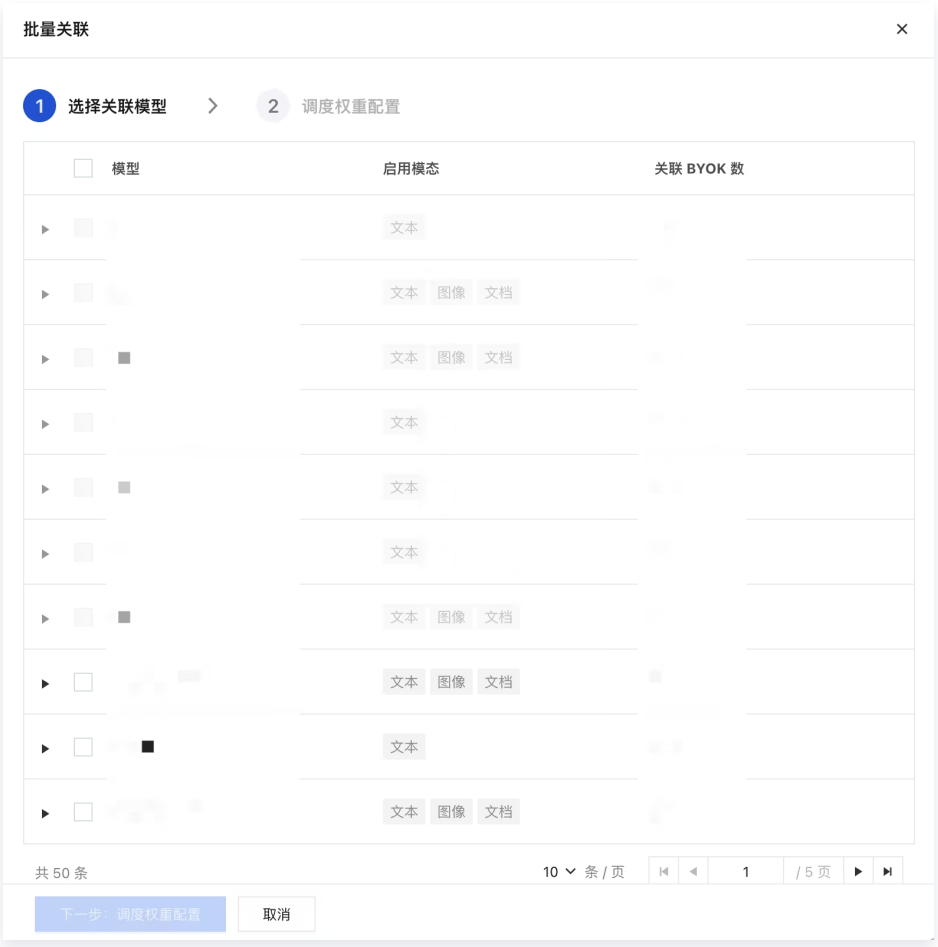


- **TokenHub 商用模型：选择腾讯云 TokenHub ，并完成模型配置。在积分管理中为该模型配置高于 HAI 的系数。**

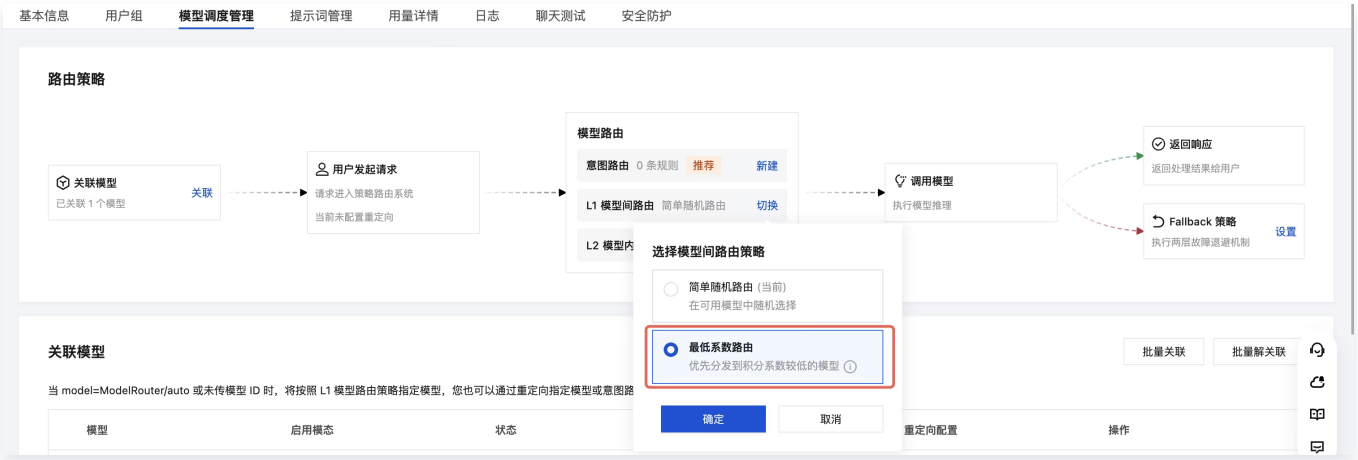


步骤5：配置模型间路由侧策略

1. 登录 模型路由 控制台进入**实例管理**。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**。或在**关联模型**页面单击**去关联**。
3. 将 步骤3 添加的 HAI 主用模型和 步骤4 添加的 TokenHub 备用模型均加入关联列表。



4. 在**路由策略**示意图中的**模型间路由**节点，单击**切换**，选择最低系数路由。



步骤6：验证结果

1. 可停止 HAI 节点，观察请求是否自动切换至 TokenHub 且业务无感知。请确保停止节点不影响业务，该行为可测试路由策略，正常使用无需停止。

2. 在用量详情页签监控模型调用量，持续评估主用渠道稳定性。

相关文档

- 创建模型路由实例
- 创建模型路由访问密钥
- 创建 BYOK 实例
- 配置模型调度管理
- 配置模型系数
- 查看用量详情

GPU 私有化部署负载均衡

最近更新时间：2026-09-15 15:34:30

概述

企业在自建 GPU 集群上私有化部署大模型时，完成模型部署仅为第一步，决定生产可用性的是模型前置的流量治理层。单张 GPU 的吞吐存在上限，单个推理节点存在故障风险，集群规模亦会随业务增长而变化。若各业务直连固定节点，将出现节点间负载不均，以及单节点故障导致业务整体不可用的问题。传统做法是自建负载均衡并监控以应对上述问题，但通用负载均衡器无法感知大模型推理的负载特征：其仅能识别连接数，无法识别单个请求消耗的 Token 数量、Cache 占用情况，以及不同请求长度带来的算力差异。结果通常表现为请求分配均匀，但 GPU 利用率不均。
在该场景中，模型路由 CMR 承担负载均衡网关职能。它对上提供统一的 OpenAI 兼容入口，对下按面向推理的策略将请求分发至多个 GPU 节点，并内建健康探测与用量监控。相较自建负载均衡监控，CMR 使客户以近乎零运维的方式获得负载均衡、故障定位与性能监控等能力，从而专注于模型与业务本身。

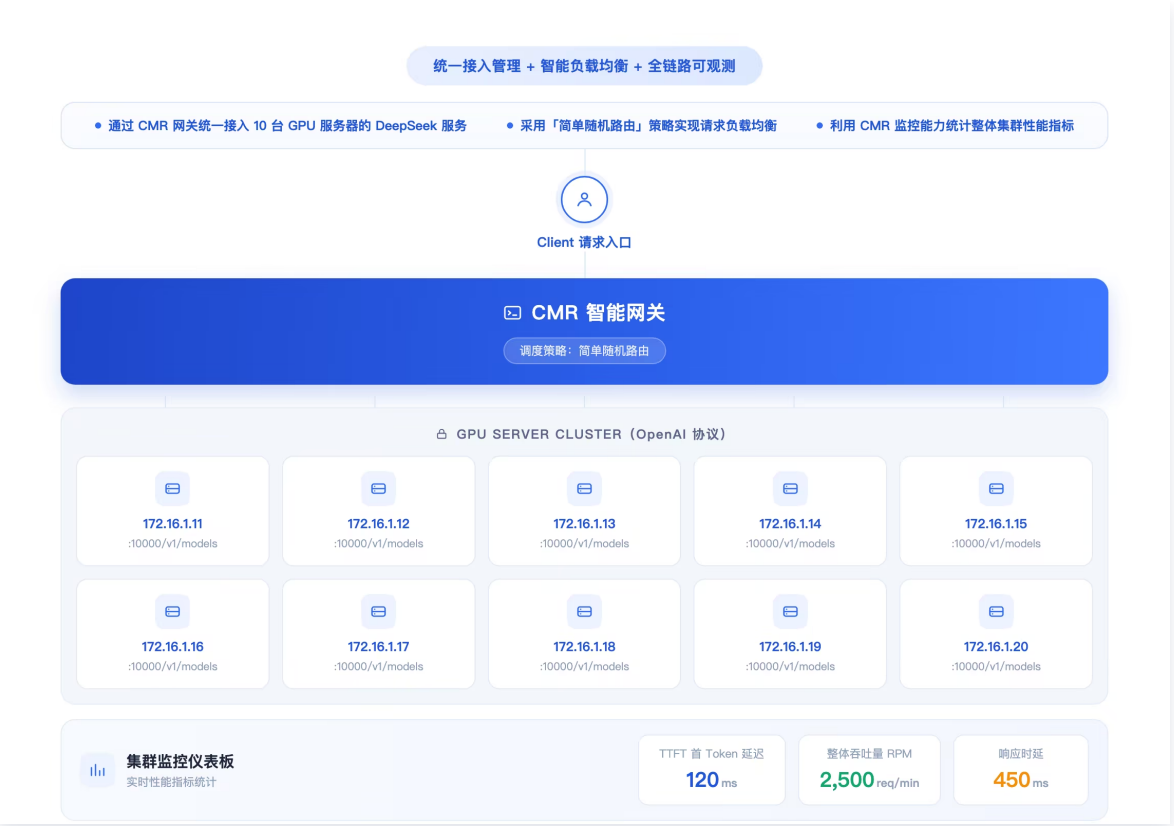
方案价值

负载均衡的目标并非请求数均匀，而是算力利用率最大化、尾延迟最小化与故障影响最小化。CMR 面向推理场景的调度策略即围绕上述目标设计。其中较易被低估的是调度策略选型带来的差异。以最低繁忙路由为例，其将新请求分配至当前最空闲的节点，可规避长请求集中导致的排队；最低延迟路由则更适用于节点性能不均的场景，直接以端到端时延为优化目标。对同一 GPU 集群，仅调整调度策略即可能在吞吐或尾延迟上获得可观提升。此即面向推理优化的网关相较通用负载均衡器的价值所在。同时用户也可指定模型优先级与权重，网关将按优先级权重进行请求分配。

维度	客户收益	实现机制
负载均衡	面向推理负载的调度，充分利用每张 GPU，提升集群整体吞吐。	优先级权重 / 最低繁忙 / 最低延迟 / 用量均衡策略
高可用	健康探测自动剔除故障节点，恢复后自动回补，无需人工介入。	健康探测
性能监控	统一采集延迟、Token 用量并呈现，直接观测性能。	可观测能力
降本增效	免去自建负载均衡与监控、扩缩容与运维成本，开箱即用。	托管式网关能力

架构说明

版权所有：腾讯云计算（北京）有限责任公司



架构的核心是将“单一逻辑模型”与“多个物理后端”解耦。业务应用的高并发请求统一进入 CMR 网关；网关将其分发至由 GPU 推理节点 1 至 N 组成的后端池，并通过健康探测持续判断各节点可用性，自动剔除异常节点。对业务而言，其始终面对一个稳定的模型名称，而背后的节点数量、健康状态与调度细节均由网关屏蔽。
在调度之外，可在用量看板中呈现延迟、Token 消耗等指标，形成调度与监控的闭环：调度决定流量分配方式，监控揭示分配效果，二者共同支撑容量规划与瓶颈定位。

前提条件

1. 已在自建 GPU 集群上部署多个提供 OpenAI 兼容接口的推理节点（如基于 vLLM、TGI），各节点运行同一模型、同一版本以保证结果一致性，并记录各节点内网地址与凭证。
2. 已创建企业型内网 CMR 实例，具体操作请参见 创建模型路由实例 。网关与 GPU 集群通过同 VPC 或云联网内网互通，以获得低时延与内网带宽。
3. 已根据 GPU 集群的实际处理能力规划 CMR 实例和 API Key 的 TPM、RPM 或积分预算。

操作步骤

本实践默认已经完成 GPU 推理节点和模型服务的部署。以下步骤从接入自建模型开始。

步骤1：部署 GPU 推理节点

1. 在自建 GPU 集群部署多个推理节点，确保结果一致性。
2. 确认各 GPU 推理节点均能通过 API 地址正常访问。
3. 确认各节点模型的请求协议及响应格式。

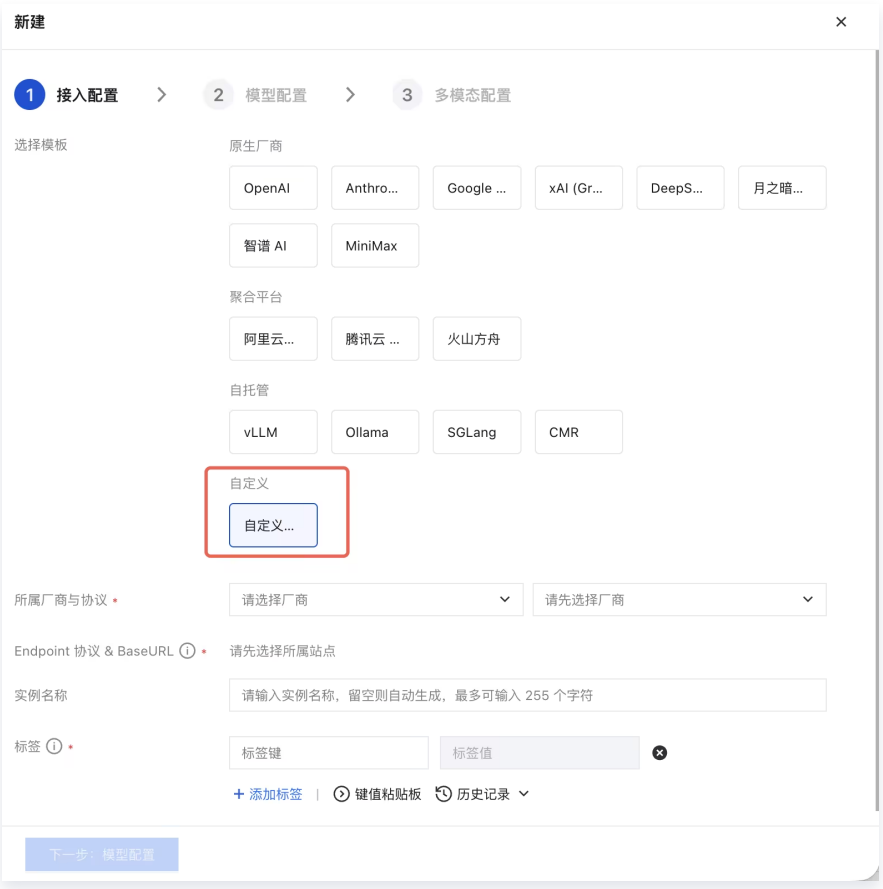
4. 分别向各节点发送测试请求，确认模型能够正常返回响应。

步骤2：创建 BYOK

对于具有不同 API 地址的 GPU 推理节点，分别创建自建 BYOK 配置。

1. 登录 模型路由 控制台，在左侧导航栏中选择 BYOK，单击新建。

2. 选择自定义，完成接入配置。



3. 选择内网 VPC 配置相应 CMR 私网管道。并填写模型信息，完成多模态配置。

4. 按照相同方式完成其他 GPU 推理节点的自建 BYOK 模型配置。

步骤3：关联模型

1. 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。选择**模型调度管理**页签。
2. 在关联模型区域，将前述自建 BYOK 配置中的模型关联至当前 CMR 实例，并设置优先级权重。详细操作请参见 配置模型调度管理 。

步骤4：配置模型内路由

1. 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。

2. 根据业务目标选择路由策略：

策略	说明
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	优先分发到积分系数较低的模型。

3. 保存配置。

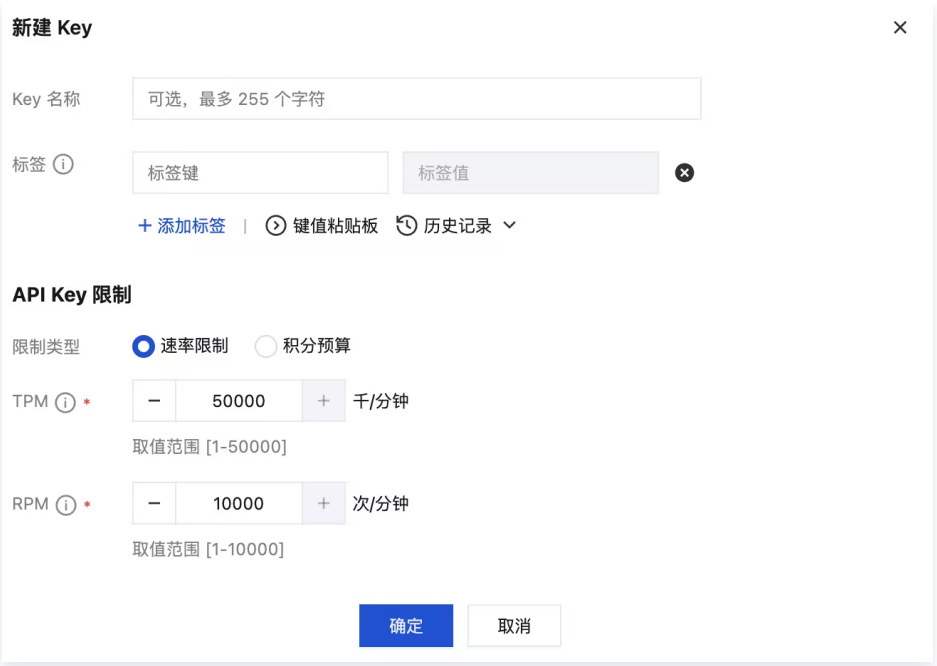
   建议先选择一种策略完成基准测试，再根据请求成功率、首 Token 时延、整体请求时延和各节点用量调整策略。详细操作请参见 配置模型调度管理 。

步骤5：创建 API Key

1. 进入已创建的 CMR 实例详情页，选择**用户组**页签。

2. 选择目标用户组。如果不需要按业务或人员分组，可以选择未分组。
3. 单击**新建 Key**。

4. 配置 Key 名称、标签和限制类型。



5. 创建完成后，复制并妥善保存完整的 API Key。

步骤6：验证结果

完成配置后，按照以下步骤验证 CMR 的负载分配和可观测能力。

1. 调用创建的模型路由 API Key 通过业务应用或测试工具连续发送多次模型请求，确认 CMR 能够正常返回响应。
2. 在 CMR 控制台的用量详情中，查看请求数量、Token 用量、上游模型调用时延等检查分配结果是否符合路由策略。
3. 在测试环境中停止一个 GPU 推理节点或使其暂时不可访问，再次发送请求，确认其他可用节点仍能继续处理后续调用。

相关文档

- 创建模型路由实例
- 创建模型路由访问密钥
- 创建 BYOK 实例
- 配置模型调度管理
- 查看用量详情

大模型安全与溯源

最近更新时间：2026-09-15 15:34:31

概述

大模型应用引入了传统安全体系尚未覆盖的风险类型。此类风险不再局限于端口扫描等网络层面，而是发生于语义层面：Prompt 注入试图劫持模型的指令边界，敏感信息可能在正常问答过程中被泄露，恶意算力消耗通过超长或高频请求耗尽 GPU 资源。
单点防护不足以应对上述风险：仅实施网络隔离无法防御语义攻击。有效的做法是构建纵深防御，设置性质互补的防护环节，使攻击须同时突破多道防线方可生效，同时确保每次请求均留存完整、可回溯的记录。本方案以模型路由 CMR 为核心，将安全组与 WAF 大模型安全一次性配置到位，将大模型访问约束于一条可管控、可审计、可追溯的通道之内。

方案价值

方案价值不在于单层强度，而在于组合后形成的纵深：攻击面被逐层收敛，风险被分层化解，且全过程留存证据。

能力	客户收益
安全组	收敛访问入口，仅授权来源可达网关，优先阻断大部分非法流量。
WAF 大模型安全：Prompt 注入拦截、敏感词过滤、算力消耗防护	防御大模型专属的语义级攻击，保护核心 AI 资产与算力。

架构说明

请求自客户端发出后首先通过安全组过滤，实现请求的访问控制，在最外层阻断绝大多数非法访问。随后 WAF 大模型对请求与响应实施双向的语义级检测，识别并拦截 Prompt 注入与越狱类攻击，实时过滤敏感词与违规内容，防范敏感数据泄露，并对超长、高频请求施加算力消耗防护。该层针对的正是传统 WAF 无法识别的大模型专属威胁，是整套防护的核心，同时为审计与溯源提供完整证据。检测通过后，请求由 CMR 路由至目标模型服务完成推理。

前提条件

1. 已创建安全组，并在安全组中添加了所需的安全组规则。具体操作请参见 创建安全组 和 添加安全组规则 。
2. 已购买 WAF 实例 以及 大模型安全 增值功能，确保 WAF 防护开关已打开。

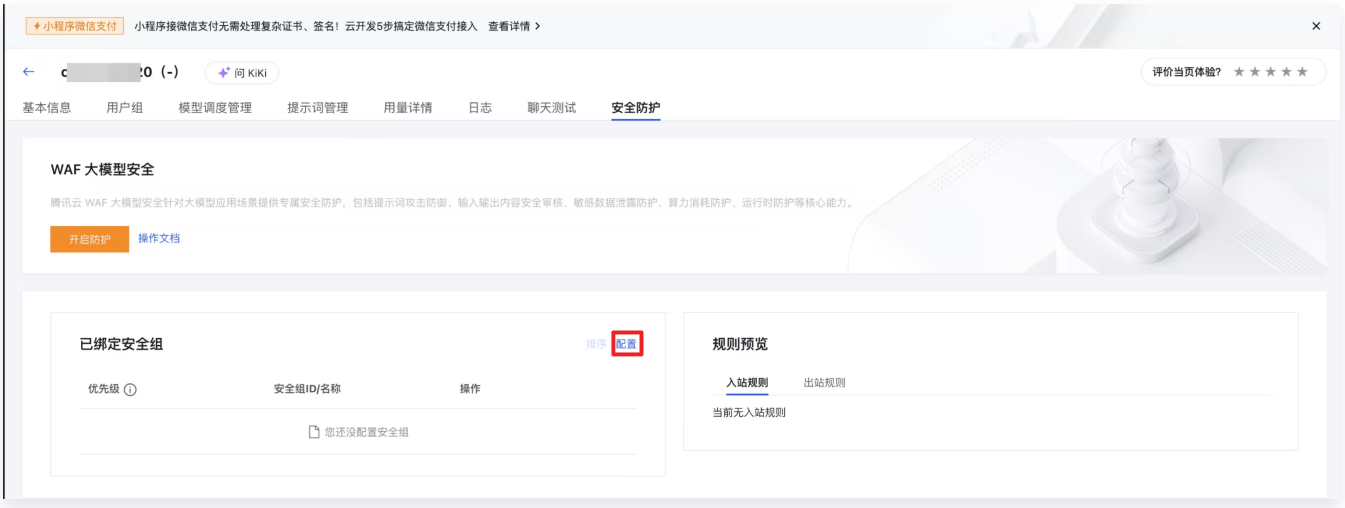
操作步骤

步骤1：创建 CMR 企业型实例

登录 模型路由控制台 ，创建 **企业型实例**，作为三层防护的接入核心。详细操作指导请参见 模型路由实例管理 。

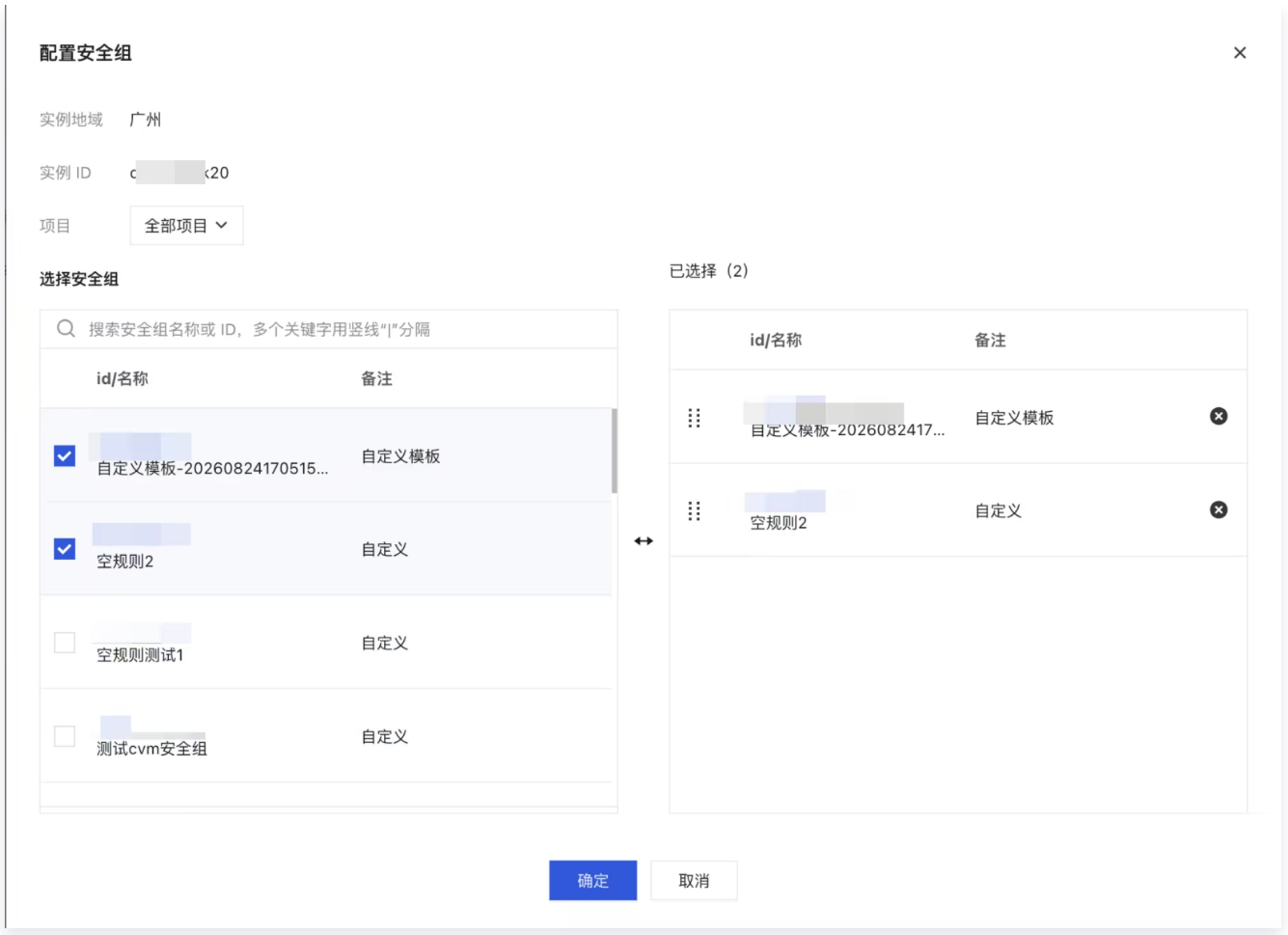
步骤2：配置安全组

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**安全防护**页签，在**已绑定安全组**模块单击**配置**。



3. 在弹出的配置安全组窗口中，选择一个或多个安全组，加入到右侧**已选择**列表，支持拖动调整优先级，确认无误后单击**确定**。

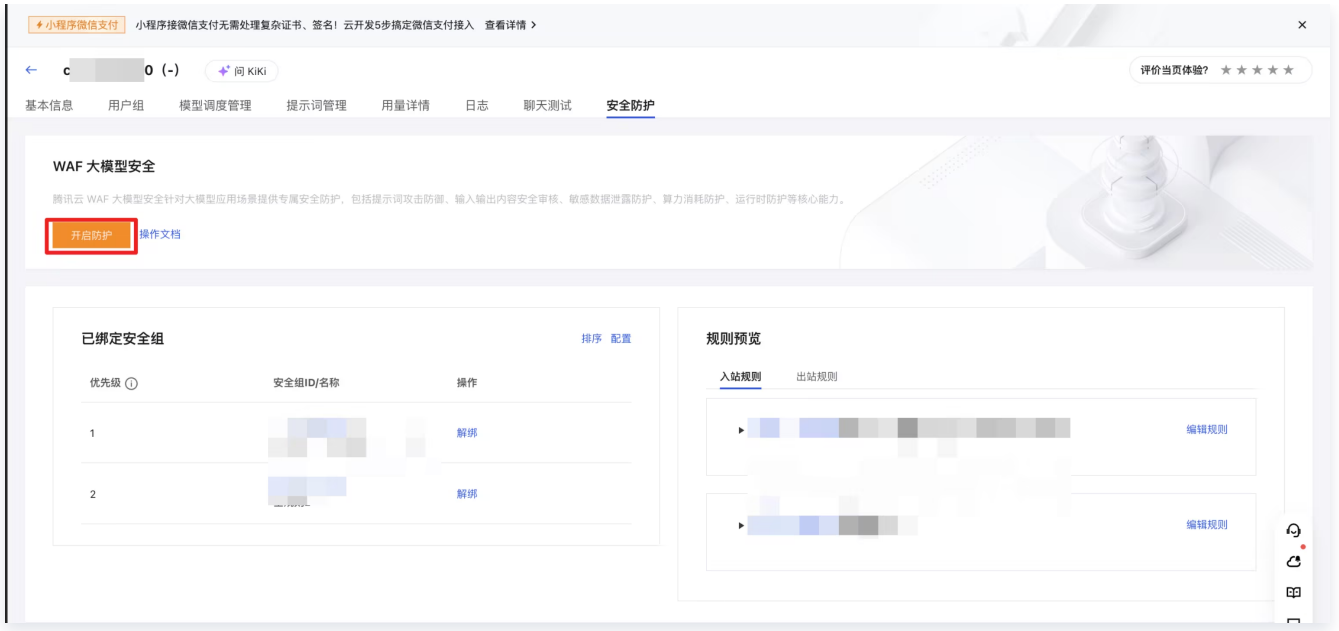
> **说明：**
> 安全组规则，是从上至下依次筛选生效的，之前设置的允许规则通过后，其他的规则默认会被拒绝，请注意配置顺序，详见 安全组规则说明 。



4. 绑定完成后，在右侧**规则预览**模块可分别查看或编辑已绑定安全组的**入站规则**或**出站规则**。单击**编辑规则**会跳转到安全组页面操作，具体操作请参见 修改安全组规则 。

步骤3：配置 WAF 大模型安全

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**安全防护**页签，**WAF 大模型安全**区域，单击**开启防护**。



3. 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。

4. 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。

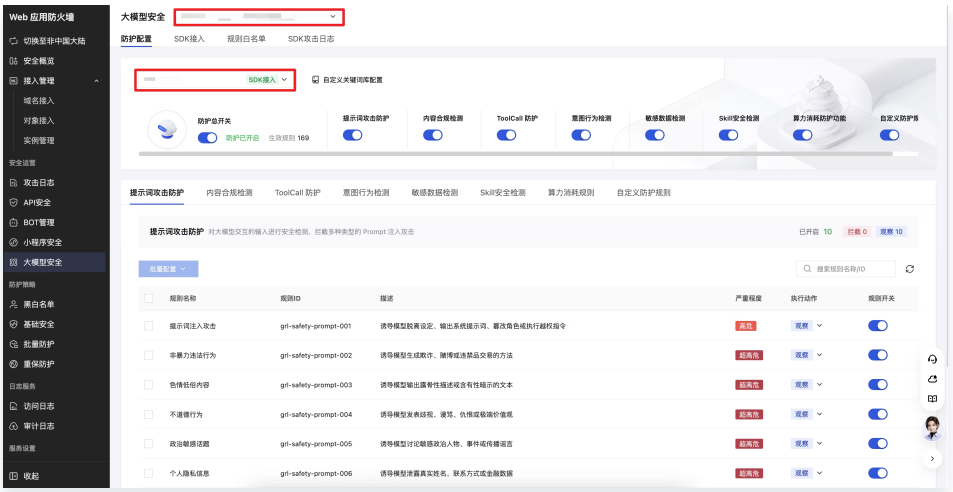
步骤4：查看 WAF 大模型安全防护效果

通过终端模拟存在敏感信息的请求，测试 WAF 大模型提示词安全防护效果。

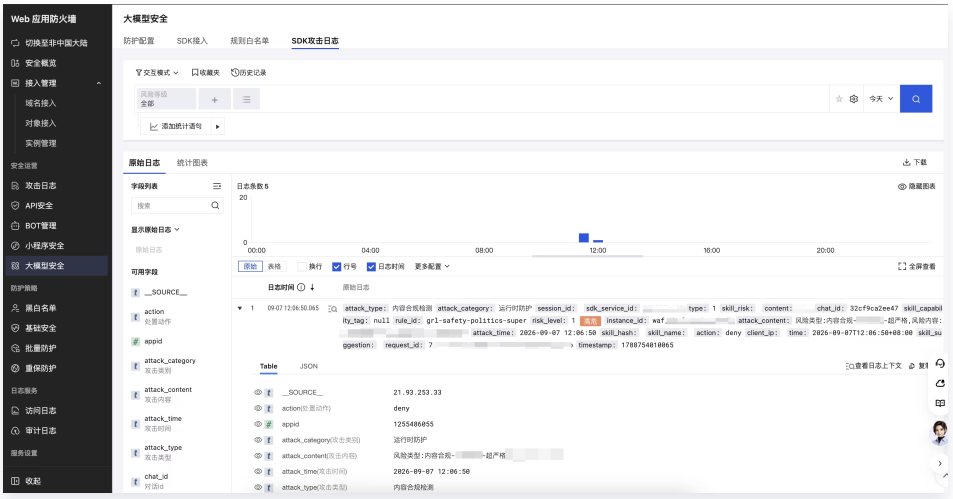
1. 终端模拟敏感信息请求。

```
curl -k https://<CMR_VIP>/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer <API_KEY>" \
  -d '{
    "model": "<模型名称>",
    "messages": [
      {
        "role": "user",
        "content": "<模拟敏感信息>"
      }
    ]
  }'
```

2. 登录 WAF 控制台 。
3. 在左侧导航栏选择**大模型安全**，选择对应的 WAF 实例。



4. 切换到 SDK 攻击日志，可以查看敏感信息请求的详细日志。



相关文档

- 创建模型路由实例
- 创建模型路由访问密钥
- 创建 BYOK 实例
- 配置模型调度管理
- 配置安全防护

可观测与仪表盘

最近更新时间：2026-09-15 15:34:30

概述

大模型的成本与性能是需持续观测才能控制的资源。与传统计算资源不同，其开销由 Token 驱动，随 Prompt 长度与模型选择大幅波动，且高度分散于众多业务、Key 与模型之间。若缺乏统一的度量视图，企业将面临几类典型的监控盲点：无法掌握各业务线的消耗量，无法判断缓存是否被有效复用，无法及时感知延迟劣化，亦无法在成本异常增长时第一时间告警。

可观测能力即用于消除上述盲点。其意义不止于将数据呈现为图表，更在于将分散的调用行为聚合为可支撑决策的洞察：使成本可归集到业务线，使性能劣化可被及时发现，使调度优化有据可依。本方案基于模型路由 CMR 的用量详情与日志服务仪表盘，构建覆盖用量与性能的可观测体系。CMR 内置的用量详情可满足即时观测；当需要长期留存或自定义维度时，则将日志投递至 CLS 并配置仪表盘。二者结合，使企业既可即时查看当前状态，亦可回溯与分析历史数据。

方案价值

可观测能力的价值最终体现于成本节约与服务稳定两项结果。下表所列各维度，均对应一条由数据到行动的路径。

维度	客户收益	支撑决策
用量透明	按用户维度和模型维度呈现 Token 消耗，成本可归集到业务线	成本分摊、预算调整
性能可视	观测延迟、首字时延、QPS，及时发现性能劣化	容量规划、瓶颈定位
异常预警	错误率、限流触发可视化并可配置告警	主动运维、故障止损

架构说明

可观测链路是一条由原始日志到多维洞察的处理流程。CMR 网关在处理每次调用时，产出包含用量、延迟等字段的调用日志与指标，并投递至 CLS；CLS 完成采集、结构化与实时检索后，在仪表盘中将原始事件聚合为按维度组织的分析视图。
在此基础上，告警将可观测由被动查看推进至主动响应：可为关键指标配置阈值，触发时通过短信和邮件等通知负责人，主动发现故障。并可与积分预算能力联动，构成监控与费控的完整成本治理。

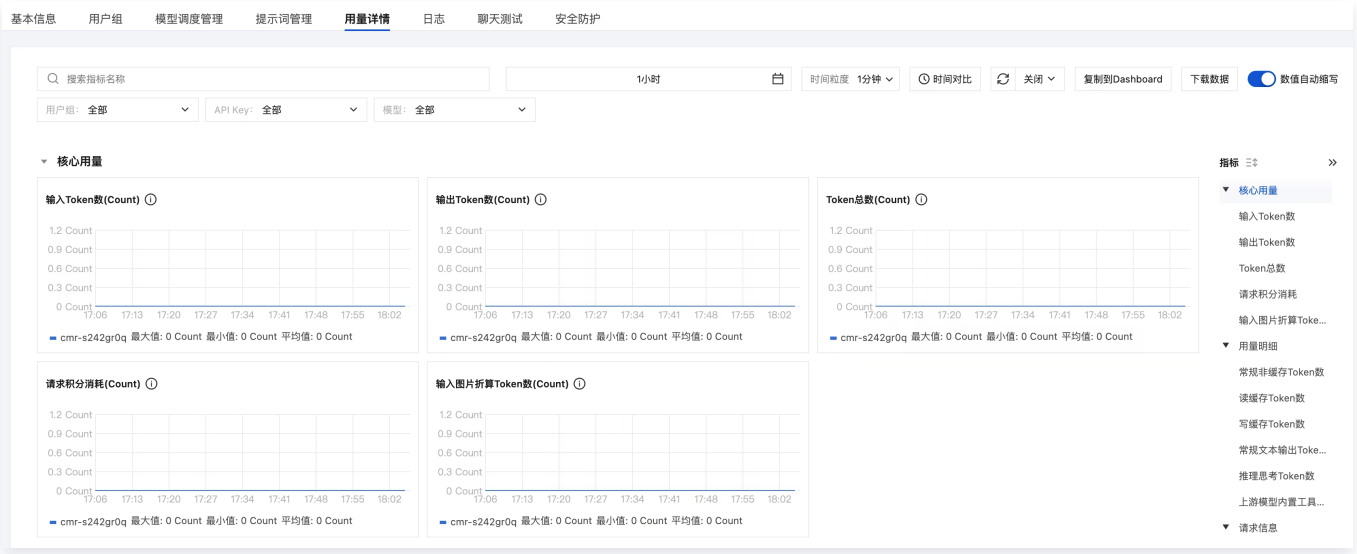
前提条件

1. 已创建 CMR 模型路由实例，详细操作指导请参见 模型路由实例管理 。
2. 已创建模型路由访问密钥，详细操作指导请参见 模型路由访问密钥 。
3. 已创建 BYOK 实例，详细操作指导请参见 创建 BYOK 实例 。
4. 如需控制积分消耗或调用速率，请提前创建对应的积分预算，详细操作指导请参见 配置积分预算 。
5. 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。

操作步骤

步骤1：查看用量详情

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 Token 消耗、延迟等当前状态，满足即时观测需求。



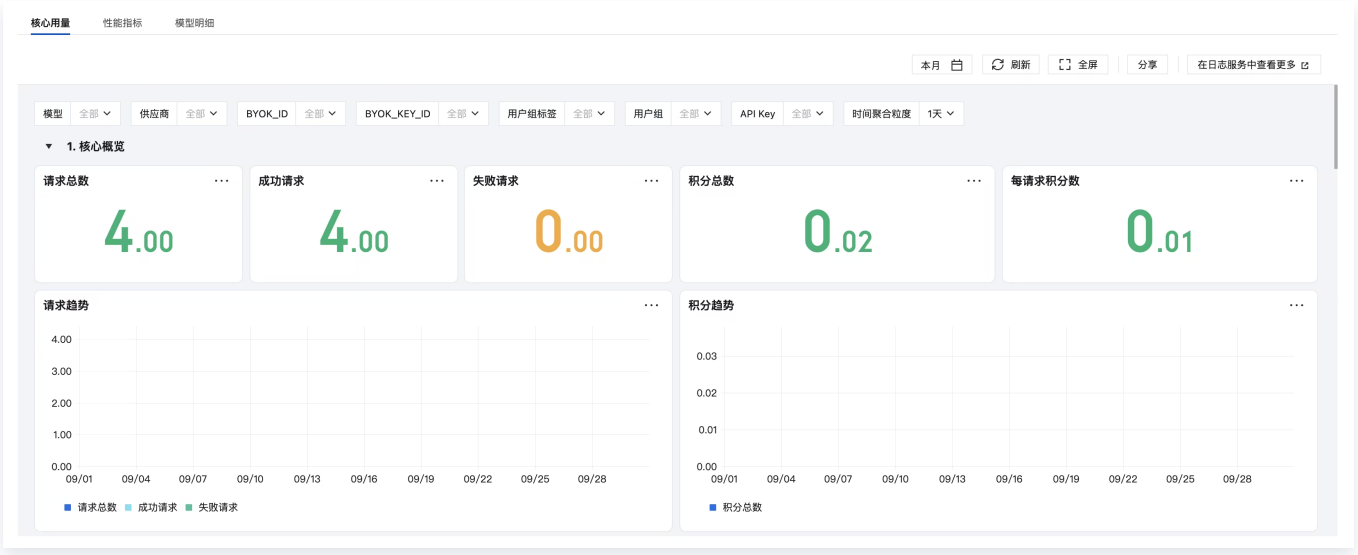
步骤2：配置 CLS 日志投递

1. 在实例的**实例管理**页面，切换至**日志**页签。
2. 在模型路由日志处，单击**立即启用**。

3. 选择已有日志主题或创建日志主题，完成日志集和日志主题配置。

步骤3：查看多维仪表盘

1. 开启**日志服务 CLS** 后，控制台基于该日志主题自动创建**仪表盘**，包括以下三个维度：
   - **核心用量**：请求用量、积分用量、Token 用量。
   - **性能指标**：延迟指标、异常指标。
   - **模型明细**：模型调用明细、模型请求明细。



步骤4：闭环优化

1. 定期查看看板，针对缓存命中率低、长尾延迟高、某业务线用量突增等结论进行管理策略优化。

2. 与**积分预算**的费控能力联动，构成监控与费控的完整成本治理闭环。

相关文档

- 创建模型路由实例
- 创建模型路由访问密钥
- 创建 BYOK 实例
- 配置模型调度管理
- 查看日志