

# 注册配置治理

## AI 网关



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

AI 网关

AI 网关概述

# AI 网关

## AI 网关概述

最近更新时间：2026-01-14 14:43:21

**AI 网关**是腾讯云智能网关推出的面向大模型与智能化场景的新一代网关产品。它专注于解决企业接入、调度和管理多种 AI 模型时面临的协议复杂、治理困难、成本不可控及存量业务改造门槛高等核心问题。

AI 网关作为企业智能化架构的流量入口与治理中枢，通过统一的协议适配、智能的路由调度与全方位的可观测能力，帮助企业高效、安全、经济地集成与使用 AI 能力，加速业务创新与智能化转型。

### ❗ 说明：

目前 AI 网关处于内测阶段。如果您有新购需求，请 [提交工单](#) 处理。

## 产品特点

- **智能模型治理**：统一接入并智能调度腾讯云混元、开源模型及第三方商业模型，通过负载均衡、熔断降级与成本优化策略，实现性能、成本与稳定性的最佳平衡。
- **存量业务快速 AI 化**：内置强大的协议转换引擎，支持 MCP、OpenAI 等 AI 生态协议与传统 HTTP/gRPC 等业务协议的双向转换，助力存量业务系统快速具备 AI 能力，有效保护企业现有 IT 投资。
- **全链路安全合规**：构建从接入认证、参数过滤到数据脱敏的多层次安全防护体系，集成 WAF、DDoS 防护等能力，保障 AI 应用合规、安全、可靠地运行。
- **企业级高可用保障**：采用多可用区高可用部署架构，支持自动故障转移与实例弹性扩缩容，服务可用性有保障。

## 业务场景

### 场景一：多模型统一治理与智能调度

- **核心问题**：企业需同时使用多个 AI 模型，但缺乏统一平台，导致模型使用混乱、资源配置低效、成本不可控。
- **解决方案**：AI 网关作为统一入口，实现多模型的可视化接入与管理。通过智能路由策略，根据请求内容、模型性能、成本等因素自动选择最优模型，并配合限流、熔断机制保障服务稳定性，实现降本增效。



场景二：存量业务系统 AI 化改造

- **核心问题**：传统企业存量系统技术栈陈旧，直接改造投入大、风险高，无法适应现代 AI 应用调用协议。
- **解决方案**：AI 网关通过协议转换引擎，将存量业务系统提供的标准 API 自动包装成 AI 应用可调用的标准化工具（MCP Tool），实现业务能力的“零代码改造” AI 化。开发人员无需关注底层集成细节，即可快速构建智能应用。



## 功能特性

功能点	说明
统一协议接入	100%兼容开源网关生态，并全面适配 AI 领域标准协议，支持 MCP、OpenAI 、SSE 等，提供传统 RESTful、gRPC 等协议的无缝转换，实现一套网关覆盖所有流量。
模型服务管理	提供模型服务的全生命周期管理，支持配置多模型供应商的密钥、API 端点。提供模型级流量控制、Fallback 容灾与精细化监控。
智能路由与编排	支持基于内容语义、成本、性能等策略的智能路由。可编排串联多个模型调用或业务 API，完成复杂任务。
精细化流量治理	提供从消费者、API 到模型等多个维度的限流、熔断、降级能力，保障后端服务与模型 API 的稳定性。
开箱即用的安全防护	集成认证鉴权、访问控制、敏感信息脱敏、防重放攻击等安全能力，提供企业级安全保障，满足合规要求。
全链路可观测性	提供从用户请求到模型响应的全链路追踪，监控 API 调用延迟、Token 消耗、模型费用等多维度指标，并支持智能诊断与告警，助力运维与成本优化。
细粒度的权限管理	通过消费者、消费者组的多级权限模型，实现 AI 能力在不同团队、项目间的安全隔离与便捷共享，支撑平台化运营。