

ZENTRIX GATEWAY

企业 AI 产品文档

配置 · 使用 · 结果核对

Zentrix Gateway 概览与快速开始

在能力中心查找已获授权的模型和工具，获取 API 密钥与接入配置，完成首次调用并核对调用记录。

操作与使用 | 权限与边界 | 结果核对

文档版本

2026.10

核对日期

2026 年 10 月 14 日

目录

1. 产品范围与入口

2. 选择模型

3. 获取和管理 API 密钥

4. 发起首次调用

5. 使用 MCP 工具与 API 服务

6. 技能、插件与外部智能体

7. 错误与限制

8. 查看记录与提交问题

9. 管理员首次开放能力

> 2026.10 · 适用于能力中心与网关调用接口；管理配置见《Zentrix Gateway 管理控制台操作指南》，完整说明见《Zentrix Gateway 用户手册》。

> 页面名称按当前源码核对；角色、许可和部署决定实际可用范围。正式操作前，以所在环境显示的配置与状态为准。

1. 产品范围与入口

Zentrix Gateway 在企业内部统一接入模型与工具能力，并在网关上执行认证、配额、路由、护栏和审计。它可以部署在客户环境内，也可以按 SaaS 方式由服务提供方托管，托管区域、套餐与服务等级以服务方案为准；模型供应商、身份源、通知渠道等外部依赖由客户配置。

平台有两个使用面：

使用面	用途	使用者
能力中心	浏览模型、MCP、API 服务、技能、插件和外部智能体目录，获取接入配置；在「首页」管理本人的 API 密钥、用量与账户安全	成员、业务系统接入人员
管理控制台	配置组织与权限、能力接入、护栏、流量规则、告警、审计和系统设置	管理员

能力中心顶部页签如下，各目录页未登录也可以打开：

页签	地址	内容
首页	`/settings`	本人的「使用情况」「模型用量」「调用日志」「模型服务」「API 密钥」「凭据管理」与个人设置；登录后显示
模型广场	`/models`	已发布的模型，含说明、计费、运行表现和调用示例
MCP 广场	`/mcp`	已发布的 MCP 服务及其工具、接入配置
API 广场	`/http-apis`	已发布的 API 服务
知识库	`/knowledge-bases`	许可证包含知识库能力且部署已安装时显示
技能广场	`/skill`	已发布的技能
插件广场	`/plugin`	已发布的插件包
外部智能体广场	`/agents/catalog`	已发布的外部智能体

目录中出现哪些条目由管理员的发布设置决定，能否调用由可用范围及其有效期决定。看到条目不代表已经获得使用权限；未获授权的 MCP 与 API 服务显示「需要管理员分发」。右上角账号菜单中的「管理控制台」只对至少能看到一个管理菜单的账号显示。

2. 选择模型

1. 打开「模型广场」，用「搜索模型」按名称搜索，或在左侧按「分类」「模型厂商」筛选，按「最新上架」「最多使用」排序。登录后列表只显示当前账号可使用的模型。
2. 点击卡片进入模型详情，页头显示服务提供方、上下文、模态、用户计费价和上架时间。
3. 按需选择「在 Playground 中打开」试用对话模型。试用产生真实调用和费用，在用量中记为“试跑”。

详情区域	读法
「首次调用 API」	Base URL、四种协议的代码示例和客户端环境变量，见第 4 节
「计费」	当前输入、输出、缓存读取、缓存写入的计费价（每百万 tokens）与价格历史；官方参考价不代表本平台扣费，实际扣费以账单为准
「运行表现」	吞吐、延迟、首字延迟、缓存命中率、工具调用与结构化输出错误率；「请求成功率」按小时显示近 3 天
「能力评测」	第三方评测得分，不代替本平台实测
「调用用量」	近 14 日的 token 用量及调用来源

3. 获取和管理 API 密钥

每把 API 密钥只绑定一个模型服务，可调用该服务下全部已启用模型，用量计入创建时选择的「API 密钥归属」。供应商的上游密钥由管理员维护，业务系统只使用网关 API 密钥。

入口	操作	有效期与原文
模型详情 → 「获取 API 密钥」	选择「API 密钥归属」，可选填「指定 API 密钥（可选）」，选择「创建」	有效期 90 天，名称自动生成；新 Key 出现在对话框下方的「API 密钥列表」中
「首页 → API 密钥」 → 「创建 key」	填写「名称」，选择「模型服务」和计费归属，可选填「指定 API 密钥（可选）」，选择「确认」	不设到期时间；结果页显示「API 密钥（仅本次可见）」

- 「API 密钥归属」的选项来自管理员的分发，例如公共、部门、职能组或个人专属；只有一个可用归属时自动选中。
- 「指定 API 密钥（可选）」用于导入业务系统已在使用的 Key，须为 32～512 位，仅含字母、数字和 `-._~+/=`。留空则自动生成。
- 创建后立即把 Key 保存到业务系统的密钥管理中，不放入截图、聊天记录或代码仓库。
- 租户在安全策略中关闭 Token 访问时无法签发 API 密钥，页面提示联系管理员。

不再使用或怀疑泄露的 Key，在「首页 → API 密钥」中选择「撤销」，使用该 Key 的调用立即失效。成员不能轮换 Key；需要保留调用连续性时，请管理员在「API 密钥管理」中执行「轮换密钥」，旧密钥在宽限期内仍可用。详见《Zentrix Gateway 用户手册》第 8 章、第 9 章。

4. 发起首次调用

Base URL 由网关的公开访问地址加固定前缀组成，模型详情的「首次调用 API」可直接复制当前环境的地址。下表中 `<网关地址>` 指该访问地址：

协议	Base URL	鉴权头
OpenAI	`<网关地址>/api/v1/agent-gateway/openai/v1`	`Authorization: Bearer <API 密钥>`
Anthropic	`<网关地址>/api/v1/agent-gateway/anthropic`	`x-api-key: <API 密钥>`，也接受 `Authorization: Bearer`
Gemini	`<网关地址>/api/v1/agent-gateway/gemini`	`x-goog-api-key: <API 密钥>` 或查询参数 `?key=`，也接受 `Authorization: Bearer`

1. 在「首次调用 API」中选择协议（OpenAI Chat Completions、OpenAI Responses、Anthropic Messages、Gemini generateContent）和写法（TypeScript、Python、cURL）。
2. 请求中的模型名填写模型广场或「列出可用模型」（`GET /models`）返回的模型名。模型不属于这把 Key 绑定的模型服务时返回 404。
3. 把 Key 注入运行环境，先发送一个最小请求，核对返回格式与响应头 `x-request-id`；再验证流式、工具调用、图像等实际使用的特性。
4. 同一请求不要同时携带两种取值不同的凭据，否则返回 400。

开发工具按协议使用以下环境变量：

工具	协议	环境变量
OpenAI SDK	OpenAI Chat Completions	`OPENAI_BASE_URL`、`OPENAI_API_KEY`
Codex	OpenAI Responses	`OPENAI_BASE_URL`、`OPENAI_API_KEY`，另附 `~/.codex/config.toml` 配置
Claude Code	Anthropic Messages	`ANTHROPIC_BASE_URL`、`ANTHROPIC_API_KEY`
Gemini CLI	Gemini generateContent	`GOOGLE_GEMINI_BASE_URL`、`GEMINI_API_KEY`

对话类接口在四种协议之间自动转换，调用方可以用任一协议调用任一对话模型；嵌入、重排序、图像、语音和批量判断只提供 OpenAI 兼容格式。流式输出期间网关每 15 秒发送一次心跳，中间代理的空闲超时应大于 15 秒。接口清单、请求体上限和会话归属请求头见《Zentrix Gateway 用户手册》第 10 章。

浏览器中能试用不代表业务服务器可达，业务侧的 DNS、代理、TLS 和出网配置需要单独验证。复制的示例可能已带上本次会话创建的 Key，分享前移除凭据。

5. 使用 MCP 工具与 API 服务

MCP 服务：

1. 打开「MCP 广场」，进入服务详情，阅读工具说明和数据边界。
2. 在「服务配置」中选择「获取连接配置」。服务要求个人凭据时，先选择「新建应用凭据」填写个人凭据，或选择「完成授权」用个人外部账号完成 OAuth 授权。
3. 在「令牌有效期」中选择有效期，选择「确认签发 Token」。可选项包括租户默认有效期、24 小时、48 小时、7 天、30 天和自定义，均受租户允许的最长有效期限制。管理员可在「全局设置 → 账号与会话安全 → 安全策略 → 外部访问」将「最长令牌有效期」设为最多 720 小时（30 天），或选择「永久有效」。允许永久有效后，MCP、HTTP API、A2A 服务的签发界面可选择「永久有效」；此类令牌没有到期时间，仍可被吊销。
4. 按客户端要求复制 Streamable HTTP 地址或 JSON 配置。Token 只显示一次，代表当前成员身份，仅限个人使用。
5. 租户开启 OAuth 时，也可选择「显示 MCP 配置」，由 MCP 客户端完成登录授权。授权页面列出回调地址和访问范围，核对后选择「允许访问」。
6. 用一次范围明确的只读调用验证结果。

工具能否调用还受管理员在「路由策略」中为 MCP 工具设置的规则限制，命中拒绝规则的调用会被拒绝。

API 服务：在「API 广场」页签打开服务详情，选择「生成 Token」签发短期访问令牌。面板给出代理地址、请求路径、请求体、访问令牌 和 curl 示例；调用地址形如 `<网关地址>/api/v1/agent-gateway/proxy/<服务 ID>/<上游路径>`，请求头携带 `Authorization: Bearer <访问令牌>`。不再使用时选择「撤销 Token」。

详见《Zentrix Gateway 用户手册》第 11 章。

6. 技能、插件与外部智能体

目录	操作	说明
技能广场	打开技能详情，选择「下载 Skill 包」，或复制安装链接与安装命令	技能在导入时已经过安装前检查；技能依赖的工具仍需单独授权
插件	打开插件详情，选择「复制个人插件源链接」用于安装	插件包不含凭据，目标端需要自行配置

目录	操作	说明
外部智能体广场	原生支持外部智能体协议的客户端：选择「令牌有效期」后「获取外部智能体配置」；MCP 客户端：选择「获取 MCP 配置」，把 `mcpServers` 条目合并到客户端配置文件	「获取 MCP 配置」只在 Agent 已发布为 MCP 工具时显示；不再使用时选择「撤销 Token」

只有在可用范围内的成员才能看到并使用这些条目。详见《Zentrix Gateway 用户手册》第 12 章。

7. 错误与限制

网关按调用方使用的协议返回错误。业务系统应按错误体中的 `reason` 处理，不解析 `message` 文本；响应头 `x-should-retry` 表示重试是否可能改变结果，可重试时可能带 `Retry-After`。

reason	HTTP 状态	处理
`UNAUTHENTICATED`	401	检查 Key 状态、到期时间与鉴权头
`PERMISSION_DENIED`	403	请管理员核对可用范围与授权有效期
`CONTENT_POLICY_BLOCKED`	403	内容被护栏拦截，调整内容后重试
`LLM_MODEL_NOT_AVAILABLE`	404	模型不存在或不在本 Key 的模型服务中
`MODEL_ENDPOINT_UNSUPPORTED`	400	模型不支持所调用的接口
`POOL_EXHAUSTED`	429	额度池用尽，错误信息说明对象、周期与恢复时间
`TRAFFIC_RATE_LIMITED`、`TRAFFIC_CONCURRENCY_LIMITED`	429	命中流量规则的限速或并发上限，稍后重试
`TRAFFIC_CIRCUIT_OPEN`	503	流量保护熔断打开，稍后重试
`LLM_RATE_LIMITED`、`LLM_UPSTREAM_SERVER_ERROR`、`LLM_OVERLOADED`	429、502、503	上游限流、错误或过载，稍后重试

同一个 HTTP 状态可能来自不同层，例如 503 可能是熔断，也可能是上游过载，以 `reason` 为准。额度按 UTC 自然周期重置，“今日”从北京时间 8 点开始。完整错误码见《Zentrix Gateway 用户手册》第 10.8 节。

8. 查看记录与提交问题

成员在「首页」左侧「用量」分组中查看本人的数据：

- 「使用情况」：最近 180 天的调用次数、消息数、累计 Token 与活跃度；

- 「模型用量」：按 API 密钥、模型服务、模型和计费归属统计消费；
- 「调用日志」：逐次核对模型和 MCP 调用，可按 Request ID 搜索，选择「导出记录」下载 CSV。缺少价格的调用显示“费用缺失”。

记录缺失时，依次检查时间范围、所用的 Key 与模型服务，以及请求是否绕过网关直连上游。

向管理员反馈问题时，提供发生时间和时区、模型或工具名称、所用协议与客户端、`reason` 与提示信息、调用日志中的 Request ID（与响应头 `x-request-id` 一致），以及能否稳定复现。不要发送完整的 Key、Token 或无关的业务内容。

9. 管理员首次开放能力

1. 在「系统设置 → 全局设置 →「授权许可」页签」确认已开通的模块，在「组织与权限」中建立部门、职能组和成员，在「角色管理」中指定系统管理员、安全管理员和审计员。
2. 在「模型供应商」接入并验证上游，在「模型服务」中选择模型，在「可用范围与配额」中先只授权试点范围。
3. 按需接入「MCP 服务」「API 服务」「外部智能体」「技能」「插件」，发布后设置可用范围。
4. 开通模型护栏和动态路由时，在「护栏」与「流量规则」中配置检测和路由，用「验证规则」核对命中顺序。
5. 由试点成员使用本人 API 密钥完成正常、无权限和额度耗尽等场景，在「概览」的「网关请求」页签与「操作审计」中核对结果，再扩大范围。

完整步骤见《Zentrix Gateway 管理控制台操作指南》和《Zentrix Gateway 用户手册》第 1 章；流量规则的字段与默认值见《Zentrix Gateway 路由策略用户手册》。