



ZENTRIX GATEWAY

企业 AI 产品文档

配置 · 使用 · 结果核对

Zentrix Gateway 用户手册

从首次配置到日常使用、安全治理、问题处理和平台维护的完整操作说明。

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. Zentrix 概览与配置流程
2. 平台对象和边界
3. 组织与权限
4. 配置模型供应商
5. 模型定价与分类
6. 创建模型服务
7. 配置模型服务的可用范围与配额
8. 成员使用模型
9. 调用方、API 密钥、认证与授权
10. 调用接口
11. 配置并分发 MCP 服务与 API 服务
12. 外部智能体、技能、插件与外部注册中心
13. 配置敏感数据（PII）检测
13.6 检测状态与总开关
14. 配置内容检测与安全围栏
15. 路由策略与流量保护
16. 请求变换、网关扩展与开放 API
17. 查看用量并控制成本
18. 用量与成本分析
19. 可观测入口、数据来源与权限边界
20. 网关请求、会话与根因定位
21. 问题诊断与处理

22. 告警、通知、留存、导出与操作审计

23. 平台运行与日常维护

24. 常用配置的维护与变更

25. 分阶段上线

26. 上线检查、稳定运行与管理交接

27. 联系支持所需信息

附录 A. 角色与入口速查

附录 B. 常用任务索引

版本说明：本手册对应 2026.10 版本，内容按代码基线 main 461d96ebd9 核对。界面截图取自 2026-10-15 的演示环境，数据为虚构示例。实际可见的菜单与功能取决于交付版本、许可证包含的模块和账号权限。

本版未在测试环境完成业务端到端验收。正式操作前，以所在环境显示的菜单、配置与状态为准。

1. Zentrix 概览与配置流程

1.1 产品范围

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。管理员在管理控制台维护组织与权限，接入模型供应商、模型服务、MCP 服务、API 服务、外部智能体、技能和插件，配置流量规则、护栏、告警与通知，再通过概览、会话诊断和操作审计检查运行结果。成员在能力中心查找已获授权的能力，在「首页」或模型详情中创建 API 密钥，查看本人的用量与调用日志。

Zentrix 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、对象存储等外部依赖由客户配置和提供；要求数据不出客户环境时，这些依赖也需要满足同样的数据边界。

平台包含两个使用面：

使用面	主要用途	常见使用者
能力中心	浏览模型、MCP、API 服务、技能、插件和外部智能体目录并获取接入配置；在「首页」管理本人的 API 密钥、凭据、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	租户管理员、系统管理员、安全管理员、审计员、运营、财务、平台管理员

功能按许可可分为基础功能和四个加购模块。许可证未包含某个加购模块时，对应菜单不显示，相关接口拒绝访问。已导入的许可模块可在「管理控制台 → 系统设置 → 授权许可」中查看。

许可模块	对应的界面
基础功能	1.2 节菜单表中未标注加购模块的全部菜单
模型护栏	「护栏」菜单：敏感数据、词库、问答、内部系统白名单
动态路由	「流量规则」菜单：规则清单、全局限额、意图集、缓存和验证规则；规则中的内容检测动作也在这里配置
全链路审计	「操作审计」菜单
FinOps	「概览 → 网关总览」中读取用量分析数据的面板（运营指标、Token、缓存、耗时、热力、调用流向与费用排行等）；没有独立菜单。「授权许可」页上显示为「用量与成本」

知识库入口另有前提：许可证包含知识库能力，且部署时安装了知识库。任一条件不满足时，能力中心的「知识库」页签和管理控制台的「知识库管理」菜单都不显示。

1.2 首页与导航

能力中心

成员登录后进入能力中心的「首页」。页面左上角显示产品名称和「能力中心」，顶部页签依次为：

页签	内容	显示条件
首页	本人的用量、调用日志、模型服务、API 密钥、凭据与个人设置	登录后显示
模型广场	已发布的模型，含说明、价格、接入地址和代码示例	始终显示
MCP 广场	已发布的 MCP 服务及其工具、接入配置	始终显示
API 广场	已发布的 API 服务	始终显示
知识库	已发布的知识库	满足 1.1 节所述知识库前提时显示
技能广场	已发布的技能	始终显示
插件广场	已发布的插件包	始终显示
外部智能体广场	已发布的外部智能体	始终显示

各目录页在未登录时也可以打开。目录中出现哪些条目由管理员的发布设置决定，成员能否调用由可用范围决定。

「首页」左侧导航分为三组：

分组	菜单	用途
用量	使用情况、模型用量、调用日志	查看本人的用量概览、按模型统计的用量和逐次调用记录
资源	模型服务、API 密钥、凭据管理	查看可用的模型服务和接入地址；创建、撤销 API 密钥；管理被模型服务引用的第三方凭据
个人设置	个人资料、偏好设置	维护显示名称、头像、邮箱、密码、两步验证、通行密钥，查看设备与登录历史；设置语言与主题

右上角账号菜单显示本人的显示名称和邮箱，下设「管理控制台」和「退出登录」。只有账号至少能看到一个管理菜单时，才显示「管理控制台」。

管理控制台

进入「管理控制台」后，侧边栏顶部显示「管理后台」和当前租户名称，其下的「返回工作台」用于回到能力中心「首页」。侧边栏按分组列出菜单，内容区顶部以「分组 > 菜单」显示当前位置。

租户管理员在 Zentrix 中看到的完整菜单如下（顺序与界面一致）：

分组	菜单	主要用途	加购模块或其他前提
概览	概览	页签：网关总览、网关请求、对话、路由策略命中、网关告警	网关总览中的用量分析面板需 FinOps
概览	会话诊断	自动识别会话中的问题，给出原因、处理建议和证据	仅租户管理员可见
搭建	插件	导入、导出和分发插件包	—
搭建	技能	导入、扫描和分发技能	—
搭建	MCP 服务	接入 MCP 服务，维护工具、凭据、权限和可用范围；页签：MCP 列表、Token 管理、分类管理	—
搭建	API 服务	以 HTTP 反向代理方式对外提供上游接口	—
搭建	外部智能体	接入外部智能体；页签：Agent 目录、任务中心	—
搭建	知识库管理	维护知识源与检索资产	知识库前提（见 1.1 节）
搭建	模型服务	配置服务包含的模型，设置可用范围与配额；页签：模型服务、API 密钥管理	—
搭建	模型供应商	接入上游模型并维护价格与分类；页签：模型供应商、模型&定价、分类管理	—
搭建	外部注册中心	从 Nacos 同步外部智能体与 MCP Server	—
组织与权限	SSO 配置	接入钉钉、飞书、企业微信、企业 IAM、LDAP 与目录推送	—
组织与权限	账号与会话安全	查看并撤销登录会话，设置安全策略	—
组织与权限	成员与部门	维护成员、部门和已离职成员	—
组织与权限	角色管理	为六个内置角色指定成员	仅租户管理员可见
组织与权限	职能组	建立跨部门成员集合，用于分发模型服务等资源	—

分组	菜单	主要用途	加购模块或其他前提
护栏	护栏	页签：敏感数据、词库、问答、内部系统白名单	模型护栏
护栏	威胁情报	接入恶意来源库，管理同步状态与命中动作	—
路由策略	流量规则	页签：规则清单、全局限额、意图集、缓存；另有「验证规则」页面	动态路由
路由策略	网关扩展	浏览、上传、安装和绑定 WASM 网关扩展	—
运行与排障	告警规则	页签：告警规则、通知渠道	—
审计与合规	操作审计	查询配置变更与访问记录，校验哈希链，导出审计证据	全链路审计
审计与合规	报表导出	导出调用与用量数据，配置定时导出	—
系统设置	授权许可	查看当前许可，导入签名许可证	—
系统设置	通知横幅	在页面顶部发布本租户公告	—
系统设置	品牌	设置产品显示名、Logo 与网站图标	—
系统设置	SMTP 配置	配置系统邮件的发送服务器	—
系统设置	全局设置	为内容安全、敏感信息识别、意图识别、问答语义匹配等功能指定默认模型	—

平台管理员（安装时创建的 root 账号和平台账号）使用平台范围的管理控制台，进入后默认打开「租户管理」。平台范围不显示上表中只属于租户的菜单（概览、会话诊断、授权许可、角色管理等），显示的菜单为：

分组	菜单	主要用途
组织与权限	SSO 配置	配置平台账号使用的「平台单点登录」（OpenID Connect）
组织与权限	账号与会话安全	页签：登录会话、管理面访问网段
组织与权限	平台账号	为平台管理人员创建个人账号，管理激活、停用与平台管理员授权
运行与排障	告警规则	页签：活跃告警、规则目录、通知渠道

分组	菜单	主要用途
运行与排障	系统运行监控、出口网络	查看平台各组成部分的运行状态；诊断平台访问公网的质量
系统设置	通知横幅、系统更新、备份与恢复	发布对所有租户生效的公告；升级平台；备份与恢复
平台扩展	部署配置、SSRF 出站防护、网关、租户管理	访问地址、对象存储、运维通知渠道、链路导出、资源容量、入口监听、数据依赖与最近一次验收；出站访问防护；公开访问地址、入口证书、网关节点、过载保护与出口判定；租户生命周期

菜单是否出现由三个条件共同决定：

1. 账号的管理范围：租户管理员只看到租户菜单，平台管理员只看到平台菜单。
2. 账号持有的角色：例如审计员只看到审计类菜单，安全管理员看到护栏、流量规则、SSO 配置等凭据与安全类菜单。角色与菜单的对应关系见 [3.4 角色管理](#)。
3. 许可证包含的模块：见 1.1 节许可模块表。

看不到某个菜单时，按以上顺序核对账号范围、角色和许可模块。直接访问未授权页面的地址时，页面显示无权限或不存在。

1.3 推荐配置顺序

1. 在「系统设置 → 授权许可」导入许可证，确认「授权用户数」「到期时间」和「高级模块」中已开通的模块。
2. 在「组织与权限」中建立部门、职能组和成员；需要企业身份登录时配置「SSO 配置」；在「角色管理」中为系统管理员、安全管理员、审计员等角色指定专人；在「账号与会话安全 → 安全策略」中设置会话时长、登录方式和密码策略。
3. 在「模型供应商」接入上游并验证模型，在「模型服务」中选择模型、设置可用范围与配额；在「全局设置」中为内容安全、敏感信息识别、意图识别等功能指定默认模型。
4. 按需接入「MCP 服务」「API 服务」「外部智能体」「技能」「插件」，或通过「外部注册中心」同步外部条目，再设置可用范围。
5. 开通模型护栏和动态路由模块时，在「护栏」维护敏感数据规则、词库和问答，在「流量规则」中配置路由、内容检测、全局限额和缓存；按需配置「威胁情报」和「网关扩展」。
6. 在「告警规则」中配置通知渠道和告警规则；使用邮件渠道时先完成「SMTP 配置」。
7. 用「流量规则」的「验证规则」模拟请求，确认命中顺序和最终目标；再由试点成员使用自己的 API 密钥完成正常、拒绝、额度耗尽和异常输入测试。
8. 在「概览」的「网关请求」「对话」「路由策略命中」页签和「会话诊断」中核对试点调用，在「操作审计」中核对配置变更，确认无误后逐步扩大可用范围。
9. 平台管理员在平台范围内确认「网关」的公开访问地址与入口证书、「备份与恢复」计划和「系统更新」方式。

1.4 主要管理职责

职责	主要对象	结果验证入口
组织与权限	部门、成员、职能组、角色、SSO 身份源、安全策略	成员与部门、角色管理、账号与会话安全的会话列表、操作审计
模型与网关	模型供应商、模型服务、API 密钥、流量规则、全局限额、缓存、网关扩展	模型服务试跑、验证规则、概览的网关请求与路由策略命中
能力接入	MCP 服务、API 服务、外部智能体、技能、插件、外部注册中心	各服务的连通性探查与调用监控、能力中心目录、操作审计
安全	护栏（敏感数据、词库、问答、内部系统白名单）、流量规则中的内容检测、威胁情报	概览的路由策略命中与网关请求、告警规则
运行与审计	告警规则、通知渠道、会话诊断、操作审计、报表导出	概览的网关告警页签、会话诊断问题列表、审计证据导出
平台维护（平台管理员）	网关、部署配置、系统更新、备份与恢复、租户管理	系统运行监控、出口网络、备份与恢复的任务记录

1.5 配置完成标准

- 试点成员能在能力中心看到获授权的能力，并使用本人 API 密钥调用成功；可用范围以外的账号调用被拒绝。
- 模型供应商、MCP 服务、API 服务和外部智能体均通过连通性验证，并在「概览 → 网关请求」中产生可查询的真实调用记录。
- 「验证规则」的模拟结果与真实调用的路由结论一致；「流量规则」的规则清单没有显示被数据面拒绝的规则。
- 配额、全局限额、敏感数据规则和 content 检测的允许与拒绝样例均符合预期，并能在请求详情的「命中记录」中查到；路由规则的命中同时在「路由策略命中」页签中可见。
- 系统管理员、安全管理员、审计员由不同人员担任；配置变更能在「操作审计」中查到操作人和时间。
- 通知渠道发送测试成功，告警接收人、备份与恢复责任人和支持联系人已经明确。

2. 平台对象和边界

Zentrix 的配置对象分为组织、能力接入、流量与安全、运行与审计四类，全部在管理控制台中维护；成员和业务系统按可用范围使用其中已发布的能力。下表只列出 Zentrix 中可见的对象，「入口」列为管理控制台中的菜单。

组织与权限

对象	用途	入口	配置建议
租户	一套独立的组织、数据与权限范围	平台范围「租户管理」	每个企业或独立组织一个租户；租户之间的数据和授权互不共享

对象	用途	入口	配置建议
部门	成员在正式组织架构中的归属，可设负责人和上级部门	「成员与部门」	按真实汇报关系建立；接入通讯录同步后由来源系统维护
职能组	跨部门的成员集合，用于分发模型服务等资源	「职能组」	按职责或项目建立，不替代部门树
成员	可以登录并使用平台的个人账号	「成员与部门」	离职人员移入「已离职成员」，其 API 密钥与登录凭据随即失效
角色	六个内置角色，决定成员能进入哪些管理菜单	「角色管理」	系统管理员、安全管理员、审计员由不同人员担任
身份源	钉钉、飞书、企业微信、企业 IAM、LDAP 与目录推送	「SSO 配置」	先完成测试登录或连接自检，再启用
安全策略	会话时长、在线人数、登录方式、密码策略与外部访问令牌有效期	「账号与会话安全 → 安全策略」	上线前确定，变更后在会话列表中确认影响

能力接入

对象	用途	入口	配置建议
模型供应商	平台连接外部或自建模型端点的方式，含凭据、接口格式、模型价格与分类	「模型供应商」	每个供应商单独维护凭据；验证通过后再被模型服务引用
模型服务	向成员和业务系统提供的一组有序模型，含可用范围、配额与限流保护	「模型服务」	按用途命名，例如“通用对话”“代码助手”
API 密钥	调用模型服务的凭据，每个 Key 绑定一个模型服务	成员在能力中心「首页 → API 密钥」或模型详情中创建；管理员在「模型服务 → API 密钥管理」检索与处置	按应用或用途分别创建，不在多人之间共享
MCP 服务	以 MCP 协议对外提供的工具服务，含工具清单、凭据与可用范围，工具级访问控制通过路由策略配置	「MCP 服务」	每个服务单独维护连接与授权，临时授权设置有效期
API 服务	以 HTTP 反向代理方式对外提供的上游接口，统一鉴权与分发，不转换为 MCP 工具	「API 服务」	适用于按上游原接口调用的应用

对象	用途	入口	配置建议
外部智能体	通过 Agent Card 接入的外部智能体，含上游凭据、调用 Token 与可见范围	「外部智能体」	接入前先测试连通；任务执行情况在「任务中心」查看
技能	从代码仓库地址或压缩包导入、按范围分发的技能包	「技能」	导入后按范围分发，在能力中心「技能广场」核对
插件	由技能和远程 MCP 组成的插件包	「插件」	导出的插件包不含凭据
外部注册中心	从 Nacos 同步外部智能体与 MCP Server 的来源	「外部注册中心」	同步得到的条目仍需单独分发
知识库	企业知识源与检索资产	「知识库管理」	仅在许可证包含知识库能力且部署已安装时可见

流量与安全

对象	用途	入口	许可模块
规则	按条件匹配请求，执行放行、拒绝、转发、改写或内容检测，按优先级依次评估	「流量规则 → 规则清单」	动态路由
全局限额	按条件限制请求次数、并发，配置熔断、上游超时与临时豁免	「流量规则 → 全局限额」	动态路由
意图集	供规则按任务类型、难度或业务场景分流的意图定义，分内置意图与自定义意图集	「流量规则 → 意图集」	动态路由
语义缓存	相似请求直接返回缓存结果，按租户和模型隔离	「流量规则 → 缓存」	动态路由
敏感数据规则	识别个人信息与密钥，设置命中动作	「护栏 → 敏感数据」	模型护栏
词库	黑名单、白名单词条	「护栏 → 词库」	模型护栏
问答	问题足够相似时直接返回预设答案	「护栏 → 问答」	模型护栏
内部系统白名单	危险工具调用检测中放行的内部系统	「护栏 → 内部系统白名单」	模型护栏
恶意来源库	外部威胁情报来源及其命中动作	「威胁情报」	基础功能

对象	用途	入口	许可模块
网关扩展	在网关请求处理过程中执行的 WASM 插件，可按模型、路由或调用方绑定	「网关扩展」	基础功能

规则中的内容检测会引用护栏中维护的敏感数据规则、词库和问答。只开通动态路由时，可以配置路由、限额和缓存；需要维护这些检测依据时，还需要模型护栏模块。

运行与审计

对象	用途	入口
通知渠道	钉钉、飞书、企业微信、Webhook 和邮件等告警投递目标	「告警规则 → 通知渠道」
告警规则	按调用质量、策略命中和费用设置的阈值告警	「告警规则 → 告警规则」
会话诊断问题	从会话数据中识别出的问题、原因与证据，可标记处理状态	「会话诊断」
操作审计记录	谁在何时修改了配置、触发了动作或访问了资源	「操作审计」（全链路审计）
导出任务	一次性或定时导出的调用与用量数据	「报表导出」
通知横幅	显示在页面顶部的公告	「通知横幅」

平台范围的对象（网关节点、公开访问地址、入口证书、备份与恢复、系统更新等）只由平台管理员维护，入口见 [1.2 首页与导航](#)。

3. 组织与权限

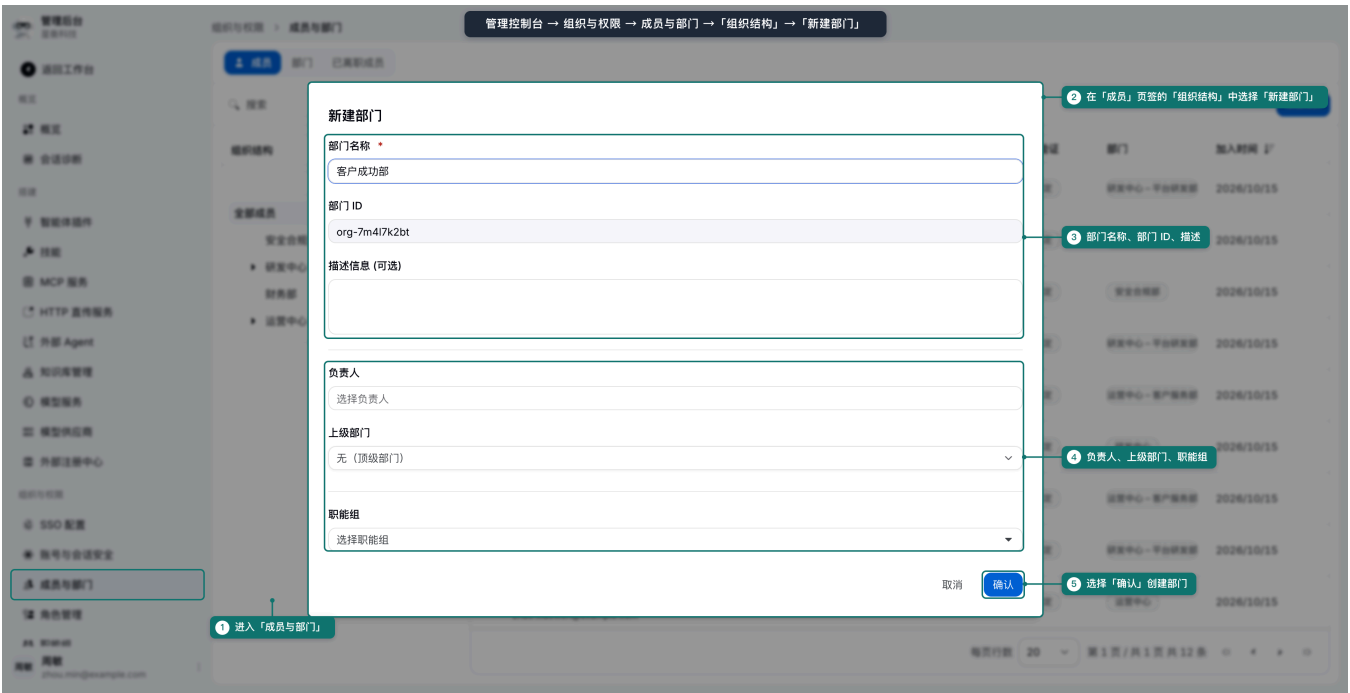
本章说明部门、成员、职能组、角色、身份源和登录安全的配置方法。组织结构应在配置模型服务和 MCP 服务的可用范围之前完成，此后即可按部门或职能组授权，不必逐人维护。

本章各菜单位于「管理控制台 → 组织与权限」，进入条件如下：

菜单	可以进入的账号
成员与部门、职能组	租户管理员、系统管理员
SSO 配置、账号与会话安全	租户管理员、安全管理员
角色管理	仅租户管理员

3.1 新建部门

- 1. 进入「管理控制台 → 组织与权限 → 成员与部门」。
- 2. 在「成员」页签左侧的「组织结构」中选择「新建部门」。
- 3. 填写「部门名称」。「部门 ID」由系统自动生成且不可修改，用于稳定识别该部门。
- 4. 按需填写「描述信息(可选)」，选择「负责人」「上级部门」和「职能组」。「上级部门」默认为「无（顶级部门）」。
- 5. 选择「确认」后，在「组织结构」中确认层级和负责人正确。



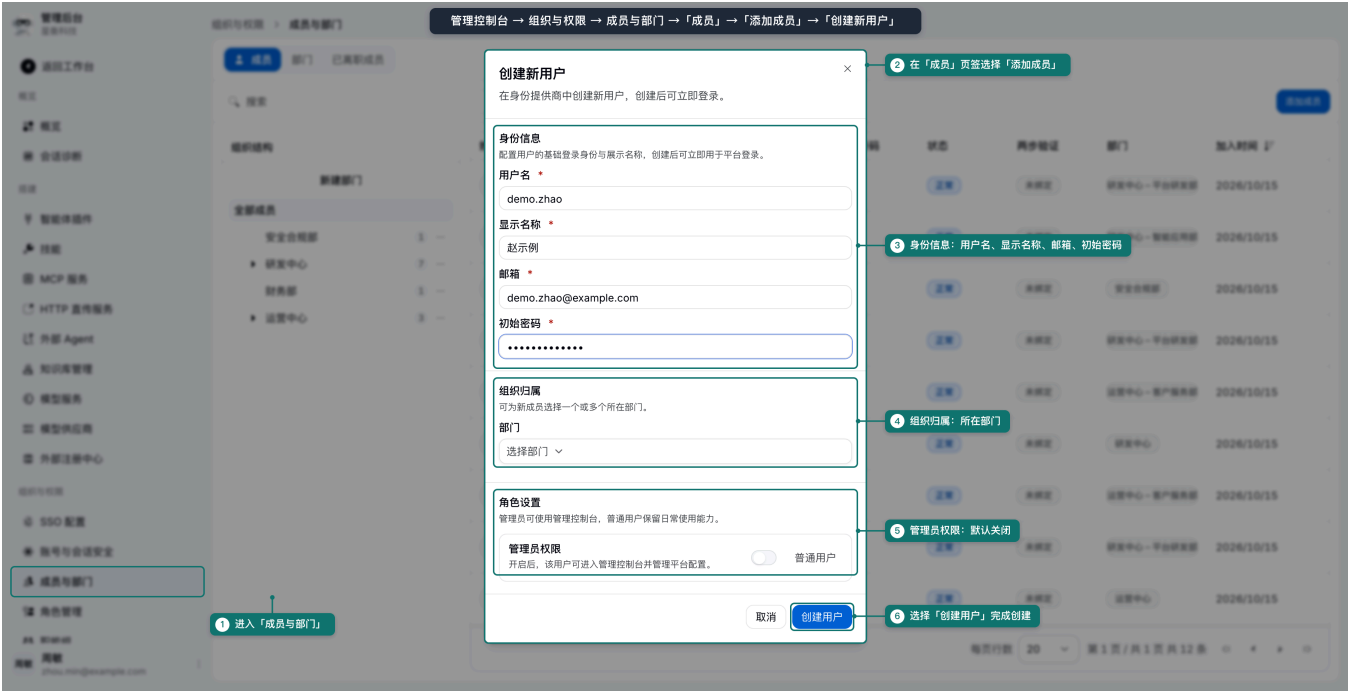
部门名称可以随组织调整而修改；部门 ID 不变，因此不要为了调整名称或层级而删除重建部门。跨部门项目优先使用职能组。

批量调整部门时，进入「部门」页签，勾选多个部门后使用「批量编辑」统一修改「上级部门」和「负责人」，或使用「批量删除」。在「部门」页签选中某个部门后，「添加成员」用于把已有成员加入该部门。

通过通讯录同步或目录推送进入平台的部门为只读，其结构与成员关系请在来源系统中维护。

3.2 添加成员

- 1. 进入「管理控制台 → 组织与权限 → 成员与部门」，在「成员」页签选择「添加成员」。
- 2. 在「创建新用户」对话框的「身份信息」中填写「用户名」「显示名称」「邮箱」和「初始密码」，四项均为必填。
- 3. 在「组织归属」中选择一个或多个「部门」。
- 4. 在「角色设置」中决定是否开启「管理员权限」。开启后该成员成为租户管理员，可以进入管理控制台并管理本租户的全部配置；关闭时为普通用户。
- 5. 选择「创建用户」。账号创建后即可登录。
- 6. 通过安全渠道把登录地址和初始密码交给本人，并要求本人登录后在能力中心「首页 → 个人设置 → 个人资料」中修改密码。



部门归属与管理权限相互独立：部门和职能组决定成员能使用哪些能力，角色决定成员能进入哪些管理菜单。只有确实需要管理整个租户的人员才开启「管理员权限」；按职责分工的管理员请在「角色管理」中授予对应角色（见 3.4 节）。

接入身份源后，成员也可以通过扫码登录、企业 IAM 自动开通账号，或通过通讯录同步、目录推送批量导入，见 3.6 节。

3.3 维护成员与已离职成员

在「成员」页签中，可按部门、关键字筛选成员；每行右侧的操作菜单提供以下操作：

操作	结果
编辑	修改成员资料
调整组织归属	修改所属部门和职能组
重置密码	为成员设置新的临时密码
重置两步验证	移除该成员的验证器、恢复码、受信任设备与通行密钥，并让其所有会话失效；对方下次登录时重新绑定
「分配 角色名」/「移除 角色名」	授予或收回账号角色，例如租户管理员权限
删除	把成员移至「已离职成员」：对方无法登录，其 API 密钥与登录凭据立即失效

限制：

- 由通讯录同步或目录推送管理的成员，资料、部门和入离职需要在来源系统中处理，操作菜单中只保留不影响来源数据的操作。
- 不能删除本租户最后一名租户管理员，也不能移除其管理员权限。

- 「重置两步验证」要求操作人在最近 10 分钟内完成过身份验证，否则先弹出「验证身份」。管理员不能重置自己的两步验证，请在本人能力中心「首页 → 个人设置 → 个人资料」中处理。
- 「已离职成员」页签列出所有已离职账号，删除操作不会物理删除账号，便于审计追溯。对已离职成员可执行：
 - 「恢复入职」：对方可以重新登录，但不会自动恢复部门、职能组和 API 密钥，需要管理员重新分配。
 - 「重新执行离职清理」：再次吊销该账号的 API 密钥、令牌和组织关系，账号仍留在已离职列表中。

3.4 角色管理

「管理控制台 → 组织与权限 → 角色管理」维护六个内置角色的成员。角色是固定编制，不能新增或删除，页面只管理“谁持有哪个角色”。租户属主（租户管理员）默认拥有全部权限；按等级保护要求部署时，建议由下列角色分工负责，不再日常使用属主账号。

角色	职责	可见的管理菜单
系统管理员	管理配置与装配，不触及凭据与审计	概览、搭建类菜单、成员与部门、职能组、网关扩展、报表导出、系统设置；不含会话诊断与角色管理
安全管理员	管理凭据与安全策略	SSO 配置、账号与会话安全、护栏、威胁情报、流量规则（其中「全局限额」「缓存」页签的接口按租户管理权限校验，由租户管理员维护）
审计员	管理审计日志与留存	操作审计
运营	值班运营：配额、观测、告警静默	告警规则
财务	只读：概览中的费用卡片，「网关请求」列表中的费用与导出；不含任何配置入口	侧边栏无菜单
只读审计	只读：请求、对话明细与导出；不能修改任何内容	侧边栏无菜单

上表「可见的管理菜单」还受许可模块限制，例如未开通全链路审计时，审计员看不到「操作审计」。仅持有财务或只读审计角色的账号，侧边栏不显示菜单，能力中心账号菜单中也不显示「管理控制台」；需要由管理员提供「概览」页面的地址后直接打开。

三员互斥：系统管理员、安全管理员、审计员三个角色两两互斥，同一账号只能持有其中一个。为已持有其中一个角色的账号再授予另一个时，系统拒绝保存。运营、财务、只读审计可以与任一角色叠加。

为角色添加成员：

1. 在页面上方选择角色，例如「安全管理员」。
2. 选择「添加成员」，在「为「安全管理员」添加成员」对话框中搜索并选择用户，确认后生效。角色只能授予单个用户，不能授予部门或职能组。
3. 需要移除时，在成员列表中选择「移除」。移除后对方立即失去该角色的全部权限。

尚未分配成员的角色可以使用「一键创建角色账号」：系统为每个空缺角色各创建一个专属账号，每个账号占用一个授权席位。创建结果中的初始密码只显示一次，请立即复制保存并交给对应负责人。

完成分配后，请用各角色账号分别登录，确认侧边栏只出现该角色应有的菜单，再在「操作审计」中核对角色授予与移除记录。MCP 工具的调用权限不由角色决定，在「路由策略 → 流量规则」中以「工具名」「组织成员」等条件配合「拒绝本次调用」动作按部门、成员、职能组配置，见第 11.5 节。

3.5 职能组

职能组是跨部门的成员集合，用于按职责分发模型服务等资源。角色管理定义成员能做什么管理操作、能进入哪些管理页面；职能组定义成员能使用哪些能力。

1. 进入「管理控制台 → 组织与权限 → 职能组」，选择「添加职能组」。
2. 填写「职能组名称」和「描述」，按需关联「模型服务」，选择「创建」。
3. 在职能组的操作菜单中选择「邀请成员」，搜索并添加成员或部门。
4. 在列表中核对「关联人数」「关联应用」「关联模型服务」。

勾选多个职能组后可以「批量编辑」或删除。删除职能组后，成员会失去仅由该职能组授予的能力。

3.6 SSO 配置

「管理控制台 → 组织与权限 → SSO 配置」用于接入企业身份源。页面首先列出「已接入的身份源」；选择「添加身份源」后，在「选择身份源」中选择以下类型之一并「开始配置」：

身份源	用途
钉钉	钉钉扫码登录、钉钉客户端内免登，以及通讯录同步
飞书	飞书扫码登录与通讯录同步
企业微信	企业微信扫码登录与通讯录同步
企业 IAM (OAuth2)	企业 OpenID Connect 或 OAuth2 单点登录
LDAP	对接 OpenLDAP 或 Active Directory，同步部门与成员
目录推送	由外部系统主动推送组织与成员

身份源启用后，登录页才会出现对应的登录入口。所有应用密钥和绑定密码加密保存，保存后不回显；编辑时留空表示保留原值。

钉钉、飞书、企业微信扫码登录

以钉钉为例，配置分四步：

1. 「准备回调」：把页面显示的回调地址登记到钉钉内部应用。
2. 「填写凭据」：填写钉钉开放平台中的「Corp ID」「AppKey」和「AppSecret」。
3. 「登录策略」：决定是否开启「首次扫码自动创建账号」（允许尚不在系统中的钉钉用户扫码后自动注册并登录），然后开启「启用钉钉登录」。启用前必须已填写 AppSecret。

4. 「组织同步」：按需把钉钉通讯录同步到部门与成员，见下文「通讯录同步」。

飞书和企业微信的步骤为「应用凭据」「登录与开户策略」「核对与启用」：填写「企业标识」「应用标识」和「应用密钥」（飞书分别为 Tenant Key 与 App ID，企业微信分别为企业 ID 与 AgentId），决定是否「允许未登记的成员自动开通账号」，自动开通的账号只能使用普通成员角色；最后把页面显示的回调地址填入服务商后台，开启「启用该身份源」。

钉钉免登：在钉钉客户端内打开 `https://<平台地址>/dingtalk-mianden?corpId=<企业 Corp ID>`，平台使用钉钉客户端提供的身份直接登录。该地址通常设置为钉钉内部应用的首页地址。不在钉钉客户端内打开、缺少 corpId 或钉钉未返回授权码时，页面提示原因并回到普通登录页。

企业 IAM 单点登录

在「企业 IAM (OAuth2)」的「选择方式」步骤中，「登入方式」有三种：

- 「OpenID Connect (推荐)」：填写「身份源地址 (Issuer)」并选择「读取配置」，平台从身份源的公开配置中读取「登录地址」与「令牌地址」；再填写「Client ID」「Client Secret」和「显示名称」。
- 「外部 OAuth2 IAM」：手工填写「授权地址 (authorize)」「Token 地址 (accessToken)」「用户信息地址 (profile)」与「用户唯一字段路径」。
- 「平台账号直接登录」：不对接外部 IAM，OAuth 授权时直接使用平台账号登录。

企业 IAM 配置启用后，同时用于外部支持 OAuth 的 Agent 经 MCP 接入时的登录。

按需设置「账号匹配字段路径(可选)」(用于匹配平台已有账号)、「允许 JIT 自动建号」(未知用户首次登录时自动创建普通用户账号)和「多因子来源」(「由本平台补做两步验证」或「以身份源的多因子认证结果为准」；选择后者时填写「允许的 acr 值」或「允许的 amr 值」)。

保存后选择「测试登录」，用身份源中的测试账号完成一次登录。测试只核对授权、令牌交换、用户信息读取和身份映射，不创建用户或登录会话。OpenID Connect 身份源只有在「验证状态」变为「已验证」后才能启用。保存后页面显示「单位登录地址」，分发给员工后，员工打开该地址即可看到本单位的登录方式。

需要让业务系统携带企业身份令牌直接调用模型时，在「业务系统调用」区域开启「允许业务系统用身份源令牌调用模型」，填写「业务系统客户端标识」，并把页面显示的「业务系统调用地址」交给业务系统开发者。调用按人员与部门计量。

平台管理员在平台范围的「SSO 配置」中只能配置「平台单点登录」，类型仅支持 OpenID Connect；只有已创建的平台账号可以通过它登录，平台不会自动建号。

LDAP / AD

LDAP 配置分为「连接目录」「检索范围」「属性映射」「启用与保存」「组织同步」五步：填写目录地址、只读绑定账号与「绑定密码」，需要时填写「私有 CA 证书 (PEM)」；填写「部门过滤器」「人员过滤器」与「分页大小」(1-1000)；在「属性映射」中映射部门与人员属性，并选择「部门成员变更模式」。保存后执行「连接自检」，成功时页面显示连接方式以及读取到的部门数和成员数。

通讯录同步

钉钉、飞书、企业微信和 LDAP 的「通讯录同步」卡片提供：

- 「立即同步」：触发一次全量对账，把授权范围内的部门、成员与成员关系同步到组织架构。
- 「自动同步」：开启「开启自动同步」，在「同步周期」中选择每 15 分钟、每小时、每 6 小时或每天，选择「保存设置」。修改周期或重新开启后，下次同步从保存时刻起算。卡片显示「最近运行」「下次同步」和连续失败次数；连续失败不会自动关闭自动同步。
- 运行状态：排队中、同步中、已完成、部分完成、待复核、已拒绝、已中止。

为防止来源系统异常导致大批账号被误停用，单次同步新增的部门、成员与成员关系合计超过 500 项，或停用成员与移除成员关系合计超过 100 项时，相应变更暂存为「待复核」，不直接生效。管理员在卡片的「Quarantine 复核」区域查看暂存变更，选择「全部批准」一次性应用本次运行的全部暂存变更，或选择「拒绝」放弃。定时同步与手动同步使用同一规则。

目录推送

适用于人事系统等外部系统主动推送组织与成员变更：

1. 在「选择身份源」中选择「目录推送」，进入「目录推送配置」。
2. 开启「启用目录推送」，填写「来源名称」「应用标识（AppKey）」和「应用密钥（AppSecret）」。
3. 已配置企业 IAM 时，可在「关联单点登录」中选择该身份源，推送的成员会按其账号匹配字段关联已有账号。
4. 选择「保存配置」，把页面显示的「目录推送地址」「连接测试地址」和凭据配置到外部系统。外部系统可先调用连接测试地址，再推送数据。

完整名单与局部更新使用同一推送地址。新增成员会使已注册账号数超出许可证席位时，本次推送整体失败。推送中标记为停用的成员，以及完整名单模式下不再出现的成员，会被移入「已离职成员」，平台同时吊销其登录会话、刷新令牌、受信任设备、模型 API 密钥和 MCP 访问令牌。停用目录推送后，平台不再接受外部推送。

3.7 账号与会话安全

「管理控制台 → 组织与权限 → 账号与会话安全」列出本租户的登录会话，包括「用户」「设备」「登录位置」「最近在线」和「状态」（活跃、已过期、已撤销）。可按用户、邮箱、ID 或设备搜索；发现异常会话时选择「撤销」，对方需要重新登录。

页面右上方的「安全策略」统一管理以下设置。打开对话框时，如本次会话的身份验证已超过 10 分钟，系统先要求「验证身份」。

分组	设置	取值与说明
登录会话	空闲超时	1-720 小时，默认 168 小时（7 天）；连续无操作超过该时长后需要重新登录
登录会话	绝对超时	1-2160 小时，默认 720 小时（30 天）；从登录时刻起的硬上限，持续活跃也必须重新认证；空闲超时不能大于绝对超时
登录会话	最多在线人数	不小于 1 的整数，默认 10000；同一租户同时在线的用户达到上限后，新用户无法登录

分组	设置	取值与说明
登录方式	允许密码登录	关闭后本租户账号只能通过单点登录登录；关闭前须已启用单点登录身份源，并经该身份源登录
登录方式	允许通行密钥登录	关闭后不能使用通行密钥作为登录方式
登录方式	强制两步验证	开启前须先为自己绑定两步验证，并以两步验证完成本次登录；开启后未绑定的成员在下次登录时须先绑定
登录方式	记住设备天数	0-30 天；0 表示不记住设备，每次登录都需要完成两步验证
密码策略	最小长度、至少包含的字符类别、禁止复用最近的密码、拒绝常见弱密码	最小长度 8-128 个字符；字符类别 0-4 类（大写字母、小写字母、数字与符号）；禁止复用 0-24 次，0 表示不限制
外部访问	默认 令牌有效期、最长 令牌有效期	1-168 小时，默认 令牌有效期默认 8 小时，且不能大于最长 令牌有效期
外部访问	模型 API 密钥轮换宽限期	0-168 小时，默认 24 小时；轮换后旧密钥仍可用的时长，0 表示立即失效
外部访问	允许类型	开启 OAuth 后，MCP 链接需用户授权后访问；开启 Token 后，可凭 Token 直接访问

平台管理员在平台范围的「账号与会话安全」中修改的是平台账号与 root 的策略：平台范围始终允许密码登录（root 只能以密码登录），且平台账号与 root 必须使用两步验证。平台范围另有「管理面访问网段」页签。

3.8 个人账户安全：两步验证与通行密钥

成员在能力中心「首页 → 个人设置 → 个人资料」维护本人的账户安全：

- 两步验证：选择「添加验证方式」，在「开启两步验证」对话框中用身份验证器应用扫描二维码并输入 6 位验证码，选择「开启验证」。随后页面显示恢复码，每个恢复码只能使用一次，且只显示这一次，保存后选择「我已保存」。已开启时可选择「关闭两步验证」。
- 恢复码：开启两步验证后显示剩余可用数量；选择「重新生成」后，原有恢复码全部失效。
- 通行密钥：选择「添加通行密钥」，按浏览器提示使用本设备的生物识别或屏幕锁定完成注册。每个账号最多 5 个通行密钥；浏览器不支持时不显示该项。
- 设备：查看当前设备和「登录历史」（时间、登录方式、IP 地址、设备/操作系统），发现陌生登录时联系管理员撤销会话。

登录时：开启两步验证的账号在输入密码后进入第二因子验证，可输入验证码，也可选择「使用恢复码」或「使用通行密钥验证」。已注册通行密钥且「允许通行密钥登录」开启时，登录页可选择「使用通行密钥登录」。

验证器和恢复码同时丢失时，由租户管理员在「成员与部门」中为其「重置两步验证」（见 3.3 节）。

4. 配置模型供应商

模型供应商记录平台如何连接一个上游模型端点：供应商类型、请求地址、模型请求格式、调用凭据（账号）和启用的模型。模型服务只能从已启用的供应商模型中选择模型，因此供应商是模型接入的第一步。

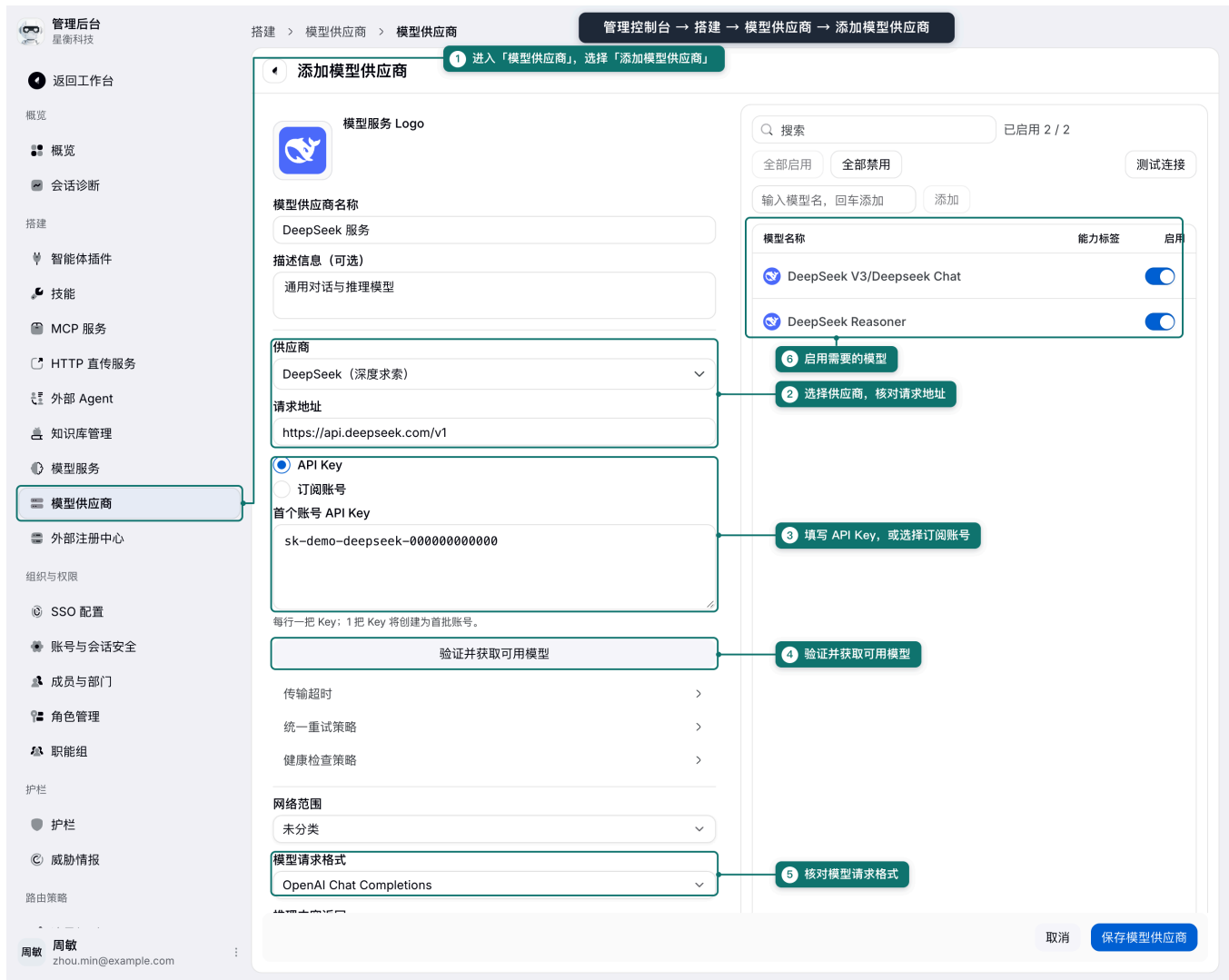
入口：「管理控制台 → 搭建 → 模型供应商」。页面顶部有三个页签：「模型供应商」（供应商列表）、「模型&定价」（全部模型的售价，见第 5 章「模型定价与分类」）、「分类管理」（模型分类）。

前提：以租户管理员或「系统管理员」角色登录。不具备权限时，「添加模型供应商」按钮置灰并提示“需要供应商管理权限才能添加模型供应商”。本章功能属于基础功能许可。

API 密钥等凭据只在页面的凭据字段中填写，不要写入工单、文档或截图。

4.1 新建并验证供应商

1. 在「模型供应商」页签选择「添加模型供应商」，进入添加页面。
2. 填写「模型供应商名称」（例如“DeepSeek 服务”）和「描述信息（可选）」。
3. 在「供应商」中搜索并选择供应商类型。内置目录覆盖国内外主流云厂商、聚合平台和本地推理服务，共两百多个类型；目录中没有的 OpenAI 兼容端点可选择「本地 OpenAI 兼容服务」后改写请求地址。选中后，「请求地址」自动填入该类型的默认地址，可按实际端点修改。
4. 核对「模型请求格式」。可选 OpenAI Chat Completions、OpenAI Responses、Anthropic Messages、Gemini generateContent 四种，默认取所选供应商类型的内置格式，没有内置格式时为 OpenAI Chat Completions。该格式决定网关以何种协议请求上游，调用方仍可用任一入站协议调用，协议转换见第 10.3 节。
5. 按需设置「推理内容返回」（默认「自动适配（推荐）」，另有「仅在 content 中返回」「单独返回 reasoning 内容」）和「网络范围」（未分类、内网、公网）。
6. 填写凭据。多数供应商类型提供两种方式：
 - 「API 密钥」：在「首个账号 API 密钥」中每行填写一把 Key，每把 Key 会创建一个账号；
 - 「订阅账号」：保存后跳转到「账号」页签完成订阅登录，见第 4.5 节。不需要凭据的本地推理服务不显示凭据区。
7. 选择「验证并获取可用模型」（右侧工具栏的「测试连接」作用相同）。该操作只探测连接，不保存。成功后右侧列出端点返回的模型；端点不返回模型列表时，在输入框中手动输入模型名并回车添加。
8. 在模型表中用「启用」开关选择需要的模型，可用「全部启用」「全部禁用」批量操作。至少启用一个模型。
9. 如需调整超时与重试，展开「高级」区域，见第 4.4 节。
10. 选择「保存模型供应商」。



保存即生效，不经过草稿审批。保存时平台会对请求地址做一次准入探测：探测失败时页脚出现错误提示，确认端点配置无误后可选择「仍要保存」；地址未通过出站安全校验或配置冲突时不能保存。提示“已保存，模型路由稍后同步”表示配置已写入，网关节点会在稍后同步。

验证结果：返回列表，确认该供应商的「可用」开关为开启，「可用模型」列显示已启用的模型，「健康」列在下一轮探查后出现结果。

4.2 可用性探查与连接排查

平台按供应商的「健康检查策略」周期探查供应商可用性，默认每 300 秒一次。探查结果显示在：

- 列表的「健康」列：最近一小时的可用性，每五分钟一个采样点；
- 列表的「耗时」列：延迟徽标，小于 300 毫秒为绿色，300~500 毫秒为黄色，超过 500 毫秒为红色，失败显示“不可用”；
- 供应商详情的「可用性探查」页签：探查历史，列为「时间」「结果」「探活后健康状态」「延迟」「失败原因」。

需要立即确认连接时，在「可用性探查」页签选择「连通性测试」。成功显示“连接正常，延迟 N 毫秒”；失败显示失败原因，例如“密钥无效或已过期”“没有访问权限”“请求太频繁，被供应商限流”“供应商额度已用完”“接口地址未通过安全校验”“网络连不上供应商”“供应商响应超时”。带上游状态码的原因会附注状态码。

连接失败时按以下顺序检查，避免同时改动多项配置：

- 1. 「请求地址」是否包含供应商要求的协议、域名和路径；
- 2. 账号的 API 密钥是否有效、是否有目标模型的权限，在「账号」页签查看账号状态；
- 3. 网关节点能否访问该端点，是否需要由平台管理员为节点配置出口策略（第 23 章）；
- 4. 「模型请求格式」是否与端点实际协议一致；
- 5. 仍失败时，记录发生时间、供应商名称和页面提示，不要记录或发送 API 密钥。

4.3 配置完成标准

同时满足以下三项，供应商才算配置完成：

- 连通性测试成功，「健康」列显示可用；
- 至少启用一个模型，并已将需要计费的模型设置售价（第 5 章「模型定价与分类」）；
- 「账号」页签至少有一个状态为“正常”的账号，页面顶部没有“账号全部不可用”提示。

只保存名称和凭据不能保证成员可以调用。成员调用还需要模型服务包含该模型，并且成员在模型服务的可用范围内（第 6、7 章）。

4.4 编辑供应商与运行参数

在供应商列表点击名称进入详情，页签依次为「基本信息」「账号」「模型定价」「上游真实性」「可用性探查」「性能测试」「操作审计」。

修改配置：在详情页选择「编辑 AI 供应商」，可修改名称、描述、「AI 平台」、「请求地址」、网络范围和模型选择。凭据不在此处修改，统一在「账号」页签维护。模型发现只提供建议，只有勾选的模型会进入配置，不会自动启用任何模型。

每次保存生成一个新的不可变配置版本，并立即替换当前生效版本。保存时同样执行准入探测；请求地址的主机变更时，平台会用新地址重新校验全部 API 密钥账号，未通过的账号标记为不可用。保存时如提示配置已被他人修改，选择「载入最新配置」后重新编辑。

「高级」区域包含三组运行参数，默认值如下：

参数组	字段与默认值
「传输超时」	连接超时 10 秒；首内容超时 120 秒；响应头超时 30 秒；非流式生成超时 600 秒；流空闲超时 60 秒；整条流上限 1800 秒
「统一重试策略」	最大重试 2 次；初始退避 250 毫秒；最大退避 2000 毫秒；可重试错误为传输错误、超时、限流和服务端错误
「健康检查策略」	「多久探一次（秒）」300 秒（上限 3600 秒）；连续失败 3 次判定不健康；连续成功 2 次恢复

停用供应商：在列表中关闭「可用」开关。停用只改变启停状态，不生成新版本。

4.5 账号与上游余额

同一供应商可以添加多个账号，网关在账号之间分配请求。每个账号可单独设置并发上限。

1. 进入供应商详情的「账号」页签，选择「添加账号」：
 - 「批量添加 API 密钥」：每行一把 Key，选择添加并验证后逐把检测，汇总可用、被上游拒绝和重复的数量；
 - 「推理令牌」或「登录 {平台} 订阅账号」：适用于 Anthropic Claude、OpenAI Codex、Google Gemini、GitHub Copilot、xAI Grok、Kimi 等订阅平台，按页面引导完成授权。授权链接 30 分钟内有效。
2. 在账号行的操作中可以测试、编辑、停用或启用、重新授权、刷新用量。编辑项包括备注、并发上限（1~100）和可用模型（留空表示跟随供应商模型列表）。
3. 订阅账号可开启「跨计费」：订阅账号全部不可用时改用 API 密钥账号按量计费。

账号状态包括“正常”“冷却中”“已停用”“异常”“需重新授权”。异常原因会标明余额不足、配额已用尽、Key 无效或已吊销等。所有账号都不可用时，详情页顶部提示调用会失败。

上游余额：「基本信息」页签的「上游余额」区域显示剩余额度、已用、总额和套餐，每 300 秒采样一次。平台只对提供余额查询接口的上游生效；上游不提供时显示“此上游未提供余额查询接口”，凭据失效时显示“上游凭据已失效，余额已停止更新”。

上游需要经代理访问时，由平台管理员在「平台扩展 → 网关」的「出口代理」页签维护代理，并在节点的出口策略中按目的地引用（第 23 章）。

4.6 模型能力用途与元数据

模型能接入哪些调用接口，由模型的能力用途决定。在供应商详情「基本信息」页签的模型列表中：

- 选择「配置能力」，在「模型能力配置」中为模型声明用途：对话、图像理解、图像生成、向量嵌入、重排序、语音合成、语音转写、判断、内容安全、敏感信息识别、轻量。能力来源可以是供应商发现、内置目录或手工配置；判断、内容安全、敏感信息识别三类用途只能由管理员显式声明。
- 选择「编辑模型元数据」修改显示名称等元数据。

供应商不提供某类接口时，对应用途不能被选中。模型或其供应商不支持所调用的接口时，网关返回 400，错误码 `MODEL_ENDPOINT_UNSUPPORTED`。全局设置中的内容安全、敏感信息识别、意图识别等默认模型（第 6.9 节），需要选用已声明对应用途的模型。

4.7 上游真实性核验

上游真实性核验用于确认上游实际提供的是所声明的模型，适用于经中转渠道接入的场景。

1. 进入供应商详情的「上游真实性」页签（页内标题「模型真伪核验」）。列表的「模型真伪」列也会显示通过、未通过、未核验的数量，点击可直达该页签。
2. 在目标模型行选择「核验」。核验只在手动触发时运行，不定期执行。
3. 核验完成后查看「核验结论」。核验分五级：接口格式兼容、计费 token 一致、能力项完整、原厂签名有效、回答风格一致；中转渠道不核验响应形态与原厂签名。
4. 对确认属于误判的级别，可选择标记误判并填写理由。

被动信号：同一页签的「被动流量信号」统计近 24 小时上游自报模型与请求模型的比对结果。出现不一致时显示“N 次上游自报模型不一致”，点击进入网关请求列表并筛选这些调用。在网关请求列表和模型服务的「调用情况」明细中，这类调用带有“上游换了模型”标记。

4.8 性能测试

性能测试按并发阶梯向上游发送流式请求，测量首字延迟、字间延迟和输出速度，用于评估上游容量。测试直连上游，不经过网关，不计入用量统计，但会消耗供应商的真实额度，并与生产流量共用该供应商的额度。

- 1. 进入供应商详情的「性能测试」页签，选择「发起测试」。
- 2. 填写参数：

字段	范围与默认值
「模型」	必选
「并发阶梯」	可选 1、2、4、8、16、32、64，默认 1、4、16，最多 8 档
「每档请求数」	1~1000，默认 64，不得少于最高并发数
「输入长度」	16~32000 词元，默认 256
「输出长度」	16~8192 词元，默认 256

- 3. 核对页面给出的预估消耗，选择「开始测试」。
- 4. 测试完成后查看报告：各档位的首字延迟、字间延迟、端到端耗时、每词元耗时（中位数、90 分位、99 分位、平均、最大），以及输出速度、请求速率和成功率；「失败原因」页签列出失败请求的分类。

限制：同一时间只能有一个进行中的测试；单请求超时 10 分钟；取消测试会立即中止在途请求，已完成的样本保留。

4.9 删除与操作审计

删除模型：在「基本信息」页签的模型行选择删除。平台先预览影响范围（模型服务、路由规则、定价记录、待吊销的 API 密钥等），确认后永久删除模型并清理关联记录。模型被全局设置的默认模型占用、或被不可变路由规则引用时，删除被阻止，需先更换或解除引用。

删除供应商：在列表的「操作」菜单中选择「删除」。供应商仍被模型服务使用时，平台拒绝删除并提示先从模型服务中移除。

操作审计：供应商详情的「操作审计」页签列出该供应商的创建、更新、删除、启停、检查连接、发现模型建议、更新模型定价、发起或取消性能测试等记录，点击记录查看变更详情。

5. 模型定价与分类

模型售价决定每次调用记入多少费用，也是模型服务金额预算和阶梯计价的基础。未定价的模型仍可调用，但其调用不计入金额统计，金额预算也会直接拒绝这类调用。

入口有两处：

- 「管理控制台 → 搭建 → 模型供应商 → 模型&定价」：按供应商查看和批量维护全部模型的售价；
- 供应商详情的「模型定价」页签：只维护该供应商的模型。

前提：租户管理员或「系统管理员」角色。售价币种为部署的结算币种；结算币种不是美元时，参照目录价需要先在 全局设置 中配置美元汇率（第 6.9 节）。本章功能属于基础功能许可。

5.1 设置模型售价

1. 在「模型&定价」页签按供应商筛选或搜索模型，在目标行选择「编辑」，打开「编辑模型定价」。
2. 在「计价模式」中选择「单个计价」「阶梯计价」或「分时计价」（分时计价见第 5.6 节）。
3. 单个计价时填写四项单价「输入」「输出」「缓存命中」「缓存写入」，单位为“结算币种 / 百万 Token”。输入或输出留空按 0 计费；缓存价留空时，按供应商模型请求格式从输入价换算，输入框中会提示换算后的值。
4. 按需设置「Token 权重」（单位“倍”）。Token 权重只把原始 Token 折算为额度消耗，不改变价格。
5. 保存。每次保存生成新的价格版本，已结算的历史调用不受影响。

在列表中可用开关单独开启或关闭某个模型的计价，也可勾选多行后「批量编辑」「开启计价」「关闭计价」。

非 Token 计价的模型按用途使用对应单位，金额填在输入价位置，输出按 0 处理，不换算缓存价，也不支持阶梯计价：

用途	计价单位
图像生成	每张
语音转写	每分钟音频
语音合成	每百万字符
重排序	每千次搜索

同一次批量编辑中选中的模型计价单位不同时，平台拒绝保存并提示分别设置。

5.2 阶梯计价

阶梯计价按累计用量分档定价，适用于与上游按用量阶梯结算的场景。

1. 在「编辑模型定价」中选择「阶梯计价」。
2. 设置「累计周期」（本小时、今日、本月或累计）和「累计方法」（租户、用户、API 密钥或会话）。
3. 选择「计价方式」：「单价」（每档填写四项单价）或「Token 倍率」（每档价格 = 基础价 × 倍率，倍率大于 0，最多 3 位小数）。

4. 选择「计算方式」：「分段累进」按档位边界把一次调用拆开计价；「全量按最终档」整笔按调用结束时所在档位计价。
5. 在「价格区间」中「添加档位」，用量单位为“万 Token”。第 1 档从 0 起，每档起点等于上一档终点，末档终点留空表示不封顶，最多 32 档。

修改档表后，已产生的账单不变。模型服务还可以针对某个授权对象单独配置阶梯价，见第 7.4 节。

5.3 参照目录模型定价

单个计价模式下，可以让模型售价跟随内置模型目录（models.dev 快照）中某个模型的公开价格。

1. 在「编辑模型定价」中找到「参照目录模型」，从下拉候选中选择。候选按与当前模型名称的相似度排序。
2. 选中后四项单价变为只读，并按“目录美元价 × 美元汇率”自动回填，不加价。
3. 保存。列表的价格来源显示“参照 · 模型名”。

限制：

- 阶梯计价模式不支持参照，切换到阶梯计价会清除参照；
- 结算币种不是美元且未配置美元汇率时，页面提示“请先在全局设置配置美元汇率”；
- 选择「清除参照」恢复为手工填写的草稿。

5.4 按目录价批量填充

需要一次为某个供应商的大量模型填写售价时，使用目录价批量填充。

1. 在「模型&定价」页签先选择一个供应商，再选择「按目录价填充」。进入填充模式后表格不可编辑。
2. 核对每个模型的建议价和跳过原因，例如“目录无价”“非对话模型”“无变化（跳过）”。目录中没有价格的模型，可点击「选参照模型」改用第 5.3 节的方式定价。
3. 选择应用，在「应用目录价建议？」对话框中设置：
 - 「填充目标」：覆盖当前范围全部、只更新来自目录价的模型、只填充四项售价都为空的模型（默认覆盖全部）；
 - 「加价系数」：售价 = 目录价 × 汇率 × 加价系数，范围 0.01~100，默认 1.00 表示按成本价；
 - 汇率：结算币种不是美元时填写 1 美元兑结算币种的汇率，预填全局设置中的美元汇率。此处输入只用于本次换算，不回写全局设置。
4. 确认后，平台为范围内每个模型写入一个新价格版本，结果摘要显示已写入的模型数。选择「完成」退出填充模式。

价格来源以徽标区分：“目录价”“参照 · 模型名”“手工”。

5.5 模型分类

模型分类决定模型广场左侧「分类」筛选中出现的分类及其包含的模型。

1. 进入「管理控制台 → 搭建 → 模型供应商 → 分类管理」。
2. 选择「新增分类」，平台生成默认名称，可直接重命名并选择分类图标。分类列表按「显示顺序」排列。
3. 选中分类，在右侧选择「添加模型」，把模型加入该分类；选择「移除」取消关联。

4. 删除分类只移除该分类及其与模型的关联，模型和模型服务不会被删除。

平台不预置分类，各租户按需建立。

5.6 分时计价

分时计价按请求所处的时段使用不同单价，适用于供应商按峰谷时段定价的模型（例如 DeepSeek 工作日 9:00–12:00、14:00–18:00 为高峰价，其余时间为空闲价）。

1. 在「编辑模型定价」中选择「分时计价」。
2. 选择「时区」，默认 Asia/Shanghai。时区随价格保存，与租户时区无关；有夏令时的时区按当地时钟换算。
3. 在「基础价格」中填写未落入任何时段时使用的四项单价。
4. 在「时段」中「添加时段」，每段填写「名称」（例如“高峰”）、开始与结束时间（24 小时制 HH:MM，结束时间可填 24:00）、「适用日」和四项单价。最多 8 段。
5. 保存。

规则：

- 按请求开始时间判定时段。一个 11:59 开始、12:01 结束的请求按 11:59 所在时段计价；该调用的费用仍计入请求结束时所在的统计周期。
- 只按星期几判定，不区分法定节假日和调休。供应商官方在节假日按空闲价收费时，平台在节假日仍按星期几对应的时段计价，与供应商账单可能不一致。
- 时段左闭右开：时段 9:00–12:00 包含 9:00，不包含 12:00。
- 结束时间早于开始时间表示跨零点（例如 22:00–02:00），凌晨部分按开始那天判定适用日：适用日为周五的 22:00–02:00 覆盖周五 22:00 至周六 02:00。
- 时段之间不能重叠，开始时间与结束时间不能相同，表单会就地报错。
- 分时只作用于四项 Token 单价；图片、音频等物理单位计价不分时。分时计价不能与阶梯计价同时使用。
- 金额预算的预占按基础价与各时段价中的最高价估算；按价格选路使用基础价。

「参照目录模型」中带有官方分时方案的模型（目前为 DeepSeek 的 deepseek-flash 与 deepseek-v4-pro）会显示方案预览。选用后按分时计价保存：基础价格为参照价，高峰时段价格为基础价格的 2 倍。

计价结果的查看位置：

- 「模型广场 → 模型详情 → 价格」列出各时段价格和基础价格，并标出当前时段；
- 调用日志的调用详情和模型服务「调用情况」的调用明细显示「计价时段」，例如“高峰（09:00–12:00）”或“基础价格”。

6. 创建模型服务

模型服务把一组供应商模型组织成面向成员和业务系统的服务。授权、配额、阶梯计价和 API 密钥都以模型服务为单位：成员获得某个模型服务的使用权后，用绑定该服务的 API 密钥调用服务内任一已启用模型。

入口：「管理控制台 → 搭建 → 模型服务」。页面顶部有两个页签：「模型服务」（服务列表）、「API 密钥管理」（第 6.7 节）。

前提：租户管理员或「系统管理员」角色；至少有一个已启用模型的供应商（第 4 章）。本章功能属于基础功能许可。

调用方在请求的 `model` 字段中指定服务内的具体模型名。平台不会因为某个模型不可用而自动改用服务内的其他模型；需要失败时转到备用模型，应在「管理控制台 → 路由策略 → 流量规则」中配置故障转移（需要动态路由许可，见第 15 章）。

6.1 新建模型服务

1. 在「模型服务」页签选择「添加模型服务」。
2. 设置「模型服务 Logo」：上传 JPG、PNG、WebP 或 GIF 图片（不超过 2 MB），或选择「生成」。
3. 填写「模型服务名称」（例如“高级 Claude 服务”）和「描述信息（可选）」。描述中写明适用任务、数据边界或成本等级。
4. 选择「添加模型」，从已启用的供应商模型中加入至少一个模型。已添加列表显示「模型名称」「供应商」「多模态」，可拖动调整顺序，用「移除」删除。
5. 选择「保存模型服务」。



保存后服务默认处于发布状态。新服务尚未分发给任何人，列表的「可见范围」列显示“未分发”，需要继续配置可用范围（第 7 章）。

模型能接入的调用接口由模型自身的能力用途决定（第 4.6 节）。同一个服务可以同时包含对话、嵌入、图像生成、语音、重排序等不同用途的模型，调用方按接口和模型名使用。

6.2 模型服务命名与拆分

- 按使用目的拆分，例如“通用对话”“代码与分析”“低成本批处理”；
- 需要不同的可用人群、费用上限或数据处理要求时，拆成不同服务；
- 不要只用供应商的模型型号作为服务名称，成员通常按任务用途选择；

- 在描述中写清适用场景和限制。

6.3 模型服务不可用排查

调用失败或成员看不到服务时，按以下顺序检查：

1. 服务详情「基本信息」页签的「发布」开关是否开启，模型列表中该模型状态是否为“正常”；
2. 模型所属供应商是否启用、健康，账号是否可用（第 4.2、4.5 节）；
3. 调用方是否在服务的可用范围内，授权是否在有效期内（第 7 章）；
4. 调用方使用的 API 密钥是否绑定该服务、状态是否为“可用”（第 6.7 节）；
5. 配额是否已用尽，额度池提示“已用尽”（第 7 章）；
6. 用「试跑」页签复现，按失败归因定位（第 6.5 节）。

6.4 服务列表与服务详情

服务列表的列为「名称」「模型数量」「可见范围」「累计消耗」「累计费用」「最近更新」，可按名称搜索、按「已启用」「已停用」筛选。行末「操作」菜单提供「编辑」和「删除」。

点击服务名称进入详情，页签依次为「基本信息」「试跑」「可用范围与配额」「调用情况」「操作审计」。右上角「编辑模型服务」用于修改名称、描述、Logo 和模型列表。

「基本信息」页签：

- 左侧：「发布」开关、「创建时间」、「服务级配额」（第 7.3 节）、「累计 Token（全部时间）」「累计费用（全部时间）」；
- 右侧：模型表，列为「排序」「模型名称」「供应商」「多模态」「状态」「当月用量」「当前已用（共计）」「当月费用」「费用（共计）」。可在此「添加模型」或移除模型。

停用或删除服务前，平台检查全局设置是否占用该服务的模型。被占用时提示受影响的功能，需先到全局设置更换（第 6.9 节）。

6.5 试跑与失败归因

「试跑」页签用真实调用链路向服务内的对话模型发起请求，用于上线前验证。试跑会真实计费，并记录在「调用情况」与操作审计中。

前提：以租户管理员身份进入（平台管理员没有租户归属，不能试跑）；服务已发布；服务中至少有一个启用的对话模型。

1. 选择要试跑的模型，默认使用服务的主对话模型。
2. 按需展开「参数」，可填写 system、Top P、「随机种子」等；留空使用模型默认行为。
3. 输入消息并发送。结果以流式返回，可选择「停止」中断（上游仍会完成生成并计费）。
4. 成功时显示摘要，例如“通 · 首字 0.8s · 共 2.1s · 356 tok · ¥0.0012”。模型请求调用工具时，可「手填返回值」「跳过」后「继续」，或选择「真执行」。

失败时，页面给出三类信息：

- 失败发生的位置：“失败发生在本平台侧：请求从未发给供应商”或“失败发生在供应商侧：请求已发出并收到错误响应”；
- 失败原因：按网关返回的稳定错误码（reason）显示中文说明，例如护栏拦截显示“内容策略已拦截此请求，请调整内容后重试。”，额度用尽显示“这把 Key 的本月Token额度已用尽，请联系管理员调整额度。”，限流显示“请求过于频繁，请稍后再试。”；没有错误码时显示“不通 · 失败原因未归类”；
- 供应商侧失败时，附带供应商 HTTP 状态码和供应商返回的信息。

错误码的完整含义见第 10.8 节。

6.6 调用情况

「调用情况」页签展示服务的运行状态和调用明细：

- 指标：「服务状态」「当前并发」「服务限流」（模型服务级的每分钟请求上限，未设置时显示“不可用”）、「1 分钟错误率」「延迟 P50」「延迟 P99」；
- 趋势：「调用量、吞吐与限流趋势」，包括调用次数、每秒请求数、每秒输入与输出 token 数、平均并发、速率限流与并发限流的拒绝次数。时间范围可选最近 1 小时、6 小时、24 小时、7 天、30 天；
- 「调用明细」：列为时间、用户、调用模型、状态、Token 数和费用。上游实际返回的模型与请求不一致的调用带“上游换了模型”标记；平台内部发起的调用会标明用途，例如“平台内部调用：内容安全检测”。

计费数据有短暂延迟。最近的调用显示“待计费”；未定价模型的用量单独提示“另有 N token 未定价”，可选择「去定价」。设置价格后，新调用开始计费，之前的调用不会补算。

6.7 API 密钥管理

「API 密钥管理」页签集中查看本租户全部网关 API 密钥。Key 由成员在能力中心创建（第 8.3 节），管理员在此检索和处置，不在此新建。

检索：按 Key 名称或前缀搜索，按「状态」「模型服务」「调用方」筛选。列表列为「名称」「调用方」「模型服务」「状态」「累计消耗」。

Key 的状态包括“可用”“已禁用”“已过期”“已删除”。

点击 Key 打开详情（「模型 API 密钥」），查看身份信息（接入凭据、计费归属、所有者）、Base URL、可用状态、额度状态、最近使用和总调用次数，并可执行：

操作	结果
「轮换密钥」	签发新密钥，新密钥只显示一次；旧密钥在宽限期内仍可用。宽限期在「管理控制台 → 组织与权限 → 账号与会话安全」页面选择「安全策略」，在「模型 API 密钥轮换宽限期」中设置，默认 24 小时，范围 0~168 小时，0 表示立即失效
「禁用」	调用立即被拒绝，且不能在此重新启用
「删除」	不可恢复，使用该 Key 的调用立即失效
额度「编辑配额」	为这把 Key 设置独立额度池，见第 7.5 节

勾选多行后可「批量禁用」「批量删除」。

6.8 操作审计

服务详情的「操作审计」页签记录该服务的创建、更新、删除、试跑，可用范围的更新、批量更新和移除，以及配额预算的创建、替换和重置。点击记录查看变更前后的内容。跨服务的配置审计查询见第 22.7 节。

6.9 全局设置：默认模型与美元汇率

入口：「管理控制台 → 系统设置 → 全局设置」。

默认模型：为平台各项功能分别指定一个模型服务与模型。选择「编辑」，在「编辑默认模型」中逐项选择，留空表示未设置。可配置的功能包括：

- 「内容安全」：护栏判断内容风险时使用，选用后还需选择「模型类型」；
- 「敏感信息识别」：识别个人信息等敏感实体；
- 「意图识别」：路由规则的意图条件使用；
- 「问答语义匹配」：问答库语义匹配使用；
- 「知识库嵌入」「知识库重排」：已获得知识库加购许可并部署知识库能力时使用。更换知识库嵌入模型会触发全部文档重建向量索引，重建期间语义检索暂停；
- 「知识页、文档摘要与日报」：后台摘要任务使用。

列表的「状态」列显示“正常”“未设置”“已停用”“已失效”，「谁能用」列说明该功能的使用范围。所选服务停用或模型失效时，页面顶部提示需要更换的功能。选用外部供应商的模型时，对话框提示用户内容会发送给该供应商。

美元汇率：在「美元汇率」区域填写“1 美元兑结算币种”的汇率（大于 0，最多 6 位小数）。汇率用于把客户端上报的美元费用换算为结算币种，也用于参照目录价定价（第 5.3 节）。不填写时不使用客户端上报的费用，改用目录估算。结算币种为美元时无需填写。

7. 配置模型服务的可用范围与配额

在模型服务详情的「可用范围与配额」页签中，管理员决定哪些对象可以使用该服务，并为每个对象设置周期预算和阶梯计价。页面说明：仅可用范围内的企业成员可使用该模型服务；每个对象一条规则，规则下每个周期各是一个额度池，用满即拒绝。

入口：「管理控制台 → 搭建 → 模型服务」，点击服务名称，打开「可用范围与配额」页签。

前提：租户管理员或「系统管理员」角色。金额预算和阶梯计价要求服务内的模型已设置售价（第 5 章「模型定价与分类」）。本章功能属于基础功能许可。

7.1 选择授权范围并设置有效期

可用范围的对象类型有四种：

对象	适用场景
所有用户	全员都应使用的基础服务
部门	部门稳定、职责清晰的场景
职能组	跨部门项目或专项能力
成员	个别成员的试点或例外授权

范围应从小到大逐步开放。新增高成本或高敏感模型时，先授权给试点部门、职能组或成员，效果与费用符合预期后再扩大。

操作步骤：

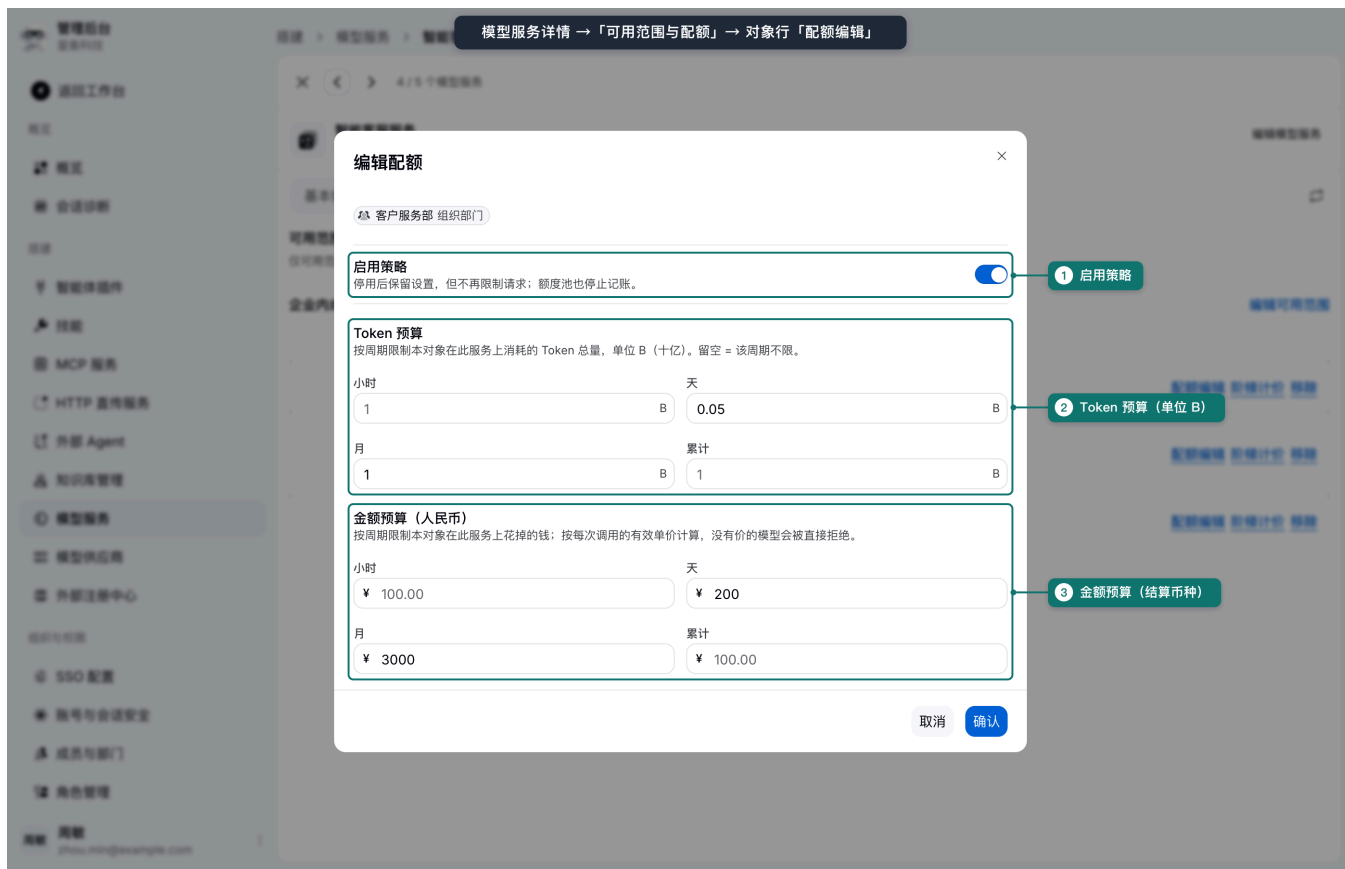
1. 在「可用范围与配额」页签选择「编辑可用范围」，打开「可用范围」对话框。
2. 在「所有用户」「职能组」「部门」之间切换，在组织架构中搜索并勾选部门或成员，或勾选职能组。右侧显示已选对象，例如“已选 2 个部门，3 位成员”。
3. 选择「下一步」，进入「配置目标有效期」。默认有效期可选 7 天、30 天、自定义或长期有效，默认为长期有效；也可以开启「为选中角色分配单独有效期」，为每个对象单独设置。
4. 保存。影响人数较多时，平台预览受影响的成员并要求再次确认。

结果：「企业内成员」列表列出已授权对象，列为「范围名称」「类型」「配额」「阶梯」。有效期未开始的对象显示“N 后生效”，进行中的显示“剩余 N”，到期后显示“已失效”。到期后请求被拒绝，历史授权记录保留用于审计。

移除授权：在对象行选择「移除」并确认。移除后该对象不能再使用此服务。

7.2 配额字段说明

在对象行选择「配额编辑」，打开「编辑配额」对话框，为该对象设置在本服务上的周期预算。



字段	说明
「启用策略」	停用后保留设置，但不再限制请求，额度池也停止记账
「Token 预算」：小时、天、月、累计	按周期限制该对象在本服务上消耗的 Token 总量，单位 B（十亿 Token），最多九位小数，例如 0.5 表示 5 亿 Token
「金额预算」：小时、天、月、累计（标题后显示结算币种，例如「金额预算（人民币）」）	按周期限制该对象在本服务上的花费，按每次调用的有效单价计算，最多六位小数

规则说明：

- 各周期、各预算类型可任意组合，每个已填写的上限都在请求准入时同时判断，任一用满即拒绝；
- 留空表示该周期不限制；填写 0 会阻断所有匹配调用；
- 启用策略时至少填写一个周期预算；
- 服务内模型都没有发布价时，金额预算不能填写；设置了金额预算后，调用未定价模型会被直接拒绝；
- 小时、天、月按 UTC 时间的自然周期重置：天从 UTC 0 点（北京时间 8 点）开始，月从每月 1 日 UTC 0 点开始；「累计」覆盖全部时间，不重置。

配额只限制当前模型服务，不是部门或成员的全部平台预算。单次请求的 Token 上限、并发数和每分钟请求数不在配额中设置，由流量规则和全局限额控制（第 15 章，需要动态路由许可）。

7.3 服务级配额

服务级配额以整个模型服务为对象，所有调用共享同一组额度池。

1. 进入服务详情的「基本信息」页签，在左侧「服务级配额」中选择「编辑配额」。
2. 字段与第 7.2 节相同，保存。
3. 用「服务配额」开关启用或停用。尚未设置任何周期预算时打开开关，会先弹出配额对话框。

左侧同时显示各额度池的用量，例如“已用 0.12B / 限额 1B”，以及「累计 Token（全部时间）」和「累计费用（全部时间）」。

7.4 阶梯计价

模型服务可以为某个授权对象设置按累计 Token 分档的价格，用于对内结算或差异化计费。

1. 在「可用范围与配额」页签的对象行选择「阶梯计价」。
2. 在「网关模型」中选择模型。模型必须已有发布价，未定价时页面提示先到价格页定价。
3. 设置累计周期：可选「月」，或该对象 Token 预算已使用的周期。累计方法固定为按该对象累计。
4. 选择计价方式：「单价」或「Token 倍率」。倍率档的价格 = 模型发布价 × Token 倍率，只影响金额，不影响额度。
5. 选择计算方式：「分段累进」按档位边界拆开计价；「全量按最终档」整笔按调用结束时所在档位计价。
6. 填写档位，用量单位为“万 Token”，第 1 档从 0 起，末档不封顶，最多 32 档。保存。

结算规则：预付费按最高档冻结、按实际档位结算；修改档表后历史账单不变，已结算的调用不重新计算。对象行的「阶梯」列显示已配置阶梯的模型数量，额度池旁显示当前所在档位，例如“第 2 档 · 再用 N Token 进第 3 档”。删除阶梯在同一对话框中选择「删除阶梯」。

7.5 单个 API 密钥的额度

除授权对象的额度池外，还可以为单个 API 密钥设置独立额度。

1. 进入「管理控制台 → 搭建 → 模型服务 → API 密钥管理」，打开目标 Key 的详情。
2. 在「额度」区域选择「编辑配额」，字段与第 7.2 节相同。只有绑定了模型服务的 Key 可以配置。
3. 保存。

这把 Key 的用量同时计入本服务和所属组织对象的额度池，任一额度池用满都会拒绝请求。未单独设置时，Key 只受服务、所属部门或职能组等共享额度池限制。

成员创建 Key 时选择的「API 密钥归属」决定这把 Key 的用量计入哪些额度池，创建结果中会列出。

7.6 配额拒绝排查

额度用尽时，网关返回 HTTP 429，错误码为 POOL_EXHAUSTED，错误信息说明是哪个对象、哪个周期、哪种预算用尽，以及何时恢复，例如“部门研发部的今日Token额度已用尽，08:00 后恢复。”。错误详情中的 scope_kind、period、budget、used、limit、window_end 可供业务系统程序化判断。

排查顺序：

1. 根据错误信息定位对象（服务、部门、职能组、个人或这把 Key）和周期；
2. 在「可用范围与配额」页签核对该对象的额度池用量，或在 Key 详情的「额度」区域核对；
3. 需要放宽时调整对应周期的上限；

4. 错误码为 `TRAFFIC_RATE_LIMITED`、`TRAFFIC_CONCURRENCY_LIMITED` 或 `TRAFFIC_CIRCUIT_OPEN` 时，拒绝来自流量保护，检查第 15 章的流量规则和全局限额。

8. 成员使用模型

成员在能力中心查找已获授权的模型，获取 API 密钥并接入业务应用或开发工具。能力中心的页签、「首页」左侧导航和右上角账号菜单见 1.2 节。

本章功能属于基础功能许可。成员能看到和调用的模型，取决于管理员为模型服务配置的可用范围（第 7 章）。

8.1 浏览模型广场

入口：「能力中心 → 模型广场」。

- 用顶部搜索框（「搜索模型」，快捷键 `⌘K`）按名称搜索；
- 左侧按「最新上架」「最多使用」排序，按「分类」和「模型厂商」筛选。分类由管理员在「分类管理」中维护（第 5.5 节）；
- 模型卡片显示模型名、分类标签、描述、上下文长度和最大输出长度。

未登录时也可浏览公开目录；登录后，列表按当前账号可使用的模型服务过滤。

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，可「复制模型 ID」。详情分为以下区域：

区域	内容
「概览」	模型说明
「首次调用 API」	Base URL、代码示例、客户端环境变量（第 8.3 节）
「计费」	当前输入、输出、缓存读取、缓存写入计费价（每百万 tokens）和价格历史；官方参考价仅供参考，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（TTFT）、缓存命中率、工具调用错误率、结构化输出错误率，可选平均值、P50、P95 和近 3 天、近 14 天；「请求成功率」按小时显示近 3 天时间线
「能力评测」	第三方评测得分
「调用用量」	近 14 日的输入、输出、推理 token 用量及调用来源

8.2 在线试用

在模型详情页头选择「在 Playground 中打开」，可直接与对话模型进行对话试用，不需要自己的 API 密钥。

前提：已登录；账号有可用的计费归属，且对应模型服务已启用并包含该模型。未登录时点击会先跳转登录。试用产生真实调用和费用，在用量中记为“试跑”。不满足条件时，页面显示具体原因，例如“该服务没有可用的对话模型”。

8.3 获取 API 密钥并发起首次调用

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。成员可通过两个入口创建 Key。

从模型详情创建：

1. 在模型详情页头选择「获取 API 密钥」，打开「获取 API 密钥」对话框。
2. 选择「API 密钥归属」。选项为管理员分发给该成员的归属，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。
3. 可选：在「指定 API 密钥（可选）」中填写已有的 Key 字符串，用于导入已有 Key。留空则自动生成。导入的 Key 须为 32~512 位，仅含字母、数字和 `-._~+/=`。
4. 选择「创建」。从此入口创建的 Key 有效期为 90 天，名称自动生成。
5. 新 Key 出现在对话框下方的「API 密钥列表」中，可显示、隐藏或复制。

从首页创建：

1. 进入「能力中心 → 首页 → API 密钥」，选择「创建 key」。
2. 填写「名称」，选择「模型服务」和计费归属。选定归属后，对话框提示这把 Key 的用量计入哪些额度池。
3. 可选：填写「指定 API 密钥（可选）」导入已有 Key，规则同上。
4. 选择「确认」。结果页显示「API 密钥（仅本次可见）」，立即复制并保存到业务系统的密钥管理中。从此入口创建的 Key 不设到期时间。

租户在安全设置中关闭 Token 访问时，无法签发模型 API 密钥，页面提示联系管理员开启。

发起首次调用：模型详情的「首次调用 API」区域提供四种协议的示例：OpenAI Chat Completions、OpenAI Responses、Anthropic Messages、Gemini generateContent，每种都有 TypeScript、Python 和 cURL 三种写法。示例中的 Base URL 由平台按网关访问地址生成；本次会话中刚创建过 Key 时，复制的示例会带上该 Key，否则为占位符 `YOUR_API_KEY`。

每种协议下方给出对应开发工具的环境变量：

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

区域中还有：

- 「会话归属（可选）」：同一对话的请求携带相同的会话请求头，可指定会话归属；不携带时由网关自动识别；
- 「列出可用模型」：一条 `GET /models` 请求列出该 Key 所属服务下全部已启用模型。

各协议的 Base URL、鉴权头和接口清单见第 10 章「调用接口」。

8.4 管理本人的 API 密钥

入口：「能力中心 → 首页 → API 密钥」。

列表列为「名称」「Key」（前缀加掩码，可复制）、「最后使用」「状态」「创建时间」和操作。状态包括“正常”“已停用”“已过期”“已撤销”。

不再使用或怀疑泄露的 Key，选择「撤销」并确认，使用该 Key 的调用立即失效。成员不能在此轮换 Key；需要保留调用连续性的轮换，请联系管理员在「API 密钥管理」中执行（第 6.7 节），或新建 Key 切换后撤销旧 Key。

同一页面下方的「访问密钥」用于签名调用平台管理接口，与模型调用无关。

8.5 查看本人用量与调用日志

入口：「能力中心 → 首页」，左侧导航「用量」分组下的「使用情况」「模型用量」「调用日志」。各分区的内容、筛选选项与统计口径见第 17.1 节。

调用日志中的 Request ID 与网关响应头 `x-request-id` 一致，向管理员反馈问题时附上该值。

9. 调用方、API 密钥、认证与授权

调用身份、访问授权与流量计数分别管理：调用方先用有效凭据通过认证，网关再按模型服务的可用范围判断是否允许调用，最后按配额、流量规则和全局限额决定是否放行以及如何计数。

9.1 理解调用对象

对象	管理位置	在授权与流量规则中的使用
平台成员	「管理控制台 → 组织与权限」下的「成员与部门」「职能组」	作为可用范围的对象（所有用户、部门、职能组、成员）；在流量规则中作为调用方
业务系统（API 密钥）	成员在能力中心创建；管理员在「管理控制台 → 搭建 → 模型服务 → API 密钥管理」统一管理	Key 绑定一个模型服务，用量计入创建时选择的归属；在流量规则中可按 API 密钥匹配和计数
业务系统（身份源令牌）	「管理控制台 → 组织与权限 → SSO 配置」中「企业 IAM (OAuth2)」身份源的「业务系统调用」区域	调用按令牌对应的人员与部门计量和授权
模型与能力	模型服务、MCP 服务等能力的管理页面	作为授权对象和流量规则的匹配条件

在流量规则中按调用方匹配时，先在条件分组「谁在调用」中选择「调用方类型」，再选择具体「主体」。调用方类型的可选项由平台按当前许可下发，例如「API 密钥」「平台用户」。调用身份只来自平台认证结果和网关的受信上下文，「验证规则」中输入的身份仅用于模拟，不构成真实授权；不要在普通请求头中自行声明其他用户身份。

9.2 API 密钥的创建与管理

API 密钥由成员在能力中心创建，管理员不在管理控制台中新建 Key。

- 成员创建：在「能力中心 → 模型广场」的模型详情选择「获取 API 密钥」，或在「能力中心 → 首页 → API 密钥」选择「创建 key」。每把 Key 绑定一个模型服务，可调用该服务下全部已启用模型，并按所选归属计入额度池。步骤与两个入口的差异（有效期、名称、原文是否再次可见）见第 8.3 节。
- 导入已有 Key：创建时在「指定 API 密钥（可选）」中填写已有 Key 字符串（32~512 位），适用于业务系统已经在使用某个 Key、需要迁移到网关的场景。
- 管理员管理：在「API 密钥管理」页签按名称或前缀、状态、模型服务、调用方检索全部 Key，执行轮换、禁用、删除和单独设置额度（第 6.7 节、第 7.5 节）。

为业务系统准备 Key 的建议流程：

1. 确定负责该业务系统的成员，确认其在目标模型服务的可用范围内（第 7 章）。
2. 由该成员创建 Key，选择与成本核算一致的「API 密钥归属」。
3. 立即把 Key 原文交付到业务系统的密钥管理系统。
4. 如需限制单把 Key 的消耗，由管理员在 Key 详情的「额度」中设置额度池。
5. 用最小请求完成一次调用，在「调用日志」或模型服务的「调用情况」中核对调用方、模型、状态和 Token。

Key 的状态包括“可用”“已禁用”“已过期”“已删除”（成员侧的“已撤销”与管理员侧的“已删除”对应同一状态）。业务系统收到鉴权失败时，应按错误码判断是 Key 状态问题还是网络问题，不要统一当作网络故障重试。

9.3 轮换、禁用和应急处理

常规轮换由管理员在「API 密钥管理」中执行：

1. 打开目标 Key 的详情，选择「轮换密钥」并确认。
2. 立即复制新密钥（只显示一次），更新到业务系统。
3. 旧密钥在宽限期内仍可用，到期自动失效。宽限期在「管理控制台 → 组织与权限 → 账号与会话安全」页面选择「安全策略」，在「模型 API 密钥轮换宽限期」中设置，默认 24 小时，范围 0~168 小时。
4. 在宽限期内确认业务系统已切换到新密钥。

成员不能轮换 Key。成员自行更换时，先创建新 Key 并完成切换验证，再「撤销」旧 Key。

发现泄露时不要使用轮换（旧密钥在宽限期内仍有效），应立即「禁用」或「删除」该 Key，再查询泄露后的调用记录。禁用后不能在管理控制台重新启用；删除不可恢复。两种操作都只阻止后续调用，不会删除已经生成的日志和用量记录。

9.4 分发能力和设置生效时间

模型服务、MCP 服务、Skill、外部智能体、API 服务等能力都通过同一个「可用范围」对话框授权，对象类型为「所有用户」「部门」「职能组」和「成员」。模型服务的操作步骤见第 7.1 节。

生效时间：

1. 在「可用范围」对话框选好对象后选择「下一步」，进入「配置目标有效期」。

- 2. 选择默认有效期：7 天、30 天、自定义或长期有效（默认长期有效）。自定义时设置「开始生效」和「结束生效」时间，结束时间必须晚于开始时间。
- 3. 需要为不同对象设置不同有效期时，开启「为选中角色分配单独有效期」。
- 4. 保存后检查对象行的状态：“N 后生效”“生效中”“剩余 N”或“已失效”，其中 N 为剩余时长，例如“2 小时 15 分”。
- 5. 使用目标成员账号验证能力列表可见、直接访问和实际调用。

已有对象的有效期限可通过「编辑有效期」单独修改，改期只作用于当前对象。授权到期后请求被拒绝，页面也不再把该授权展示为有效；历史授权记录保留用于审计。

9.5 认证方式支持边界

方式	适用范围	配置位置与说明
API 密钥	模型调用接口	成员创建，管理员在「API 密钥管理」中管理；鉴权头见第 10.1 节
身份源令牌	业务系统以终端用户身份调用模型	租户管理员或「安全管理员」在「SSO 配置」中打开「企业 IAM (OAuth2)」身份源，在「业务系统调用」区域开启「允许业务系统用身份源令牌调用模型」，并填写「业务系统客户端标识」。业务系统携带从该身份源取得的用户令牌直接调用，调用按人员与部门计量。令牌须使用非对称签名算法，签发方、受众、签名或有效期不符时返回 401
OIDC 单点登录	管理控制台与能力中心的登录	「管理控制台 → 组织与权限 → SSO 配置」。登录会话不能用于调用模型接口
客户端 IP 网段	按来源地址限制调用	在流量规则中使用条件「客户端 IP 网段」（需要动态路由许可）。该条件只对直连网关的调用生效

外部连接器（如 MCP 服务）使用的上游认证方式，在对应连接器的鉴权设置中配置和测试，不属于模型调用的认证方式。

9.6 创建、启停并维护调用规则

入口：「管理控制台 → 路由策略 → 流量规则」，需要动态路由许可，并以租户管理员或「安全管理员」角色登录。页面页签为「规则清单」「全局限额」「意图集」「缓存」。完整说明见第 15 章。

- 1. 在「规则清单」中搜索目标规则，或选择「新建规则」，填写名称、调用类型和匹配条件。
- 2. 按业务选择转发、放行或拒绝，或增加检测、限额、改写等动作。核对增强规则和终止规则的区别。
- 3. 选择「保存规则」，在「确认变更」中核对字段差异、关联对象和命中顺序后选择「确认并应用」（新建与编辑都经过这一步，见第 15.3.1 节）。规则清单中的启停开关、拖动排序和删除在确认后直接生效。
- 4. 核对启用状态和下发提示，使用「验证规则」检查正例与反例，再通过实际入口验证目标、权限和执行结果。

- 5. 停用使用「停用」，修改使用「编辑」。删除前准备替代路由和保护，删除后不能在页面直接恢复。
- 6. 预览提示配置已被修改时，按提示重新生成预览后再确认，不要重复提交过期的确认。

授权的起止时间在可用范围中设置（第 9.4 节）。流量规则中的临时阈值和临时豁免只调整相应保护，不延长授权或凭据有效期。

9.7 调用身份、计数与配额

授权范围决定“是否可以调用”，匹配条件决定“是否适用某条规则”，计数维度决定“如何分别累计”。

- 1. 在模型服务的「可用范围与配额」中核对目标调用方的授权与有效期。
- 2. 在「规则清单」中选择真实入口提供的调用方身份作为条件。
- 3. 在「全局限额」中「新增限额」，「限制类型」可选「请求次数」「并发限制」「熔断」等（见第 15.4.2 节）。分别填写「匹配条件」和「计数维度」，例如匹配客服模型、计数维度选择「按 API 密钥」，使每把 Key 各自累计。
- 4. 配置阈值和处理方式，保存后核对下发状态。
- 5. 用范围内、范围外及不同计数对象的请求分别验证，确认没有把多方流量累计到同一对象。
- 6. 请求被拒绝时按错误码区分来源：

错误码	来源
POOL_EXHAUSTED	模型服务、部门、职能组、个人或单把 Key 的额度池用尽，错误详情的 scope_kind 指明对象类型（第 7.6 节）
MODEL_SERVICE_QUOTA_HARD_REJECT	超出模型服务可用额度
TRAFFIC_RATE_LIMITED	流量规则或全局限额的次数限速
TRAFFIC_CONCURRENCY_LIMITED	同时进行的调用数已满
TRAFFIC_CIRCUIT_OPEN	熔断打开

普通请求携带的任意身份或标签不能替代受信认证信息。调整组织成员、凭据或模型授权后，应重新验证路由条件与计数范围。

9.8 鉴权故障定位

现象	检查顺序
返回未认证（401）	鉴权头名称和格式、Key 状态、过期时间、Base URL；使用身份源令牌时核对签发方、受众和签名算法
同时携带两种凭据返回 400	只保留一种凭据，或确保两者取值相同
已认证但无权调用（403）	调用方是否在模型服务可用范围内、授权是否在有效期内、流量规则是否拒绝

现象	检查顺序
返回模型不存在（404）	请求的模型名是否属于 Key 绑定的模型服务，模型是否已启用
某部门成员看不到能力	部门归属、可用范围对象、有效期状态和当前账号权限
临时授权提前失效	有效期起止时间、时区、成员身份和请求时的有效性判定
禁用或删除后仍有请求成功	是否使用了另一把 Key、是否刚轮换且旧密钥仍在宽限期内、调用记录中的真实身份
只在某个入口失败	该入口使用的是登录会话、身份源令牌还是 API 密钥；登录会话不能调用模型接口

10. 调用接口

业务应用和开发工具通过网关的模型调用接口使用模型。网关在入口完成鉴权、授权、配额与流量规则判断，把请求转换为上游供应商的接口格式，并把响应和错误按调用方使用的协议返回。本章面向接入网关的开发人员和技术负责人，功能属于基础功能许可。

10.1 Base URL 与鉴权

Base URL 由网关的公开访问地址加固定前缀组成。成员在模型广场的模型详情「首次调用 API」区域可直接复制当前环境的地址（第 8.3 节）。下表中 <网关地址> 指该访问地址：

协议	Base URL	鉴权方式
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

规则：

- 同一请求同时携带两种凭据且取值不同时，网关返回 400，不会择一使用；
- 凭据可以是成员创建的 API 密钥，也可以是启用「业务系统调用」后的身份源令牌（第 9.5 节）；
- 请求中的模型名（OpenAI、Anthropic 为请求体的 model 字段，Gemini 为路径中的模型名）填写模型广场或模型列表接口返回的模型名。模型必须属于这把 Key 绑定的模型服务，否则返回 404，错误码 LLM_MODEL_NOT_AVAILABLE；
- 可选请求头 x-zenova-conversation-id 用于指定会话归属；不携带时由网关自动识别会话，识别顺序与取值要求见第 10.9 节；

- 请求体上限为 32 MiB，语音转写文件上传为 25 MiB。

10.2 接口清单

OpenAI 兼容接口（路径相对于 OpenAI Base URL）：

方法与路径	用途	流式
POST /chat/completions	对话，支持工具调用	支持
POST /responses	OpenAI Responses	支持
POST /completions	文本补全	不支持
GET /models、GET /models/{model}	列出或查询本 Key 可用的模型	不适用
POST /embeddings	向量嵌入	不适用
POST /rerank	重排序	不适用
POST /images/generations	图像生成	不适用
POST /audio/speech	语音合成	不适用
POST /audio/transcriptions	语音转写（文件上传）	不适用
GET /audio/transcriptions/realtime	实时语音转写（WebSocket）	不适用
POST /decisions、POST /systemone	批量判断	不适用

Anthropic 原生接口：POST /v1/messages（支持流式），GET /v1/models。

Gemini 原生接口：POST /v1beta/models/{model}:generateContent、POST /v1beta/models/{model}:streamGenerateContent，同时接受 /v1/models/... 路径；GET /v1beta/models 列出模型。不支持 countTokens 等其他方法，调用返回 404。

模型必须具备对应用途（第 4.6 节）。模型或其供应商不支持所调用的接口时返回 400，错误码 MODEL_ENDPOINT_UNSUPPORTED；批量判断接口在所选上游不支持判断时返回 400，提示“该模型不支持此接口”。

10.3 协议自动转换

对话类接口支持四种协议：OpenAI Chat Completions、OpenAI Responses、Anthropic Messages、Gemini generateContent。调用方可以用其中任一协议调用任一对话模型，不必与供应商的「模型请求格式」一致：

- 入站协议与上游格式相同时，网关原样转发请求与响应；
- 不同时，网关把请求转换为上游格式，再把响应（包括流式事件）转换回入站协议；
- 转换中有字段无法在上游格式中表达而被丢弃时，响应头 x-zenova-conversion-dropped 列出被丢弃的部件，多个标识以逗号分隔，例如 anthropic.tool.web_search,responses.store。原样转发的请求不带该响应头。

嵌入、重排序、图像、语音和批量判断接口只提供 OpenAI 兼容格式，不做协议转换。

因此，Claude Code 可通过 Anthropic 接口、Gemini CLI 可通过 Gemini 接口、Codex 可通过 Responses 接口使用网关上的任意对话模型。各工具的环境变量见第 8.3 节「获取 API 密钥并发起首次调用」。

10.4 流式输出、重试与心跳

开启流式：OpenAI 与 Anthropic 接口在请求体中设置 "stream": true；Gemini 使用 :streamGenerateContent 方法。

重试与切换：

- 网关只在上游返回首个内容之前重试或切换目标，内容一旦开始输出就不再重试，避免重复内容；
- 同一目标的重试次数与可重试错误类型取供应商的「统一重试策略」（默认最多重试 2 次，第 4.4 节）；
- 只有在流量规则中配置了故障转移时，网关才会切换到备用档位的其他模型（第 15 章，需要动态路由许可）。

心跳：流式请求在 10 秒内仍未得到上游首个内容时，网关先返回 200 响应头并开始发送心跳，此后的失败以流内错误帧交付；流式输出过程中，网关每 15 秒插入一条心跳。Anthropic 协议使用 event: ping 事件，其他协议使用 SSE 注释行 :keepalive，标准客户端会忽略注释行。中间代理或负载均衡的空闲超时应大于 15 秒。

流内错误：流式输出开始后发生失败时，网关发送一条与所选协议错误体同形的流内错误帧，再正常结束流。已输出的内容与用量照常计费。护栏在流式输出中途拦截时，流以协议的终结帧结束，并在终结帧中带 safetyIntervention 字段（kind 为 stream_cut）。

10.5 嵌入与重排序

嵌入：POST /embeddings，请求体包含 model 和 input。按输入 Token 计费。

```
{"model": "your-gateway-model", "input": "把这段文字变成向量。" }
```

重排序：POST /rerank，请求体包含 model、query 和 documents。按搜索次数计费：上游返回搜索次数时采用该值，否则按“1 次查询加每 100 个文档 1 次”计算；价格按每千次搜索设置。

```
{"model": "your-gateway-model", "query": "退款进度", "documents": ["订单已发货。", "退款将在三个工作日内到账。"] }
```

10.6 图像与语音

接口	关键参数	计费
POST /images/generations	model、prompt、n	按返回的图片张数
POST /audio/speech	model、input、voice	按 input 的字符数（Unicode 字符），价格按每百万字符设置
POST /audio/transcriptions	multipart/form-data，字段 model、file	按音频时长，价格按每分钟设置

语音转写只读取音频文件头计算时长，不解码整段音频；无法识别格式时返回 400，提示“无法识别音频格式”。

实时转写使用 WebSocket：

1. 连接 GET /audio/transcriptions/realtime?model=<模型名>，可选 language（ISO-639-1 语种代码，如 zh，缺省自动识别）。

- 子协议为 `zenova.transcription.v1`。凭据放在 `Authorization` 头；浏览器无法设置该头时，把凭据作为第二个子协议携带，服务端不会回显。
- 以二进制帧发送 16 kHz、单声道、16 位小端 PCM 音频。网关按停顿切段，每段作为一次转写请求独立鉴权、计费，并以 `{"type": "transcript", "text": "...", "final": true}` 文本帧返回。
- 客户端可发送 `{"type": "input_audio_buffer.commit"}` 立即切段，发送 `{"type": "session.finish"}` 结束会话；服务端推完全部段后发送 `{"type": "finished"}` 并关闭连接。出错时发送 `{"type": "error", "code": "...", "message": "..."} 并关闭。`

10.7 批量判断

批量判断对一段内容同时回答多道是非、选择或评分题，适用于分类、审核等场景。调用前需在模型的能力配置中声明「判断」用途。

`POST /decisions` 请求体包含 `model`、`input`（被判断的内容）和 `questions`。每道题包含 `name`、`instructions` 和 `type`：

- `predicate`：是非题，答案为概率；
- `choice`：选择题，`choices` 至少 2 项，每项包含 `value` 和 `description`；
- `score`：评分题，`levels` 至少 1 级，每级包含 `label` 和 `description`。

```
{
  "model": "your-gateway-model",
  "input": "这笔订单被扣了两次款。",
  "questions": [
    { "name": "is_complaint", "instructions": "这是否是一条投诉？", "type": "predicate" },
    { "name": "department", "instructions": "应该交给哪个部门？", "type": "choice",
      "choices": [ { "value": "billing", "description": "付款、发票和退款。" },
                  { "value": "other", "description": "以上都不是。" } ] }
  ]
}
```

`POST /systemone` 是同一用途的另一种请求格式：用 `state` 放被判断的内容，`questions` 为按题名索引的对象。

限制：

- 一次请求最多 256 题，题名不能重复；
- 超过所选上游的单批上限时，网关按上限拆分并发送，再按题名合并答案，每一批单独记一条用量；
- 图片只能以 `data:` 地址放入内容，不接收图片的上游返回 400；
- 缺题、多题或答案不在选项内时，网关拒绝并指出具体题目；
- 不同上游给出的判断分数不能直接比较。

计费按 Token，同时计输入和输出。

10.8 错误格式与错误码

错误按调用方使用的协议返回：

协议	错误体形状
OpenAI	<code>{"error":{"message","type","code","reason","request_id"}}</code>
Anthropic	<code>{"type":"error","error":{"type","message","reason"},"request_id"}</code>
Gemini	<code>{"error":{"code","message","status","details"}}</code>

`reason` 是稳定的错误码，HTTP 状态和是否可重试都由它决定，业务系统应按 `reason` 处理错误，不要解析 `message` 文本。常见错误码：

reason	HTTP 状态	含义与处理
UNAUTHENTICATED	401	凭据缺失、无效或已过期，检查 Key 状态与鉴权头
PERMISSION_DENIED	403	已认证但无权执行此调用，检查可用范围与授权有效期
CONTENT_POLICY_BLOCKED	403	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
VALIDATION_FAILED	400	请求格式或参数错误
POOL_EXHAUSTED	429	额度池用尽，错误详情说明对象与周期（第 7.6 节）
MODEL_SERVICE_QUOTA_HARD_REJECT	429	超出模型服务可用额度
MONTHLY_TOKEN_EXCEEDED	402	租户本月 Token 上限已用尽，由平台管理员在「租户管理」中调整（第 23.8 节）
BILLING_UNAVAILABLE	503	网关暂时无法确认租户额度状态，可稍后重试
TRAFFIC_RATE_LIMITED	429	命中流量规则的限速，可稍后重试
TRAFFIC_CIRCUIT_OPEN	503	流量保护熔断打开，可稍后重试
LLM_RATE_LIMITED	429	上游限流，可稍后重试
LLM_UPSTREAM_SERVER_ERROR	502	上游服务错误，可稍后重试
LLM_OVERLOADED	503	上游过载，可稍后重试

响应头：

- `x-request-id`: 网关侧请求标识, 与 OpenAI、Anthropic 错误体中的 `request_id` 相同, Anthropic 协议还同时返回 `request-id`。向管理员反馈问题时附上该值;
- `x-should-retry`: `true` 或 `false`, 表示重试是否可能改变结果;
- `Retry-After` (秒) 与 `retry-after-ms` (毫秒): 仅在错误可重试且网关知道等待时长时返回。可重试的 429 若没有其他等待时长, 返回 `Retry-After: 1`; 网关数据面尚未就绪而返回 503 时, 也带 `Retry-After`。

错误体中的 `message` 是随版本可能调整的说明文字, 多数为英文; 管理控制台和能力中心按 `reason` 显示中文说明。完整的错误码清单、各协议的错误体与流式终结帧形态, 见随产品发布的文档站中的[错误参考页](#)。

10.9 会话识别

网关把同一场对话的多次调用归入同一个会话, 供会话列表与会话详情 (第 20.3 节)、会话诊断 (第 20.4 节)、按会话计数的限额 (第 15.4.3 节) 和会话切换保护 (第 15.7.4 节) 使用。会话与请求标识 `x-request-id` 无关: 请求标识每次调用各不相同, 不影响会话归属。

网关按以下顺序识别会话, 命中一项即停止:

1. 请求头 `x-zenova-conversation-id`: 调用方显式指定的会话编号。
2. 客户端自带的会话编号: Claude Code、Codex、OpenCode 等客户端在请求中携带的会话编号, 无需额外配置。
3. Responses 续接与会话: 使用 OpenAI Responses 接口并携带 `previous_response_id` 时, 续接请求归入上一轮所在的会话, 并沿用上一轮的识别方式; 未携带 `previous_response_id` 但请求体带有 `conversation` 时, 按该值识别。
4. 首条用户消息: 以上信息均未提供时, 网关按对话中的首条用户消息识别同一对话的多轮请求。

能够自行控制请求的调用方 (例如自建的 Agent 平台或业务系统) 可以携带请求头 `x-zenova-conversation-id`, 精确指定会话归属。取值要求如下:

- 使用 UUID, 或使用由字母、数字、`-`、`_` 组成且不少于 22 个字符的编号。
- 最长 128 个字符; 取值不符合上述格式或超过长度上限时, 请求返回 400。
- 同一对话的全部请求携带相同取值, 不同对话使用不同取值。

使用时注意以下限制:

- 按首条用户消息识别的会话只用于会话归并与展示, 不参与按会话计数, 也不启用会话切换保护。需要这两项功能时, 应携带 `x-zenova-conversation-id`, 或使用自带会话编号的客户端。
- 按首条用户消息识别要求对话历史保持完整。调用方截断、摘要或改写早期消息 (包括首条用户消息) 后, 后续请求可能被识别为新的会话; 同一调用方、同一终端用户发起的首条用户消息完全相同的两场对话, 可能被归入同一会话。对会话归属有精确要求时, 应显式携带请求头。
- 使用 Responses 续接时, `previous_response_id` 必须原样回传网关下发的响应编号 (形如 `resp_zr2...`)。回传其他取值, 或回传其他凭据获得的响应编号, 请求返回 400, 错误码 `invalid_previous_response_id`。
- Responses 续接依赖上一轮所在供应商保存的响应。上一轮所在的供应商或供应商账号当前不可用 (错误码 `provider_account_continuation_unavailable`), 或上一轮响应经协议转换生成、上游供应商未保存该响应 (错误码 `response_continuation_unrestorable`) 时, 无法凭 `previous_response_id` 续接。此时请求的 `input` 若已携带完整上下文 (包含此前的助手回复、工具调用或工具结果), 网关去掉 `previous_response_id` 后照常转发; 否则返回 409, 调用方应在 `input` 中携带完整上下文后重试, 或开启新对话。

11. 配置并分发 MCP 服务与 API 服务

网关通过两类服务向成员和业务系统提供外部工具与接口：

服务类型	入口	调用方式	适用场景
MCP 服务	「管理控制台 → 搭建 → MCP 服务」	MCP 客户端按工具调用，网关承担 MCP 协议	远程 MCP 服务，或需要把已有 HTTP 接口转为 MCP 工具
API 服务	「管理控制台 → 搭建 → API 服务」	应用携带访问令牌 按上游原接口调用，网关整体转发	业务系统需要按原接口访问上游，不需要转为工具

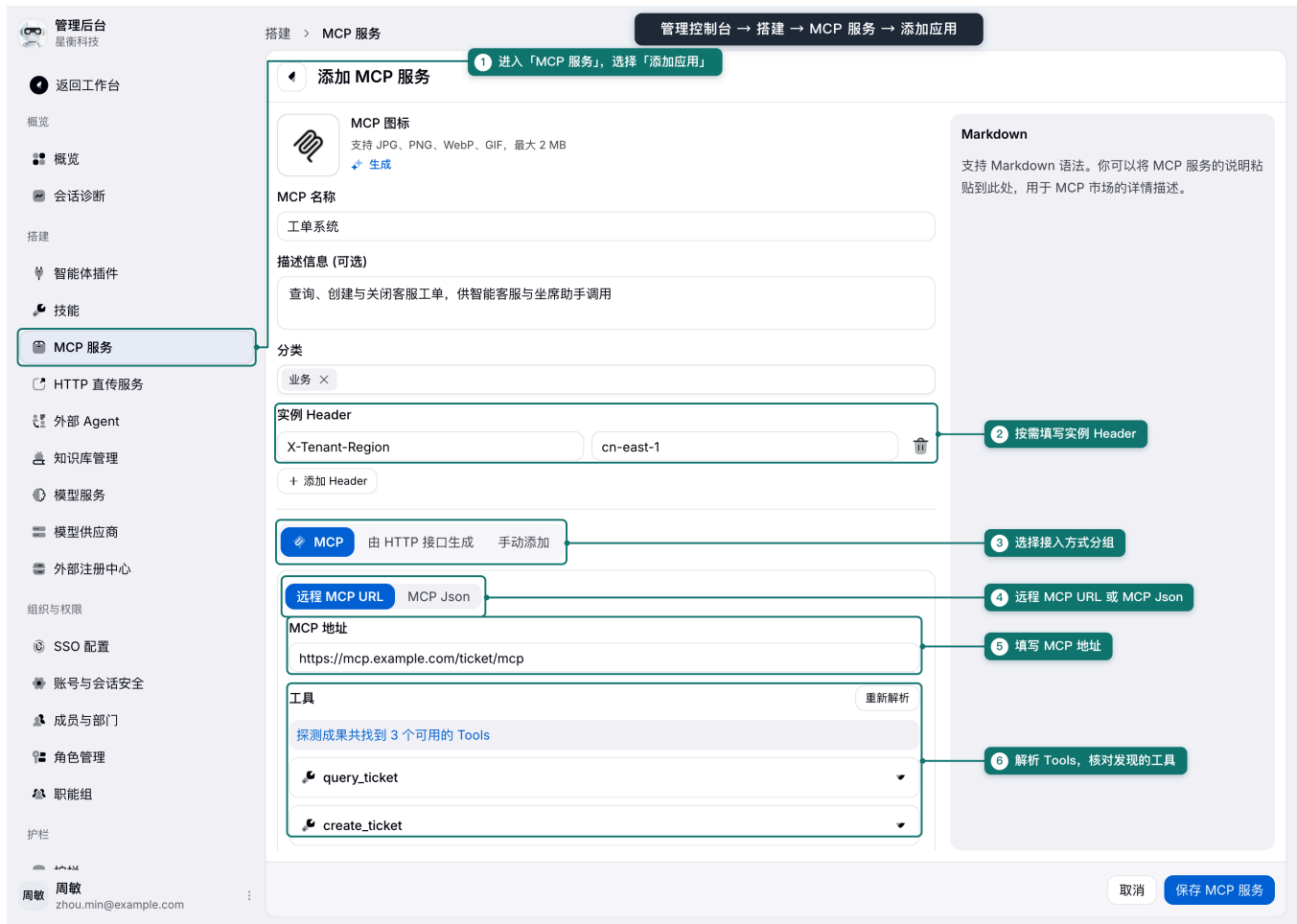
两类服务都属于「基础功能」，由租户管理员或「系统管理员」维护。完整配置包括连接、鉴权、发布状态和可用范围。仅保存服务不代表成员或业务系统已经获得使用权限。

11.1 创建 MCP 服务

在「MCP 服务」页面的「MCP 列表」标签选择「添加应用」，进入「添加 MCP 服务」。先填写公共信息：「MCP 名称」、「MCP 图标」（JPG、PNG、WebP、GIF，最大 2 MB）、「分类」、「描述信息 (可选)」，以及右侧「Markdown」中供成员阅读的用途、数据边界和使用示例。

再选择接入方式。页面顶层分为「MCP」「由 HTTP 接口生成」和「手动添加」三组，共六种方式：

顶层分组	方式	需要提供	结果
「MCP」	「远程 MCP URL」	「MCP 地址」，可选「实例 Header」	代理已有的远程 MCP 服务
「MCP」	「MCP Json」	mcpServers 格式的 JSON 配置	按 JSON 中的服务地址代理远程 MCP 服务
「由 HTTP 接口生成」	「OpenAPI」	粘贴 OpenAPI 3.0/3.1 JSON 或 YAML，或填写文档 URL、上传文件；按需填写「Base URL (可选)」	所选接口逐个发布为 MCP 工具
「由 HTTP 接口生成」	「Postman Collection」	粘贴或上传 Collection	同上
「由 HTTP 接口生成」	「cURL」	粘贴 cURL 命令	同上
「手动添加」	手动定义接口	「MCP 工具名」「方法」「路径」「MCP 工具描述」，选择「添加接口」逐个添加	手动定义的接口发布为 MCP 工具



由 HTTP 接口生成的工具由网关承担 MCP 协议，上游保持普通 HTTP 接口，凭据由网关注入。导入 OpenAPI 时，页面在「导入提示」中列出需要处理的问题，例如文档声明了多个服务地址、API 密钥位于查询参数、接口数量超过导入上限或缺少 operation ID。

远程 MCP 的传输协议限制如下：

- 支持 Streamable HTTP。通过「远程 MCP URL」接入的上游仅支持旧版 HTTP + SSE 时，平台仍可连接，详情页显示弃用提示。
- 「MCP Json」中的传输类型只接受 Streamable HTTP 的各种写法，声明为 `sse` 的条目不能导入。
- 不支持 stdio 方式（以 `command` 启动）的本地 MCP。

保存前发现工具：

1. 填写连接信息后，离开连接配置区或选择「解析 Tools」。页面提示「正在连接 MCP 服务并解析 Tools…」，完成后显示发现的工具数量。
2. 在工具预览中核对工具名称、描述和参数。预览会标出每个工具的副作用：「读取」「写入」或「破坏性」。
3. 选择「保存 MCP 服务」。工具探测失败时仍可先保存，之后再重试。

上游在列出工具时也要求鉴权，而平台尚未保存凭据时，可在 MCP 详情的「工具」标签选择「临时凭据」，在「使用临时凭据获取工具」中填写「Bearer 令牌」或「附加 Header」，然后选择「获取工具列表」。该凭据只用于本次请求，不保存在平台上，也不能用于后续调用；平台只保存发现的工具。

已保存服务的「工具」标签列出「已发现工具」及其「输入参数」，并显示可调用、漂移和过期的工具数量。上游工具变化后，使用该标签右上角的刷新按钮重新发现工具。

MCP 地址的出站访问受平台级「SSRF 出站防护」约束。该防护默认关闭，关闭时只校验地址格式与协议，MCP 地址可以指向内网、回环等任意地址；开启后，指向私网、回环、链路本地或云元数据的地址一律被拒绝。接入不受信的 MCP 前，应由平台管理员开启该防护，或通过出口代理、网络策略限制可访问的目标。

11.2 配置鉴权与请求头

在 MCP 详情「基本信息」的「鉴权」区域选择「编辑」，打开「MCP 鉴权」。

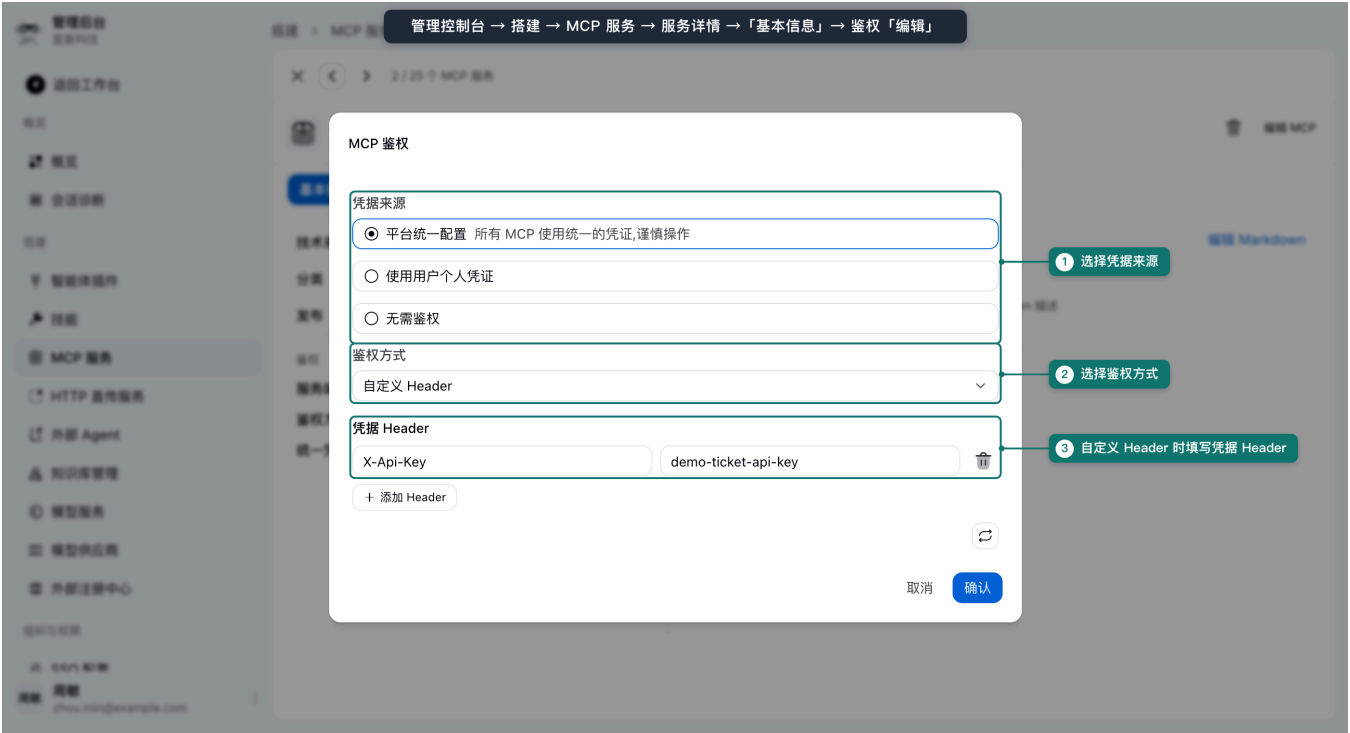
「凭据来源」有三种：

选项	含义	适用场景
「平台统一配置」	全部调用使用管理员保存的同一份企业凭据	公共数据源、企业统一账号
「使用用户个人凭据」	每位成员使用自己的外部账号或密钥，成员在能力中心完成授权（见第 11.8 节）	个人邮箱、网盘等需沿用外部系统权限的服务
「无需鉴权」	调用时不注入凭据	上游本身不要求认证

「鉴权方式」可选「设备码」「OAuth」「扫码授权」「Bearer 令牌」「API 密钥」「Basic Auth」「Cookie」「自定义 Header」和「统一连接」。OAuth 类方式需要上游支持相应授权流程。

请求头分两层配置：

- 「实例 Header」：在创建或编辑连接时配置，随连接固定携带，适合租户实例级的固定参数。
- 「凭据 Header」：鉴权方式选择「自定义 Header」时在鉴权表单中配置，按凭据保存，适合真正用于认证的信息。



同名 Header 以鉴权方式中的配置为准。Header 值和企业凭据按密文字段加密保存，保存后不会明文回显，编辑时留空表示保留原值。截图、培训材料、邮件和工单不得包含真实值。

11.3 发布与分发

「发布」开关位于 MCP 详情的「基本信息」中，列表中的状态显示为「已发布」或「未发布」。发布与分发的职责不同：

操作	作用	不代表什么
保存 MCP 服务	保存连接、工具和说明	不代表已经发布或授权
「发布」	服务出现在能力中心的「MCP 广场」中	不授予使用权限；未分发的成员看到「需要管理员分发」
编辑可用范围	向所有用户、部门、职能组或成员授予使用权	服务未发布时不能添加分发目标

服务未发布时，「可用范围」标签提示先在「基本信息」中发布。分发步骤如下：

1. 打开 MCP 详情的「可用范围」标签，选择「编辑可用范围」。
2. 在「可用范围」中按目标类型（「所有用户」「部门」「职能组」）筛选，或搜索成员，把需要授权的目标加入已选列表。部门行上的「下级」用于展开子部门；授权给部门时，其下级部门的成员同样获得授权。
3. 在「配置目标有效期」中设置「默认有效期」：「长期有效」「7 天」「30 天」或「自定义」。需要时勾选「为选中角色分配单独有效期」，为个别目标单独设置有效期，或选择「跟随默认」。
4. 核对已选目标与有效期后确认。

第一步：选择授权目标

可用范围

当前仅可搜索和添加个人可见的部门组织架构和人员信息。如需添加列表中未展示的部门，请输入部门的完整名称进行精确搜索

职能组

搜索

职能组

☐ 数据安全工作组
 ☒ 智能客服项目组
 ☐ 大模型应用专项组

选中全部

已选 1 个部门, 1 个职能组

运营中心

智能客服项目组 4 位成员

清空

1 按类型筛选：所有用户、职能组、部门

2 勾选目标，右侧核对已选列表

3 选择「下一步」

取消

下一步

每页行数 20

第 1 页 / 共 1 页 共 3 条

<<

<

>

>>

第二步：配置目标有效期

配置目标有效期

默认有效期

☒ 长期有效
 ☐ 7 天
 ☐ 30 天
 ☐ 自定义

4 选择默认有效期

☐ 为选中角色分配单独有效期

5 需要时为个别目标单独设置

取消

上一步

保存

6 确认后保存

每个目标显示「生效中」「待生效」或「已失效」状态。有效期到期后授权自动失效，平台约每 30 秒扫描一次到期授权。

新服务应从试点部门或专用职能组开始分发。临时项目使用有截止日期的授权，避免项目结束后授权继续保留。成员的实际可见性和使用权限以平台权限校验结果为准，不要假定某类角色天然绕过可用范围。

11.4 MCP 连接验证与排查

发布到目标范围前，至少完成以下验证：

- 「工具」标签中的工具数量和名称符合预期，没有大量漂移或过期工具。
- 使用无效参数时，工具返回明确错误而不是静默失败。
- 试点成员能够在「MCP 广场」获取连接配置并完成调用，非授权成员无法签发 Token。
- 需要限制的工具已在「路由策略」中设置调用规则（见第 11.5 节），范围外的调用被拒绝。

- 可用范围到期后，目标成员不再能够调用该服务。
- 「调用情况」中出现真实调用记录，「可用性探查」显示连接可用。

成员看不到或无法使用 MCP 时，依次核对发布状态、可用范围、授权有效期和成员组织归属。工具不可用时，核对服务端点、请求头、上游权限和发现结果，不得在排查材料中记录密钥。

11.5 设置工具调用权限

MCP 服务的工具调用权限在「路由策略」中配置。

1. 打开 MCP 详情的「基本信息」，在左侧信息栏的「路由策略」一行选择「查看适用的路由策略」。页面跳转到「路由策略 → 流量规则」的规则清单，并预填「调用类型」和「能力」两个筛选，列出显式命中该服务的规则和未限定能力的规则。
2. 选择「新建规则」，添加条件（条件之间为“且”关系）：
 - 「调用类型」选择「MCP 调用」，「具体能力」选择该 MCP 服务；
 - 「工具名」：比较方式为「属于」或「不属于」。已选择该 MCP 服务时，可在「选择已登记的工具」中勾选工具，也可在「通配模式」中输入带 * 的模式，例如 `delete_*`；
 - 「MCP 方法」：可选「调用工具」「列出工具」「资源」「提示词」，新建时默认「调用工具」；
 - 「组织成员」：在「部门（含下级）」和「职能组」中选择，两类取并集，比较方式为「属于」或「不属于」；选择上级部门时同时覆盖其全部下级部门；
 - 调用方是具体成员时，再添加「调用方」条件，比较方式同样可选「属于」「不属于」。
3. 添加动作。不允许调用时选择「拒绝本次调用」，调用方收到 403 和所填的拒绝原因码；需要按匹配规则或判断模型检查调用参数与返回结果时，选择「内容检测（护栏）」的「工具调用防护」，见第 14.3 节。
4. 保存规则，在「确认变更」中核对后选择「确认并应用」，再用「验证规则」和范围内、范围外的账号分别验证（第 15.3 节）。

限定“只有某部门可以调用某个工具”时，写成一条拒绝规则：「调用类型」为「MCP 调用」，「工具名」属于该工具，「组织成员」不属于该部门；需要例外放行的成员，在同一条规则中增加「调用方」「不属于」条件。仅开放指定工具时，写一条拒绝规则，「工具名」不属于这些工具。含「组织成员」条件的拒绝规则，在平台无法解析调用者的部门成员关系时按条件满足处理，即拒绝调用，不会放行。

路由策略在服务的可用范围之上生效：调用方须先在 MCP 服务的可用范围内（第 11.3 节），再满足规则。「角色管理」中的管理角色只决定管理控制台各功能的查看与管理，不决定谁可以调用工具。

升级前在 MCP 详情「安全策略」中配置的工具权限、参数内容规则、默认处置和请求体上限，升级时转换为名称为「MCP 护栏迁移 · 服务名 · 项」的路由策略，排在租户原有策略之前。名称末项为「需人工改写」的策略按更严格的方式转换，拒绝该服务的相应调用，原配置全文写在策略说明中；请据此改写规则后停用或删除该策略。

11.6 查看调用情况与可用性

调用情况：MCP 详情的「调用情况」标签可按「最近 1 小时」「近 24 小时」「近 7 天」「近 30 天」查看「调用趋势」，以及「调用量」「成功率」「平均耗时」「P95 耗时」「当前并发」和「限流」「熔断」状态。「最近调用」列出每次调用的「时间」「调用来源」「用户」「工具」「类型」「状态」「耗时」和「结果」，可打开「调用载荷」查看单次调用的请求与返回内容。

- 查看调用载荷需要部署时已配置报文归档存储；归档开启之前的调用没有留存报文。每次查看载荷都会写入一条审计记录。
- 限流、并发上限和熔断阈值在流量规则中配置（需要「动态路由」许可模块），MCP 作为一类调用与模型调用共用同一套规则，见第 15 章。

可用性探查：「可用性探查」标签记录服务端点的连通性。系统每 5 分钟自动检查一次，也可选择「检查连接」手动触发。记录列出「时间」「端点」「触发方式」（「系统检查」或「手动检查」）、「结果」和「延迟」，探查记录保留 30 天。

操作审计：「操作审计」标签记录对该服务的创建、修改、授权和凭据变更操作。

11.7 管理访问令牌与分类

Token 管理：「MCP 服务」页面的「Token 管理」标签集中列出成员为各 MCP 服务签发的访问令牌，列包括「名称」「类型」「用户」「MCP 服务」「状态」「签发时间」「到期时间」「最后使用」。

1. 在「搜索名称 / 用户 / MCP 服务」中搜索，或按状态筛选「有效（未过期、未撤销）」「已过期」「已撤销」。
2. 打开某个 Token 查看状态原因、来源、成功率、使用记录和最近活动。
3. 需要回收时，单个选择「删除」，或勾选多个后选择「批量删除」。删除即撤销该 Token，此后不能再用于调用。

签发用户失去签发权限后，Token 会显示不可用原因。

分类管理：「分类管理」标签用于维护 MCP 服务在目录中的分类。选择「添加分类」新建分类，为分类选择图标、修改「分类名称」，在右侧通过「添加应用」「移除」调整分类下的服务。选择「删除分类」时，该分类下的服务回到原始目录分类或未分类状态，服务本身不会被删除。

11.8 成员在能力中心获取 MCP 接入配置

「MCP 广场」列出本租户已发布的 MCP 服务。成员不在可用范围内时，卡片显示「需要管理员分发」，无法签发连接配置。获取连接配置的步骤如下：

1. 打开 MCP 服务详情，在「服务配置」中选择「获取连接配置」。未登录时显示「登录后获取连接信息」。
2. 服务使用「使用用户个人凭据」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，然后选择「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」和「自定义」，均不超过租户允许的最长有效期。
4. 按客户端要求复制 Streamable HTTP 地址或 JSON 配置。Token 仅显示一次，调用时使用当前成员身份，仅限个人使用。

租户可在「管理控制台 → 组织与权限 → 账号与会话安全」的「安全策略」中设置「允许类型」（「OAuth」「Token」）以及「默认令牌有效期」「最长令牌有效期」（1 至 168 小时）。开启 OAuth 后，成员可选择「显示 MCP 配置」，由 MCP 客户端按 Authorization Code + PKCE S256 完成登录授权；两种方式都关闭时，页面提示租户安全设置已关闭能力访问。

授权同意页：MCP 客户端发起 OAuth 授权时，网关作为授权方向成员展示「授权请求」页面，标题为「{客户端名称} 请求连接」，列出「回调目标」「完整回调 URI」「账号」和「访问范围」，并提示客户端名称由客户端自行提供、未经平台验证。成员核对回调地址和访问范围后选择「允许访问」或「拒绝」。请求超时后页面显示「授权请求已过期」，需回到客户端重新发起连接。

11.9 通过 MCP 管理网关

平台内置一个平台管理 MCP 服务，把管理控制台中的常用管理操作提供为 MCP 工具。管理员可在 Claude Code、Codex 等支持 Streamable HTTP 的 MCP 客户端中，通过对话或自动化任务完成日常配置。

该服务按调用者的真实身份执行：账号在页面上能执行的操作，在 MCP 中同样可以执行；页面上无权执行的操作，MCP 中返回同样的拒绝结果。每次调用经过与页面相同的认证、许可、变更控制和审计。

获取接入配置：

- 1. 在「MCP 广场」打开平台管理 MCP 服务的详情（描述以「平台管理 MCP」开头），选择「获取连接配置」。
- 2. 在「工具集」中勾选本次需要的功能分组。页面只列出当前账号和许可可用的工具集。
- 3. 选择「令牌有效期」后选择「确认签发 Token」，复制生成的配置写入 MCP 客户端。接入地址中的 `?toolsets=` 参数表示已选工具集，访问令牌以 `zg_mcp_` 开头。

当前开放的工具集如下：

工具集	适用角色	涵盖操作
「组织与用户」	租户管理员	部门、职能组、成员归属以及租户内用户账号与角色
「模型」	租户管理员	模型供应商、供应商账号或密钥、模型列表刷新以及模型服务
「MCP」	租户管理员	MCP 服务的连接探测、创建、导入、更新、删除与工具同步
「Skill」	租户管理员	Skill 来源分析，导入、修改、重新同步和删除
「外部智能体」	租户管理员、成员	外部智能体的注册、探测、审核和管理，或通过桥接发起对话与任务查询
「分发」	租户管理员	模型服务、MCP、Skill、外部智能体等能力的上下架与可用范围
「调用方」	租户管理员、成员	调用方应用的增删改与启停；网关模型 Key 的签发、停用、轮换、删除、转交所属人与绑定应用；消费组及其成员维护，以及按消费组批量轮换和批量设置到期时间
「配额」	租户管理员	租户内各对象的模型预算与并发配额、用量重置，以及模型 Key 路由与额度自检

工具集	适用角色	涵盖操作
「租户」	平台管理员	租户的创建、修改、启停和删除，租户管理员与租户总配额；仅对平台管理员显示，默认不勾选

使用限制如下：

- 调用未勾选工具集中的工具会被拒绝。
- 删除、停用和密钥轮换类工具采用两步确认：第一次调用不传 `confirm` 时只返回影响范围和确认提示，不执行写入；向用户展示影响并获得同意后，再以 `confirm: true` 调用同一工具才执行。
- 凭据只写不读：上游供应商密钥、外部 MCP 或 Agent 凭据、网关模型 Key 均无法通过 MCP 读回明文。创建或轮换模型 Key 时，明文只在该次响应中返回一次。

11.10 API 服务

API 服务以反向代理方式对外提供上游 HTTP 接口。网关统一完成调用方鉴权、授权分发、上游凭据注入和流量控制，请求和响应按原接口转发，不转换为 MCP 工具。需要供 MCP 客户端按工具调用时，应在「MCP 服务」中选择「由 HTTP 接口生成」。

页面包含「API 服务列表」「Token 管理」和「分类管理」三个标签。列表列包括「名称」「分类」「状态」「健康」「可见范围」和「添加时间」。

创建服务：

1. 选择「新建服务」，在「新建 API 服务」中填写「名称」「上游端点 URL」，按需添加「实例 Header」并选择「分类」，然后保存。
2. 需要向上游注入凭据时，在服务详情「基本信息」的「鉴权」区域选择「编辑」配置凭据。
3. 在「基本信息」中打开「发布」，再在「可用范围」标签按第 11.3 节的方法分发。

详情页标签包括「基本信息」「可用范围」「调用情况」「可用性探查」和「操作审计」，含义与 MCP 服务相同。可用性探查检查上游 HTTP 端点的连通性，每 5 分钟自动检查一次，记录保留 30 天。

调用方接入：成员在能力中心的「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」「HTTP 方法」「请求路径」「请求体」「访问令牌」和「curl 示例」。调用地址格式为：

<网关地址>/api/v1/agent-gateway/proxy/<服务 ID>/<上游路径>

调用方在请求头中携带 `Authorization: Bearer <访问令牌>`。不再使用时选择「撤销 Token」。租户在安全策略中关闭 Token 访问时，成员无法获取 API 服务的连接配置。

流量控制：API 服务在流量规则中是独立的调用类型「API 服务」，可配置限流和熔断（需要「动态路由」许可模块），见第 15 章。

12. 外部智能体、技能、插件与外部注册中心

本章说明「管理控制台 → 搭建」下的四个入口：「外部智能体」「技能」「插件」和「外部注册中心」。四项均属于「基础功能」，由租户管理员或「系统管理员」维护；成员在能力中心的「外部智能体广场」「技能广场」和「插件广场」页签中使用已授权的对象。

12.1 接入外部智能体

「外部智能体」用于接入由其他系统提供、遵循外部智能体（A2A）协议的 Agent 服务。页面包含「Agent 目录」和「任务中心」两个标签。「Agent 目录」列出「名称」「来源」「状态」「协议版本」「健康」「调用方」「接入时间」和操作。

- 1. 选择「接入外部智能体」，打开「接入外部智能体」向导。
- 2. 在「来源」步骤选择「名片 URL 拉取」并填写 Agent Card 地址（如 `https://agent.example.com/.well-known/agent-card.json`），或选择「粘贴 AgentCard JSON」。
- 3. 「凭据绑定」步骤只作提示，上游凭据在注册后于 Agent 详情中配置（见第 12.2 节）。
- 4. 在「连通测试」步骤选择「开始测试」，确认提示「连通正常」及延迟。
- 5. 在「完成」步骤核对「名片名称」「协议版本」和「签名校验」（「有签名」或「无签名」），然后完成注册。注册后 Agent 处于「待审批」状态。
- 6. 在「Agent 目录」对该 Agent 选择「批准」或「拒绝」。只有已批准的 Agent 才能签发调用方 Token 和分发给成员。

网关为每个外部智能体提供独立的调用入口，调用方无需知道上游地址：

项目	地址
调用入口（RPC）	<code><网关地址>/api/v1/agent-gateway/a2a/<Agent ID></code>
改写后的 Agent Card	<code><网关地址>/api/v1/agent-gateway/a2a/<Agent ID>/.well-known/agent-card.json</code>

改写后的 Agent Card 只保留展示字段，服务地址全部替换为网关地址，上游地址、文档与图标链接、扩展声明和签名不对外暴露。Agent 详情中并排展示「原始 Card」和「改写后 Card（网关入口）」，便于核对。

Agent 详情的「名片地址」旁可重新拉取名片。上游名片中的端点变更后，已保存的凭据绑定会暂存，需重新审批后才会再次注入。

从外部注册中心同步的 Agent 在列表中只能「查看Card」，不能单独删除，需通过注册源统一管理（见第 12.7 节）。手动接入的 Agent 删除后，其网关调用凭据同时失效。

12.2 配置上游凭据与调用方 Token

上游凭据：在 Agent 详情的「上游凭据」区域设置「凭据归属」：「企业共享」或「无需凭据」。「鉴权方式」由 Agent Card 声明决定，页面不能自选，支持「API 密钥」「基础认证」和「Bearer 令牌」。保存后，网关调用上游时注入该凭据；切换为无需凭据时，保存后解除绑定，调用时不注入任何凭据。

以下情况不能通过该区域配置：

- Agent 声明 OAuth2/OIDC 鉴权时，页面提示需单独配置；
- mTLS、Cookie 或查询参数中的 API 密钥不受支持。

调用方 Token：管理员可为外部调用方签发访问令牌。

1. 在 Agent 详情的「调用方 Token」区域选择「签发」。
2. 在「签发调用方 Token」中填写「标签」（如 `acme-workbuddy`），选择「有效期」：1 天、7 天、30 天、90 天或「永久有效」，默认 30 天。
3. 复制「Token 明文」。明文仅此一次展示，关闭后无法找回；令牌以 `zg_a2a_` 开头。
4. 列表显示「标签」「前缀」「过期时间」「最近使用」「状态」和「签发时间」。不再使用时选择「吊销」，吊销后最多 15 秒内全站生效，不可恢复。

改写后的 Agent Card 声明两种调用鉴权：网关签发的 Bearer 令牌，以及授权码加 PKCE S256 的 OAuth 授权。

12.3 分发外部智能体与发布为 MCP 工具

发布与可用范围：Agent 详情的标签包括「详情」「可用范围」「调用情况」和「操作审计」。在「详情」中打开「发布」，把 Agent 上架到能力中心的「外部智能体广场」；再在「可用范围」中按「所有用户」「部门」「职能组」或成员授权，并设置「开始生效」「结束生效」，预设为「7 天」「30 天」「自定义」「长期有效」。「外部智能体广场」只向已获授权的成员展示已批准的 Agent。

外部智能体的部门授权按成员的直属部门匹配。需要覆盖子部门成员时，应同时添加子部门或改用职能组授权。

发布为 MCP 工具：在 Agent 详情打开「发布为 MCP 工具」，状态显示为「已发布」。发布后，成员可在 MCP 客户端中通过网关提供的外部智能体 MCP 桥接调用该 Agent。关闭后，已有的 MCP 配置立即不可调用。

成员接入：成员在「外部智能体广场」打开 Agent 详情，按客户端类型选择接入方式：

区域	适用客户端	操作
「原生协议接入」	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」，配置包含 Card 地址、调用地址和一次性展示的 Bearer 令牌；可「复制配置」「下载配置」或「撤销 Token」
「WorkBuddy / MCP 接入」	MCP 客户端；仅在 Agent 已发布为 MCP 工具时显示	选择「获取 MCP 配置」，把生成配置中的 <code>mcpServers</code> 条目合并到客户端配置文件，不要覆盖其他已有条目

租户在安全策略中关闭 Token 接入时，成员不能生成外部智能体或 MCP 调用配置。

12.4 监控外部智能体与查询任务

健康与探活：Agent 详情的「健康」显示「正常」「不可达」或「未知」。需要立即确认时选择「立即探活」。

调用情况：「调用情况」标签展示「调用趋势」，以及「调用量」「成功率」「平均耗时」「P95耗时」「当前并发」和「限流」「熔断」状态；「最近调用」列出每次调用的时间、调用来源、用户、状态、耗时和结果。外部智能体调用的限流与熔断在流量规则中配置（需要「动态路由」许可模块），见第 15 章。

操作审计：「操作审计」标签记录该 Agent 的注册、审批、凭据和分发变更。

任务中心：「任务中心」标签列出经网关发起的外部智能体任务。

- 1. 按「全部 Agent」「全部状态」筛选，或在「按 taskId 搜索」中输入任务标识。
- 2. 列表显示「Task ID」「目标 Agent」「调用方」「状态」「开始时间」「耗时」和「最近更新」。状态包括已提交、执行中、待输入、待授权、已完成、失败、已取消和被拒绝等。
- 3. 展开任务查看「状态节点」，核对状态变化历史。
- 4. 需要终止时选择「取消任务」。网关会向目标 Agent 发送取消请求，能否终止取决于对端实现。

外部智能体接入完成的标志是：名片可读取、已批准、探活正常、凭据与端点匹配、可用范围正确，并且「调用情况」和「任务中心」能看到真实调用。某一步成功不能替代其他步骤。

12.5 导入与分发技能

「技能」页面管理本租户可分发给成员的技能（Skill）。列表可按「状态」「分类」筛选，或在「搜索技能或分类」中搜索。

导入技能：选择「添加技能」，在同一对话框中选择一种来源：

来源	操作
公共目录搜索	在「搜索 registry」中输入关键词，搜索对接 skills.sh 公共目录
来源坐标或地址	在「技能来源」中填写名称或地址，支持 skills.sh 名称（如 react-best-practices）、owner/repo@skill、gitlab:group/repo、github.com/owner/repo#v1.2、ClawHub 地址、站点发布的 .well-known 技能清单地址或压缩包下载地址
本地上传	在「或上传技能包」中选择 .zip 文件，压缩包内须包含 SKILL.md

- 1. 填写来源后选择「分析来源」。上传 .zip 只触发风险分析，不会立即安装。
- 2. 页面列出发现的候选技能，以及每个技能的风险等级、检查清单和准入结论。可选择「查看内容」预览技能文件。
- 3. 对允许安装的技能选择「确认安装」。

单个下载包上限为 20 MiB。内网环境无法访问公共目录时，需由部署管理员把搜索与下载地址配置为私有镜像（部署参数 LP_SKILLS_SEARCH_API_BASE、LP_SKILLS_DOWNLOAD_BASE、LP_CLAWHUB_API_BASE）。

安装前检查：所有来源在安装前经过同一套检查，覆盖 SKILL.md 与附件中的提示注入、凭据外泄、破坏性命令、路径穿越，以及可执行文件与脚本、依赖安装、Shell 执行、外部网络访问、宽泛权限与敏感数据等十五项内容。检查基于规则匹配，不做语义分析。

检查结果按六级风险判定：「纯指南」「知识资产」「受管集成」「本地执行」「外部 / 高敏」「禁止级」。准入结论如下：

结论	含义
「允许安装」	风险等级在租户准入线内
「企业策略拒绝」	风险等级超出租户准入线
「安全规则阻断」	命中提示注入、凭据外泄、破坏性命令或路径穿越等禁止项，判为「禁止级」，一律阻断
「扫描不完整，按高风险处理」	检查未能完整执行
「分析失败」	来源无法解析或下载

页面以风险标尺展示「你所在企业的准入线」和该技能超出准入线的档数。租户准入线默认为「外部 / 高敏」，不能设为「禁止级」，管理员导入同样受准入线约束。调整准入线通过开放 API。

分发：技能详情的标签包括「信息」「可用范围」和「下载记录」。在「信息」中打开「发布」，再在「可用范围」中按所有用户、部门、职能组或成员授权，并设置生效与失效时间，到期自动失效。授权给部门时，其下级部门的成员同样获得授权。

下载记录：成员在能力中心「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。管理员在技能详情的「下载记录」中查看每次下载的「时间」「下载方式」（Registry 拉取或手动下载 zip）和「下载人」。

维护：在技能详情选择「编辑」修改「名称」「描述」「分类」，或选择「从来源重新导入」获取来源的最新版本。选择「删除」时，页面列出已分发目标数量及引用关系，需二次确认；删除后技能无法恢复，相关分发一并移除。

12.6 导入与导出插件

插件把技能和远程 MCP 服务打包为一个插件包，供外部智能体工具体安装。「插件」页面列出「插件」「状态」「更新时间」和操作，工具栏提供「导入」「导出为插件」和「刷新」。

导入插件包：

1. 选择「导入」，在「导入插件包」中选择「插件包(.zip)」。需要参考格式时选择「下载示例包」。
2. 选择「预览」，核对将要导入的组件及其是否受支持。
3. 选择「导入」。包内的技能组件与技能导入使用同一套安装前检查，超出租户准入线或命中禁止项的技能会使导入被拒绝。

导出为插件：

1. 选择「导出为插件」，在「导出为 Agent 插件」中填写「插件名」「版本」「描述」和「许可证」。
2. 勾选要打包的「技能」和「MCP 服务」。只有远程（Streamable HTTP 或 SSE）MCP 服务可导出，其他服务显示「不可导出」；名称含非 ASCII 字符的技能需手动填写「技能 slug」。
3. 选择「预览」核对内容，然后选择「导出」下载 zip，或选择「保存到平台」把同一份包保存到本租户的插件列表。

导出的插件包不含任何凭据：MCP 服务只带服务地址，请求头为空，凭据只以需求声明的形式写入，目标端需自行配置。打包前平台还会检查请求头名称和取值，发现疑似凭据时拒绝导出。

分发与发布：插件详情的标签包括「概览」「组件」「来源与供应链」和「分发」。

- 在列表中对插件选择「分发」，按可用范围向成员授权。
- 在详情的「分发」标签中选择「申请发布」，把插件发布到内部分发目录，状态依次为「审核中」「已发布」。发布申请需经平台审核；插件在发布后又更新时状态为「待重新发布」，需「重新申请发布」。不再提供时选择「下架」，之前发出的安装链接随即失效。

发布本身不授予使用权限，成员能否使用仍由可用范围决定。成员在能力中心「插件广场」页签打开插件详情，可「复制个人插件源链接」用于安装。

12.7 同步外部注册中心

「外部注册中心」用于接入 Nacos 命名空间，按轮询间隔把其中的外部智能体和 MCP Server 同步到网关。当前仅支持 Nacos。

1. 选择「新建源」，在「新建外部注册源」中填写「名称」「地址」「命名空间」和「轮询间隔（秒）」。
2. Nacos 需要认证时填写「用户名」和「密码」，两者必须成对填写；匿名访问可留空。
3. 选择同步对象：「同步外部智能体」「同步 MCP Server」。
4. 按需打开「自动批准（信任源）」。
5. 选择「保存」，然后选择「测试连接」。
6. 在列表中选择「立即同步」，或等待按轮询间隔自动同步。

同步规则如下：

- 同步进来的外部智能体和 MCP 服务仍需在各自的可用范围中显式分发，成员才能看到。「自动批准」不改变这一点。
- 同一注册源的同一条目重复同步时更新原记录。被拒绝的条目在下次同步时回到待审批状态。
- 上游条目连续两次轮询都不存在时才视为删除：外部智能体被停用，MCP 服务被归档并吊销其网关令牌。单次同步结果为空或删除超过一半时，平台暂停删除并标记为降级。
- 删除注册源会吊销其同步进来的 MCP 网关令牌，并移除已同步的外部智能体和 MCP 服务；之后可重新接入。

13. 配置敏感数据（PII）检测

敏感数据检测分两处配置：

- 「管理控制台 → 护栏 → 敏感数据」维护可识别的敏感数据规则。「护栏」是侧栏「护栏」分组下的同名菜单，页面包含「敏感数据」「词库」「问答」「内部系统白名单」四个页签。
- 「管理控制台 → 路由策略 → 流量规则」中的「内容检测」动作决定哪些调用检测哪些类型、命中后如何处置。

护栏页只定义能识别什么。没有任何流量规则引用时，敏感数据规则不会对调用生效。

前提	说明
角色	租户管理员或「安全管理员」
许可模块	「护栏」页需要「模型护栏」；在「流量规则」中启用检测还需要「动态路由」
依赖配置	识别人名、地址、机构名时，需在「管理控制台 → 系统设置 → 全局设置」的默认模型中设置「敏感信息识别」模型

13.1 敏感数据页签与内置规则

进入「管理控制台 → 护栏 → 敏感数据」。工具栏提供搜索、筛选（「状态」「分组」）、「导入」「导出」「试运行」和「添加PII策略」。列表列为「规则名称」「正则」「类型」「来源」：规则名称后显示所属分组；「来源」为「内置」或「自定义」。



平台内置二十多条规则，分为四个分组：

分组	覆盖内容
通用	身份证号、手机号、银行卡号、电子邮箱、固定电话、护照号、车牌号
金融	统一社会信用代码、纳税人识别号、银行账号、IBAN、证券账户
医疗	病历号、医保卡号、患者编号
密钥	API 密钥、Access Token、密码、私钥，以及 sk-style API 密钥、AWS Access Key ID、GitHub PAT 等常见密钥格式

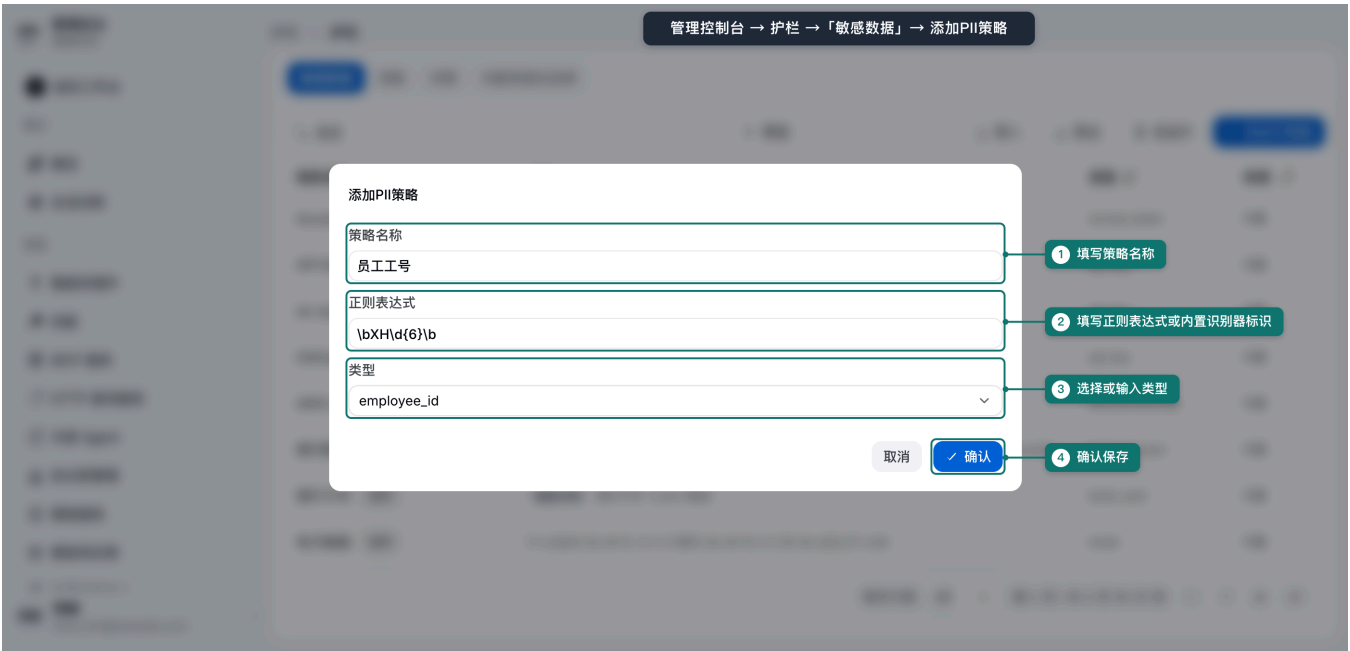
身份证号、手机号、银行卡号三条规则在「正则」列标为「智能识别」，分别使用「身份证号（校验位+出生日期）」「手机号（号段+长度校验）」「银行卡号（Luhn 校验）」算法校验，用于减少误报。纳税人识别号、银行账号、证券账户、病历号、医保卡号、患者编号等格式松散的规则，要求号码附近出现中文或英文标签（例如“病历号”“MRN”）才会命中。

内置规则由平台维护并随版本更新。选择内置规则打开「查看PII策略」弹窗，内容只读，不能修改或删除。本页不提供单条规则的启停开关：一条规则是否参与检测，取决于流量规则选中的类型（见 13.4 节）。

人名、地址、机构名没有固定格式，不提供正则规则，由敏感信息识别模型识别（见 13.4 节）。

13.2 新增、导入与导出自定义规则

1. 选择「添加PII策略」，打开同名弹窗。
2. 在「策略名称」中填写能说明数据含义的名称，例如“客户编号”。
3. 在「正则表达式」中填写匹配规则。也可以填写内置识别器标识复用算法校验，例如 `lp-recognizer:cn_mobile@v2`（手机号）、`lp-recognizer:cn_id_card@v2`（身份证号）、`lp-recognizer:bank_card@v2`（银行卡号）。
4. 在「类型」中选择已有类型，或输入新的类型标识。流量规则按类型引用规则，同一类型可以对应多条规则。
5. 选择「确认」。



保存时服务端校验正则表达式，不合格的规则不会保存，页面提示失败原因。限制如下：

- 不能为空，长度不超过 500 个字符；
- 不支持前后行断言（look-around）；
- 重复量词不超过 10 个，重复嵌套不超过 2 层；
- 编译后体积超出执行引擎上限的表达式会被拒绝。

选择自定义规则所在行打开「编辑PII策略」，可修改名称、正则表达式和类型，或选择「删除」。使用内置识别器标识的规则在弹窗中不提供「删除」。

「导出」生成 pii-rules.json，包含当前规则。「导入」只接受同格式的 JSON 文件。导入时与平台版本不一致的内置规则会被跳过，页面列出「这几条内置规则与平台不一致，已跳过，平台规则未改」及不一致的内容；导入中断时，可在同一弹窗选择「重试」从中断处继续。

规则上线前，应使用脱敏样本覆盖“应命中、边界值、不应命中”三类文本，并通过 13.3 节的试运行核对。表达式过宽会造成误报和业务阻断，过窄会造成漏检。

13.3 试运行

选择「试运行」，打开「敏感数据试运行」：

- 1. 在「要检测的类型」中选择一个或多个类型。
- 2. 在「检测到敏感信息」中选择处置方式：「仅记录，放行」「脱敏后继续处理」或「拦截」。
- 3. 在「要检查的文本」中粘贴样例，选择「检查」。

结果显示结论（例如「已拦截」）和命中说明，与请求详情中的护栏结论使用同一种展示。试运行使用与网关相同的规则求值，只读，不修改配置，不影响线上调用。选择人名、地址或机构名时，试运行会调用敏感信息识别模型，未设置该模型时无法得到这几类结果。

13.4 在流量规则中启用检测与命中处置

在「管理控制台 → 路由策略 → 流量规则」中新建或编辑规则，选择「+ 添加动作」→「内容检测」，将卡片中的「检测方式」设为「PII 敏感数据」。规则的匹配条件、执行顺序和发布方法见第 15 章。

检测范围。「检测范围」分为「模板」和「单独类型」两页，可同时选择。模板是一组类型，单独选择的类型追加到检测范围，重复类型只检测一次。平台提供四套模板：

模板	包含的类型
通用	通用分组与密钥分组的类型
证券	通用、金融、密钥分组的类型
医疗	通用、医疗、密钥分组的类型
研发安全	电子邮箱与密钥分组的类型

命中处置。卡片中的「命中后的处理」与规则出口「检测到敏感信息」是同一设置，可选「仅记录，放行」「脱敏后继续处理」「拦截」。处置按以下顺序确定：

- 在「单独类型」中点名的类型，按本规则选择的处置执行；
- 只通过模板带入的类型，按规则自身的默认动作执行。内置规则和新建的自定义规则默认动作均为脱敏，这类类型命中后按「脱敏后继续处理」处置；
- 人名、地址、机构名没有规则行，按本规则选择的处置执行。

需要对某个类型拦截或仅记录时，应在「单独类型」中逐个选中该类型，不能只依赖模板。

一次请求命中多条带内容检测的规则时，检测类型合并，同一类型取更严格的处置（拦截高于脱敏，脱敏高于仅记录）。

人名、地址、机构名。这三类由「全局设置 → 默认模型」中的「敏感信息识别」模型识别，需选用支持敏感信息识别用途的模型服务。未设置时，规则编辑器中这三类显示「需要先选模型」，不可选；仍有规则选用它们时，全局设置不允许清空该模型。模型识别结果与正则结果合并后按上面的顺序处置。模型调用失败时按规则的「组件不可用时」设置处理，见 14.4 节。

检测阶段。内容检测默认同时检测请求输入和模型输出，阶段设置见 14.2 节。

13.5 脱敏结果与审计

- 脱敏。处置为「脱敏后继续处理」时，命中内容在发往上游前替换为租户绑定的加密占位符（形如 [PIIREF:v1:...]），原值不发送给模型。业务系统经 API 密钥调用时，网关不会把模型回复中的占位符还原为原值。
- 审计记录。命中审计只记录数据类型、命中次数、策略来源和处置动作，不包含请求原文、规则名或命中值。
- 原文证据。命中的原文以加密形式保存。租户管理员可在请求详情的「护栏」区选择「查看原文」，在「护栏原文证据」中查看；每次查看都会写入操作审计。
- 验证。用测试调用触发规则后，在请求详情的「护栏」区核对结论，例如已脱敏的类型与次数，或已拦截的类型。

13.6 检测状态与总开关

敏感数据规则只定义“能识别什么”。一次调用是否执行敏感信息检测、检测输入还是输出、命中后拦截还是脱敏，都在路由策略规则的内容检测段中设置；只建规则、没有路由策略引用时，规则不参与任何调用。

检测状态。页面顶部始终显示本租户的检测状态：

- 生效中：至少有一条已启用的规则。被路由策略引用的规则会参与判定。
- 未生效：没有已启用的检测规则，同一位置给出该原因。添加或启用规则后检测立即生效。

状态行下方显示规则数（含启用数）、自建词库数，以及“被 N 条路由策略引用”。“被 N 条路由策略引用”统计的是护栏段引用了敏感信息检测或词库、且处于启用状态的路由策略规则，点击即可跳转到路由策略列表并只显示这些规则。

没有检测总开关。敏感数据检测是否生效只取决于是否有已启用的规则：规则启用即参与判定，停用即不参与。需要“先建规则、暂不生效”时，保存后停用该规则即可。

特别注意：请版本 2026.10 的检测规则启用状态升级前关闭了总开关的租户，升级后只要有已启用的规则，检测即生效。

14. 配置内容检测与安全围栏

内容检测是「流量规则」中的一个动作：规则匹配到调用后，按动作中选择的检测方式检查请求输入和模型输出，并按配置拦截、记录或放行。检测所用的词库、问答和内部系统白名单在「管理控制台 → 护栏」中维护，敏感数据规则见第 13 章。本章另外说明平台级的 SSRF 出站防护和入口 Web 应用防护。

前提	说明
角色	租户管理员或「安全管理员」；「SSRF 出站防护」仅平台管理员可见
许可模块	「护栏」页及其检测数据需要「模型护栏」；在「流量规则」中添加内容检测需要「动态路由」；SSRF 出站防护和 Web 应用防护属于「基础功能」
依赖配置	风险类目与提示注入检测需要「内容安全」模型；必答需要「问答语义匹配」模型；人名、地址、机构名需要「敏感信息识别」模型；三者都在「管理控制台 → 系统设置 → 全局设置」的默认模型中设置

14.1 维护词库、问答与内部系统白名单

词库。进入「管理控制台 → 护栏 → 词库」，页面按「内置黑名单」「自定义黑名单」「白名单」分页签显示词库。

- 内置黑名单按六类组织：政治敏感与国家安全、违法暴恐与虚假谣言、色情低俗与不良价值观、歧视偏见与仇恨言论、隐私侵权与商业合规、高风险领域应用限制。中文词条共两万多条，类目参照 GB/T 45654 对齐扩充；另附除政治类以外的英文词库。列表名称显示为“类别(中文)”或“类别(英文)”。内置词库随版本更新，可以查看、搜索和导出，不能修改。
- 选择「新建词库」，填写「名称」，在「类型」中选择「自定义黑名单」或「白名单」，在「分类」中选择六类之一。
- 打开词库后选择「添加词条」，填写「词条」和「别名」。别名用逗号分隔，与词条命中同一条规则。
- 「导入」接受 UTF-8 编码的 .txt 文件：每行一个词条，词条后可接一个制表符和逗号分隔的别名；单次最多 20,000 条，同一文件中的重复词条会被跳过，导入后页面提示导入与跳过的条数。「导出」生成同格式的 .txt 文件。
- 白名单中的词条用于豁免黑名单命中，降低正常业务用语的误报。

问答。进入「管理控制台 → 护栏 → 问答」，维护必答题库。命中时网关直接返回预设答案，不再调用模型。

- 列表列为「标准问」「分类」「分组」「相似问」「来源」，可按分组和分类筛选。
- 选择「新建问答」，填写「标准问」「答案」，按需「添加相似问」，并填写「分类」「分组」「来源」。未分类的问答对所有分类生效。
- 「导入」接受 .jsonl 或 .txt 文件，「导出」生成 safety-qa-items.jsonl。
- 平台内置七千多条问答，带「平台内置 · 随版本更新」标签，只读。
- 问答阈值随流量规则设置，见 14.2 节。识别意思相近的问法依赖「问答语义匹配」模型。

内部系统白名单。进入「管理控制台 → 护栏 → 内部系统白名单」，选择「新建名单」。在「新建白名单」中填写「名称」，并按需填写「内部域名」「代码仓库」「存储桶」「内网主机」「镜像源」「敏感数据位置」，每行一条。列表显示各栏条目数、「被引用」和「最近更新」；被规则引用的名单不能删除，需先在规则中取消引用。

名单供 14.3 节工具调用防护的模型判断使用：判断模型把名单中的目标视为公司内部系统，不在名单中的一律视为外部；从「敏感数据位置」向外拷贝视为外泄。名单是判断模型的依据，不是确定性的放行规则。

14.2 在流量规则中添加内容检测

在「管理控制台 → 路由策略 → 流量规则」中新建或编辑规则，选择「+ 添加动作」→「内容检测」。动作卡片标题为「内容检测（护栏）」，在「检测方式」中选择一种方式；需要多种方式时再次添加「内容检测」，每种方式一张卡片。规则编辑器的通用操作见第 15 章。

检测方式	主要配置	依赖
PII 敏感数据	「检测范围」（模板与单独类型）、「命中后的处理」，见 13.4 节	人名、地址、机构名需要「敏感信息识别」模型
词库	在「模板」中选择引用的词库（可多选），以及 14.3 节的规避写法设置	无模型依赖
必答	在「必答条目」中选择引用的问答（可多选）	「问答语义匹配」模型，未设置时不可选，保存会被拒绝
风险类目	在「模板」中选择风险分类（六类，可多选）、「提示注入模式」	「内容安全」模型，未设置时不可选
工具调用防护	见 14.3 节	规则的「调用类型」须限定在「模型调用」「MCP 调用」「API 服务」之内；其中模型判断只作用于模型调用

内容安全模型。在「管理控制台 → 系统设置 → 全局设置」的默认模型区域选择「编辑」，在「内容安全」一行选择模型，并在「模型类型」中选择「通义千问安全护栏」「阿里安全推理护栏」「结构化判断模型」或「自定义内容安全检测服务」。保存时平台用一句违规样例和一句安全样例实际调用核对，核对失败不保存。所选模型不识别某些风险分类或提示注入时，规则编辑器会在对应字段下方说明，选择这些分类不会命中。

风险类目与提示注入。

- 「提示注入模式」用于越狱与提示词注入检测，默认「拦截」，可改为「仅记录，放行」或「放行，不做处理」。
- 风险分类命中后的处置在规则出口「检测到风险内容」中选择：「拦截」「仅记录，放行」或「放行，不做处理」。新加入的分类默认启用，语义判定阈值为 0.5，明确命中与疑似命中均为拦截。
- 需要逐类设置处置或阈值时，在规则编辑器的「YAML 编辑」中修改 `action.guard.riskCategories` 各项的 `exactAction`、`suspectedAction`、`semanticThreshold`。阈值越低，检测越敏感，误报也越多。
- 内容安全模型只给出“安全、有争议、不安全”三档判断时，阈值不高于 0.5 拦截有争议及以上的内容，高于 0.5 只拦截不安全的内容。

×

客服内容安全检测

草稿

流量规则 → 规则编排 → 添加动作「内容检测（护栏）」

有未保存的修改

✓ 配置正常

▷ 试运行

取消

保存规则

🏠 规则编排

从条件到动作，看清每一次请求的去向

YAML 编辑

🔄

📄

收起节点

○ 内容检测（护栏）

动作类型

内容检测（护栏）

动作参数

检测方式

风险类目

提示注入模式

拦截

检测到风险内容

拦截

执行阶段

before_upstream

1 检测方式选择「风险类目」

2 选择风险分类

3 提示注入模式的处置

4 检测到风险内容时的处置

模板

+3

请选择

当前内容安全模型不识别：高风险领域应用限制，选择这些类别不会命中

+

处理结果

按配置完成处理，返回响应

○ 异常处理

组件不可用时

放行，继续网关路由

仅在执行组件不可用时触发。内容检测命中、触发限额时，按对应动作中设置的策略处理。

5 护栏检测出错的处置

100%

▼

流程总览

调试与预览

▷ 试运行

📄 配置检查

在本地预览条件与动作，保存后可验证实际执行。

请求事实

示例请求

模型服务

输入本次请求的事实值

模拟情景

正常请求

▷ 运行测试

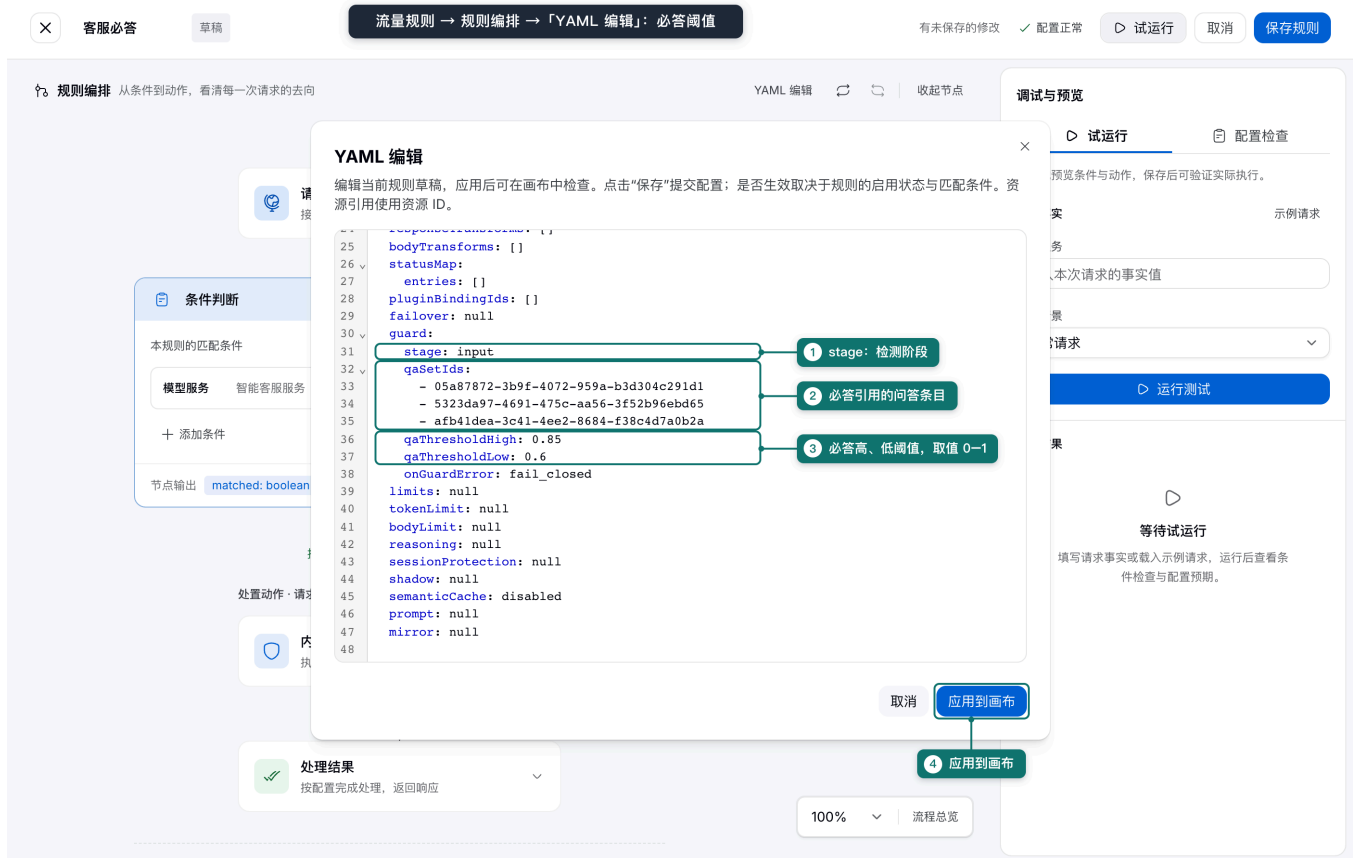
运行结果

▷

等待试运行

填写请求事实或载入示例请求，运行后查看条件检查与配置预期。

必答阈值。相似度达到高阈值时返回预设答案；介于低、高阈值之间时记录审计并继续后续检测；低于低阈值时不代答。规则未设置阈值时，高、低阈值均为 0.75。阈值在「YAML 编辑」中设置 `action.guard.qaThresholdHigh` 与 `action.guard.qaThresholdLow`，取值 0-1。必答只作用于请求输入。



检测阶段。内容检测默认检测「输入+输出」。需要只检测输入或只检测输出时，在「YAML 编辑」中将 `action.guard.stage` 改为 `input` 或 `output`。

图片内容与多轮对话。在「YAML 编辑」中配置 `action.guard.image` 后，输入中的图片先由图片理解模型转写为描述和图中文字，再与上下文一起按词库、敏感数据和风险类目等检测项判定。`model.gatewayModelId` 选择支持图片输入的模型，`allowedFetchHosts` 填写允许网关拉取图片的内网主机名，未填写时只拉取公网图片。所选模型为外部服务时，图片会发送至该服务。`action.guard.multiTurn` 提供三项检测：`reassembleShortMessages`（拆分输入检测，将连续短消息按用户消息顺序拼接后检测）、`cumulativeOutput`（多轮回答合并检测）和 `sessionEscalation`（会话多次命中限制：`window` 取 `m10`、`m30`、`h1`、`h6`、`h24`，`threshold` 取 2 到 20，`response` 取 `reject_session` 或 `treat_suspected_as_exact`，`duration` 的取值同 `window`）。以上均默认不开启。验证规则时，可在「验证规则」的「消息」中为每条消息「附图」并选择「检测方向」。无法读取内容的文件（PDF 等文档类附件、以文件编号引用而非直接携带字节的图片）不做检测，照常放行并在「路由策略命中」中留「未检测」记录。判定细节见《Zentrix Gateway 路由策略用户手册》第 5.2 节。

多条规则叠加。一次请求命中多条带内容检测的规则时，检测项合并执行：检测阶段取并集，同一分类或类型取更严格的处置与更低的阈值，必答阈值取较低者。

14.3 规避写法识别与工具调用防护

规避写法。词库匹配前，网关对词条和待检文本执行相同的归一化，始终开启：转小写、全角转半角、繁体转简体、字母变体与同形字归一、连续三次及以上的重复字符压缩为一个，并跳过空白、零宽字符和常见分隔符号。拼音全拼和声母缩写按疑似命中记录。

词库卡片上另有四项设置：

设置	作用	默认值
「字符插桩命中后」	只有经过归一化才命中时的处置	「拦截」
「编码解码」	识别 Base64、Hex 和 URL 编码片段，解码后再匹配	关闭
「解几层」	解码层数，可选 1~3 层，层数越多 CPU 开销越大	2 层
「解码后命中」	只有解码后才命中时的处置	「拦截」

×

客服词库检测

草稿

流量规则 → 规则编排 → 「内容检测（护栏）」 → 检测方式「词库」

有未保存的修改

✓ 配置正常

▷ 试运行

取消

保存规则

🔗 规则编排

从条件到动作，看清每一次请求的去向

YAML 编辑

收起节点

○ 内容检测（护栏）

动作类型
内容检测（护栏）

动作参数
检测方式
词库
模板
商业机密词库 ×
隐私侵权与商业合规(中文) ×

字符插桩命中后
拦截
编码解码
2层

解几层
2层
解码后命中
拦截

执行阶段
before_upstream

1 检测方式「词库」与引用的词库

2 字符插桩命中后的处置；编码解码开关

3 解几层与解码后命中中的处置

+

✓ 处理结果
按配置完成处理，返回响应

异常处理
组件不可用时放行，继续网关路由

100%

▼

流程总览

调试与预览

▷ 试运行

📄 配置检查

在本地预览条件与动作，保存后可验证实际执行。

请求事实

示例请求

模型服务

输入本次请求的事实值

模拟情景

正常请求

▷ 运行测试

运行结果

▷

等待试运行

填写请求事实或载入示例请求，运行后查看条件检查与配置预期。

两项处置均可选「拦截」「仅记录，放行」或「放行，不做处理」。业务通过网关传输 Base64 附件或签名时，开启「编码解码」可能产生误报，建议先在受控范围选择「仅记录，放行」观察，核对误报后再改为「拦截」。

工具调用防护。工具调用防护对模型发起的工具调用、MCP 调用和 API 服务调用生效，规则的「调用类型」须限定在这三类之内。它由「匹配规则」和「模型判断」两部分组成，可单独或同时配置：

- 「匹配规则」（对模型调用、MCP 调用、API 服务均生效）：先于模型判断执行，只做字面匹配，不调用模型。选择「添加匹配规则」后填写：「名称」；「工具」（支持 * 通配，如 *_file）；「检查位置」（「模型发出的调用参数」或「工具返回的结果」，API 服务分别检查请求体与响应体）；「字段」（留空表示检查全部字段，多层用点号分隔，如 args.content）；「匹配方式」（「正则」或「关键词」，关键词不区分大小写与全角、半角）和「匹配内容」；「例外」（只豁免与例外重叠的那一处命中）。每条规则可填写示例后选择「测试」，查看「命中」「未命中」及命中的字段与片段。每个工具调用防护最多 32 条匹配规则，匹配内容不超过 4096 字节。代码执行类场景建议同时启用模型判断。规则中的工具名在近 30 天没有出现过时，保存前提示核对工具名。

- 「模型判断（仅作用于模型调用）」：由判断模型判定工具调用是否危险，策略的调用类型不含模型调用时不显示。配置项：「检测范围」（「仅检测执行命令类工具」为默认，或「全部工具调用」）；「额外工具名」（追加需要审查的工具名，支持 * 通配）；「拦截类别」（共十一类：从网上下载脚本直接运行、把敏感数据发到公司外部、大批量或不可恢复的删除、改生产环境、改共享基础设施、扩大权限、暴露密钥或令牌、逃出隔离环境、破坏审计或监控、修改流量目标、改写历史记录）；「自定义类别」（选择「添加自定义类别」，填写「名称」「判定标准」，并各至少填写一条「正例」「反例」，各最多 8 条，可选择「测试」核对判对数；最多 32 个）；「判断模型」（选择了拦截类别或自定义类别后必须选择；选择「评测」可运行内置评测用例，查看准确率与误拦数，每条用例会实际调用一次所选模型）；「内部系统白名单」（引用 14.1 节的名单）。
- 「命中后」：「仅记录」（默认）或「拦截」。
- 「向调用方返回拦截说明」：默认勾选，「命中后」为「仅记录」时不可用。

「命中后」选择「拦截」时：模型发出的调用命中，网关去掉该调用并通知模型改用其他方式；MCP 调用参数或 API 服务请求体命中，拒绝本次调用；工具返回结果或 API 服务响应体命中，替换为「该内容已被拦截」。对模型发出的调用，网关向同一目标重新请求并提示模型换一种做法；同一次调用被拦截 3 次后不再重试，网关移除被拦截的工具调用，并提示联系管理员。命中记录在「概览 → 路由策略命中」中显示依据（匹配规则名称或类别名称），匹配规则命中可在「更多」中选择「在策略中添加例外」。

14.4 失效处理、拒答话术与流式输出

失效处理。规则出口「组件不可用时」的「护栏检测出错」决定检测无法完成时的处理，可选「放行，继续网关路由」或「拒绝本次调用」。以下情况按该设置处理：

- 网关上的护栏配置已过期；
- 规则引用的词库、问答条目或敏感数据模板在网关上缺失；
- 内容安全模型或敏感信息识别模型未设置或暂时不可用；
- 送检内容超过单次检测上限，有一部分未送检；
- 命中原文证据未能保存。

问答语义匹配模型不可用时只退回原句匹配，不按此项处理。该项由规则逐条设定，保存规则前逐条确认；高风险、强监管或包含敏感数据的正式业务设为「拒绝本次调用」。多条规则叠加时，只要有一条为拒绝，就按拒绝处理。按拒绝处理时，非流式调用返回 HTTP 503，输入检测阶段出错时不再请求模型；流式输出在出错处中止。

拒答话术。拦截时返回给调用方的拒答话术默认为“很抱歉，我无法回答您的问题。”。需要自定义时，在「YAML 编辑」中设置 `action.guard.refusal.default`，并可在 `action.guard.refusal.byCategory` 中按风险分类覆盖，分类话术优先。拒答内容应说明请求无法继续及内部支持渠道，不应暴露检测规则、阈值或绕过方法。非流式请求被拦截时，网关返回 HTTP 403。

流式输出。检测阶段包含输出时，流式响应同样受检，网关边检测边放行：

- 词库、敏感数据正则等规则层检测逐段进行。网关只扣住每段末尾可能与后续文字组成命中的一小段，长度取决于规则引用的词库中最长的词条和敏感数据规则的最大匹配长度，匹配长度不受限的正则会把扣留范围延伸到句末；其余文字检测通过即放出。
- 风险类目、人名等需要模型判断的检测按句进行，每句经模型判定通过后放出，判断时只携带提问和最近一段前文。模型调用失败时重试一次，仍失败则按「护栏检测出错」处理。
- 每路流式响应同一时刻至多有一次模型判断在途，判断期间到达的句子合并到下一批送检。启用模型判断时，判断模型的并发约等于同时进行的流式响应数，可据此预留判断模型的容量。

- 命中脱敏时，扣住的片段替换后放出，前后文字不变。
- 命中拦截时，网关在命中处中止，按调用协议发送终结帧：Chat Completions 为 `finish_reason: content_filter` 加 `[DONE]`，Responses 与 Anthropic Messages 为 `event: error`；拒答信息放在终结帧根上的 `safetyIntervention` 字段（`kind: stream_cut`）。已放出的内容保留。
- 中止后网关继续读完上游输出，按上游实际用量计费。
- 单段连续扣留达到约 64 KiB 时，网关对扣留内容做一次完整检测后放出；跨度超过该上限的命中词在流式下检测不到，非流式没有这一限制。
- 模型的思考内容和工具参数只检测、不改写，需要改写时按中止处理。

14.5 上线验证与命中通知

内容检测上线前至少覆盖以下样例：允许内容、各风险分类样本、敏感数据样本、零宽与同形字规避、编码内容、必答问题、检测组件不可用、输入命中、输出命中，以及流式与非流式两种调用。

1. 先用「敏感数据试运行」（13.3 节）和流量规则的「验证规则」（第 15.3 节）核对判定。
2. 用测试调用触发规则，在请求详情的「护栏」区核对结论与命中说明，例如「已拦截」及命中的词库名称。
3. 验证记录至少包含规则名称、检测方式、阈值、预期处置、实际处置、请求标识和责任人。

输入、输出两个方向在 OpenAI 兼容入口与网关原生入口都执行；只挂内容检测、不指定转发目标的增强规则同样作用于输出方向。请求阶段命中或非流式回答命中时返回 403，`error.message` 为规则配置的拒答话术，`error.details` 中 `reason` 为 `safety_intervention` 的条目带命中类别、检测方向与规则标识；流式回答命中时以终结帧结束，拒答话术位于 `safetyIntervention`。检测服务不可用且规则选择阻断时同样返回 403，但原因码与说明不同于命中。引用的租户词库、问答集尚未发布成功时，保存或配置发布会失败并在路由策略页指出缺少的对象。完整说明见[路由策略用户手册 5.2 节与错误参考](#)。

命中通知不在内容检测中配置，入口为「管理控制台 → 运行与排障 → 告警规则」。需要逐次通知时，在「告警规则」页签目录的「内容护栏」分类中，为需要的「内容风险类别 · 入口」行绑定通知渠道（「通知渠道」列），再打开行末「操作」列的通知开关；词库、问答与内容风险判定的命中按所属类别通知，敏感数据命中不外发。需要按统计窗口告警时，选择「新建告警规则」→「网关策略事件：拦截、故障转移等」，设置统计窗口与阈值。接收渠道在「通知渠道」页签维护。详见第 22.2 节。



14.6 SSRF 出站防护与 Web 应用防护

SSRF 出站防护。平台管理员进入「管理控制台 → 平台扩展 → SSRF 出站防护」，页面为「平台 SSRF 出站防护」，包含三项设置：

设置	默认	说明
「SSRF 出站防护」	关闭	开启后，平台发起的出站请求先检查目标地址：私网、回环、链路本地、云元数据地址（如 169.254.169.254 及常见元数据主机名）、CGNAT 与 IPv6 本地地址等不可访问。地址在建立连接时重新解析并检查，用于防御 DNS 重绑定
「HTTP 出站防护」	关闭	开启后出站只允许 HTTPS
「内网出站放行名单」	关闭	名单有独立开关。关闭时，企业 IAM 换票等平台代发请求可访问普通内网地址；开启后只允许访问「部署内置放行（只读）」中的地址和手动添加的地址

放行名单的每一项必须写成 host:port，例如 172.16.0.1:8443；回环、链路本地和云元数据地址不能加入名单。添加后选择「保存放行名单」。各项设置保存后立即生效，无需重建服务；运行时尚未完成更新时，页面显示「目标设置已保存，运行时仍在收敛。」。

防护作用于模型供应商调用、MCP 服务地址（注册探测与工具调用转发）、外部智能体地址、告警通知发送以及 IAM 等平台代发请求。MCP 地址被拦截时，页面提示「出于安全策略，平台拒绝访问这个地址（指向内网或被禁止的网段）。请改用可对外访问的地址。」。「内网出站放行名单」只作用于企业 IAM 换票等平台代发请求；开启「SSRF 出站防护」后，位于内网的模型供应商与 MCP 服务地址会被拒绝。接入不受信的模型供应商或 MCP 服务前，平台管理员应先确认现有模型端点与 MCP 服务是否位于内网，再开启防护。

Web 应用防护。控制台入口和网关入口可启用 OWASP 核心规则集（CRS 4.x，偏执等级 1）。网关入口不检查请求体，只检查 URI 与请求头，避免提示词和模型响应被误判；控制台入口检查请求体。该功能没有控制台页面，由部署人员通过部署配置项 `LP_WAF_ENGINE_MODE` 设置：

取值	行为
Off（默认）	不检测
DetectionOnly	仅检测，命中只记录
On	拦截命中的请求

命中记录归属到部署指定的边缘租户（`LP_WAF_TENANT_ID`，未指定时为最早创建的启用租户）。当前 Helm 部署不读取该配置项。

15. 路由策略与流量保护

「流量规则」集中维护模型与能力调用的路由规则、全局限额、意图集和语义缓存，并提供「验证规则」页面模拟整套规则的判定结果。本章功能均属于动态路由许可模块。规则中「内容检测」动作的检测项、模板与判断模型配置见第 14 章，本章只说明它在规则中的位置。

15.1 入口与生效顺序

15.1.1 入口、许可与权限

从「管理控制台 → 路由策略 → 流量规则」进入。页面顶部有「规则清单」「全局限额」「意图集」「缓存」四个页签；「规则清单」页签右上方的「验证规则」按钮打开独立的验证页面。

- 许可：「流量规则」页面及其中的意图集、判断模型状态、内置意图试判接口属于动态路由模块。许可未包含该模块时，侧栏不显示「流量规则」。
- 权限：页面要求凭据管理权限，内置角色中租户管理员和安全管理员具备。看不到入口时，先核对「系统设置 → 授权许可」中的模块，再核对账号角色。「全局限额」与「缓存」两个页签的接口另按租户管理权限校验，只持安全管理员角色的账号可以进入页面，但这两个页签的加载与保存会被拒绝，需由租户管理员维护。
- 同一分组下的「网关扩展」维护「挂载插件」动作引用的插件绑定，见第 16.2 节。

页签或页面	管理对象	主要任务
规则清单	按判定顺序排列的路由规则	新建、导入、复制、排序、启停、删除规则

页签或页面	管理对象	主要任务
全局限额	按调用类型独立维护的流量保护配置	请求次数、并发限制、熔断、非法尝试保护、上游调用超时、调用方临时豁免
意图集	内置意图目录与自定义意图集	查看内置意图、试判一句话、维护自定义意图集，见第 15.9 节
缓存	租户级语义缓存参数与统计	调整参数、查看命中占比、清除缓存，见第 15.10 节
规则详情	单条规则	查看画布与实时流量、编辑、启停、查看命中的请求、修改记录、删除
验证规则	整套规则的模拟判定	输入请求事实，查看命中链、最终目标、候选目标和判定原因

「全局限额」独立于单条路由规则维护，但仍按自身的匹配条件和计数维度生效，并不表示所有请求共用一份额度。模型服务配额、API 密钥配额与这里的限额分别计数。

15.1.2 规则如何参与一次调用



图 1：配置验证步骤示意。保存、下发、模拟和真实调用分别回答不同问题。

请求通过身份认证后，网关按「优先级」数字从小到大检查已启用的规则。一条规则的全部条件都满足才算命中；不满足时跳过本条，继续检查下一条。规则清单的「命中后」列标明规则类型：

规则类型	「命中后」列	命中后的处理
增强规则：只配置检测、限额、改写、缓存等动作，没有转发目标	继续匹配	叠加本条配置，继续评估后续规则
终止规则：配置了转发目标，或动作为「放行」「拒绝」	结束匹配	后续规则不再评估

- 「放行」结束规则判定，请求按网关既有路由处理，仍受身份权限、额度和目标可用性检查约束。

- 「拒绝」不向上游派发，网关返回 403，错误响应中带有配置的拒绝原因码。
- 清单为空或没有规则命中时，请求按网关既有路由处理，这不保证一定存在可用目标。平台不内置路由规则，清单中的规则均由本租户维护。
- 多条规则同时命中时，内容检测按整条规则链合并，同一检测项冲突时取更严格的配置，整次请求只送检一次；「单次请求 Token 上限」和「请求体大小上限」取最严的值。
- 「解析后的网关模型」条件在选定目标后复核。复核时命中的规则要求的目标与已选目标不一致时，网关返回 503。

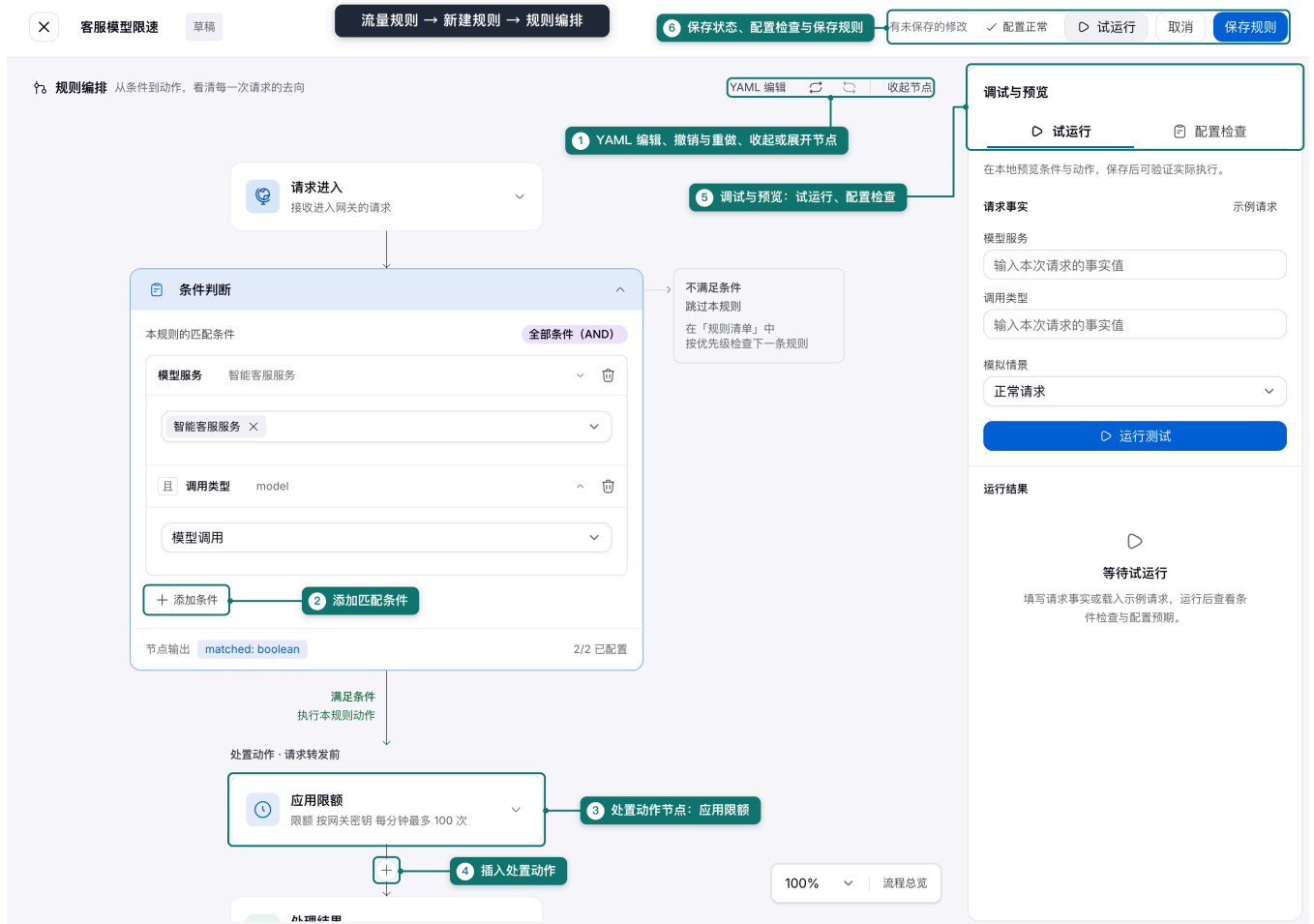
15.1.3 如何阅读画布

规则编辑器以「规则编排」画布展示一条规则，节点从上到下依次为：

- 「请求进入」：说明流量入口；请求路径、调用身份和模型信息作为条件判断的输入。
- 「条件判断」：本规则的匹配条件，全部条件均需满足（AND）。右侧「不满足条件 / 跳过本规则」表示按优先级检查下一条规则。
- 「处置动作 · 请求转发前」「处置动作 · 收到上游响应后」：一个动作一个节点，节点上标注「执行顺序 n / m」。同类改写动作可在本阶段内调整先后，其他动作的执行顺序由网关确定。
- 「处理结果」：显示「命中后的配置预期」，不表示已有请求执行成功。
- 「异常处理」：按「系统固定行为」「命中处理」「组件不可用时」分组列出可配置的处置方式，随动作增减变化。

画布工具与面板：

- 顶部「YAML 编辑」以 YAML 编辑当前草稿，选择「应用到画布」后回到画布检查。内容不超过 256 KiB，资源引用使用资源 ID。
- 「撤销修改」「重做修改」作用于本次编辑中尚未保存的修改，最多回退 50 步。
- 「收起节点」「展开节点」只改变显示，不会停用动作；底部的「画布缩放」和「流程总览」用于定位较长的流程。
- 右侧「调试与预览」包含「试运行」和「配置检查」两个页签。「试运行」在本地按填写的请求事实预览条件与动作，需要网关求值的条件标为「需要网关验证」，应保存后用「验证规则」核对；「配置检查」列出草稿中的问题，点击问题即定位到对应配置。
- 底部依次显示保存状态（如「有未保存的修改」）、配置检查结论（「配置正常」，或「配置异常」及问题数）、「试运行」「取消」和「保存规则」。仍有待完善的配置时「保存规则」不可用。



15.2 创建规则与配置条件

15.2.1 新建规则的方式

方式	入口	说明
从空白新建	「规则清单」右上方「新建规则」	打开空白画布
套用模板	「导入规则模板」	在「规则模板」中选一条，填写高亮的必填字段后选择「导入这条」。模板有六十多条，覆盖路径、请求头、查询参数、模型名、模型服务、数据来源、敏感信息、调用类型、能力、意图等条件与转发、放行、拒绝动作的组合，以及加权轮询、时延优先等示例
从文件导入	「从文件导入」	读取 .yaml 或 .yml 文件，可包含多条规则；同名占位对象只需选择一次。规则按文件顺序创建在清单末尾，文件中的资源 ID 原样保存
复制已有规则	规则行「更多操作 -> 复制」	以「原名称 的副本」预填编辑器
从内置意图建规则	「意图集 -> 内置意图」中某个意图的「用它建规则」	预填该意图条件和规则名，见第 15.9 节

每个租户最多 500 条规则，清单右上方显示「已有条数 / 上限条」。达到上限后新建、导入和复制不可用；导入的条数超过剩余名额时，页面提示还能创建的条数。

15.2.2 创建一条模型路由

1. 在「模型供应商」和「模型服务」中完成目标接入与调用验证，为测试调用方准备有效凭据和模型使用权限。
2. 选择「新建规则」，在画布顶部填写能说明用途的名称，例如“客服问答主路由”。名称为 1-120 个字。
3. 在「条件判断」中选择「添加条件」，从「选择匹配条件」中添加「客户端声明的模型名」，比较方式选「等于」，填入调用方实际提交的模型名，例如 customer-support。这是请求里的名称，不是模型 ID。
4. 选择「添加处置动作」，再选「转发到」。条件中还没有调用类型时，编辑器自动补上「调用类型」为「模型调用」。
5. 在转发节点选择「目标方式」，先选模型服务，再选该服务中的模型。需要分流时在本档继续添加目标并设置权重；需要故障转移时添加备用档位并开启「启用故障转移」，见第 15.7 节。
6. 查看「处理结果」，修复「配置检查」中的问题后选择「保存规则」，在「确认变更」中核对后选择「确认并应用」，见第 15.3.1 节。
7. 用「验证规则」检查应命中和不应命中的请求，再用实际业务入口发送受控请求，见第 15.3 节。

本章示例名称和阈值仅用于说明操作，不代表推荐容量。不要把示例模型名直接当作已接入模型使用。

15.2.3 创建一条增强规则

只想给客服模型增加限速、不改变去向时，添加相同的模型名条件，再添加「应用限额」并设置请求速率、计数维度和阈值，不添加转发目标。这条规则在清单中显示为「继续匹配」。

将增强规则排在需要受它约束的终止规则之前。前面的规则已经转发、放行或拒绝时，后面的增强规则不会参与判定。要实现“允许名单 + 拒绝其他请求”，用「放行」动作放行名单内的请求，再在其后放置兜底的「拒绝」规则；没有转发目标的普通规则是增强规则，不会结束判定。

15.2.4 条件之间的关系与数量限制

- 同一条规则内的条件之间是“且”关系；同一条件内的多个取值满足任意一个即可。否定比较方式（如「不等于」「不以…开头」「不匹配正则」）表示与全部取值都不符合时才满足。
- 一条规则最多 16 个条件。多数条件可填 1-32 个取值，「调用方」条件可选 1-42 个对象。
- 同一字段上互相矛盾的条件无法同时成立，保存时会被拒绝。“A 或 B”应写成一个条件的两个取值。
- 正则表达式由数据面引擎执行，不支持零宽断言和反向引用。要表达“不以…开头”，改用比较方式中「排除」组的选项。
- 未添加条件的规则在清单中显示「匹配所有请求」，保存前确认它是有意设置的兜底规则并检查其位置。

15.2.5 条件类型

「选择匹配条件」按以下分组列出条件，可用搜索框按名称查找：

分组	条件	填写与比较方式	注意事项
身份与来源	调用方	先选「调用方类型」，再按名称搜索并选择对象，比较方式为「属于」「不属于」	只列出当前入口能提供的身份种类；手动填写的 ID 不能代替选择器
身份与来源	标签	选择标签键与取值，比较方式为等于、属于、不等于	标签键须是已定义的调用标签；没有已定义标签时本条件无可选项
身份与来源	组织成员	在「部门（含下级）」和「职能组」中选择，两类取并集；比较方式为「属于」「不属于」	调用方所属部门无法解析时，含本条件的「拒绝本次调用」规则按条件满足处理，其他动作的规则按不满足处理
模型与能力	调用类型	模型调用、MCP 调用、A2A 调用、连接器调用、API 服务	转发、限额和多数改写动作要求将调用类型限定为一种
模型与能力	具体能力	先选调用类型，再选具体能力	不能把模型 ID 当作 MCP 工具填写
模型与能力	工具名	比较方式为「属于」「不属于」；已选 MCP 服务时可选择已登记的工具，也可输入含 * 的通配模式，例如 delete_*	用于 MCP 调用、连接器调用和 API 服务（取接口名称）
模型与能力	MCP 方法	「调用工具」「列出工具」「资源」「提示词」，可多选，新建时默认「调用工具」	用于 MCP 调用
模型与能力	客户端声明的模型名	等于、以…开头、匹配正则及对应的排除式；可从模型目录选择或直接输入	不接受网关模型 ID；同一名称可能跨多个供应商命中，需精确锁定时改用「解析后的网关模型」
模型与能力	模型服务	选择模型服务	请求须能确定服务归属
模型与能力	解析后的网关模型	选择网关模型	只在选定目标后复核，不能用来决定选哪个模型
请求内容	数据来源	「MCP 工具返回的数据」	由可信通道识别；不能与限额同用
请求内容	个人信息类型	选择一个或多个类型，可选置顶的「任意敏感信息」	类型来自本租户的个人信息类型目录（见第 13 章），依赖请求经过敏感信息检测

分组	条件	填写与比较方式	注意事项
意图与场景	意图	先选「意图维度」（任务类型、任务难度、业务场景、自定义意图集（高级）），再勾选意图	只在模型调用中产生；需要设置意图识别模型，见第 15.9 节
请求基础	请求路径	以 / 开头；等于、以…开头、匹配正则及排除式	反向代理前缀和不同入口可能改变可见路径
请求基础	URL 参数	参数名（区分大小写，最多 64 字节）和取值	只对直连网关的调用生效
请求基础	请求头	头名称（小写字母、数字和连字符，最多 64 个字符）、比较方式和取值	凭据类、转发来源类、传输类和网关内部请求头会被数据面剥离，不能作为条件
时间与来源	时段	选择时区（IANA 名称），添加一个或多个 HH:MM-HH:MM 时间段，允许跨零点	按所选时区判定，与管理员电脑的显示时间无关
时间与来源	客户端 CIDR	合法的 IPv4 或 IPv6 网段，例如 10.0.0.0/8	只对直连网关的调用生效；经平台转发的流量不携带可信客户端地址
时间与来源	灰度 Cookie	选择 Cookie 名称并填写取值	取值只写不回显，编辑规则时须重新输入原始值；不能与限额同用

15.2.6 调用身份与事实缺失

真实调用身份来自凭据认证、登录账号和受信请求上下文。在「验证规则」中填写的调用方只是判定输入，不会为真实调用授予权限；把对象 ID 复制到普通请求头也不能替代认证。

不同入口提供的匹配信息不同。模拟结果中条件显示「请求未提供该信息」时，应补齐真实请求能提供的信息；显示「此调用类型不提供该匹配信息」时，应调整条件或调用类型。配置意图、个人信息类型等条件前，先确认意图识别模型或敏感信息检测已经可用。未检测、检测失败与已检测但值不匹配是不同情况，不能统一理解为“规则配置无误”。

15.3 验证规则与真实调用

15.3.1 保存与确认变更

新建和编辑规则都先打开「确认变更」：在编辑器中选择「保存规则」，或在规则详情编辑后选择「保存」。核对「字段差异」「关联对象」（变更前引用、变更后引用）和「命中顺序」；配置了故障转移时另显示「故障转移链」。本次变更修改敏感请求头时，须勾选确认后才能保存。核对无误后选择「确认并应用」；范围不符时选择「取消」返回编辑。

确认变更

×

保存前请确认本次变更对命中顺序和关联引用的影响。

字段差异

名称	变更前	变更后
	客服问答流量保护	客服问答调用保护

关联对象

变更前引用

无关联对象

变更后引用

无关联对象

命中顺序

变更前

1. 第 1 位 · 客服问答流量保护

变更后

1. 第 1 位 · 客服问答流量保护

影响预览仅分析静态配置，不执行上游调用。目标在实际请求中的可用性须通过运行状态和实际调用验证。

确认后，命中的请求会按新规则走。发布前请先看命中顺序。

取消

确认并应用

图 2：确认变更。确认前核对实际变化、关联对象和命中顺序。

- 应用成功后提示「规则「名称」已保存，正在下发到数据面。」，表示配置已提交，运行端可能仍在加载。
- 规则清单中的启停开关、拖动排序和删除在确认后直接生效，不再打开「确认变更」；规则详情中的启停开关仍会打开「确认变更」。
- 数据面拒绝某条规则时，清单顶部提示「有 N 条规则被数据面结构性拒绝，未参与判定」并列出现因，例如「规则编译失败」「规则尚未同步到数据面」；规则详情的状态显示「数据面未采纳」。修正后重新保存，或删除该规则。

- 规则清单、规则详情和「验证规则」都不显示各网关节点是否已加载新版本。保存后以真实调用验证生效情况。

15.3.2 使用「验证规则」

进入方式：「规则清单」右上方的「验证规则」；或规则详情中的「验证这条规则」，结果默认定位到该规则。两种方式都按整套规则的先后顺序评估。验证页面没有调用记录选择器，请求事实需手动填写。

- 选择「调用类型」，填写「请求的模型」（与规则中「客户端声明的模型名」是同一个值）；非模型调用改选「能力」。
- 选择「调用方」，或选择「凭据身份」（API 密钥），两者二选一。
- 按需填写「请求路径」「试运行时间」（用于时段条件）。内容检测测试运行时，选择「检测方向」（「输入」或「输出」），在「消息」中为每条消息选择「角色」（「用户」「助手」）并填写「消息内容」，可选择「附图」添加图片，用「添加消息」增加多轮；「检测方向」为「输出」时另填「本轮回答」。命中的规则配置了工具调用防护时，可填写「工具名」「参数或命令」，需要检查工具返回结果时填写「返回结果」，并选择「工具调用防护」。
- 需要时展开「更多请求事实」，填写「指定模型服务」「意图提示词」「客户端 IP」「查询参数」「灰度 Cookie」「调用标记」「数据来源」「PII 类型」「请求头」。这里填写的是假设值，真实流量中由平台自动判定。
- 需要查看后备顺序时勾选「模拟主档失败」；需要按带金额预算的调用方演练时勾选「按带金额预算的调用方演练」，链上未定价的档位会被跳过。
- 必填事实齐全后页面自动计算，「模拟请求」可重新计算。修改输入后等待新结果出现再读取。

结果	正确理解	下一步
「命中」	终止规则命中，后续规则不再评估	确认它的顺序和动作符合预期
「增强规则命中，继续评估」	本条叠加了配置，后续规则仍会评估	查看完整命中链和最终目标
「不匹配」	已评估，但条件不成立或所需信息缺失	展开条件，查看「值不匹配」「请求未提供该信息」等原因
「未评估」	前序规则已结束判定	核对前序规则的顺序与动作
「未命中任何规则，按当前发布的默认配置路由」	由网关既有路由决定去向	确认默认目标和权限可用

结果区还列出「本次模拟命中的规则链」、模拟选定的路由目标、各目标的期望占比；按排名选目标的策略另显示候选目标的名次、质量分、价格、预计首字时延和是否被排除。勾选「模拟主档失败」时显示后备目标所在档位与转移链；未配置更高档位时提示主档失败后无后备目标。验证按控制面已保存的规则计算，刚保存的规则可能尚未在网关生效。

模拟不会调用上游模型，也不能证明限额已计数、改写已执行或目标可连通。填写「意图提示词」「消息」或执行「工具调用防护」时，页面会调用相应的判断模型或检测服务，附图会发送至所选图片理解模型，请使用允许送检的测试内容。加权轮询的模拟使用独立于生产的轮询进度，连续模拟不会消耗真实流量的轮询顺序。

15.3.3 用真实调用完成验证

使用业务实际采用的入口、凭据和请求格式。测试请求应使用脱敏数据和明确的测试对象，避免影响其他业务。

验证项目	操作	应保留的结果
路由	发送应命中和不应命中的请求	实际目标、状态码、请求标识、命中规则
速率限制	在所选计数维度内受控增加请求，再更换维度取值	达到阈值后的处理，以及不同计数对象是否分别计数
并发	保持若干请求同时执行，再增加一个请求	同时执行数、拒绝结果和结束后的释放情况
改写	发送带可观察测试字段的请求	上游实际收到的值和调用方实际收到的响应
内容检测	发送正常文本和预设命中样例	放行、记录、拦截或脱敏的实际结果
熔断	注入受控失败或延迟，等待半开恢复	上游接收次数、拒绝结果和恢复过程，见第 15.6 节
故障转移	让主要目标满足已配置的触发条件	备用目标是否收到请求、总尝试次数和耗时

规则详情的「查看命中的请求」打开「概览」的「网关请求」视图，并按本规则筛选；也可在「概览」的「路由策略命中」中核对命中记录并跳转到规则详情。请求被拒绝时，保留调用方收到的完整状态码、错误码与请求标识。状态码与错误码的对应关系见第 10.8 节。

部分失败请求可能缺少完整调用记录。查不到日志不能单独证明请求没有执行，应结合原始响应、上游接收记录和时间范围排查；记录缺失时提交运行管理员处理。

15.4 配置全局限额

15.4.1 列表与来源

进入「流量规则 → 全局限额」，先在页签右侧选择调用类型（模型调用、MCP 调用、A2A 调用、API 服务），列表只显示该调用类型的限额。可按「限额类型」筛选。

「来源」列	含义	维护方式
全局限额	在本页签创建的独立限额	行内「更多操作」中编辑、启停、删除、创建临时豁免
来自路由规则	路由规则中「应用限额」动作派生的限额，显示规则名称	只读，点击规则名称到来源规则中编辑；来源规则已删除时可「删除残留限额」
平台强制限额	平台预置的硬下限	只读

调用类型为 MCP 调用时，列表顶部说明上游调用超时口径：未被已启用的超时限额覆盖的调用，使用系统默认值（连接 5000 毫秒、请求 30000 毫秒）。限额的「计数维度」不被当前数据面支持时，该行显示「计数维度不可用」及原因。限额保存后若提示「限额已保存，尚未同步到模型网关」，系统会自动重试，同步完成前网关可能仍使用

之前的配置。

15.4.2 新建限额

1. 选择「新增限额」，打开「创建全局限额」画布（「限额编排」）。
2. 在「请求计数入口」中选择「限制类型」：请求次数、并发限制、熔断、非法尝试保护、上游调用超时。切换类型会把阈值重置为该类型的默认值，名称和匹配条件保留。「名称」可留空，留空时使用限额摘要作为名称；限额编号由系统分配。
3. 在「计数对象」中设置匹配条件，请求次数和并发限制还需选择「计数维度」，见第 15.4.3 节。
4. 在「限额配置」中填写阈值，在「超限处理」中选择处理方式。
5. 在右侧「调试与预览」中，「试运行」可模拟「未触发限额」与「触发限额」两种情景（不发送请求、不读取实时计数），「配置检查」列出待完善项。
6. 选择「启用全局限额」发布；编辑已有限额时选择「保存全局限额」。

编辑已有限额时不能更改限制类型；需要换类型时新建一条。删除限额后无法恢复，需要时重新创建。

15.4.3 匹配条件与计数维度

匹配条件决定哪些请求受这条限额约束；计数维度决定这些请求如何分别累计。例如只匹配客服模型、按 API 密钥计数，表示每个 API 密钥对客服模型的请求分别计数。改变匹配条件不会自动改变计数维度，编辑时分别核对。

匹配条件有两层：

- 统一条件：与规则清单相同的条件类型（不含灰度 Cookie），见第 15.2.5 节。按网段限流使用其中的「客户端 CIDR」。
- 保留的匹配条款：按调用方身份、客户端 IP、路由、模型服务、网关模型、能力、能力类型、应用、所属部门、职能组、依赖、端点、MCP 服务或工具、服务名与服务命名空间、部署环境与部署区域等字段匹配，比较方式为「等于」（只能一个值）、「属于其中之一」「不等于」。条款中的「按客户端 IP」按完整地址精确比较，不做网段包含。

已添加统一条件时，保留的匹配条款暂不参与匹配；删除全部统一条件后条款重新生效。两层都为空时，限额对所选调用类型的全部请求生效。每层最多 16 个条件，跨条件全部满足才匹配；同一字段上两条肯定条件的取值必须有交集，否则页面阻止保存。

计数维度可多选，选择多个维度时按取值组合计数。可选项以所选调用类型为准，包括按 API 密钥、按 API 密钥（直连调用）、按用户、按调用方分组、按内部服务、按客户端 IP、按会话、按路由、按网关模型、按模型服务、按应用、按所属部门（逐部门计数）、按职能组（逐组计数）、按 MCP 服务、按 MCP 工具和按标签等。会话只能作为计数维度，不在匹配条款中提供。按会话计数时，按首条用户消息识别会话和无法识别会话的请求不计入此维度，Responses 续接请求沿用上一轮的识别方式；识别顺序与取值要求见第 10.9 节。选择计数维度前确认真实入口会提供该信息。

15.4.4 请求次数、超限处理与临时阈值

在「限额配置」中为每个时间窗填写「周期内请求数」和「即时突发量」。时间窗可选秒、分钟、小时，最多 3 个，同一周期不能重复，数值须为正整数。这是按时间持续补充的滚动限速，不是到整点清零的配额；多个时间窗同时生效，请求需满足每个时间窗。即时突发量用于容纳短时峰值，平时不占用额度。

「超限处理」可选：

选项	行为	参数
拒绝请求	返回限流错误与重试提示，模型调用返回 429	—
排队等待	等待可用额度，超时或队列已满时仍返回限流错误	「最大排队深度」 1-512、「最大等待（毫秒）」 1-30000、「每个组织的排队请求上限」 1-128、「每个组织的队列容量（字节）」 1-33554432
返回固定应答	不再调用模型，直接返回预设内容	「告诉调用方的原因代号」：小写字母、数字、下划线，1-64 字节；「助手回复内容」：必填，不超过 64 KiB

「额度用完时」另有两个选项：「直接拒绝」，或「先占用公共额度，公共额度也不够才拒绝」。公共额度指同一调用类型下只按租户计数的请求次数限额；不存在这样的限额时仍按拒绝处理。

开启「启用临时阈值」时：

1. 为每个时间窗填写临时请求数和临时突发量。临时值都不得低于基础额度，且至少一项高于基础额度。
2. 填写开始和结束时间。结束时间须晚于开始时间且晚于当前时间，窗口最长 30 天。
3. 到期后数据面自动恢复基础额度，无需再次发布。临时阈值只放宽流量阈值，不延长凭据有效期，也不授予模型访问权。

15.4.5 并发限制

并发限制统计同时执行的调用数，不是一段时间内的调用总数。

- 「并发上限」：正整数，达到后新请求立即被拒绝，不进入队列。模型调用返回 429。
- 「恢复阈值（可选）」与「并发硬上限（可选）」：须同时填写，并满足“恢复阈值 < 并发硬上限 ≤ 并发上限”。达到硬上限后拒绝新请求，并发数低于恢复阈值后恢复接收。
- 「为优先请求保留的名额（可选）」：0 到并发上限减 1。繁忙时普通请求最多用到“上限减保留名额”，余下名额只放行优先请求。
- 「繁忙时的优先级」：0-3 的整数，0 为普通，3 为最高。一个请求同时命中多条带优先级的并发限额时取最高档。档位只决定繁忙时谁能用保留名额，没有保留名额的限额中优先级不起作用。

15.4.6 非法尝试保护与上游调用超时

非法尝试保护需要填写「统计窗口（秒）」「非法尝试次数阈值」和「暂停时长（秒）」，均为正整数。统计窗口内非法尝试达到阈值后，暂停该匹配对象的请求，暂停期间返回 429。当前由 MCP、A2A、API 服务等能力调用在下游凭据校验失败时计入，不统计普通的 HTTP 失败；验证时确认请求确实经过对应入口且事件被识别。

上游调用超时包含「连接超时（毫秒）」和「请求超时（毫秒）」，范围均为 50-600000，连接超时不得大于请求超时。连接超时限制建立上游连接的等待时间，请求超时限制整个上游请求（含响应体）的时长。达到超时后终止本次上游调用；配置了故障转移时，后续处理由故障转移规则决定。熔断中的「慢调用阈值」只用于熔断统计，不会中断慢请求。

15.4.7 创建调用方临时豁免

在独立限额行的「更多操作」中选择「创建调用方临时豁免」。系统新建一条独立的豁免配置，不修改原限额；匹配字段沿用原限额的第一个匹配字段，取值需重新确认。

- 1. 核对调用方、匹配范围和「豁免类型」。可选类型为请求次数、并发限制、熔断和非法尝试保护，不包含上游调用超时。
- 2. 填写「原因」和「到期时间」。普通豁免最长 7 天，到期时间须晚于当前时间。
- 3. 开启「豁免平台强制限额」时，有效期最长 2 小时。只在确有授权的应急场景使用。
- 4. 保存后分别用豁免范围内与范围外的调用验证；到期后再次检查原保护恢复。

豁免按类型生效：它适用于指定类型的所有匹配限额，并非仅适用于发起豁免的那条限额。豁免不等于身份认证、模型授权或其他配额的豁免。

15.5 处置动作与异常处理

15.5.1 动作清单

在画布中选择「添加处置动作」，或在已有动作节点后选择插入按钮，从列表中选择动作。不满足依赖的动作显示为不可选，悬停可查看原因；动作节点内的「动作类型」可替换为其他动作。

动作（画布名称）	阶段	适用调用类型	要点
转发到	请求转发前	限定为一种	模型调用选网关模型，其他调用选能力端点；见第 15.7 节
放行	—	不限	结束判定，交给网关既有路由；不能与检测、限额、改写或插件动作共存
拒绝本次调用	—	不限	终止请求，返回 403 和「拒绝原因码」；不能与检测、限额、改写、插件或镜像共存
内容检测（护栏）	请求转发前（输出检测在收到响应后执行）	按检测项	检测项包括 PII 敏感数据、词库、必答、风险类目、工具调用防护，配置见第 14 章
应用限额	请求转发前	限定为一种	规则内的请求速率、并发、熔断和上游调用超时，见第 15.5.5 节
单次请求 Token 上限	请求转发前	模型调用	见第 15.5.4 节
请求体大小上限	请求转发前	模型调用	见第 15.5.4 节
语义缓存	请求转发前	模型调用	见第 15.10 节
推理指令	请求转发前	模型调用	见第 15.5.4 节

动作（画布名称）	阶段	适用调用类型	要点
会话切换保护	请求转发前	模型调用	见第 15.7.4 节
影子分派	请求转发前	模型调用	需先有转发目标，见第 15.7.5 节
提示词	请求转发前	模型调用	见第 15.5.2 节
改写请求头、改写请求字段、改写请求正文	请求转发前	请求头改写仅模型调用	见第 15.5.3 节
改写响应头、改写响应字段、状态码映射	收到上游响应后	响应头与状态码仅模型调用	响应字段改写只对非流式响应生效
挂载插件	由插件绑定配置	模型调用	选择已启用的插件绑定，一条规则最多 8 个，见第 16.2 节
镜像复制	收到上游响应后	MCP、连接器、A2A 或 API 服务中的一种	见第 15.5.6 节

「放行」与「拒绝本次调用」不能与其他动作共存。一条规则中的动作须要求同一种调用类型；需要不同调用类型的动作请拆成两条规则。画布按网关固定的执行阶段排列动作，不支持任意连线。

15.5.2 内容检测与提示词

「内容检测（护栏）」在规则中决定哪些请求送检、检测哪些项目以及命中后的处置。检测项、模板、判断模型和处置方式的配置见第 14 章。在规则层面注意：

- 区分输入、输出和双向检测，输出检测不能替代输入防护。
- 「仅记录，放行」「拦截」「脱敏后继续处理」等处置的结果不同，须分别用样例验证。
- 多条规则命中时，检测类型合并，同一类型取最严格的处置，整次请求只送检一次；应检查完整规则链。
- 风险类目和提示注入检测需要在「系统设置 → 全局设置」中选择内容安全模型，未设置时这两项不可选。

「提示词」动作引用提示词库中的提示词，字段为「选择提示词」「版本标签」（production 或 latest）和「注入位置」（「前置」或「后置」）。租户中没有提示词时该动作无可选项。

15.5.3 改写请求与响应

- 「改写请求头」「改写响应头」：操作为「设为」「追加」「删除」，头名称最多 128 个字符，同一规则内不能重复。Host、Content-Length、Content-Type、连接控制类等请求头不允许改写；网关管理的响应头不能改写。修改敏感请求头时保存前需二次确认。
- 「改写请求字段」：用 JSON Pointer 定位字段，例如 /metadata/channel，操作为「新增」「覆盖」「删除」「复制」「搬移」；复制和搬移须填写「来源位置」，值按 JSON 格式填写，文本需带双引号。
- 「改写请求正文」：在请求体中复制、搬移字段，或把字段「转成文本」「转成数字」「转成是否」。
- 「改写响应字段」：调整最终非流式响应，可用「只保留列出的字段」保留 1–16 条路径。

- 「状态码映射」：把指定的上游状态码换成另一个状态码返回调用方，只能配在会改变目标或拒绝请求的规则上。不要把上游失败不加区分地包装为成功。

每类改写一条规则最多 16 项。路径以 / 开头、不超过 160 字节，字段名中的 / 和 ~ 按 JSON Pointer 规则转义。请求改写不能改动路由目标模型或网关保留字段，也不能通过改写 /model 或受保护请求头绕开模型选择和授权。真实密钥只在受保护的凭据入口配置，不要写进规则名、普通值或截图。

保存后分别测试正常字段、缺字段、空值、类型不符、上游失败和流式调用。模拟命中只证明规则判定，改写结果须从上游请求或调用方响应核对。

15.5.4 Token、请求体与推理指令

- 「单次请求 Token 上限」：填写「最大输入 Token」和「最大输出 Token」中的至少一项，范围 1–10,000,000。输入超限时在发给供应商前拒绝，并告知调用方这是平台限制；输出上限会改写请求的 max_tokens。
- 「请求体大小上限」：单位 KB，范围 1–32768（即 32 MiB 全局硬顶）。按入站原始请求体在鉴权、模型解析和正文改写之前判断，超出时返回 413。
- 「推理指令」：「推理」可选「开启」或「不开启」；开启时「推理强度」可选「默认」「低」「中」「高」，默认按中档补充。调用方已指定推理参数时以调用方为准；「不开启」只表示平台不补充，不会删除调用方已带的参数。目标模型不支持所选档位时自动折到最接近的一档。多条规则命中时取优先级最高的那条。

多条规则同时设置 Token 上限或请求体上限时取最严的值。

15.5.5 规则内限额

「应用限额」只对命中本规则的请求生效，并作为「来自路由规则」的派生限额出现在「全局限额」列表中。节点内先是「请求速率」开关（计数维度、「额度用完时」、时间窗和「临时阈值（可选）」），其下的「并发、熔断与超时（可选）」中有「并发」「熔断」和「上游调用超时」开关。字段含义与范围同第 15.4 节和第 15.6 节，其中「繁忙时的优先级」在规则内以「最高」「较高」「普通」三档选择。

- 超出规则内限额时只能「拒绝本次调用」，并返回配置的原因码；排队、固定应答、非法尝试保护和临时豁免需要在「全局限额」中创建独立限额。
- 规则须将「调用类型」限定为一种，且不能使用灰度 Cookie 或数据来源条件。
- 同一请求可能同时受规则内限额和全局限额约束。

15.5.6 镜像复制

「镜像复制」只用于一种明确的 MCP、连接器、A2A 或 API 服务调用，按「复制比例」（1–100%）把请求复制到另一个能力端点。主请求不等待复制请求，复制结果也不返回调用方；复制请求按正常调用计费与限流，会额外占用目标端点的配额。

开启前依次选择能力与能力端点作为「复制目标」，并勾选同主机和跨域名两项流量去向确认。默认不复制可能产生副作用的请求；开启「允许复制可能产生副作用的请求」前，确认复制环境能安全接收创建、删除等调用。

15.5.7 异常处理

「异常处理」节点只列出当前动作可配置的项目：

- 「系统固定行为」：条件不满足时，不做处理，交由下一条规则判定。
- 「命中处理」：内容检测命中、超出限额、输入超出 Token 上限、请求体超限等情况的处置，在对应动作中设置或固定为拒绝。
- 「组件不可用时」：检测组件不可用时选择「拒绝本次调用」或「放行，继续网关路由」；模型入口的请求字段、请求正文改写失败时选择拒绝或保留原内容继续；MCP、A2A、API 服务入口的改写失败选择「阻断请求」或「保留原内容继续调用」。请求头、响应头与状态码改写失败时固定拒绝本次调用。
- 转发目标全部不可用时返回错误；启用故障转移后，按触发条件尝试备用目标，备用目标用尽或达到尝试上限后停止转移。

异常处理不是所有故障的统一兜底：权限不足、配额耗尽、护栏拦截、目标不可用和客户端取消可能在不同阶段结束请求。应测试实际会发生的失败路径，而不只确认异常节点存在。

15.6 配置并验证熔断

15.6.1 熔断解决什么问题

熔断根据已完成的调用样本判断上游是否持续失败或变慢。样本数达到下限且失败率或慢调用率任一达到阈值后，在熔断持续时间内拒绝后续调用、不调用上游；到期后进入半开状态，允许有限数量的探测调用。任一探测失败重新进入熔断，全部探测成功完成后恢复正常处理。

熔断与并发、超时是不同的控制：并发限制同时执行数，超时限制等待时间，熔断依据近期结果暂停后续调用。熔断本身不切换上游；只有配置了故障转移的规则，才会跳过熔断中的候选目标并尝试转移链上的下一个目标。



图 3：熔断状态示意。半开期间只允许有限数量的探测调用，仍可能拒绝其他请求。

15.6.2 从哪里配置

可在规则的「应用限额」中展开「并发、熔断与超时（可选）」并开启「熔断」，也可在「全局限额」中新建限制类型为「熔断」的限额。两处参数含义一致，匹配范围和所属配置不同。对同一业务同时使用两处配置前，明确各自的保护对象并分别验证。

全局限额中，前五个参数位于「限额配置」，「熔断持续时间（秒）」和「半开探测次数」位于「超限处理」。规则内配置时，先将「调用类型」限定为一种。

×

创建全局限额

草稿

管理控制台 → 路由策略 → 流量规则 → 全局限额 → 新增限额

有未保存的修改

配置正常

试运行

取消

保存

限额编排

收起节点

配置检查

请求计数入口

名称
智谱轻量模型熔断

限制类型 *
熔断

检查匹配条件

计数对象

应用限额配置

限额配置

阈值配置 * ①

统计窗口 (秒)
60

最小调用样本数
20

失败率阈值 (%)
50

慢调用阈值 (毫秒)
3000

慢调用率阈值 (%)
60

触发限额时执行

超限处理

超限处理 * ②

不可修改

熔断期间拒绝请求，不调用上游；30 秒后进入半开状态，最多允许 3 个探测请求，根据探测结果决定恢复或继续熔断。

熔断持续时间 (秒)
30

半开探测次数
3

① 限制类型选择「熔断」

② 限额配置：统计窗口、样本数与失败率、慢调用阈值

③ 超限处理：熔断持续时间与半开探测次数

100%

▼

流程总览

调试与预览

▶ 试运行

⌵ 配置检查

在本地模拟一次限额触发情景，确认网关如何处理请求；不会发送请求或读取实时计数。

模拟情景
未触发限额

▶ 运行测试

运行结果

▶

选择模拟情景并运行，查看配置预期。

检查当前配置是否完整。实际限额以保存后的网关执行结果为准。

参数	含义	可填写范围	规则内默认值
统计窗口（秒）	汇总近期调用结果的时间范围	1-3600	30
最小调用样本数	至少累计多少次调用后才按比例判定	1-1000000	10
失败率阈值（%）	失败样本比例达到该值触发熔断	1-100	50
慢调用阈值（毫秒）	耗时达到该值的调用记为慢调用	1-1800000	2000
慢调用率阈值（%）	慢调用比例达到该值触发熔断	1-100	50
熔断持续时间（秒）	熔断后暂停调用的时长	1-3600	30

参数	含义	可填写范围	规则内默认值
半开探测次数	半开阶段允许同时执行的探测调用上限	1-1000	5

以上均为整数。「慢调用阈值」用毫秒，其余时间参数用秒。失败类别由入口的实际调用结果归类，不能把所有业务 4xx 都当作上游故障。熔断期间，OpenAI 兼容协议的模型调用返回 503（错误码 `TRAFFIC_CIRCUIT_OPEN`），Anthropic 协议返回 529；实际排查以所用入口的响应为准。

15.6.3 用受控失败验证

准备独立的测试调用方与可控制返回结果的测试上游，不要用正常业务制造故障。以下数值只用于验证行为，不能直接当作生产容量配置。

- 1. 设置统计窗口 60 秒、最小样本 4、失败率 50%、慢调用阈值 30000 毫秒、慢调用率 100%、熔断 8 秒、半开探测 2 次。
- 2. 保存并确认应用，重新打开详情逐项核对七个字段。
- 3. 用同一保护对象在窗口内依次发送成功、失败、成功、失败四次请求。前 3 个样本不足 4，不应提前触发；第 4 次后失败率为 50%。
- 4. 紧接着继续请求，核对熔断拒绝和上游接收计数。拒绝期间被保护的上游不应再收到这些请求。
- 5. 等待 8 秒后发送探测。让探测失败，应重新进入熔断；再等待并让探测成功，检查普通请求恢复。
- 6. 保持 2 个半开探测同时执行，再发第 3 个请求，检查探测上限；释放前两个后核对恢复结果。

15.6.4 用慢调用验证

设置统计窗口 30 秒、最小样本 2、失败率 100%、慢调用阈值 200 毫秒、慢调用率 50%、熔断 8 秒。先发送一次快速成功请求，再发送一次耗时超过 200 毫秒的成功请求。

即使两次都是成功响应，慢调用比例达到 50% 后，下一次请求也可能被熔断。核对实际耗时、响应和上游接收次数，不要只看 HTTP 200 判断不会触发。

停用熔断配置后，等待配置同步并用实际调用确认恢复，不要把停用开关当作即时复位按钮。上线前针对实际拓扑、调用类型和流式请求补充验证。

15.7 流量分配与故障转移

15.7.1 目标方式、档位与权重

「转发到」节点的「目标方式」有两种，切换时会清空已选目标：

- 「固定模型」：每个目标先选模型服务，再选该服务中的网关模型。
- 「按请求模型对应」：每个目标选一个供应商，再在「按请求模型对应」表中为条件里的每个模型名选择该供应商下承接它的模型，可用「按名称自动匹配」生成建议。要求规则恰有一条非否定的「客户端声明的模型名等于」条件；主要目标须覆盖全部模型名，备用档位的空格表示「本档跳过」；单个供应商最多 32 个映射。不能与其他目标种类混用，也不支持「时延优先」。

非模型调用的转发目标是能力端点：先选能力，再选端点。

一条规则最多 8 个转发目标。第一档为「主要目标」，之后为「备用档位 1」「备用档位 2」等。

- 主要目标按目标选择策略和权重选出一个，未选中的主要目标不会自动作为备用。
- 同一档位的权重（0-100 的整数）合计为 100%，调整一个目标时其余目标按比例分配剩余权重；「平均分配权重」把 100% 按整数尽量均分。权重为 0 的目标不参与分配，同档至少一个目标权重大于 0。
- 同一档位的转发目标不能重复。停用或不可用的目标不会因提高权重而恢复。
- 配置了备用档位但没有启用故障转移时，保存会被拒绝。

例如主要目标中 A、B 分别为 80、20，表示主要目标内的期望分配；备用档位的权重只在进入该档时使用，不与主要目标合计。

15.7.2 目标选择策略

「目标选择策略」	行为	前提与验证
按调用方稳定分配	按请求特征和权重做加权哈希，同一调用方总是落到同一目标；权重描述许多调用方之间的分配比例	用多个实际调用方观察分布，不能只重复同一请求
加权轮询	在同一档位内按权重比例轮流选择；每个网关副本有独立的轮询进度	统计足够数量的实际派发，排除提前拒绝和不可用目标的影响
时延优先	按权重抽两个主要目标，选预计首字时延更快的一个；新目标先试后评	主要目标至少 2 个，且只支持模型目标；预计时延来自各网关副本的本地统计
质量优先	先比质量分，差距在容差内算打平，再比价格、再比时延，仍打平按加权哈希选	质量分取自内置模型目录的基准指数，价格取自己发布的价目；缺少数据的目标按与最优者并列处理
成本优先	先比输入与输出单价，差距在容差内算打平，再比质量分、再比时延，仍打平按加权哈希选	同上；未定价的目标在规则编辑器中显示「未定价」

「质量优先」「成本优先」可填写「容差」（0.1%–50%，不填按默认 5%）。按排名选目标时，不满足请求需要的工具调用、图片、JSON 输出、推理或上下文长度的候选会被排除；全部候选都被排除时改为按权重选一个。只有一个可选目标时，策略不影响目标分配。

15.7.3 配置故障转移

1. 将调用类型限定为「模型调用」，在转发节点配置主要目标。
2. 选择「添加备用档位」并添加备用目标。没有备用目标时「启用故障转移」不可开启。
3. 开启「启用故障转移」，选择触发条件（任一满足即转移，最多 16 个）：「连不上上游」「上游超时」「上游返回指定状态码」「响应内容命中指定条件」。默认勾选连不上上游、上游超时和状态码 429、500、502、503、504。

4. 指定状态码时每个值须在 100–599 之间；响应内容条件需填写以 / 开头的「响应内容位置」（如 /error/type）和期望值。
5. 设置「总尝试次数」（1–10，包含首次调用，默认 3）和「首内容前故障转移总预算（秒）」（1–600，默认 120）。达到任一上限后不再尝试备用目标。
6. 保存并核对应用状态，在「验证规则」中勾选「模拟主档失败」查看转移链，再制造受控故障检查实际的备用调用。

故障转移按备用档位顺序尝试，同一备用档位内权重较高的目标优先。只有启用了故障转移，失败时才会尝试备用目标；目标排列在后面，或同档有多个目标，都不代表自动转移。

「响应内容命中指定条件」只对非流式请求生效；连接失败、超时和状态码条件也适用于流式请求。转移只发生在向调用方输出首个内容之前，已开始输出的流式响应不会被撤回后改用其他模型。用量与额度计入请求进入的模型服务，费用按实际执行档位的单价结算。

15.7.4 会话切换保护

「会话切换保护」让同一会话的后续请求继续使用上一轮的模型。只有新候选在当前策略的首要指标上领先超过切换裕度、且用户轮次达到下限时才切换；正在续传工具调用，或提示词前缀与上一轮一致时不切换。

- 只能配在「质量优先」或「成本优先」策略上，且主要目标至少有两个转发目标。
- 字段：「切换裕度」0.1%–50%（默认 5%）、「最少用户轮次」1–100（默认 1）、「空闲超时」30–86400 秒（默认 300）。
- 会话按第 10.9 节的顺序识别。按首条用户消息识别的会话和无法识别会话的请求不启用保护。

15.7.5 影子分派

「影子分派」把被采样的请求（经过请求变换和脱敏之后的请求体）再复制一份，异步发给影子模型，只用于评估。影子请求的失败、超时或丢弃不影响主响应；影子跳过限流，但照常计费，用量按用途「影子」单独记账，不进入主流量统计。响应体不保存，只记录大小、状态、耗时与 Token。

- 需先配置转发目标，影子模型不能与转发目标相同。
- 字段：「采样比例」0.1%–100%（默认 1%）、「并发上限」1–64（默认 2）、「队列上限」1–256（默认 8）、「超时」1–300 秒（默认 30）、「响应读取上限」1 KiB–8 MiB（默认 1 MiB）。
- 删除该动作节点即停止影子分派。配置影子分派的规则，详情页提供「影子对比」，按请求数、延迟 p50、p95 和成本对比主目标与影子模型。

15.7.6 实时流量与转移结果核对

分流规则的全部目标都经模型派发时，规则详情的查看态提供「规则编排」和「实时流量」两个视图。「实时流量」按时间窗显示调用次数、成功率、故障转移次数和每千词元成本，并按目标列出配置权重与实际占比、调用次数、成功率、耗时（中位与慢请求）和故障转移转出、转入次数。

核对转移结果时，记录主要目标与备用目标的接收情况、最终响应、耗时和请求标识。客户端超时应覆盖网关的尝试预算；存在外部副作用的请求要评估重复调用的影响。故障转移仍受权限、额度和目标归属检查。

15.8 维护、校验与故障定位

15.8.1 查找、排序、启停、删除与修改记录

- 搜索：按规则名、规则 ID 或目标模型搜索，至少输入 2 个字符。可按「状态」筛选，并在「筛选」中按「调用类型」多选，或按「能力」选择具体服务（MCP 服务详情的「查看适用的路由策略」即按本服务预填这两项）；结果包含显式限定了该调用类型、能力的规则和未限定的规则。
- 导出：规则清单右上方选择「导出」，把全部规则导出为 YAML 文件。认证请求头值与灰度 Cookie 值不写入文件，文件注释和提示列明被剔除的规则，导入前须补齐这些字段；规则清单未加载完整或没有规则时不可导出。
- 排序：按优先级排列且未搜索、未筛选时，可拖动行首手柄，或使用「移到最前」「上移」「下移」「移到最后」调整判定顺序，调整即生效。调整后验证完整规则链，尤其是增强规则和兜底拒绝规则。
- 启停：使用行内状态开关；勾选多行后可「启用所选」「停用所选」「删除所选」，确认后生效。停用后配置保留，重新启用前核对关联对象是否仍可用。
- 删除：行内「更多操作 → 删除」，在确认框中选择「从清单摘掉」后立即删除；规则详情中为「删除规则」。删除无法撤销，删除前确认替代路由和保护已经就绪。
- 修改记录：规则详情的「修改记录」按时间列出每次变更的操作人和字段变化。「载入此版本」把该版本载入编辑器，保存并确认应用后才作用于请求。
- 未保存就关闭编辑器时，在「退出规则编辑？」中选择「继续编辑」或「放弃编辑」，放弃不会保存当前草稿。

15.8.2 常见校验与处理

提示或限制	处理方式
名称为空或超长、取值为空	填写 1-120 个字的名称和完整取值，不要只输入空格
条件或取值数量超限	拆分为可解释的多条规则并重新核对顺序，不要用互相冲突的条件合并
模型名填成了网关模型 ID	填入客户端声明的模型名；按对象匹配改用「模型服务」或「解析后的网关模型」
目标缺少模型服务、不存在或已停用	先选模型服务再选模型，确认供应商与模型均已启用并可调用
所有权重为 0、目标重复	按档配置有效权重，去除同一档位的重复目标
调用类型与动作不匹配	按提示把「调用类型」限定为一种，或拆成两条规则
「时延优先」需要两个目标	在主要目标中再加一个模型目标，或改用其他策略
配了备用档位但未启用故障转移	启用故障转移，或删除备用档位
正则无效、CIDR 无效、时区或时间段无效	按字段格式修正，并验证正反例及时区边界
请求头或 JSON 路径受保护	使用平台提供的路由、凭据或目标配置，不改写受治理字段
临时阈值低于基础额度、窗口已过期	同时核对临时数值和起止时间，保证至少一项放宽

提示或限制	处理方式
熔断、并发或超时参数无效	检查整数、范围、单位及字段间关系
灰度 Cookie 需要重新填写	重新输入原始取值，不要复制页面中的脱敏回显
目录、模板或关联对象加载失败	等待服务恢复后重试；目录不可用时保存会暂停

15.8.3 保存冲突和运行异常

错误码的含义与是否可重试见第 10.8 节。下表按现象给出排查顺序：

现象	排查顺序与处理
保存提示版本冲突或预览无效	保留待修改内容，刷新读取最新配置，核对他人的变更后重新预览；不要重复确认旧预览
页面保存成功，实际行为未变化	重新打开确认字段；检查启用状态、顺序、清单顶部的结构性拒绝提示和调用时间；稍后再用真实入口验证
模拟命中，真实请求未命中	核对调用类型、入口、认证身份、模型名、路径、请求事实与时区；URL 参数和客户端 CIDR 只对直连网关的调用生效
返回 403	结合错误码区分规则拒绝、凭据权限、模型范围或额度问题；检查拒绝规则和授权，不能只调整路由目标
返回 413	检查规则中的「请求体大小上限」
返回 429	区分请求次数、排队超时、并发限制和非法尝试保护；分别核对全局限额、规则内限额、模型服务与 API 密钥配额以及计数维度
返回 503 或熔断提示	核对失败比例、慢调用比例、样本量和熔断时长；检查是否处于半开状态、上游是否仍异常；使用「解析后的网关模型」条件时，还要核对选路后复核是否失败
返回 504 或上游超时	检查连接超时、请求超时、故障转移总预算和客户端超时；比对上游耗时和网络情况
改写或响应格式不符	确认动作阶段、调用类型、JSON 路径、字段类型以及是否流式；核对改写失败时的处置是否保留了原内容
备用目标没有收到请求	检查备用档位、「启用故障转移」、触发条件、尝试预算、目标授权与健康；确认请求在首个内容输出前失败
查不到失败调用记录	扩大时间范围并核对规则与身份；结合原始响应和上游记录排查，不以缺少日志认定请求未执行

问题反馈至少包括：规则或限额名称及编号、操作时间及时区、调用类型与入口、脱敏后的请求信息、响应状态和错误码、请求标识、预期结果与实际结果。不要提交原始 API 密钥、Cookie、个人信息或上游凭据。

15.9 意图集

15.9.1 意图识别模型

意图条件依赖意图识别模型。在「系统设置 → 全局设置」中为「意图识别」选择模型；「意图集 → 内置意图」页顶部显示当前状态，例如「意图识别模型：名称」，或「还没有设置意图识别模型，意图条件不会命中」。模型所在服务停用或模型不可用时，页面同样提示，意图条件不会命中。

请求命中带意图条件的规则时，网关调用一次意图识别模型，内置维度与自定义意图集合并在同一次调用中判断，单次判断超时为 8 秒。判断不可用（未设置模型、服务停用、调用失败或请求中没有可判断的文字）时意图条件不命中。意图只在模型调用中产生，MCP、A2A 和 API 服务请求不会带有意图信息。

15.9.2 内置意图

「意图集 → 内置意图」按三个维度列出平台内置的意图目录：

维度	内容
任务类型	写作、实用指导、技术帮助、多媒体、查找信息、自我表达、其他 / 未知七组，共二十多个细类，例如翻译、编程、数据分析
任务难度	「难题」与「非难题」两档；具体性、领域知识、复杂度、问题求解、创造性、技术准确性、实际应用七条判定标准中满足至少 6 条判为难题
业务场景	订单、收货地址、发票、付款、退款、账户、配送等十多个客服类别，共二十多个意图，例如申请退款、跟踪订单

每行显示「意图」和「说明」，「用它建规则」以该意图为条件打开新规则，名称预填为「维度：意图」。

页面中的「试一句」可粘贴一句真实提问，选择「试一下」后显示「难度」「类型」「场景」和「内容风险」四行结果，难度另附命中的判定标准。内容风险一行需要已设置内容安全模型。试判会调用判断模型，请使用允许送检的文本。

15.9.3 自定义意图集

在「意图集 → 自定义意图集」中选择「新建意图集」，填写：

- 「名称」：1-120 个字，不能与已有意图集重名。
- 「说明」：不超过 2000 个字。
- 「识别灵敏度」：「保守」（宁可漏判）、「均衡」（默认）、「激进」（宁可多判）。
- 「意图」：至少一个，最多 64 个。每个意图填写「显示名」（1-120 个字，不能重复）和「一句话描述」（说明什么样的提问属于这个意图，判断模型按这句描述识别）。

在规则条件中选择「意图」，将「意图维度」切换为「自定义意图集（高级）」，再选择意图集和意图。被流量规则引用的意图集不能删除，被引用的意图不能移除，页面会列出引用它的规则。

15.10 语义缓存

15.10.1 开启方式与边界

语义缓存只对命中「开启了语义缓存」的规则模型请求生效：在规则中添加「语义缓存」动作即开启。命中这类规则请求先查缓存，问题足够相似时直接返回缓存答案，不再调用模型；未命中时照常调用模型，并把答案写入缓存。没有规则开启语义缓存时，请求不查也不写缓存。缓存只对 OpenAI Chat Completions 格式、消息中只有一条用户消息（可另带系统或开发者消息）且请求体不超过 256 KiB 的单轮请求生效，流式与非流式请求均适用。

- 缓存服务与中文向量模型（bge-small-zh-v1.5）及重排模型随安装包部署在本地，Zentrix 安装默认包含该能力，不调用外部向量服务。部署未配置缓存服务时，「缓存」页签提示「本部署未配置语义缓存服务，无法清除」。
- 语义缓存只在网关接入层执行，并按本次命中的整条规则链判断能否参与。规则链中出现以下任一配置时，这次请求不查也不写缓存：转发到指定目标、应用限额、单次请求 Token 上限、挂载插件、响应改写（改写响应头、改写响应字段、状态码映射）、输出检测、改写请求头。因此语义缓存适合配在不指定转发目标的增强规则上，请求随后按网关既有路由处理。
- 「放行」「拒绝本次调用」规则不能开启语义缓存。
- 规则编辑器允许在带转发目标的规则上添加「语义缓存」，但按上述判断这类请求不会走缓存。上线前用「缓存」页签的命中占比确认缓存确实生效。

15.10.2 缓存参数与容量

「缓存」页签的设置是租户级的一份配置，留空表示使用默认值，保存后各实例最迟 5 秒内生效：

字段	范围	默认值	说明
匹配置信度（%）	0-1000	500	重排得分达到该值才算命中，越高越严格
有效期（秒）	正整数	3600	缓存条目的保留时长
隔离粒度	全租户共享、按工作空间、按用户、按智能体	按用户	「全租户共享」命中率最高，但不同人可能看到彼此的答案；「按用户」只命中本人问过的答案
最多缓存条目数、最多占用空间	正整数	不限制	「每租户上限」，留空即不限制

缓存始终按租户和模型分开。Zentrix 形态没有工作空间与智能体的管理入口，通常选用「按用户」或「全租户共享」。

15.10.3 命中占比与清除

- 「缓存概况」按模型列出「条目数」「占用空间」「最早写入」「最近写入」。还没有启用的规则开启语义缓存时，页面提示并提供「去规则清单添加」。

- 「命中占比」分「近 1 小时」「近 24 小时」「近 7 天」显示，计算方式为命中次数 ÷（命中次数 + 实际调用模型的次数）；另显示未开启缓存的调用占比，只按已上报的调用计算。
- 「清除缓存」中选择「清除全部」并「确认清除」，删除本租户的全部缓存条目。修改答案口径或模型后，可先清除缓存再验证。

16. 请求变换、网关扩展与开放 API

本章说明业务系统接入网关时的两类配置：安装网关扩展并绑定到流量，以及获取开放 API 文档。请求和响应改写属于流量规则的动作，见第 15 章。

16.1 请求和响应变换

请求和响应变换在「管理控制台 → 路由策略 → 流量规则」的规则动作中配置，需要「动态路由」许可模块。可用的变换动作包括「改写请求头」「改写请求字段」「改写请求正文」「改写响应头」「改写响应字段」和「状态码映射」，改写失败时的处理方式在「异常处理」中设置。动作的适用调用类型、字段约束和验证方法见第 15.5 节，本章不重复说明。

16.2 安装、配置和绑定网关扩展

网关扩展是运行在网关请求处理过程中的 WASM 插件，用于在模型调用或 MCP 调用中执行准入、改写等逻辑。入口为「管理控制台 → 路由策略 → 网关扩展」，属于「基础功能」，需要租户管理员或「系统管理员」权限。

页面为插件目录，列出本租户可用的网关扩展，列包括「插件」「类型与能力」「当前版本」「状态」「绑定数」「凭据绑定数」「安装时间」。可按「安装状态」筛选，也可在「搜索插件…」中输入 2 至 128 个字符、不超过 8 个词进行搜索。本租户上传的私有插件带「本租户」标记，平台发布的插件在详情中显示为「平台共享」。

插件详情包含以下标签：

标签	内容
「概览」	身份信息、安装状态、当前版本、能力（「插件类型」「插入阶段」「读取请求体」）、供应链标识与摘要
「配置」	安装插件、填写插件配置并「保存配置」
「网关策略」	请求屏蔽、防重放和并发上限，见第 16.3 节；仅对已安装、非 MCP 上游用户身份类的网关扩展显示
「版本」	版本历史、「插件版本管理」
「绑定」	流量绑定和密钥绑定
「权限」	插件声明的「网络目标」「密钥要求」「安全与数据要求」「PII 要求」
「执行统计」	插件运行结果与耗时

安装并启用插件的步骤如下：

1. 在目录中选择「安装」，或打开详情后在「配置」标签操作。
2. 查看「权限」标签，确认插件声明的网络目标、密钥、安全与 PII 要求符合本单位要求。
3. 在「配置」标签按表单填写配置后选择「安装（停用）」；插件无需配置时选择「以空配置安装（停用）」。插件以停用状态安装，配置发布后再启用。
4. 等待状态依次经过「等待发布」「等待网关加载」并显示「已生效」，再在详情页打开「启用」开关。
5. 在「绑定」标签确认绑定范围；插件已停用时，绑定不会生效。
6. 用命中和不命中的请求分别验证，再到「执行统计」核对结果。

绑定插件到流量

流量绑定决定插件对哪些请求执行。绑定范围包括「全部流量」「模型」「路由」「调用方」「消费者」「MCP 服务」和「所有 MCP 服务」。每条绑定显示执行方式（「强制」或「观察」）、采样比例和优先级，可在「绑定」标签中启停或「删除」。

不同类型插件的绑定方式不同：

- MCP 上游用户身份类插件（「类型与能力」显示「MCP 上游用户身份」）：在「绑定」标签选择「创建绑定」，在「MCP 服务」中选择一个服务，或选择「所有 MCP 服务」。选择「所有 MCP 服务」时，插件对当前及后续新增的所有已启用、通过网关访问的 MCP 服务的全部调用生效。该类绑定始终以「强制」方式执行全部调用。
- 模型调用类插件：「绑定」标签展示和维护已有绑定；新建绑定通过开放 API `POST /api/v1/plugins/bindings`，`scopeType` 取 `global`、`model`、`route` 或 `principal`；未指定执行方式时默认为观察模式。创建并启用绑定后，可在流量规则详情的「插件绑定」标签中通过「添加插件」挂载到规则；插件仅在调用类型限定为「模型调用」的规则上生效，每条规则最多挂载 8 个插件绑定，详见第 15 章。

MCP 上游用户身份类插件与 MCP 服务自带的上游用户身份配置不能同时作用于同一服务：

- 目标 MCP 服务已启用上游用户身份原配置时，创建或启用插件绑定前需关闭原配置。对话框提供「关闭原配置并创建绑定」，需要填写「变更原因」。
- 租户内任一 MCP 服务仍保留上游用户身份原配置时，不能绑定「所有 MCP 服务」。
- 同一 MCP 服务只能有一条启用的此类插件绑定；已存在「所有 MCP 服务」绑定时，不能再为单个服务启用绑定，反之亦然。
- IAM 身份来源未配置时，缺少用户身份的调用会被拒绝，页面提供「前往 IAM 提供者设置」。

维护插件版本与卸载

1. 在「版本」标签的「插件版本管理」中选择「上传新版本」提交完整插件包，或对已有版本选择「切换到此版本」。
2. 切换前在「目标版本配置」中填写目标版本的完整配置，选择「继续预览」。预览会列出将被删除的配置键，需勾选确认。
3. 在「确认插件变更」中核对影响后选择「确认」。确认后才发布到网关；并发变化会阻止本次操作。
4. 页面显示发布和加载进度。新版本在网关运行失败时，平台自动回滚到原版本并提示「自动回滚完成」。
5. 卸载前选择「卸载」查看影响：将清理的绑定和配置、将被清理的网络凭据绑定，以及将停用并移除插件引用的流量规则。本租户私有插件可同时勾选「同时删除这个租户私有插件包及全部版本」。
6. 操作完成后在插件目录、「绑定」标签、「执行统计」和操作审计中核对版本与状态。

上传私有插件与下载开发包

- 上传插件：在插件目录选择「上传插件」，提交 .zplugin 插件包。插件包不超过 140 MiB，平台校验压缩包结构、清单、制品摘要、WASM 组件、接口类型和配置结构，校验未通过时拒绝上传。上传成功后作为本租户私有插件出现在目录中。
- 下载 SDK：选择「下载 SDK」，在「下载网关扩展 SDK」中选择「开发语言」（「Rust：编译为 WASM 组件」或「Lua：只需编写 plugin.lua」），填写「插件标识」，选择插件类型（「请求准入：模型调用前放行或拒绝」「请求转换：修改请求正文或请求头」「响应转换：处理模型流式响应」）。下载包包含对应 SDK、WIT 接口定义、清单、示例代码以及构建、校验和发布说明。Lua 无需配置 Rust 或 WASM 工具链，但解释执行可能慢于 Rust；对延迟或吞吐敏感时优先选择 Rust。
- AI 开发说明：同一对话框的「下载 AI Skill」提供一份 SKILL.md，供 AI 编程助手按本网关的接口编写插件。

查看执行统计

「执行统计」仅适用于已安装的网关 WASM 插件。启用安装并至少有一条启用的流量绑定后才开始采集。

1. 选择时间范围，自定义区间不超过 7 天。
2. 查看「执行次数」「成功」「拦截」「观察命中」「失败」「p95 耗时」和「各阶段执行次数」。
3. 按「阶段」和「执行结果」筛选执行记录。阶段包括「P0 模型改写」「P1 请求准入」「P1' 路由感知准入」「P2 请求改写」「P3 响应改写」。
4. 在记录中核对「发生时间」「原因」「耗时」「绑定」和「版本」；需要查看对应调用时，在「调用链 / 请求」列选择「请求」，在弹出的对话框中直接查看该次调用的报文，选择「调用链」进入调用链。

页面提示存在投递缺口时，对应时间段内的部分执行记录未能采集，不能据此判断插件未运行。统计仅涵盖当前安装，重新安装后不包含此前的数据。

16.3 配置插件网关策略

已安装的网关扩展（MCP 上游用户身份类除外）在详情中提供「网关策略」标签。策略由网关侧执行，随该插件的绑定生效，保存后选择「保存网关策略」发布。

策略	配置项	限制与说明
「请求屏蔽规则」	「规则 ID」「原因码」「路径前缀」「路径包含」「请求头名称」「请求头包含」「请求头等于」「提示信息」	最多 200 条；每条至少设置路径前缀、路径包含或请求头名称之一；填写请求头名称时需同时设置“包含”或“等于”条件。命中即拒绝请求
「防重放（Nonce + 时间戳）」	「启用执行」「强制要求 nonce 头（缺则拒绝）」「Nonce 请求头」「时间戳请求头」「最大时间偏差（秒）」「Nonce 保留时间（秒）」	偏差与保留时间范围均为 1 至 86400 秒；重复 nonce 被拒绝。需与调用方的时间戳和 nonce 生成方式联调
「并发上限」	「启用执行」「最大并发数」「计数维度」「超限行为」「租约 TTL（秒）」	最大并发数 1 至 100000；计数维度可选 consumer、model、org、consumer_model；超限行为当前只有 reject；请求结束时释放槽位

单次请求 Token 上限在流量规则中配置，见第 15.5.4 节。

插件声明了网络目标和密钥要求时，「绑定」标签的「密钥绑定」列出已有的网络凭据绑定及其「允许目标」。每条绑定以逻辑名称引用一个密钥，并限定该密钥只能用于指定目标。新建密钥绑定通过开放 API POST `/api/v1/plugins/secret-bindings`（参数为 `installationId`、`logicalName`、`secretReference`、`allowedDestination`）。

插件的「执行统计」用于观察运行状态，「概览」的用量面板和「网关请求」用于核对消耗，操作审计用于追溯配置变更。

16.4 获取开放 API 文档

开放 API 规范文件（OpenAPI）用于查看平台管理接口的路径、参数和响应结构。规范文件不经访问入口对外提供，需要时向交付方案索取与平台同版本的文件。规范中列出的接口不代表当前身份有权调用。

调用接入流程如下：

1. 取得与目标平台同版本的规范文件，生成或更新客户端。
2. 按接口所属身份类型配置平台会话、API 密钥或服务 Token。
3. 在隔离环境发送最小请求，验证认证、权限、幂等和错误结构。
4. 在「概览 → 网关请求」或「操作审计」中核对调用方、时间、耗时、状态码和关联 ID。
5. 把实际开放的路径、身份、模型范围和限流约束记录到业务接入单。

需要通过 MCP 客户端执行管理操作时，可使用平台管理 MCP，见第 11.9 节。

16.5 集成上线检查

- ☐ 真实业务调用经过预期的流量规则。
- ☐ API 密钥或服务 Token 已存入密钥管理系统，日志中不可见原文。
- ☐ 请求和响应变换已分别用流式、非流式请求验证。
- ☐ 网关扩展的权限声明、绑定范围、执行方式和网关策略已经确认，「执行统计」中有预期记录。
- ☐ 开放 API 规范与平台版本一致，错误响应和限流行为已纳入客户端处理。
- ☐ 调用记录可以关联调用方、模型、路由、插件和最终状态。

17. 查看用量并控制成本

本章说明成员查看本人用量、管理员查看全租户用量与费用的入口。逐项分析方法见第 18 章，费用与配额告警的配置见第 22.2 节。

17.1 查看个人用量

入口：能力中心「首页」，在左侧导航「用量」分组中选择分区。所有成员均可使用，页面只显示本人的数据，不需要加购许可模块。

分区	内容
「使用情况」	「最近 180 天使用情况」：「模型调用次数」「MCP 调用」「消息数」「累计 Token 数」及与前 180 天的对比；「模型使用率」「MCP 使用率」等排行；「最近一年活跃度」热力图；以及「今日 Token 消耗」「当月 Token 消耗」等汇总
「模型用量」	按「今天」「最近 7 天」「本月」统计（默认「本月」），可按「API 密钥」「模型服务」「模型」「计费归属」筛选；显示「可用模型服务」「可用 API 密钥」数量，以及「消费金额」「模型调用次数」「总 Token」「平均单次成本」，「消费趋势」与「模型占比」图表；「消费明细」按「日期」「API 密钥」「模型服务」「模型」「计费归属」列出「调用次数」「Token」「金额」
「调用日志」	「按调用」或「按回答」查看逐次模型与 MCP 调用；时间范围可选「最近 24 小时」「最近 7 天」「最近 30 天」或「自定义范围」，可按「调用类型」「状态」「计费归属」筛选，在搜索框中按模型、调用方、Request ID、Trace ID 或错误搜索，并可「导出记录」为 CSV

使用说明：

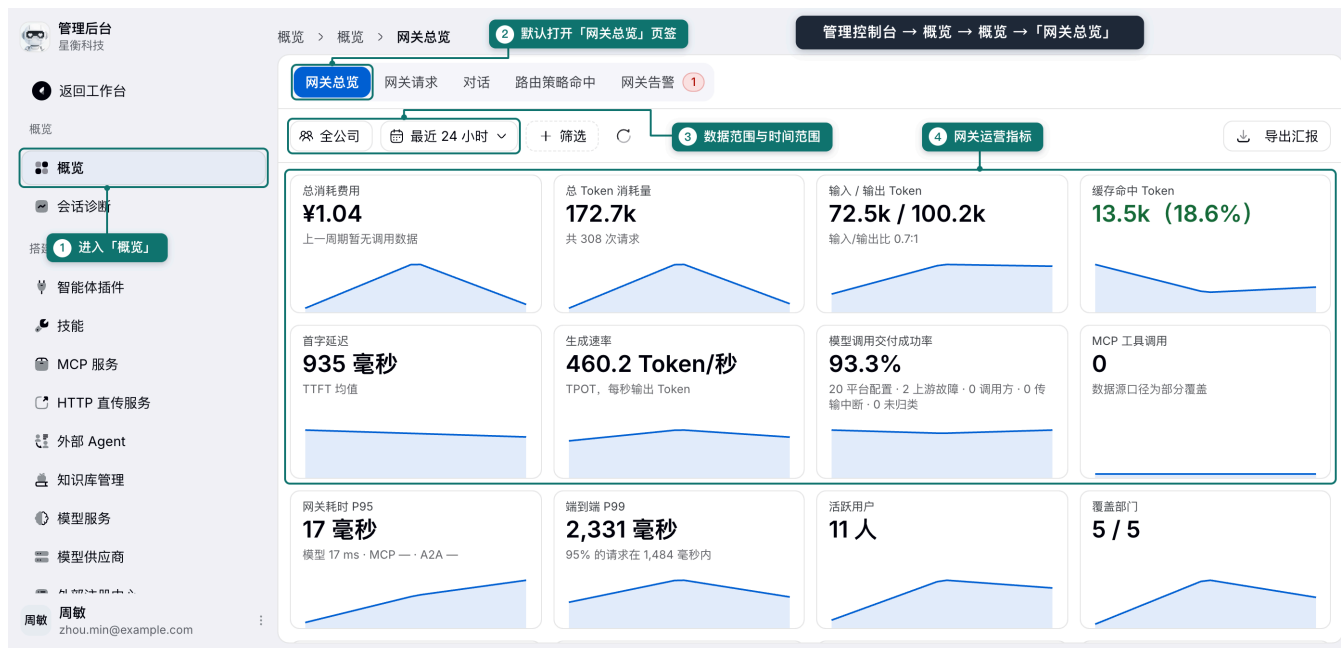
1. 核对近期消耗时，先在「模型用量」选择统计范围，再按 API 密钥或模型服务筛选，确认消耗来自哪一把 Key、哪一项服务。
2. 金额按调用发生时的单价计算。部分调用缺少定价时，页面显示「费用不完整」或「费用缺失」，金额与图表只统计已记录费用的调用。
3. 需要核对单次调用时，进入「调用日志」，按时间和状态筛选后查看调用详情。

限制：「模型用量」的「今天」「最近 7 天」「本月」由服务端按账号档案中记录的时区划分自然日，未记录时区时按 UTC 计算。「偏好设置」中的「时区」只影响本浏览器的时间显示，不改变该统计窗口。

17.2 管理员查看用量与成本

入口：「管理控制台 → 概览 → 概览」，默认打开「网关总览」页签。该页按所选时间范围和数据范围汇总全租户的网关调用，常用内容如下：

- 「网关运营指标」：「总消耗费用」「总 Token 消耗量」「输入 / 输出 Token」「缓存命中 Token」「首字延迟」「生成速率」「模型调用交付成功率」「MCP 工具调用」「网关耗时 P95」「端到端 P99」「活跃用户」等指标，附与上一周期的对比。
- 「各模型费用消耗走势」「Token 消耗趋势（输入 vs 输出）」「Prompt 缓存命中走势」「请求耗时构成」「请求结果归因」「小时热力」「调用流向」等趋势与归因面板。
- 「模型费用排行」「部门费用排行」「人员费用排行」三张费用排行，以及「客户端」「工具调用」「预算」面板。



前提与限制：

- 租户管理员与「运营」「系统管理员」「安全管理员」「审计员」角色可查看；「财务」角色只能看到运营指标、各模型费用走势与三张费用排行；「预算」面板只对租户管理员与「系统管理员」显示。侧栏的「概览」菜单对租户管理员与「系统管理员」显示，其他角色通过概览页面地址进入。
- 读取用量分析数据的面板（包括运营指标、Token、缓存、耗时、热力、调用流向与费用排行）需要许可包含「FinOps」模块。未授权时这些面板整格隐藏，页面仅保留告警、工具调用与 MCP 相关面板。
- 费用依赖模型定价。未配置定价的调用在排行中计为「未定价次数」，不计入金额。

建议每周核对：费用与 Token 是否突增，「模型调用交付成功率」是否下降，「端到端 P99」是否恶化，某一模型、部门或人员的费用占比是否异常，新上线的模型服务是否有实际调用。发现异常时，点击排行或归因中的行，页面会切换到「网关请求」页签并带上对应筛选，用于核对单次调用。

18. 用量与成本分析

管理员在「管理控制台 → 概览 → 概览」的「网关总览」页签分析用量、成本、可靠性与调用流向，再到同页的「网关请求」页签核对单次调用。各面板使用同一批调用事实，切换面板时不需要重新换算口径。

前提：角色、许可模块与定价要求见第 17.2 节。以美元计价的模型按「全局设置」中的「美元汇率」折算为结算币种（第 6.9 节）；未定价的调用不计入金额。

统计范围：

- 时间范围：「最近 24 小时」「最近 7 天」「最近 30 天」「本年度」，或「自定义时间范围」（不超过 366 天），默认「最近 7 天」。时间边界与图表时间轴按浏览器当前的显示时区计算，成员可在能力中心「首页 → 个人设置 → 偏好设置」的「时区」中调整。「小时热力」只支持 30 天以内的范围。
- 数据范围：点击「数据范围」，在「组织架构」中选择部门或成员（部门按每笔调用的归属部门统计，含下级部门），或在「职能组」中选择职能组；「全公司」表示不限。
- 调用条件：可再按「客户端」「模型」「接入凭据」筛选。

- 每个页签分别记住自己的筛选条件，手动切换页签不会带走当前条件；从面板下钻时，目标页签使用下钻带入的条件并保留时间范围。

18.1 分析 Token 和调用量

「网关运营指标」一行给出所选范围内的总量与对比，Token 相关指标包括「总 Token 消耗量」「输入 / 输出 Token」「缓存命中 Token」；调用相关指标包括「模型调用交付成功率」「MCP 工具调用」「活跃用户」「覆盖部门」「网关实时并发」等。每项附与上一周期的对比；没有样本时显示原因（例如「时间窗内没有耗时样本」），不显示为 0。

核对 Token 消耗及其归属：

1. 设定时间范围和数据范围，确认范围覆盖需要分析的业务时段。
2. 在「网关运营指标」记录「总 Token 消耗量」与「输入 / 输出 Token」，判断消耗由输入还是输出驱动。
3. 在「模型费用排行」「部门费用排行」「人员费用排行」中找出占比最高的对象；排行同时显示调用次数。
4. 点击排行中的行，页面切换到「网关请求」页签并带上模型、部门或人员条件，逐条核对调用的 Token 与费用。

部门归属规则：API 密钥记在部门账上时计入该部门；否则计入使用人所在部门，使用人属于多个部门时计入这些部门最深的共同上级。

核对调用量、成功率与调用主体

1. 保持同一时间范围，查看「模型调用交付成功率」，并记录其下方按「平台配置」「上游故障」「调用方」「传输中断」「未归类」拆分的失败数。
2. 在「请求结果归因」中查看各类结果的请求数与占比，类别包括「成功」「策略 · 限流 / 排队超时」「策略 · 配额 / 预算」「策略 · 护栏拦截」「上游 5xx / 超时 / 429」「上游提前断流」「客户端取消」「网关错误」；无法判定归因的调用另计为「未记录」。
3. 在「调用流向」中查看失败集中在哪个接入凭据、客户端、模型或供应商（见 18.2）。
4. 点击归因或流向中的行，进入「网关请求」页签，打开代表性调用，核对状态码、路由结论、命中规则与上游尝试。
5. 修正配置后，使用同一主体发送测试请求，并用相同筛选条件确认成功率恢复。

在到达网关数据面之前被拒绝的请求不计入「请求结果归因」；平台探针与系统调用也不计入。

分析 Token 消耗趋势

1. 选择能同时覆盖正常时段和异常时段的时间范围。
2. 在「Token 消耗趋势（输入 vs 输出）」中确认突增发生的时间点；输入 Token 含缓存读写。
3. 在「Prompt 缓存命中走势」中查看缓存命中率是否在同一时段下降。命中率 = 缓存命中 Token / 输入 Token，分母为空时显示「—」。
4. 在「小时热力」中查看星期 × 小时的请求分布，区分周期性高峰与偶发突增。
5. 依次加上「模型」「客户端」「接入凭据」条件对比，区分「请求变多」与「单次请求变长」。
6. 调整模型、配额或调用方式后，用相同粒度继续观察，确认变化持续存在。

18.2 识别高消耗接入凭据

「调用流向」面板按「接入凭据 → 客户端 → 模型 → 供应商」四层显示调用分布，用于识别调用量最大的接入凭据及其去向。

- 口径：按调用计数，每笔调用只计一次，四层合计相等，与「网关运营指标」中的请求数同源。
- 每层显示请求量最高的 8 项，其余合并为「其他 N 个」；缺少对应信息的调用分别归入「未经接入凭据」「未标识客户端」「未知模型」「未分配供应商」。
- 接入凭据节点显示为「所属人 · 凭据名」，没有所属人时只显示凭据名。
- 悬停连线或节点可查看「请求数」「失败数」「失败率」「换路后成功」与「占本列」。

操作步骤：

1. 设定时间范围，在「调用流向」中找到请求量最高的接入凭据节点。
2. 点击该节点，页面切换到「网关请求」页签并按该凭据筛选。「其他」与未标识节点不可点击。
3. 回到「网关总览」，在「接入凭据」条件中选择该凭据，查看「总消耗费用」「总 Token 消耗量」和费用走势，确认其消耗量级。
4. 将结果与该凭据对应的业务系统、模型服务和预算核对。
5. 需要完整组合时，在面板右上角选择「导出 CSV」，导出全部「接入凭据 × 客户端 × 模型 × 供应商」组合及其请求数、失败数、失败率与换路后成功次数。

18.3 成本预算和尖刺

费用相关面板：

面板	用途
「网关运营指标」中的「总消耗费用」「平均单次费用」	所选范围的总费用与单次均值
「各模型费用消耗走势」	各模型的费用按天拆分
「模型费用排行」「部门费用排行」「人员费用排行」	按费用排列，显示「消耗」「调用次数」「未定价次数」「占比」；部门按每笔调用的归属部门统计，一笔只计一次，不向上级汇总
「预算」	各配额与预算的当前用量、上限和触顶次数，点击行进入对应对象
「单场会话花费」	按模型与会话轮数分档的单场会话费用
「提示词缓存 · 按客户端」「语义缓存节省」	缓存命中与节省情况

成本尖刺发生后按以下顺序分析：

1. 确认模型定价或供应商映射近期是否调整；排行中「未定价次数」增加时，先补齐定价。
2. 在「各模型费用消耗走势」定位突增的模型和日期。
3. 在三张费用排行中确认突增来自哪个部门或人员，再按「接入凭据」条件确认来源凭据。
4. 对照「Prompt 缓存命中走势」与「语义缓存节省」，确认缓存失效没有放大实际消耗。

5. 下钻到「网关请求」，打开代表性调用，检查重复调用、长上下文和失败重试。

预算的额度在模型服务的「可用范围与配额」页签中设置（见第 7 章）。需要在费用接近预算时收到通知，在「告警规则」中启用「月度成本饱和度」规则并填写「月度预算」（见 22.2）。页面中的费用用于观察趋势与归属；涉及结算时，以供应商或企业计费系统的账单为准。

18.4 分析可靠性与耗时

可靠性与耗时相关的面板：

- 「网关运营指标」中的「首字延迟」（TTFT 均值）、「生成速率」（每秒输出 Token）、「网关耗时 P95」「端到端 P99」。
- 「请求耗时构成」：把一次调用拆为「排队」「接入层处理」「网关处理」「失败重试」「上游（流式含生成）」五段，每段是均值，可以相加。只统计耗时分段齐全的调用。
- 「生成吞吐速率」：模型生成速度随时间的变化。
- 「供应商」：各供应商的「请求」「状态」「探活延迟」「最近探活」，以及「上游 429」和「故障转移兜住」比例。
- 「上游业务系统响应耗时（P95 性能瓶颈）」与「业务工具 / MCP 调用频率与吞吐」：MCP 上游系统的耗时与调用量；耗时面板只对租户管理员显示。

分析步骤：

1. 选择异常发生的时间范围，记录请求量、成功率与各耗时指标。
2. 在「请求耗时构成」判断耗时增长落在排队、网关处理还是上游段。
3. 上游段变慢时，在「供应商」中查看探活状态、「上游 429」与「故障转移兜住」，判断是否为上游限流或故障。
4. 按「模型」条件比较正常时段与异常时段，确认是否只影响特定模型。
5. 下钻到「网关请求」，打开慢请求和失败请求，核对排队、路由、上游尝试与插件耗时。
6. 调整模型、配额或路由后，使用相同筛选条件验证耗时与成功率是否恢复。

「生成吞吐速率」「请求量」「供应商」与 MCP 耗时面板需要运维管理权限（租户管理员或「运营」角色）。样本量很小时，百分位与比例波动较大，不宜据此做重大调整。故障转移的实际行为由已发布的流量规则决定（见第 15 章），面板只反映已发生的结果。

分析接入凭据的耗时和调用明细

1. 在「接入凭据」条件中选择目标凭据，查看该凭据的请求量、成功率、Token、「首字延迟」与「端到端 P99」。
2. 去掉凭据条件，与平台整体对比，判断该业务系统是否单独变慢。
3. 查看「请求耗时构成」，确认差异集中在哪一段。
4. 切换到「网关请求」页签，保持该凭据条件，打开慢请求与失败请求，核对路由、上游尝试和耗时。
5. 调整调用方式、模型或配额后，用同一凭据发送测试请求，确认明细和汇总指标均恢复。

18.5 切换统计范围与导出

同一问题可从费用、Token、耗时和调用流向等面板观察。比较前固定时间范围和主要筛选条件，确保比较的是同一批数据。

1. 先明确问题与统计周期，例如「某接入凭据昨日失败率升高」。
2. 设定时间范围，再通过「数据范围」与「客户端」「模型」「接入凭据」条件缩小范围。
3. 核对各面板的单位与口径说明（标题旁的提示），口径不同的数字不直接相减。
4. 从面板下钻到「网关请求」，用实际调用解释汇总趋势。
5. 需要留存分析结果时，选择工具栏中的「导出汇报」，导出与指标行同一范围的 CSV 汇报；「调用流向」面板另可单独「导出 CSV」。
6. 需要定期取得用量汇总或完整调用记录时，使用「报表导出」（见 22.6）。

「导出汇报」要求账号具备运维查看权限，且许可包含「FinOps」模块。

18.6 处理负载与容量异常

当调用量、错误率或耗时明显变化时，按同一时间范围完成以下检查：

1. 在「网关运营指标」查看「网关实时并发」，在「请求量」面板查看成功与失败叠加的请求趋势，确认负载变化的时间与幅度。
2. 在「调用流向」和三张费用排行中确认变化集中在哪些接入凭据、模型或部门。
3. 在「供应商」面板确认上游是否出现 429 或探活异常。
4. 进入「网关请求」核对状态、排队、重试和代表性请求，确认实际返回由模型服务配额、API 密钥配额还是流量规则触发。
5. 需要调整时，一次只修改一类配额或保护规则，并记录原值和恢复值。
6. 使用正常业务请求验证恢复情况；持续异常时，再检查模型健康与上游服务状态。

容量问题不能只依据单个时间点或平均耗时判断，应结合时间范围、请求量、失败类型和调用明细确认影响范围。

19. 可观测入口、数据来源与权限边界

Zentrix 的可观测数据保存在客户环境内，在管理控制台中查看。经过网关的调用由网关自动记录，客户端无需改造。本章说明各入口的位置与可见角色、数据从哪里来、依赖哪些组件，以及数据安全边界。具体操作见第 20 章。

19.1 选择入口

入口	路径	可见角色与前提	用途
「网关总览」页签	管理控制台 → 概览	租户管理员与「系统管理员」「运营」「安全管理员」「审计员」，「财务」只看到费用类面板（见 17.2 节）；用量分析面板需要 FinOps 模块	整体请求量、成功率、耗时与失败归因
「网关请求」页签	管理控制台 → 概览	租户管理员、「运营」「财务」「只读审计」；「运营」只看到本人发起的请求	检索单次调用，查看结果、路由、改写、上游尝试和报文
「对话」页签	管理控制台 → 概览	租户管理员、「只读审计」	按会话查看多轮调用、耗时、Token、费用与诊断
「路由策略命中」页签	管理控制台 → 概览	租户管理员。页签对「安全管理员」也显示，但命中数据接口只对租户管理员与「系统管理员」放行，安全管理员无法加载数据；页签属于基础功能，汇总的规则来自「流量规则」	按规则查看路由策略的命中次数、放行与拦截
「会话诊断」	管理控制台 → 概览 → 会话诊断	租户管理员	自动识别会话中的问题，给出原因、证据与修复建议，处理后复查
「操作审计」	管理控制台 → 审计与合规 → 操作审计	租户管理员、「审计员」；需要全链路审计模块	配置变更、敏感数据检测与访问记录
「报表导出」	管理控制台 → 审计与合规 → 报表导出	租户管理员、「系统管理员」	下载调用与用量数据、配置定时导出
「系统运行监控」	管理控制台 → 运行与排障 → 系统运行监控	平台管理员	网关节点、控制面服务与平台依赖的运行状态
「出口网络」	管理控制台 → 运行与排障 → 出口网络	平台管理员	平台访问公网的分层时延、丢包与逐跳诊断
Prometheus 指标抓取	无界面	具备运行查看权限的账号身份；见 19.3	把按租户过滤的技术指标接入客户自有监控

「网关告警」页签与告警规则的配置见第 22 章。上述页面只显示当前账号权限与租户范围内的数据；角色由「管理控制台 → 组织与权限 → 角色管理」分配。

19.2 数据来源与依赖

数据	来源	依赖与限制
网关请求	网关自动记录每次模型、MCP、A2A 和 API 服务调用	无需客户端改造。请求列表、请求详情、「网关总览」的用量面板、用量分析与调用日志读取分析存储（ClickHouse），报文正文读取对象存储；分析存储是必装组件。分析存储暂时不可用时，这些页面提示「分析数据库读取超时，请稍后重试」，恢复后可查；网关节点在此期间把调用明细保留在本地，恢复后自动补齐，页头的「数据截至」时间停留在分析存储最后一次写入的时刻。模型调用本身不受影响
请求与响应报文	网关按「客户端请求」「发往上游的请求」「上游响应」「返回客户端的响应」四段采集	需要平台对象存储：单机裸机部署可使用自带对象存储、本机目录或现有的 S3 兼容存储，多机与 Helm 部署只能使用现有的 S3 兼容存储。报文先在网关节点暂存，再补传到中心存储；留存规则见 22.5。单段超过 32 MiB 时截断
策略命中	网关在每次请求中记录命中的路由、流量限额、内容护栏（含工具调用防护）和站点防护等策略	存放在分析存储中。分析存储未就绪时，「路由策略命中」显示「策略命中数据暂不可用（观测存储未就绪），下方为空。」，请求详情显示「命中记录暂不可用（ClickHouse 未配置或查询失败）」
会话	网关把同一会话的多轮调用串联	会话列表的「来源」列显示「网关」「客户端上报」或「网关与客户端上报」。只有网关数据时，看不到客户端本地执行的命令与工具步骤
会话诊断结论	平台按会话数据定期运行规则检查；开启模型评审后，再由所选模型评审会话	模型评审按所选模型计费，可设置月预算，见 20.4
平台运行状态	各服务实例与网关节点定期上报运行状态	实例的处理器与内存历史依赖分析存储；不可用时显示「历史数据暂不可用」。网关节点详情的本机负载、过载保护与入库状态来自节点自报，不依赖分析存储

19.3 Prometheus 指标抓取

平台在 `/ops/metrics/{service}` 路径上提供按租户过滤的 Prometheus 文本格式指标，供客户自有的 Prometheus 抓取。`{service}` 取值为 `zenova-control-plane`、`zenova-edge`、`zenova-worker` 和 `zenova-runtime-store`；模型路由的指标随 `zenova-edge` 一并返回。未部署的服务返回 404。

返回内容只包含两类序列：

- 带租户标签且属于当前租户的序列；
- 不带租户维度的技术指标，包括 `process_*`、`lobsterpool_http_server_request_duration_seconds`、`db_pool_*`、`lobsterpool_db_pool_*` 和 `gen_ai_server_*`。

限制：抓取请求必须以具备运行查看权限的账号身份认证，租户管理员、「运营」「系统管理员」「安全管理员」「审计员」具备该权限。机器调用使用「访问密钥」，每个请求单独签名。采用该接口前，请与支持团队确认采集方式；客户需要自备 Prometheus 与看板。

19.4 数据安全

观测数据可能包含用户输入、模型输出、工具参数和请求头。平台对报文采取以下控制：

- 每次读取报文正文或差异都写入一条审计记录；审计写入失败时不返回正文，页面显示「报文访问审计写入失败，未返回报文」。
- 请求头中的凭据在采集时抹除，页面显示「已抹除 · 长度 N」或「平台凭据」，不显示原值。
- 请求详情的「请求改动」只记录改写的位置与方式，不记录请求内容与请求头的值。
- 留存策略为不留存内容时，正文在网关节点上即不产出，页面显示「按租户日志策略未留存报文」；超过留存时长的报文被清理，页面显示「报文已超过租户日志策略的留存时长，已清理」。
- 会话诊断证据的原文超过保留期后显示「原文已过保留期」。

接入前应约定：哪些输入输出允许留存原文、敏感数据与内部地址如何脱敏、各角色可查看的范围、导出与归档的保留期限。「已脱敏」必须由实际报文样本证明，不能只根据配置开关的状态判断。

20. 网关请求、会话与根因定位

本章说明如何在管理控制台中查询单次网关请求、按会话查看多轮调用、处理会话诊断给出的问题、核对路由策略命中，以及平台管理员如何查看平台运行状态和出口网络。入口与权限见 19.1。

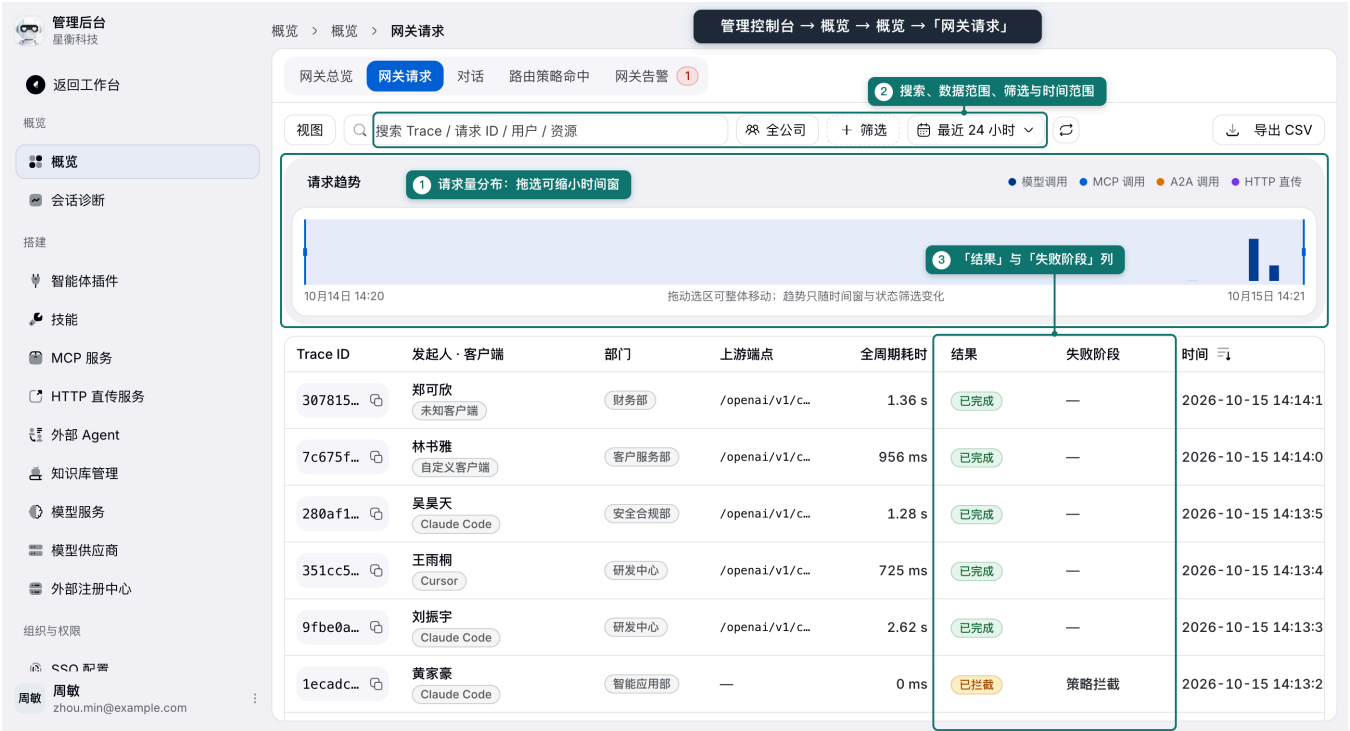
20.1 查询网关请求

入口：「管理控制台 → 概览」，选择「网关请求」页签。

页签顶部是时间范围与请求量分布条，在分布条上拖选可缩小时间窗；下方是请求列表。工具栏提供：

- 搜索框「搜索 Trace / 请求 ID / 用户 / 资源」；
- 顶栏实体筛选：「客户端」「目标模型」「接入凭据」，以及部门和成员范围；
- 列表筛选：「结果」「调用类型」「命中策略」「HTTP 状态码」「耗时区间（毫秒）」「供应商」「请求 ID」「MCP 工具」等；
- 「导出 CSV」：导出当前筛选范围内的请求。

列表默认显示调用记录，主要列包括「Trace ID」「发起人 · 客户端」「部门」「上游端点」「全周期耗时」「结果」「失败阶段」「请求造型」「耗时构成」「模型」「输入 / 输出 Token」「费用」和「时间」。「结果」取值为「已完成」「已拦截」「失败」「已取消」。



选择「结果归因」「错误类型」「责任方」「渠道」「网关模型」「发出模型」或「上游模型一致性」任一筛选后，列表切换为交付记录，按上游交付的结果展示「渠道」「网关模型」「网关 / 上游状态码」「发出模型」「上游自报模型」「一致性」「重试次数」「责任方」「交付结论」等列。「操作」列对「平台配置」类失败给出「配置路由策略」，对「上游故障」类失败给出「查看上游健康」。点击交付记录打开「交付详情」，其中「终止原因」列出「结果归因」「终止原因码」「失败阶段」「截断预算」和「供应商」，点击「查看该请求」进入请求详情。

「上游模型一致性」用于发现上游替换模型：取值「一致」「不一致」「上游未声明」。「不一致」表示上游响应中自报的模型与网关发出的模型不同。

限制：

- 「运营」角色只看到本人发起的请求，页面显示「仅显示你本人发起的请求。查看全公司的请求需要管理员、财务或只读审计权限。」
- 具备「财务」角色、但不是租户管理员或「只读审计」的账号可以浏览列表，不能打开请求详情。
- 「处理节点」筛选仅平台管理员可用。

20.2 阅读请求详情

点击调用记录进入请求详情。页头显示 Trace ID、发起人、请求造型和结果，提供「看该模型流水」「看该会话」（有会话读取权限时显示）和「刷新」；会话尚在整理时，会话位置显示「会话整理中」并自动重试，超过 1 分钟显示「会话整理较慢，请稍后刷新」，并可用上一条、下一条在当前列表中切换（「第 N / M 条」）。详情分为五个页签：

页签	内容
「请求」	「上游请求路径」（MCP 与 A2A 显示「派发目标」）、「响应状态码」，以及原始报文

页签	内容
「概要」	「身份归属」（Trace ID、请求 ID、发生时间、保留期等）、「发起人与凭据」（发起方类型、发起人、API 密钥指纹、客户端 IP 等）、「调用对象」、「模型解析」（「客户端声明的模型」「实际路由模型」「模型服务」）、「标签」
「Token 与用量」	「输入 Token」「输出 Token」「总 Token」，以及按实际执行模型单价计算的「费用」和「计入服务」
「时延与成败」	「成败」（「状态」「错误类别」「拦截原因」「失败阶段」「本窗口内被拒次数」）、「时延分段」（首字延迟、鉴权、护栏、路由分发、上游响应等耗时）、「耗时分解」瀑布图（接入层处理、排队、网关处理、失败重试、上游、端到端）
「路由与改写」	「请求去向」「上游尝试」「请求改动」「命中记录」

「路由与改写」页签按请求经过网关的顺序展示：

- 「请求去向」：「客户端声明的模型」、接管请求的「规则链」（「第 N 条规则」，最后一条标注「终局规则」）、「最终执行」的供应商与模型、「路由结论」（「已改路由」「未改路由」「已拦截」）及「原因」。本次请求未触发规则时显示「本次请求没有触发任何路由规则」。
- 「上游尝试」：每一次尝试的「序号」「档位」「供应商 / 模型」「账号 · 选号理由」「结果」「原因」「状态码」「耗时」「首字时间」。结果包括「成功」「响应被拒」「超时」「客户端中止」「流中断」等；故障转移与重试在这里逐次可见。
- 「请求改动」：「模型名」的改写（「原名 → 新名」）和命中规则对「请求头与请求体」的改写位置；「原始报文」一行提供「对比原始请求与上游请求」。
- 「命中记录」：该请求命中的全部策略，列出「时间」「规则」「处置」「结果」和「故障转移」，覆盖路由策略、流量限额、内容护栏（含工具调用防护）和站点防护。

命中语义缓存的请求不访问上游，「请求」页签中上游各段显示「命中语义缓存，未请求上游」。

对比客户端请求与上游请求

需要确认网关实际改写了什么时，在「请求」页签查看原始报文：

1. 在报文区选择段落：「客户端请求」「发往上游的请求」「上游响应」「返回客户端的响应」。发生重试或故障转移时，上游各段按「第 N 次」分开。模型与 MCP 调用显示这四段；A2A 与 API 服务调用只显示「客户端请求」与「返回客户端的响应」两段。
2. 在「正文」「请求头」或「响应头」之间切换查看；「复制全文」复制当前段。
3. 选择「差异」查看客户端请求与发往上游的请求之间的改动。差异按「模型」「系统提示词」「第 N 条消息」「工具」「请求头」等对象分组，已抹除的请求头只显示「已抹除」，不显示值，标注「新增」「删除」「改写」；跨协议转发时页面注明入站与上游的协议族。网关未改写时显示「网关未改写这次请求」。
4. 「返回客户端的响应」与上游响应完全相同时，页面显示「与上游响应完全一致（正文与响应头都相同）」。

报文不可读时，页面给出原因：「该请求发生在报文归档开启之前，没有留存报文」「报文归档中，稍后可查看」「按租户日志策略未留存报文」「报文已超过租户日志策略的留存时长，已清理」「归档存储尚未就绪，报文暂不可读」等。超过采集上限的段标注「已截断」，不能参与差异对比。每次读取正文或差异都会写入审计记录，见 19.4。

20.3 查看对话与会话详情

入口：「管理控制台 → 概览」，选择「对话」页签。该页签对租户管理员和「只读审计」角色可见。

列表按会话列出「标题」「客户端」「接入」「成员」「来源」「轮次」「费用」和「最近活动」，「客户端」显示客户端标识与图标，「接入」显示该会话所属的 Agent 观测接入名称；可按「客户端」「来源」和「诊断问题」（「有诊断问题」「无诊断问题」）筛选。点击会话进入会话详情，页头显示「会话信息」（调用方、密钥、最终用户、组织、活跃时段）和「安全策略：命中 N · 脱敏 N」。会话详情的页签按数据来源显示：

页签	内容
「概要」	来源、客户端、成员、轮次、费用与起止时间
「对话」	按轮展示用户输入与模型回复；用户回退或重新生成后被替换的轮次默认隐藏
「客户端轨迹」	客户端上报的步骤与统计；只在会话包含客户端上报数据时显示
「网关调用」	「影响到用户的故障」（分「故障」「预期拦截」「辅助调用」）、「网关兜住的异常」（重试、换渠道或排队后成功）、「调用耗时构成」，以及按服务汇总的调用数、首字中位、完成中位、成本和「命中规则」；每条失败可点击「请求详情」
「诊断」	该会话的诊断卡，见 20.4

「网关调用」页签顶部给出结论，例如「故障影响 N 轮（其中 M 轮最终完成）· 网关兜住 K 次 · 根因：…」，并提供「定位到第 N 轮」。页面提示「覆盖不全」时，表示部分轮次没有网关记录；提示「本会话没有经过网关的请求」时，客户端可能直连上游，这部分成本、失败与策略都看不到。

会话包含网关调用时，可使用「导出 JSONL」（「Hermes 训练格式」或「Turn 运维明细」）；「导出报告 PDF」生成该会话的报告。

20.4 处理会话诊断问题

入口：「管理控制台 → 概览 → 会话诊断」。该页面对租户管理员可见，页签依次为「总览」「成员」「任务场景」「最佳实践」「问题」，页签行右侧是「诊断设置」。尚未收到会话数据时，页面显示「还没有收到会话数据」。

平台对近期会话自动运行诊断规则，按「问题」汇总同类命中。当前规则包括：

规则	说明
「上游模型服务报错」「请求超出上下文上限」	上游报错与上下文超限

规则	说明
「提示词缓存反复失效」	缓存前缀不稳定，按失效原因分组（如「工具定义不稳定」「系统提示词不稳定」）
「被额度或限流拒绝」	按拒绝原因分组：「额度不足」「预算不足」「触发限流」「达到月度 token 上限」
「上游换了模型」	上游实际返回的模型与网关发出的模型不同，按渠道分组
「护栏介入」「工具调用失败」「连接器连续失败」	策略介入与工具、连接器失败
「出错后原样重试」「辅助调用带着整段上下文」「命令被钩子反复拦下」「输入只有一句「继续」「提交前没有验证」「测试先过后挂」「文档反复改写」	代理客户端的使用方式问题，依赖会话的操作轨迹

其中「提示词缓存反复失效」「被额度或限流拒绝」「上游换了模型」等规则直接使用网关记录；使用方式类规则依赖客户端上报的操作轨迹，只有网关数据时这类规则的识别范围有限。

处理一个问题：

1. 选择「问题」页签。清单列出「严重度」「问题」「负责方」「影响」「费用」、近期每日次数和「状态」，可按「负责方」「严重度」和客户端筛选。「负责方」取值包括「员工」「运行框架」「技能」「平台配置」「模型」「环境」「客户端配置」「模型与外部服务」「待判定」。
2. 点击问题进入详情。「概况」给出「概要」（费用、影响会话、次数、最近出现）、「每日发生次数」「分布特征」「原因分析」和「修复建议」；「证据」列出命中的会话，可点击「打开会话」；「处理记录」列出发现、标记与复查的历史。
3. 按修复建议完成调整后，在「标记为」中选择「已处理，开始观察」，填写「做了什么」（必填，最多 500 字）。
4. 第 7 天系统自动复查：好转则显示「已改善，问题已解决。」；未好转则显示「未改善，问题已回到未处理。」；样本不足时继续观察。「已解决」只能由系统复查判定，不能手动标记。
5. 判定不成立或决定不修复时，选择「忽略」并选择原因：「判定错误」的问题不再提醒；「接受不修」的问题在发生率翻倍时再次提醒。已忽略或观察中的问题可以「恢复为未处理」。

问题状态为「未处理」「观察中」「已解决」「已忽略」。多人同时处理时，页面提示「已被某人改为…，已刷新」。

会话详情「诊断」页签中的诊断卡显示单场会话的结论：「命中 N 个问题」「这场会话没有命中任何判定。」；诊断尚未完成时显示「这场会话尚未完成诊断，请稍后再查看。」；诊断无法完成时显示「无法完成诊断：」及原因（「诊断结果未通过数据校验」「分析库多次拒收诊断结果」「多次读取会话数据失败」）。会话有新数据时，诊断重新排队。

使用总览、成员、任务场景与最佳实践

「总览」「成员」「任务场景」「最佳实践」四个页签依赖模型评审。模型评审在「诊断设置」中开启：

设置项	说明
「评审范围」	「全部会话」「有问题的会话全部评审，其余每人每天评审 3 场」或「不评审」

设置项	说明
「会话评审模型」	本租户网关中已启用的对话模型；每场会话最多评审一次，建议选择可用的较低价模型
「月预算」	留空表示不限；达到上限后本月新会话只做规则检查，次月 1 日恢复
「预计月度费用」	按所选模型单价估算；首次开启时另行显示补评审的一次性费用
「周报推送渠道」	每周一 09:00 推送判断句与简报链接，不含成员和会话内容；所选通知渠道须订阅「会话诊断周报」事件
「采集助手回复正文」	会话的上报来源本身也允许采集正文时才生效

点击「保存设置」后，下一轮诊断起生效，已评审的会话不重评。评审调用所选模型并按其单价产生费用，「本月已用」显示当月已用金额。

开启评审后：

- 「总览」：按「统计周期」（近 7 / 15 / 30 / 90 天）、部门和客户端显示「采用」「效果」「效率」「用法」「支撑」五张指标卡，以及「需要关注」「趋势」「未达成原因」「任务类型构成」「部门对比」。
- 「成员」：按「用得最好（相对效果）」「需要帮助」「尚未用起来」分组列出成员，点击成员查看画像；「导出表格」导出成员名单，导出操作写入审计。
- 「任务场景」：按任务类型和复杂度查看「达成组与未达成组的做法差异」、按部门对比和「会话示例」。
- 「最佳实践」：由模型根据评审样本自动生成并自动发布，不需要手动创建。每条实践随机保留一部分成员暂不推送，经 14 天验证期对比两组达成率，结论为「判定生效」「未达到预期，系统自动撤下」「样本不足，系统自动撤下」等；生效满 90 天后到期。管理员可在实践详情中「撤下」，撤下后 60 天内不再生成同一做法的实践。

限制：未开启评审时，上述页签显示「未开启评审」，「问题」页签照常可用。成员侧查看最佳实践与本人诊断的页面位于观测工作台，当前不在 Zentrix 的菜单中。

20.5 查看路由策略命中

入口：「管理控制台 → 概览」，选择「路由策略命中」页签。该页签对租户管理员和「安全管理员」显示，但命中数据接口只对租户管理员与「系统管理员」放行，安全管理员打开后无法加载数据；页签属于基础功能；汇总的规则来自「流量规则」，未开通动态路由时没有可汇总的规则。

列表按规则汇总当前时间窗内的路由策略命中，列出「策略规则」「匹配判定条件」「网关执行动作」「命中次数」「7 天走势」「最近命中」，可按「处置」（「放行」「拒绝」「改路由」「改写」「降级」「限速」）和「命中情况」（「命中」「未命中」「未求值」「错误」）筛选，搜索框为「搜索规则名称 / ID」。时间范围可选「最近 24 小时」「最近 7 天」「最近 30 天」或自定义区间，自定义区间不超过 31 天。

- 点击规则行打开详情，查看「规则身份」「匹配判定条件」「网关执行动作」和「命中事实（当前时间窗）」（命中次数、放行次数、拦截次数、去重来源数、「P95 决策时延」），点击「打开策略规则」进入编辑。
- 点击「下钻命中请求 →」，切换到「网关请求」页签并按该规则筛选。
- 规则已删除但仍有历史命中时，显示「规则已删除，以下为历史命中」。

本页签只汇总路由策略。流量限额、内容护栏等其他策略的命中，在单次请求详情的「命中记录」中查看。

追踪请求中的治理动作

请求被拦截、改写、降级或改路由时，按以下顺序判断策略是否在预期位置生效：

1. 在「网关请求」中定位目标请求，记录时间、发起人、接入凭据和请求 ID。
2. 打开详情的「路由与改写」页签，在「命中记录」中按时间核对各策略的「处置」与「结果」先后顺序；在「请求去向」中核对「路由结论」与「原因」。
3. 涉及改写时，在「请求改动」中核对改写位置，并在「请求」页签用「差异」核对实际发出的内容。
4. 对照「上游尝试」，区分「策略按预期阻止」与「上游自身失败」。
5. 调整规则后，用正例、反例和边界样例分别发起请求，确认目标请求的结果改变、非目标请求保持不变；规则的命中次数在「路由策略命中」中同步变化。
6. 在「操作审计」中核对规则变更记录，并在事件记录中写明验证结果与负责人。

20.6 平台运行状态与出口网络

以下两个页面只在平台管理员的管理控制台中显示，位于「运行与排障」分组。

系统运行监控

入口：「管理控制台 → 运行与排障 → 系统运行监控」。页面每 30 秒自动刷新，也可点击刷新按钮。页面顶部汇总异常与需要关注的数量，下方分为三组：

- 「数据面 · N 个节点」：每个网关节点一张卡片，点击进入该节点的实例页。尚无节点时显示「还没有网关节点」和「前往网关节点」。
- 「控制面」：控制面服务与「后台任务」等服务状态、实例数与版本是否一致。点击服务查看实例列表：「落点」「版本」「运行态」（「备用」「接流」「已排空」）「主」「已运行」「最后上报」「状态」；「查看资源占用」显示该实例的处理器与内存趋势，可选 15 分钟、1 小时、6 小时。
- 「依赖」：「系统数据库」「权限服务」「身份提供方」「ClickHouse」「对象存储」「密钥服务」「语义缓存」。点击依赖查看「提供方式」（「平台托管」「外部提供」）、「接入地址」「连通情况」「延迟」「最近探测」「影响范围」和「处理建议」。

状态取值为「正常」「需要关注」「异常」「未启用」，原因包括「没有实例在运行」「有实例未就绪」「有实例停止上报」「实例版本不一致」「没有主实例」「无法连接」「部分实例无法连接」「响应偏慢」。实例未就绪时列出具体原因，如「配置尚未装载」「与控制面失联」「路由策略尚未装载」。

依赖异常的影响以页面「影响范围」为准，例如控制面不可用时管理后台与配置下发不可用，已下发到网关节点的配置继续生效；「ClickHouse」一项的「影响范围」写明调用明细、用量统计与日志的写入和查询不可用：「网关请求」列表与详情、「网关总览」的用量面板、策略命中、模型服务运行指标和实例历史样本都读取分析存储，期间这些页面不可查；模型调用不受影响，网关节点在本地保留调用明细，恢复后自动补齐。

出口网络

入口：「管理控制台 → 运行与排障 → 出口网络」。探测由控制面主节点发出，页面显示「探测节点」名称。

- 1. 打开右上角「启用出口网络探测」。约 1 分钟后出现第一批结果。关闭后停止发送探测包，历史数据保留，逐跳诊断仍可手动使用。
- 2. 查看分层结论。五层依次为「局域网网关」「运营商接入」「国内公网」「海外公网」「常用网站」，每层显示最近 24 小时的健康状态：「正常」「异常」「故障」「无数据」「不回应探测」。页面顶部给出结论，例如「出口网络正常」或「某层异常：丢包 X，从某时开始，已持续 N 分钟」。
- 3. 在「趋势」中选择目标与时间范围（1 小时、24 小时、7 天、30 天），查看延迟与丢包率。出网经过代理时，趋势同时显示「业务路径」与「直连对照」。
- 4. 在「探测目标」中查看各目标的「层级」「探测方式」「来源」「当前状态」「延迟」「丢包」。目标来源包括「自动识别」「内置」「模型供应商」和「自定义」；已启用的模型供应商的公网地址自动加入探测。
- 5. 需要探测其他站点时，点击「添加目标」，填写「名称」（1-32 个字）、「地址」（公网 IPv4 地址或域名）、「层级」（「国内公网」「海外公网」「常用网站」）和「探测方式」（「网站访问」「建连探测」「延迟探测」）；「建连探测」必须填写「端口」（1-65535），「网站访问」默认 443。最多添加 50 个自定义目标，只能添加公网地址。
- 6. 某个目标异常时，点击该行的「诊断」执行逐跳追踪，约需 10 秒。结果列出「跳数」「地址」「丢包」「平均延迟」，并给出「已到达目标，路径上没有持续丢包」或「从第 N 跳（地址）开始持续丢包」。历次结果保存在「诊断记录」中。

判定规则在页面「判定说明」中列出，当前版本为固定值：1 分钟内丢包不低于 5%，或延迟超过平时 3 倍且不低于 50 毫秒，记为异常；丢包不低于 50% 记为故障；连续 2 分钟成立才改变结论。同层只有一个目标出问题时，该层最多记为异常。

限制：本机不允许 ping 时，公网目标改用建连测量，「局域网网关」与「运营商接入」两层无法测量；部署不支持逐跳追踪时，「诊断」不可用。

20.7 根因定位顺序

现象	建议顺序
首字慢	「耗时分解」中的排队与网关处理 → 「时延分段」中的护栏与路由分发 → 「上游尝试」的「首字时间」 → 「出口网络」
总耗时高	「耗时分解」最长段 → 「上游尝试」中的重试与故障转移 → 供应商与出口网络
请求被拦截	「拦截原因」 → 「命中记录」中的「处置」 → 对应护栏或流量规则 → 「操作审计」中的近期变更
上游返回异常	「责任方」与「结果归因」 → 「上游尝试」的「结果」与「原因」 → 「查看上游健康」 → 「出口网络」
输出与预期不符	「请求改动」 → 「差异」 → 「上游模型一致性」 → 会话诊断中的「上游换了模型」
成本突增	「输入 / 输出 Token」 → 会话详情的「网关调用」按服务汇总 → 会话诊断中的「提示词缓存反复失效」「辅助调用带着整段上下文」
只在部分成员失败	发起人与部门筛选 → 接入凭据与可用范围 → 「失败阶段」中的「认证」「鉴权」

现象	建议顺序
平台整体异常	「系统运行监控」中的异常组 → 依赖的「影响范围」与「处理建议」 → 「出口网络」

21. 问题诊断与处理

本章按现象列出排查入口，并说明如何从一次失败的网关请求定位到原因。主要使用的页面是「管理控制台 → 概览」中的「网关请求」「对话」「路由策略命中」页签、「会话诊断」和「操作审计」；平台级问题使用平台管理员的「系统运行监控」与「出口网络」。各页面的完整说明见第 19 章和第 20 章。

现象	优先核对	进一步核对
看不到管理控制台或目标入口	当前账号是否为管理员或具备对应职责角色；许可是否包含该页面所属模块	是否登录了正确的租户；「系统运行监控」「出口网络」只在平台管理员的管理控制台中显示
登录后首页与预期不同	Zentrix 登录后进入个人首页的「使用情况」	管理员从右上角账号菜单进入「管理控制台」
供应商验证失败	地址、API 密钥、供应商类型和网络	供应商侧的模型权限；「出口网络」中该供应商目标的时延与丢包
模型服务已启用但不能用	服务中是否有已启用的模型	「网关请求」中该模型请求的「责任方」；「平台配置」表示渠道凭据没有该模型的访问权，或该模型没有可用渠道
请求被额度或速率拒绝	「网关请求」的「结果归因」是否为「策略 · 配额 / 预算」或「策略 · 限流 / 排队超时」	模型服务配额、API 密钥配额和全局限额；「会话诊断」中的「被额度或限流拒绝」问题
用量页面一直为 0	所选时间范围是否包含实际调用	调用是否发生在其他模型服务、接入凭据或成员下
成员看不到 MCP 服务	MCP 是否已发布，成员是否在可用范围内	授权有效期和组织归属
MCP Tools 为空或调用失败	端点是否可达、Tools 是否解析成功	在「网关请求」中按「调用类型」筛选「MCP 调用」，查看失败记录；排查材料中不得记录密钥
敏感数据误报或漏报	流量规则选中的检测类型、正则表达式和命中处置	用「敏感数据试运行」和脱敏样本复测 (13.3 节)
护栏未按预期拦截	流量规则中是否有启用的「内容检测」动作覆盖该调用	请求详情「路由与改写」页签的「命中记录」中是否有该次请求的命中
模型调用慢或失败	「网关请求」中按时间和「结果」筛选	打开请求详情，查看「时延与成败」和「路由与改写」页签

现象	优先核对	进一步核对
多轮会话反复失败或费用异常	「对话」页签中的会话详情	「会话诊断」中的问题清单
配置被意外修改	「操作审计」中选择时间范围和操作	在搜索框中查询操作人、动作或资源

21.1 网关请求：定位失败调用

「网关请求」记录经过网关的每一次模型、MCP、A2A 和 API 服务调用，包括结果、失败阶段、耗时、命中规则和每次上游尝试。排查运行期失败时，先在这里找到目标请求，再按详情中的线索进入对应的配置页面。

前提：

- 租户管理员、「运营」「财务」「只读审计」角色可以打开该页签。
- 「运营」角色只看到本人发起的请求；查看全租户的请求需要租户管理员、「财务」或「只读审计」权限。具备「财务」角色、但不是租户管理员或「只读审计」的账号不能打开请求详情。

操作步骤：

1. 进入「管理控制台 → 概览」，选择「网关请求」页签。
2. 在工具栏选择覆盖问题发生时间的时间范围。
3. 使用筛选器缩小范围：「结果」选择「失败」或「已拦截」；需要按失败原因查看时，选择「结果归因」「错误类型」或「责任方」，列表切换为交付记录，「操作」列给出对应的处理入口。
4. 已知请求 ID、Trace ID 或成员时，在搜索框「搜索 Trace / 请求 ID / 用户 / 资源」中直接输入。
5. 点击目标行打开请求详情；交付记录中先打开「交付详情」，查看「终止原因」后点击「查看该请求」。按下表判断原因类型。

详情中的线索	原因类型	处理入口
「时延与成败」中「拦截原因」有值， 「命中记录」中有内容护栏的命中	护栏	「护栏」中对应的敏感数据、词库或问答规则
「结果归因」为「策略 · 限流 / 排队超时」或「策略 · 配额 / 预算」，「失败阶段」为「限流」或「配额」	限流与配额	「流量规则」中的全局限额、模型服务配额、API 密钥配额
「路由与改写」中「路由结论」为「已拦截」，或「失败阶段」为「路由选择」	路由	「流量规则」中「规则链」所列的规则
「失败阶段」为「认证」或「鉴权」	调用身份	API 密钥、调用规则与可用范围
「责任方」为「上游故障」，「上游尝试」中出现「超时」「流中断」等结果	上游	「操作」列的「查看上游健康」进入「模型供应商」；平台管理员查看「出口网络」
「责任方」为「平台配置」	渠道配置	「操作」列的「配置路由策略」进入「流量规则」；同时核对模型服务的渠道凭据

详情中的线索	原因类型	处理入口
「责任方」为「调用方」	请求本身不合法，或指名了不存在的模型	通知调用方修正请求
「责任方」为「传输中断」	已开始返回，中途断开	客户端网络与平台出口网络

6. 修正配置后，由同一调用方重新发起一次请求，在列表中确认新记录的「结果」为「已完成」。

已购 FinOps 模块时，也可以从「网关总览」页签的「请求结果归因」图表进入：点击某一类结果，页面切换到「网关请求」并带上相应的「结果归因」筛选。

21.2 操作审计：定位配置变更

配置变更后出现资源不可用或行为变化时，进入「管理控制台 → 审计与合规 → 操作审计」。该页面需要全链路审计模块，租户管理员和「审计员」可以使用。

工具栏提供时间范围、搜索框「搜索操作人、动作或资源...」和筛选器；筛选项包括「按操作筛选」「按结果筛选」「数据等级」「策略来源」「策略动作」「模型」等。「操作人员」和「资源」没有独立筛选项，需要在搜索框中查询。

1. 选择覆盖问题发生时间的范围，并设置「按操作筛选」与「按结果筛选」。
2. 在搜索框中输入操作人员、动作或资源名称；需要时继续选择「数据等级」「策略来源」或「策略动作」。
3. 打开目标记录，核对操作人员、对象、变更时间、结果和页面提供的前后状态。
4. 将记录时间与「网关请求」中异常开始出现的时间对齐。
5. 确认误操作后回到对应管理页面恢复配置，再执行一次正常请求和一次无权限请求。
6. 在「操作审计」中确认恢复操作已有记录，资源状态与预期一致。

追溯 AI 调用与合规决策

需要说明一次 AI 请求被放行、拦截、改写或路由到某个模型的原因时，按同一请求 ID 和时间把以下记录串联起来：

1. 在「网关请求」中定位目标请求，记录请求 ID、Trace ID、发起人、模型、结果和发生时间。
2. 打开请求详情的「路由与改写」页签：「请求去向」给出客户端声明的模型、接管请求的规则链和最终执行的供应商与模型；「命中记录」列出该请求命中的全部策略及动作；「上游尝试」列出每次尝试的结果与原因。
3. 需要核对网关实际发出的内容时，在「请求」页签对比客户端请求与发往上游的请求。每次读取报文正文都会写入一条审计记录。
4. 涉及敏感数据时，进入「操作审计」，在「按操作筛选」中选择「敏感数据」，按相同时间范围查找检测记录。
5. 请求属于多轮会话时，点击详情页头部的「看该会话」，在会话详情的「网关调用」和「诊断」页签中查看同一会话内的其他失败与诊断结论。
6. 配置近期有调整时，在「操作审计」中查询模型、路由规则、配额、MCP 或护栏的变更记录。
7. 将请求、策略命中、会话和配置变更按时间顺序整理到同一事件记录中，区分请求失败、策略拒绝和上游失败。

8. 修正配置后，用相同入口发送一条不含真实敏感信息的验证请求，确认「网关请求」中出现新的记录且结果符合预期。

审计记录用于说明平台实际执行了什么，不替代业务系统中的结果核对。报文已按租户日志策略清理或未留存时，应以页面可见的证据完成判断，不通过其他渠道索取或传播敏感原文。

22. 告警、通知、留存、导出与操作审计

告警把调用质量、策略命中和费用变化转为可处置的事件，并通过通知渠道外发；报表导出与操作审计用于取得用量数据和合规证据。

入口与权限：

功能	入口	角色	许可模块
告警规则、通知渠道	「管理控制台 → 运行与排障 → 告警规则」	租户管理员、「运营」	基础功能
告警事件处置	「管理控制台 → 概览 → 概览」的「网关告警」页签	租户管理员、「运营」	基础功能
报表导出	「管理控制台 → 审计与合规 → 报表导出」	租户管理员、「系统管理员」	基础功能
操作审计	「管理控制台 → 审计与合规 → 操作审计」	租户管理员、「审计员」	全链路审计

「告警规则」页对租户管理员显示「告警规则」「通知渠道」两个页签；平台管理员看到的是平台告警目录，见 22.8。

22.1 创建通知渠道

从「告警规则」页的「通知渠道」页签进入。

- 选择「新建渠道」，填写「名称」，在「类型」中选择「钉钉」「飞书」「Webhook」「企业微信」或「邮件」。
- 机器人渠道与 Webhook 填写「Webhook URL」；钉钉可另填「加签密钥」。邮件渠道填写「收件人」，多个邮箱用逗号分隔。
- 在「事件类型」中勾选该渠道接收的事件，默认只接收「告警」。
- 保存后，新渠道处于「未验证」状态。在列表中选择「发测试消息」。
- 在接收端确认收到测试消息后，回到列表选择「确认收到」，状态变为「已验证」。只有已验证的渠道可以绑定到告警规则。

前提与限制：

- 邮件渠道使用「管理控制台 → 系统设置 → SMTP 配置」中的发件设置，未配置 SMTP 时邮件无法送达。
- Webhook 地址受平台出站地址策略约束，指向本机回环等受限目标时，测试返回「目标地址被禁止」。
- Webhook 渠道填写密钥时，请求附带 HMAC 签名头。

- 费用与配额通知使用告警规则中的「月度成本饱和度」「Token 额度饱和度」等规则（见 22.2），渠道的「事件类型」选择「告警」。
- 「最近投递」显示测试或真实告警中最近一次投递的时间与结果（「已送达」「失败」「死信」「未送达」）。测试成功只说明当时可达；渠道上线前应再触发一条真实规则，在接收端与告警事件中核对。

删除渠道会一并删除该渠道的投递记录，已发出的告警无法撤回。

22.2 配置告警规则

「告警规则」页签是一张规则目录，合并了四类规则，可按「分类」筛选：

来源	分类	说明
内置阈值规则	「模型调用质量」「API 服务质量」「网关与能力调用」「用量与计费」「供应商健康」「审计与会话风险」「会话健康」	五十多项内置指标，例如「模型调用成功率」「上游 5xx 比例」「首响 P95 (ms)」「月度成本饱和度」「Token 额度饱和度」「供应商健康」
护栏命中	「内容护栏」	按「内容风险类别 · 入口」逐行列出，触发条件为「命中即通知」
网关策略告警	「策略治理」	管理员新建，针对拦截、限流、故障转移等策略事件
质量告警	「质量事件」	管理员新建；在 Zentrix 中可用的触发源为故障转移次数与故障转移占比

目录列为「规则」「分类」「级别」「触发条件」「阈值」「通知渠道」「最近命中」「最近更新」。

启用内置阈值规则：

1. 在目录中找到规则，阅读「触发条件」，例如「5 分钟内平均值高于」。内置规则的统计窗口与聚合方式固定，不能修改。
2. 在「阈值」列修改数值。显示「(不可调)」的规则由后台任务直接产生事件，只能开关通知。
3. 「月度成本饱和度」一行另需填写「月度预算」，该规则按本月累计费用占月度预算的百分比判断，默认阈值为 80%。「外部 MCP QPS 基准占比」「外部 MCP 平均响应时间」两行需填写「持续 / 恢复秒数」，前者另需填写「压测基准 QPS」；这两条规则只对采集完整的统计窗口判定，通知默认关闭。
4. 为规则绑定至少一个已验证的通知渠道，再开启通知。未绑定渠道时无法开启通知。
5. 需要批量处理时，勾选多条规则后选择「配置并启用」统一设置通知渠道，或选择「停用所选」。

新建网关策略告警：

1. 选择「新建告警规则」→「网关策略事件：拦截、故障转移等」。
2. 填写「规则名称」，在「触发条件」中选择「命中次数」「拦截次数」「拦截占比」「涉及的调用方数」「命中突然归零」「故障转移次数」或「故障转移占比」。
3. 填写「统计窗口（分钟）」（不少于 5）与「阈值」。占比类按百分比填写（0 到 100），并须填写「最少样本」（不小于 1），样本不足时不判定。

4. 按需选择「策略类型」限定范围。故障转移类条件必须绑定一条流量规则；建议在「质量下降」对话框中选择这两类触发源，从「流量规则」下拉中选择规则。
5. 选择「通知渠道」后创建。网关策略告警的条件在新建时确定，之后只能开关通知和更换渠道。

配置质量告警与护栏命中通知

质量告警：

1. 选择「新建告警规则」→「质量下降：评分跌破阈值、护栏误杀上升」。
2. 在「触发源」中选择「故障转移次数」或「故障转移占比」，选择「流量规则」并填写「窗口（分钟）」与阈值。
3. 选择已验证的通知渠道，选择「保存质量规则」。质量规则可在目录中编辑或删除。

护栏命中通知：目录中「内容护栏」分类按六类内容风险（如「政治敏感与国家安全」「隐私侵权与商业合规」）与入口（如「客户模型网关」「内部调用」）逐行列出，触发条件为「命中即通知」。为需要的行绑定渠道并开启通知。风险分类、阈值与处置在流量规则的「内容检测」动作中配置（见第 14.2 节），告警规则只负责通知。

22.3 处置告警事件

租户的告警事件在「管理控制台 → 概览 → 概览」的「网关告警」页签处理；页签上的徽标显示未恢复的条数。「网关总览」中的「网关告警动态」面板列出最近触发且尚未恢复的告警。

操作	含义
「确认」	记录已知悉，状态变为「已确认」；源信号可能仍在触发
「静默」	暂停该事件的通知，状态变为「已静默」；事件仍保留
「标记已解决」	确认问题已处理，状态变为「已解除」，事件移出待处理
「指派」	指定负责人；可「取消指派」

处理时先打开「命中详情」核对规则、时间与关联对象，再到路由、配额、模型或护栏等配置入口处置。处置前后可在「网关请求」页签按同一时间范围核对调用结果。

22.4 处理投递故障

投递失败时，渠道的「最近投递」与测试结果显示失败原因：

失败原因	处理方式
「渠道未确认」「渠道已停用」	完成测试与「确认收到」，或重新启用渠道
「网络不可达」「出网被拦截」	检查地址、网络、代理与出网策略
「目标地址被禁止」「目标不合规」	改用可访问的公网 HTTPS 地址或经批准的内部通知出口，地址须以 http:// 或 https:// 开头
「凭据无效」「凭据无法解析」	检查加签密钥、Bearer token 等凭据

失败原因	处理方式
「供应商拒绝」	按提示中的供应商返回码检查机器人配置、权限与限流
「投递服务不可用」	平台投递服务暂不可用，稍后查看重试结果

投递失败后由平台按计划重试，重试耗尽的记录标记为「死信」。站内告警事件始终保留，外部通知失败不会删除事件记录。

22.5 调用记录与内容的留存

各租户使用系统默认策略：留存调用的请求与响应内容，调用内容与调用记录均永久保留。

- 清理任务按该策略执行，覆盖网关调用记录、MCP 等能力调用记录、网关会话、报文对象与会话正文。调用明细、策略命中与能力调用活动存放在分析存储，到期时间按租户的调用记录保留期计算；缩短保留期后，超期数据由清理任务删除。默认策略下不会删除数据，分析存储与对象存储的容量需要按调用量和报文大小规划。
- 操作审计记录固定在线保留两年，不受留存策略影响。
- 审计或合规要求限定留存期时，请联系技术支持确认交付版本的配置方式。

22.6 导出调用与用量数据

从「管理控制台 → 审计与合规 → 报表导出」进入。每次导出都会保留一份产物，可在任务详情中随时下载或重跑。

- 选择「新建导出」，填写「名称」（不超过 64 个字符）。
- 在「数据」中选择「用量汇总」或「完整调用记录」。选择完整调用记录时，可在「终端用户（可选）」中限定范围。
- 在「频率」中选择「跑一次」「每天」「每周」或「每月」。「跑一次」需要填写起止日期，结束日期须晚于开始日期；定时任务每次分别导出前一天、上一周或上一个月的数据。
- 在「去向」中选择：
 - 「收件箱」：文件保留在任务中，在页面提示的天数内下载；
 - 「对象存储（S3）」：填写「存储桶」和可选的「对象前缀」，并填写已有「凭据引用」（以 secretref:// 开头），或填写「对象存储地址」「区域」「访问密钥 ID」「访问密钥」及寻址方式，新凭据会先验证连接再保存；
 - 「Webhook」：填写以 https 开头的「Webhook 地址」与「凭据引用」。
- 保存后，「跑一次」的任务立即排队执行；定时任务按频率执行。列表显示「数据」「频率」「范围」「去向」「行数」与「最近一次」的状态（「排队中」「进行中」「已完成」「失败」「已过期」）。
- 打开任务查看运行历史，对已完成的运行选择「下载」，或选择「再跑一次」。可对定时任务「停用」「启用」「编辑导出」或「删除」。

验证与限制：

- 下载文件后核对行数、时间范围和数据范围。每次运行的时间窗在提交时确定并记录在运行历史中。
- 所选范围部分已进入归档时，页面提示「这段数据要从归档取回，提交后需要几分钟」；超出调用记录留存期的范围无法导出。

- 删除任务会一并删除运行历史，已导出到外部存储的文件不会被删除。
- 编辑或新建导出需要租户管理员或「系统管理员」权限。

22.7 查询操作审计并导出证据

「操作审计」记录谁在何时修改了配置、触发了动作或访问了资源。按变更定位问题的步骤见第 21.2 节，本节说明证据检索、校验与导出。

1. 进入「操作审计」，设定时间范围，在搜索框中按操作人、动作或资源搜索。
2. 按需添加筛选条件，例如「按操作筛选」「按结果筛选」「数据等级」「策略来源」「策略动作」「应用」「模型」。
3. 点击记录打开「审计记录」详情，查看操作人、资源、变更前后内容与结果；详情页可按当前列表顺序前后翻看。
4. 选择「导出审计证据 CSV」，按当前时间范围与操作筛选导出（未设时间范围时取最近 30 天）；或选择「证据 PDF」生成 PDF 证据包。PDF 在后台生成，完成后自动下载，并记入「导出历史」，之后可从「导出历史」再次下载。
5. 需要临时整理时，可选择「复制当前可见行 CSV」，只复制当前页可见的记录。

归档与完整性校验：选择「归档与校验」打开对话框。

- 「完整性校验」：显示「运行状态」「最近一次校验时间」「结果」与「覆盖范围」。平台每 10 分钟对新增审计记录做一次增量哈希链校验；选择「立即校验」可对本租户全部审计记录发起一次全量校验，识别缺失或被篡改的记录。
- 「归档批次」：审计日志与调用日志按天封存为只读批次，列表显示「类别」「封存日期」「记录时间范围」「记录条数」「保护状态」「复核结果」，可「下载批次数据」与「下载批次清单」。归档需要部署时配置审计归档对象存储，未配置时不产生归档批次。

操作审计记录在线保留两年。

22.8 平台管理员的告警目录与通知出口

平台管理员进入「管理控制台 → 运行与排障 → 告警规则」后，页面显示「活跃告警」「规则目录」「通知渠道」三个页签，用于处理部署本身的运维告警。

- 「活跃告警」：列出正在发生的平台告警，显示「场景」「对象」「严重度」「触发时间」「持续时长」「通知状态」，点击告警行打开详情。尚未配置已验证渠道时，页面顶部提示「尚未配置已验证的通知渠道。告警触发后将无法外发通知。」
- 「规则目录」：内置运维规则按场景分组（如「服务进程不可用」「主机资源耗尽」「网关过载」「数据丢失风险」），显示「严重度」「已加载」与「近 7 天命中对象」等，用于确认规则是否已加载生效。
- 「通知渠道」：页签顶部的「运维与心跳渠道」为「一般运维告警」「严重运维告警」「心跳」三类路由各提供一个选择框，从平台渠道中选择；新装环境需要三类都绑定，未绑定时页面提示「运维或心跳渠道尚未全部绑定」。只有平台范围、已启用、已验证、已勾选「告警」事件类型且目标地址以 http:// 或 https:// 开头的渠道可选，不满足的渠道置灰并说明原因，例如「该渠道尚未验证，请先在下方渠道列表里发送测试消息并确认收到。」。此外，除平台渠道列表外，「告警通知出口」可填写「平台 Webhook URL」，选择「认证」方式（「无」或「Bearer」），并「发送测试」。启用后，未配置专属 Webhook 的规则使用平台出口，站内告警始终保留。

前提与限制：

- 「活跃告警」与「规则目录」依赖「运维观测」扩展（安装时选择，或在「系统设置 → 系统更新」中启用）；未安装时，规则加载状态显示「未安装运维观测扩展」。

23. 平台运行与日常维护

运维范围：本章标注为平台管理员专属的页面只在平台管理台显示，由私有部署的平台管理员（root，或被授予“平台管理员”的平台账号）操作；SaaS 方式下由服务运营方执行，普通租户看不到这些入口。标注为租户管理员的页面在各租户的管理控制台中操作，只影响本租户。命令行操作需要在安装主机上以 root 执行。

本章说明系统升级、备份恢复、日常巡检、运行参数、部署配置、网关节点、租户以及授权许可等维护工作。升级、恢复和入口变更前，应按组织流程确认影响范围、维护窗口和回退方案，并保留操作记录。

维护入口分布如下：

一级菜单	页面	适用账号
系统设置	“系统更新”“备份与恢复”	平台管理员专属
系统设置	“通知横幅”	平台管理员发布全平台公告；租户管理员发布本租户公告
系统设置	“授权许可”“品牌”“SMTP 配置”	租户管理员
平台扩展	“部署配置”“网关”“租户管理” “SSRF 出站防护”	平台管理员专属
运行与排障	“系统运行监控”“出口网络”	平台管理员专属，需要运维管理权限

升级、回滚、备份和恢复只由 `zenova` 程序及其常驻的升级服务执行：控制台页面与命令行调用同一个执行器，产生同一份记录。

23.1 执行系统升级

平台管理员专属。从“管理控制台 → 系统设置 → 系统更新”进入，页面包含“当前版本”“升级包”“升级”和“升级历史”四个区域。

- 在“当前版本”核对“版本”“部署方式”和各能力状态。能力状态包括“已启用”“本机已有，未启用”“需要获取”和“许可证未包含此能力”。
- 获取升级包：在“升级包”选择“复制获取链接”（安装介质声明了获取地址时显示），在可访问互联网的电脑上打开链接下载升级包，或由交付方提供升级包，再回到本页选择“选择升级包”上传。页面只接受 `.zpkg` 升级包；上传中断后可选择“继续上传”。
- 上传完成后系统自动执行升级前检查。检查项包括“许可证”“磁盘空间”“平台匹配”“部署配置”“基础设施版本”“数据库变更耗时”“能力包齐全”“网关节点兼容性”“智能体目标版本”“可直接升级的最低版本”和

“版本顺序”，结果为“通过”“未通过”或“不适用”。未通过时按提示处理，例如“前往部署配置”“前往许可证”“重新选择升级包”，处理后选择“重新检查”。

4. 核对页面给出的影响说明。快速切换时控制台切换约数秒，期间操作稍慢、不会报错；不支持快速切换的部署在切换期间控制台暂不可用。两种情况下网关调用均不受影响。
5. 开始升级前，按第 23.2 节完成一次备份并记录备份标识。
6. 选择“升级到 {版本}”，在弹窗中输入当前账号的密码，再选择“开始升级”。升级开始后由系统自动完成，期间不要关闭服务器。
7. 在进度中依次查看各阶段：“准备”“控制台切换”“界面更新”“中心网关更新”“能力更新”“智能体滚动更新”“网关节点滚动更新”“数据整理”“收尾”。数据库变更在“准备”与“控制台切换”阶段内完成，不单独列为一个阶段。不适用的阶段会被跳过。
8. 完成后在“升级历史”核对“版本”“结果”“用时”“切换用时”和“操作人”，再验证登录、模型调用和常用管理操作。然后在“部署配置 → 最近一次验收”中选择“立即验收”。

升级结果的含义与处理：

结果	含义	处理
已完成	全部阶段成功	执行业务验证
已完成，部分项需处理	界面、中心网关、能力或网关节点中有一项换包失败，其余已完成	按失败项说明处理，例如检查对应网关节点
未成功，已恢复到原版本	新版本未能就绪或数据库变更失败，系统已保持原版本运行	保留原因说明，修正后重新升级
需要人工处理	切换或执行被中断，需要管理员介入	按提示选择“重试此步骤”；仍失败时保留升级记录并联系技术支持

同一时间只能有一个升级在进行；已有升级时页面提示“已有升级正在进行”。

使用命令行升级与回滚

在安装主机上执行：

```
sudo zenova upgrade ./<升级包>.zpkg
```

命令把升级包交给本机升级服务，依次完成上传、升级前检查和切换，并持续显示进度。它与控制台升级是同一个流程、同一条升级记录。

本机升级服务的版本早于升级包时（例如控制台上传后提示“发布清单版本不受支持”），请解压与升级包同一版本、同一产品版本的安装包，用其中的程序执行 `sudo ./zenova upgrade ./<升级包>.zpkg`。命令先核对升级包与安装程序、本机已安装的产品版本一致，再由新程序接管本机升级服务，随后照常升级；接管任一步失败时恢复原升级服务，本机不发生变化。这一次升级完成后，以后的升级恢复在控制台进行。

需要退回上一次升级之前的版本时执行：

```
sudo zenova rollback
```

新版本的数据库变更已经落库时，回滚会被拒绝；此时只能升级到修复版本，或按第 23.3 节恢复升级前的备份。

升级前检查提示“基础设施版本不满足要求”且平台数据库、分析数据库由本平台部署时，在维护窗口执行 `sudo zenova infra upgrade <基础设施包>`，完成后回到“系统更新”选择“重新检查”。使用现有数据库时，由数据库管理员负责升级。

当前版本早于升级包要求的“可直接升级的最低版本”时，页面提示“需要先升级到中间版本，请复制获取链接获取”；按提示先升级到中间版本，再升级到目标版本。早期以容器方式部署的安装不能升级，需按《Zentrix Gateway 部署与服务选型指南》的清空重装说明重新安装。

启用或停用可选能力

在“当前版本”的能力列表中，“本机已有，未启用”的能力可直接选择“启用”；已启用的能力可选择“停用”；“需要获取”的能力选择“复制获取链接”取得能力包。Zentrix 网关版本可选的能力包括“知识库”“语义缓存”“运维观测”和“数据高可用”，能否使用同时受许可约束。Kubernetes 集群部署的能力由集群配置管理，本页不受理启停。

23.2 备份平台数据

平台管理员专属。从“管理控制台 → 系统设置 → 备份与恢复”进入。页面顶部显示当前版本和备份目录可用空间。

- 1. 选择“立即备份”。页面提示“备份已开始，完成后会出现在列表中。”
- 2. 在任务信息中查看“当前步骤”，依次为“准备”“安装状态”“平台数据库”“分析数据库”“对象存储”“校验”。
- 3. 完成后在“备份列表”核对“创建时间”“版本”“发起人”“大小”和“包含内容”。未纳入的组件显示为“(未纳入)”及原因。
- 4. 需要把备份保存到安装主机以外时，选择“下载”取得 `.zbackup` 归档，转存到独立的存储位置。

备份内容与未纳入的情况：

组件	纳入条件
安装状态	始终纳入，包含部署配置与主密钥，需按密钥材料保管备份文件
平台数据库	由本平台部署时纳入；使用现有 PostgreSQL 时不纳入，由数据库自身的备份机制负责
分析数据库	已配置时纳入
对象存储	自带对象存储或本机目录时纳入；使用现有 S3 兼容存储时不纳入，由该存储自身保护

密钥服务的密钥目录 `/etc/zenova/secret-service/keys` 纳入主机级备份，并与备份归档分开保管；换主机恢复时需要该目录。

备份由页面或命令行手动触发。定期备份、异地转存和过期备份清理由部署管理员按组织流程安排。

命令行使用同一套备份：

命令	用途
<code>sudo zenova backup</code>	立即备份，完成后打印备份标识
<code>sudo zenova backup list</code>	列出本机备份，最新的在前
<code>sudo zenova backup export <备份标识> --output <文件></code>	导出为 .zbackup 归档
<code>sudo zenova backup import <文件></code>	导入归档，校验每个文件的摘要后登记为本机备份
<code>sudo zenova backup delete <备份标识></code>	删除一个备份

不再需要的备份在列表中选择“删除”，删除后不可找回。已有备份、恢复或升级在进行时，备份操作会被阻止。

Kubernetes 集群部署时，本页提示数据层备份由集群侧负责，不提供本机备份。

23.3 恢复和演练

平台管理员专属。恢复会用备份覆盖本机数据，备份之后写入的数据将丢失。只能恢复同一版本、同一部署形态与平台的备份；版本不一致时页面提示需先安装为同一版本再恢复。

1. 恢复优先在隔离环境演练：在演练主机安装与备份相同的版本，选择“上传备份”导入 .zbackup 归档。
2. 在“备份列表”找到目标备份，选择“恢复”。
3. 在“从备份恢复”中核对备份标识与时间，按提示输入 RESTORE，选择“开始恢复”。
4. 恢复期间控制台服务会停止，页面显示“正在恢复”；不要关闭页面。任务依次经过“停止控制台服务”、各组件恢复和“启动控制台服务”。
5. 页面提示“恢复已完成。”后，验证登录、模型调用、MCP、路由、权限、审计和告警。
6. 记录可恢复到的时间点、恢复耗时、缺失内容和需要重建的外部依赖数据。

命令行恢复时，先停止控制台服务组，再执行恢复：

```
sudo zenova stop zenova-worker
sudo zenova stop zenova-runtime-store
sudo zenova stop zenova-control-plane
sudo zenova restore <备份标识或 .zbackup 文件>
```

命令要求输入确认词 RESTORE。参数为 .zbackup 文件时先导入再恢复。完成后依次执行 `sudo zenova start zenova-control-plane`、`sudo zenova start zenova-runtime-store`、`sudo zenova start zenova-worker`。

恢复过程中旧数据在新数据完整就位前保留。在另一台主机恢复时，还需具备原主机 `/etc/zenova/secret-service/keys` 目录的主机级备份，见第 23.2 节。使用现有 PostgreSQL 或 S3 兼容存储时，这些数据不在备份中，需由对应管理员按其备份机制同步恢复到同一时间点。覆盖生产环境前，应确认完整备份、近期恢复演练、业务暂停安排和恢复后验证方案均已就绪。

23.4 日常巡检与服务日志

日常巡检建议按以下顺序进行：

检查对象	入口	关注内容
平台组成部分	平台管理员：“管理控制台 → 运行与排障 → 系统运行监控”	“控制面”“数据面”“依赖”各组的状态；“需要关注”“异常”的项目及其影响范围
网关节点	平台管理员：“管理控制台 → 平台扩展 → 网关 → 概览”	“节点总数”“在线节点”“入口池可服务”
公网出口	平台管理员：“管理控制台 → 运行与排障 → 出口网络”	探测目标的延迟与丢包趋势、诊断记录
部署验收	平台管理员：“管理控制台 → 平台扩展 → 部署配置 → 最近一次验收”	是否“已通过”，阻断与警告项
业务调用	租户管理员：“管理控制台 → 概览”	网关总览、网关请求和网关告警等页签（按权限和许可显示）
备份	平台管理员：“备份与恢复”	最近一次备份时间与包含内容

“系统运行监控”中，“控制面”包括“控制面”“运行时存储”“后台任务”；“数据面”列出各网关节点；“依赖”包括“系统数据库”“ClickHouse”“对象存储”“权限服务”“密钥服务”“身份提供方”和“语义缓存”，并标明“平台托管”或“外部提供”。打开某一项可查看实例的“版本”“运行态”“最后上报”和影响范围。外部提供的依赖出现异常时，应联系其维护方并检查网络。

“出口网络”按探测目标展示平台访问公网的延迟与丢包趋势，用于定位变慢或断开的层次。可选择“添加探测目标”（只能添加公网地址），并对目标执行逐跳诊断。关闭探测后停止发送探测包，历史数据保留。

控制台不可用时，在安装主机上执行：

```
sudo zenova status
```

该命令不依赖控制台，列出各服务运行状态、当前版本、各实例以及进行中与最近一次升级；加 `--json` 输出机器可读结果。对未运行的服务执行 `sudo zenova restart <服务名>`，服务名以 `zenova status` 的输出为准。Linux 主机使用 `journalctl -u <服务名>` 查看服务日志。控制面不可用期间，网关节点按最近一次下发的配置继续处理调用。

23.5 运行参数边界

平台管理员专属。网关过载保护在“管理控制台 → 平台扩展 → 网关”的“过载保护”页签设置，控制面数据库连接上限在“平台扩展 → 部署配置”的“资源容量”中设置：

位置	内容	生效方式
“网关 → 过载保护”	“处理器占用 (%)”和“内存占用 (%)”各设“高水位”与“低水位”。网关所在机器的占用超过高水位时拒绝新请求，降到低水位以下后恢复；低水位必须小于高水位，对全部网关节点生效	保存后下发到在线节点，无需重启。页面的生效状态行显示“正在同步到 N 个节点”或“已在全部 N 个在线节点生效”；没有在线节点时显示“当前没有在线网关节点，节点上线后自动生效”

位置	内容	生效方式
“部署配置 → 资源容量”	“控制面数据库最大连接数”，可填 12-200	保存后控制面重启，期间控制台暂时不可用，网关请求不受影响；影响本控制面下的所有租户

在“网关 → 节点”的节点详情中，每个实例显示“本机负载”（处理器与内存占用）和“过载保护”状态（“正常”“过载中”“恢复中”“负载数据不可用”）。节点健康原因显示“本机负载过高，正在拒绝新请求”时，表示占用已超过高水位；显示“本机负载数据不可用，过载保护未生效”时，应检查该节点实例是否在线上报。

其他限制分别在对应页面设置，不能用某一处配额代替其他层级的限制：

限制	入口
模型服务的可用范围与专属限制	“模型服务”详情 → “可用范围与配额”
全局限额	“路由策略 → 流量规则 → 全局限额”（需要“动态路由”许可模块）
租户每月 Token 上限	平台管理员：“租户管理” → “编辑配额”
同时在线人数、会话时长	“账号与会话安全” → “安全策略”

调整步骤：

1. 记录原值、目标值、调整原因、影响范围和计划恢复时间。
2. 保存后查看页面给出的生效状态，例如“当前生效值”或发布结果。
3. 使用受控请求验证正常范围和超限范围，并在调用记录中确认拒绝来源。
4. 观察同一时间范围内的请求量、错误率和耗时；出现异常时恢复原值。
5. 具备“全链路审计”许可模块时，在“操作审计”中核对调整与恢复记录。

23.6 维护部署配置与入口监听

平台管理员专属。从“管理控制台 → 平台扩展 → 部署配置”进入。页面列出可由管理员修改的部署配置、入口监听、数据依赖和最近一次验收结果。

区域	内容
访问地址	“访问域名”
对象存储	“对象存储方式”、服务地址、区域、存储桶、寻址方式、访问密钥，以及“报文归档存储桶”“审计归档存储桶（不可删除、不可覆盖）”
运维通知渠道	飞书告警、严重告警与心跳通知地址

区域	内容
链路导出	“链路导出地址（可选）”，即外部 OTLP 接收端地址。只填写地址不会开启导出，导出开关不在本页，启用前请与支持团队确认；开启后模型调用的链路数据默认带截断后的提示词与回复内容
入口监听	“业务入口”与“管理入口”的“端口”和“绑定网卡”
数据依赖	“平台数据库”“分析数据库”“对象存储”的部署方式（“由本平台部署”或“使用现有服务”）及管理方
最近一次验收	安装时或手动触发的部署验收结果，可选择“立即验收”

修改配置项：

1. 在目标项选择“修改”，填写“新值”并保存。敏感项只显示“已设置”或“未设置”，不回显原值。
2. 页面提示“已保存，正在让新值生效”，相关服务在后台重新加载或重启。生效完成后显示已重启的服务数量；控制台需要切换时会无中断切换到新配置。
3. 显示“有服务未能重启”或“控制台未能切换到新配置”时，在安装主机执行 `sudo zenova status` 检查。

“对象存储方式”不支持在线修改，页面提示“不支持在线修改：请备份后按新的选择重新安装，再恢复数据”。升级进行中或上一次修改仍在生效时，所有修改按钮不可用。使用现有服务的数据依赖由客户自行维护，本页不提供其凭据轮换。

在安装主机上可以用命令修改同一组配置项，例如：

```
sudo zenova config set PUBLIC_URL https://gateway.example.com
```

敏感项不接受写在命令行上，需加 `--value-stdin` 从标准输入读取，或在终端提示时输入。

修改入口端口与绑定网卡

“业务入口”提供模型、MCP、A2A 等对外服务，“管理入口”提供控制台。两个入口默认均为 80 端口，端口相同时合并为一个入口，改为不同端口后拆分监听。

1. 在“入口监听”选择对应入口的“修改”，填写端口并选择绑定网卡（“全部网卡”或某个本机地址）。
2. 保存前确认影响：修改业务入口后，外部调用方与网关节点需改用新端口；修改管理入口后，当前页面所在地地址失效，页面给出新的管理访问地址并提供“前往新地址”。
3. 两个入口使用相同端口时绑定网卡必须相同；端口已被本机其他程序或平台其他服务占用时保存会被拒绝。
4. 生效后从新地址访问控制台，并使用外部客户端验证业务入口。

安装时可通过 `--business-entry-port`、`--admin-entry-port`、`--business-entry-bind`、`--admin-entry-bind` 预置这些值。Kubernetes 集群部署的入口由集群配置管理，本页不受理修改。修改入口后，同步检查防火墙、负载均衡和“网关 → 入口与证书”中的公开访问地址，见第 24.8 节。

23.7 管理网关节点

平台管理员专属。从“管理控制台 → 平台扩展 → 网关”进入，页面包括“概览”“入口与证书”“节点”“过载保护”“出口判定”“出口代理”六个页签。控制面所在主机本身就是中心网关节点，随安装一并启动；远端网关节点用于在其他主机上就近处理请求。

添加远端节点：

1. 在“节点”页签选择“添加节点”，在“添加网关节点”中填写“节点名”（小写字母、数字和连字符，2-31 个字符，创建后不可修改）、“节点地址”（客户端访问该节点用的域名或 IP，入口证书需要覆盖它）、“地域（可选）”、“控制面回连地址”（节点主动连接控制面的地址，页面要求 https:// 开头）和“安装方式”（“Linux (systemd)”或“macOS (launchd)”。控制面使用自签或内部 CA 时粘贴“控制面 CA 证书”。
2. 按需设置“启用的面”（模型服务为必选）、“副本数”、“本地暂存配额（GB）”和“暂存满时”的处理方式。与控制面断连期间，请求记录先暂存在节点本地，恢复后补传。
3. 选择“创建并生成安装命令”。安装命令只显示一次，有效期 24 小时；过期后在节点详情选择“重新签发安装命令”。
4. 把安装器 `zenova` 和节点离线包 `zenova-node.zpkg` 放到目标主机的同一目录，在该目录以 root 运行页面给出的命令。命令形如 `sudo ./zenova install node --package zenova-node.zpkg --control-plane <控制面回连地址> --token <一次性令牌>`，令牌只能使用一次。
5. 页面显示“上线进度”，节点首次上线后显示“节点已上线”。超过时限仍未上线时，按页面列出的顺序排查。

日常维护：

- 在节点列表核对“状态”“版本”“配置落后版本数”“最近心跳”“5 分钟请求数”“5 分钟错误率”“回传积压”和“本地暂存”。
- 在节点详情的实例信息中核对“入库状态”（“正常”“明细积压”“控制面不可达”“已暂停接收”“已按自报用量结清”“未知”）、“待确认计费”“待写入明细”“最早一条已等待”和“本地剩余空间可支撑时长”。“明细积压”表示调用明细尚未写入分析数据库，节点继续转发并在本地保留，恢复后自动补齐；“已暂停接收”表示节点本地存储已满，新调用被暂缓。
- 系统升级的“网关节点滚动更新”阶段会逐台更新远端节点；节点更新失败时升级结果为“已完成，部分项需处理”。
- 节点已离线且确认不再恢复时，可在“更多操作”中选择“强制退役”，输入节点名确认。强制退役会立即吊销该节点的全部实例身份，并永久丢弃尚未回传的请求记录和本地暂存的报文，无法恢复。

“出口判定”和“出口代理”用于查看和配置节点访问外部服务的出口方式，变更前应与网络管理员确认允许的出口目标。

23.8 管理租户

平台管理员专属。从“管理控制台 → 平台扩展 → 租户管理”进入。列表可按“活跃”“已禁用”“已删除”筛选，并显示名称、状态、结算货币、管理员和本月 Token 等信息。

创建租户：

1. 选择“创建租户”，在“创建新租户”中填写“企业信息”（“租户名称”“标识”）和“默认管理员”（“管理员姓名”“管理员邮箱”“管理员密码”）。“标识”用作 IDP 组织标识，只能使用小写字母、数字和连字符。

2. 选择“结算货币”（人民币或美元）。该租户的售价、额度和账单都使用这个货币，创建后不能更改。
3. 选择“创建”，使用默认管理员账号登录该租户，完成后续成员与配置工作。

维护租户：

- “编辑租户”：修改租户名称与标识；标识变更会影响 IDP 组织映射。
- “编辑配额”：Zentrix 网关形态使用“每月 Token 上限”和“存储上限 (GiB)”。留空表示不限制，填 0 表示本租户不能调用模型或不允许上传。Token 按自然月统计，保存后十几秒内在网关生效，可能少量超出。
- “查看用量”：查看租户在所选时段的用量详情。
- 打开租户详情，在“管理员”页签选择“添加管理员”，从企业已有成员中授予管理员权限；“移除管理员”不能移除最后一个管理员。需要时可为管理员“重置两步验证”。
- 在租户详情选择“导出数据”，确认“开始导出”后异步导出该租户的全部数据，就绪后选择“下载”。
- 删除租户后可在“已删除”中选择“恢复”。

23.9 授权许可、通知横幅、品牌与 SMTP 配置

授权许可

租户管理员。从“管理控制台 → 系统设置 → 授权许可”进入。许可按租户生效，导入和移除只影响当前租户。

1. 在“升级商业授权”中选择“导出授权请求 (.json)”，导出的请求文件包含租户与部署的绑定信息。
2. 将请求文件发给供应商，换取签名后的 .lic 文件。
3. 选择“导入 .lic 文件”。许可证在本地验签，校验通过后立即生效。
4. 核对“授权用户数”“授权天数”“到期时间”“高级模块”“License ID”“授权对象”和“导入时间”。

“高级模块”显示“模型护栏”“动态路由”“全链路审计”“用量与成本”的开通状态；“用量与成本”对应 FinOps 模块。未开通高级模块时显示“未开通高级模块（仅主许可）”。未导入许可证时，网关以评估模式运行，页面显示“当前为 Zentrix 评估模式（离线包预置）。不限期试用；导入正式 License 后立即覆盖评估状态。”，“高级模块”中的四个模块均显示为已开通。选择“移除 License”后回落到评估模式。

通知横幅

平台管理员发布对所有租户生效的平台公告；租户管理员发布只对本租户生效的公告。从“管理控制台 → 系统设置 → 通知横幅”进入。

1. 选择“发布新公告”，选择级别（“信息”“警告”“严重”“系统维护”），填写内容（最多 500 字）。
2. 选择“受众”：“全体成员”或“使用指定模型服务的用户”（需选择模型服务）。
3. 设置“开始时间”“结束时间”，开始时间留空表示立即生效。
4. 需要预告模型服务下线时，填写“事件时间”，将“到期动作”设为“停用模型服务”，并选择“提醒节点”。
5. 按需勾选“发送站内信”“发送通知渠道”，核对预计覆盖人数后选择“发布”。
6. 在列表核对“状态”（“待生效”“生效中”“已结束”“已撤回”）。不再需要时选择“撤回”，撤回后不能恢复启用。

品牌

租户管理员。从“管理控制台 → 系统设置 → 品牌”配置本租户的“产品显示名”“Logo”和“Favicon”。图片支持 PNG、SVG、WebP、JPEG，最大 2MB。选择“恢复默认”清除本租户的自定义品牌。

SMTP 配置

租户管理员。从“管理控制台 → 系统设置 → SMTP 配置”进入。SMTP 用于密码重置、邮箱验证等系统邮件，需客户自备 SMTP 服务。

1. 选择“添加 SMTP”，在“发件人”步骤填写“发件人名称”“发件人邮箱”和“站点 URL”。
2. 在“服务器”步骤填写“主机”“端口”“用户名”“密码”和“连接安全”，并保持“启用自定义 SMTP”开启。建议使用支持 TLS 的 587 端口。
3. 保存后在“测试”步骤选择“发送测试邮件”，测试邮件发送到当前账号的邮箱，到收件箱确认收到。
4. 在“连接状态”核对“上次测试状态”“上次测试时间”和“上次测试耗时”。

已保存的密码不回显，编辑时留空表示保持不变。选择“禁用”后，即使已保存凭据，平台也不会发送事务邮件。

24. 常用配置的维护与变更

管理员不仅需要完成首次创建，还需要持续查询、调整、停用和清理平台对象。每次变更都应确认影响范围，保存后检查页面状态，并核对变更记录。“管理控制台 → 审计与合规 → 操作审计”页面需要“全链路审计”许可模块；模型供应商、模型服务、MCP 服务和外部智能体的详情中另有“操作审计”页签，记录该对象自身的变更。

24.1 维护管理角色、平台账号与登录会话

管理权限按职责分配，不把日常成员账号直接授予全部管理权限，也不多人共用同一个管理员账号。

租户管理角色

租户管理员从“管理控制台 → 组织与权限 → 角色管理”维护本租户六个内置角色的成员。角色职责、可见菜单、三员互斥规则和添加成员的步骤见第 3.4 节。日常维护按以下顺序进行：

1. 授予或调整角色后，使用该账号重新登录，检查管理菜单和允许动作，并确认无关页面不可访问。
2. 职责调整时先由接任人员验证权限，再对原人员选择“移除”。移除后该成员立即失去该角色的全部权限。

MCP 工具的调用权限在“路由策略”中以“工具名”“组织成员”等条件配置，不由管理角色决定。

平台账号

平台管理员专属。从“管理控制台 → 组织与权限 → 平台账号”为平台管理人员创建个人账号。平台账号为可选项，root 始终可以管理平台。

1. 选择“新建账号”，填写“姓名”“邮箱”，选择“登录方式”：“密码登录”会生成一次性激活链接，本人打开后设置密码并绑定两步验证；“仅单点登录”不生成激活链接，本人首次通过平台单点登录登录时按已验证邮箱自动绑定。

2. 需要授予全部平台管理权限时，选择“授予平台管理员”；收回时选择“撤销平台管理员”。
3. 人员暂时离岗时选择“停用”，其全部会话立即失效；确认不再使用后选择“删除”。激活链接过期时选择“重新生成激活链接”。

平台账号与 root 必须使用两步验证，不可关闭。

平台管理员从平台管理台右上角账号菜单打开「设置」，其中有「账号资料」「偏好设置」「账号与安全」「访问密钥」四个分区。「访问密钥」用于创建与撤销访问密钥，以签名方式调用平台管理接口与平台 MCP，权限与本人相同；密钥明文只在创建时显示一次。

登录会话与安全策略

从“管理控制台 → 组织与权限 → 账号与会话安全”进入。租户管理员看到本租户的会话，平台管理员看到全平台的会话。会话检索与撤销、“安全策略”各项设置及其取值范围见第 3.7 节。

平台管理员还可在“管理面访问网段”页签开启“仅允许以下网段访问管理控制台”，保存约 10 秒内所有实例生效。误配置导致无法进入控制台时，在安装主机执行 `sudo zenova admin-network show` 查看配置，用 `sudo zenova admin-network allow --cidr <网段>` 追加网段，或用 `sudo zenova admin-network disable` 关闭白名单；从安装主机本机回环地址访问始终放行。

登录事件、角色授予与移除记录在“操作审计”中查看（需要“全链路审计”许可模块）。

同步组织、维护部门并管理成员会话

租户管理员从“管理控制台 → 组织与权限 → 成员与部门”维护组织。页面包括“成员”“部门”“已离职成员”三个页签。

1. 在“成员”中按账号状态和部门筛选目标成员，核对现有归属。
2. 修改部门名称、负责人或上级部门前，查看该部门已获得的模型服务、MCP 服务、技能和外部智能体授权。
3. 调整成员所属部门或职能组后，让该成员重新登录，验证能力列表和管理菜单是否符合新职责。
4. 确认离职后删除成员，账号移至“已离职成员”，平台同时撤销其登录会话、刷新令牌与受信任设备，并吊销其 API 密钥与 MCP 访问令牌。成员丢失两步验证设备时，可为其“重置两步验证”，该操作同时让其所有会话失效。
5. 通过单点登录或目录同步维护组织时，从“管理控制台 → 组织与权限 → SSO 配置”的“已接入的身份源”选择“添加身份源”。可选的来源包括“钉钉”“飞书”“企业微信”“企业 IAM (OAuth2)”“LDAP”和“目录推送”。
6. 单次同步的变更量超过阈值时，变更暂存为“待复核”，由管理员“全部批准”或“拒绝”（见第 3.6 节）。LDAP 的分页大小在“检索范围”步骤设置。
7. 在“登录会话”和“操作审计”中检查账号状态、会话处理及组织变更。

同步来源的部门结构和成员关系在“成员与部门”中为只读，修改应在来源系统完成。组织调整完成的标志是：部门树正确、成员归属正确、授权范围随新归属生效，并且原范围不再可用。同步冲突时不要重复创建账号，应根据页面给出的用户名、邮箱或外部身份标识处理匹配关系。

24.2 维护模型供应商

从“管理控制台 → 搭建 → 模型供应商”进入供应商列表，打开供应商详情。详情包括“基本信息”“账号”“模型定价”“上游真实性”“可用性探查”“性能测试”和“操作审计”页签。

1. 在“基本信息”核对“请求地址”“模型请求格式”和已启用的模型；修改连接信息时选择“编辑 AI 供应商”，重新核对模型列表，只启用需要向当前组织提供的模型（见第 4.4 节）。
2. 在“账号”维护上游账号与 API 密钥。停用账号后立即不再接收流量；轮换密钥时先添加并验证新账号，再停用旧账号。
3. 在“可用性探查”选择“连通性测试”，查看结果、延迟和失败原因；连续失败时核对网络、凭据、协议兼容性和上游权限。
4. 在“模型定价”维护计价信息，在“操作审计”核对编辑和启停记录。
5. 删除模型前查看删除影响：页面列出引用该模型的模型服务、路由规则、API 密钥、定价记录等。被不可变路由规则引用时删除被阻止，需先修改这些规则（见第 4.9 节）。
6. 最后使用最小请求验证真实调用，并在调用记录中确认供应商、模型和状态。

保存成功只说明配置已经保存；连通性测试和真实调用均成功，才表示新的连接信息可以投入使用。凭据字段留空是否代表保留原值，以页面提示为准。

24.3 维护模型服务

从“管理控制台 → 搭建 → 模型服务”查询和维护成员实际使用的模型服务。详情包括“基本信息”“试跑”“可用范围与配额”“调用情况”和“操作审计”页签。

1. 在“基本信息”核对描述、模型列表和“发布”开关。
2. 编辑模型顺序或更换主模型前，检查目标模型已启用且供应商连接正常；修改后在“试跑”中发送一条测试请求。
3. 修改“可用范围与配额”后，分别使用获授权与未获授权账号验证列表和真实调用。
4. 暂停服务时关闭“发布”开关；恢复后执行一条最小请求，在“调用情况”确认状态和用量记录恢复。
5. 删除服务前处理 API 密钥、能力分发和路由中的引用。出现“暂不能删除模型服务”时，根据页面列出的引用对象完成迁移。
6. 变更后在能力中心检查服务名称、可见范围和实际可选状态。

需要提前告知成员模型服务下线时，可在“通知横幅”发布受众为“使用指定模型服务的用户”的公告，并设置到期动作“停用模型服务”，见第 23.9 节。

服务内某个模型不可用时，网关不会自动改用服务内的其他模型。需要失败时转到备用模型，在流量规则中配置故障转移（见第 15.7.3 节），并再次验证合规和调用结果。

24.4 维护 MCP 服务

从“管理控制台 → 搭建 → MCP 服务”进入，在“MCP 列表”中维护服务。

1. 使用名称或连接信息筛选服务，打开详情检查“发布”状态、连接方式、工具和“可用范围”。
2. 修改服务地址、Header 或鉴权方式后，重新执行“解析 Tools”。

3. 通过 OpenAPI、Postman Collection 或 cURL 更新由 HTTP 接口生成的工具时，检查新增、更新、冲突和未导入项，再保存工具目录。
4. 对工具名称、参数或说明进行调整后，使用有效参数和无效参数各执行一次测试。
5. 停止向成员提供服务时关闭“发布”或移除“可用范围”；两者承担不同作用，应分别核对。工具级的调用限制在“路由策略”中设置。
6. 删除服务前解除其他能力对它的引用，并确认没有仍在运行的调用。
7. 在详情的“调用情况”和“操作审计”中核对调用结果、耗时、使用者和变更记录。

连接验证成功但成员仍不可用时，依次检查发布状态、可用范围、有效期、成员组织归属、分发策略和凭据状态。

24.5 维护外部智能体服务

从“管理控制台 → 搭建 → 外部智能体”进入，页面包括“Agent 目录”和“任务中心”两个页签。接入、审批、上游凭据、调用方 Token、可用范围、发布为 MCP 工具和任务查询的完整步骤见第 12.1–12.4 节。日常维护按以下顺序进行：

1. 选择“立即探活”检查当前连通性；来源为名片 URL 时可重新拉取名片，检查端点、能力或鉴权声明是否变化。端点变更后凭据绑定会被暂存，需重新审批后才能再次注入。
2. 修改上游凭据或可用范围后，使用范围内与范围外的账号各验证一次。
3. 在“任务中心”按 Task ID 搜索任务，核对“目标 Agent”“调用方”“状态”和“耗时”。需要取消时选择“取消任务”；能否终止取决于对端实现，应确认任务实际进入“已取消”。
4. 删除前检查活动任务、调用方、路由及可用范围引用。删除后该 Agent 的网关调用凭据随之失效，此操作不可撤销。

接入完成的标志见第 12.4 节。

24.6 核对配置变更

对组织、模型、MCP 和外部智能体的重要变更使用同一套检查方法：

1. 记录变更对象、原因、影响范围、原值和预期结果。
2. 在页面中保存变更，确认状态不再处于等待、失败或未发布。
3. 使用一个获授权账号完成正向操作，再使用一个未获授权账号验证拒绝结果。
4. 在“概览”的“网关请求”页签或对象详情的“调用情况”中检查真实执行结果。
5. 在对象详情的“操作审计”页签或“操作审计”页面核对操作人员、对象、结果和时间。
6. 结果异常时恢复原配置，并保留请求标识和页面错误供排查。

24.7 接入新的模型与外部能力

新增模型供应商、MCP 服务、API 服务或外部智能体时，使用一致的接入顺序：

1. 确认能力用途、负责人、网络目标、认证方式和数据边界。
2. 在对应管理页面创建连接，凭据只填写在受保护字段中。
3. 执行页面提供的验证、模型获取、Tools 解析或连通测试，确认平台能够读取预期能力。
4. 配置模型服务、发布状态和可用范围，从试点部门或职能组开始授权。

5. 使用获授权账号完成真实调用，使用未获授权账号验证拒绝。
6. 在调用记录和变更记录中核对使用者、目标能力、结果与耗时。

接入多个同类能力时，应使用能够区分环境、用途和数据边界的名称，避免成员在能力中心选错。

24.8 变更访问地址与入口证书

平台管理员专属。平台的对外访问地址和入口证书从“管理控制台 → 平台扩展 → 网关”的“入口与证书”页签维护；入口端口与绑定网卡在“部署配置”中修改，见第 23.6 节。

修改公开访问地址

1. 在“公开访问地址”中填写“业务访问地址”。该地址供模型、MCP、A2A 等对外服务使用，网关入口域名与证书签发名由它的主机名派生。填写 `http://` 或 `https://` 开头的主机名或 IP，不带路径。
2. 按需填写“管理访问地址”。留空时使用业务访问地址的主机加管理入口端口。
3. 选择“保存”，在“确认修改公开访问地址”中核对变更前后的地址，选择“确认保存”。保存后立即生效：对外服务的端点与回调使用业务访问地址，控制台链接使用管理访问地址。
4. 需要回到安装时的地址时选择“恢复安装地址”。

修改访问地址前，应先在 DNS、负载均衡和防火墙中准备好新地址，并通知调用方更新端点和回调地址。

签发或上传入口证书

入口证书是覆盖入口域名的通配符证书，下发到所有网关节点入口。业务访问地址是 IP 地址时不能签发证书。

自动签发：

1. 在“自动签发证书”中开启“启用自动签发与续签”。“入口域名”由公开访问地址派生，不能单独修改。
2. 填写“DNS 托管域”，选择“DNS 服务商”（“Cloudflare”或“阿里云云解析”）并填写对应凭据。凭据需要该托管域的 DNS 编辑权限，保存后不再显示。
3. 按需填写“联系邮箱（可选）”，保存后选择“立即签发”。
4. 状态变为“已签发”后，核对“当前证书有效期至”。证书在到期前 30 天自动续签。

签发失败时，按页面原因处理：写入 DNS 验证记录失败时检查凭据权限与托管域；连不上证书颁发机构时检查控制面出网；域名验证未通过时等待 DNS 记录生效后重试。

手动上传：

1. 不使用自动签发时，在“手动上传证书”中粘贴“证书链 (PEM)”和匹配的“私钥 (PEM)”，选择“上传 / 替换证书”。
2. 在“当前证书”核对“主体备用名称”“过期时间”“SHA-256 指纹”和“节点入口下发”进度。

检查入口自检

在“入口自检”中核对“入口域名”“部署级通配符证书”和“来源 IP 规则可信代理配置”等项。有项目未就绪时按提示处理后重新自检；全部就绪后，从外部客户端通过新地址发送最小请求，在调用记录中核对结果。

25. 分阶段上线

新能力应先在试点范围内验证，再逐步向全员开放。建议按以下顺序上线：

1. 选择一个试点部门或职能组，在「授权许可」中确认本次上线涉及的许可模块已经开通。
2. 配置一个模型供应商和一个面向明确任务的模型服务，在模型服务详情的「试跑」页签确认调用成功，并在「可用范围与配额」中只授权试点范围。
3. 为试点范围设置保守的配额和限流保护；开通动态路由时，按需在「流量规则 → 全局限额」中设置次数与并发限额。
4. 开通模型护栏时，在「护栏」中维护敏感数据规则、词库和问答，在「全局设置」中为内容安全、敏感信息识别和问答语义匹配指定默认模型；再在流量规则中添加内容检测，确定「组件不可用时」中「护栏检测出错」的处理方式和拒答话术（见第 14.4 节）。
5. 需要外部工具时，配置 MCP 服务的连接、工具权限，并为可用范围设置有效期。
6. 在「流量规则」的「规则清单」中选择「验证规则」，输入试点请求的模型、密钥和请求内容后选择「模拟请求」，核对命中的规则、执行顺序和最终目标；勾选「模拟主档失败」，确认后续候选目标符合预期。保存规则时，在「确认变更」中核对本次变更对命中顺序和关联引用的影响，再选择「确认并应用」。
7. 加入少量试点成员，由成员在能力中心「首页 → API 密钥」创建 Key，并使用获授权的模型服务或 MCP 服务完成调用。
8. 完成正常任务、敏感数据、风险内容、无权限资源、异常输入和额度耗尽等测试。
9. 确认规则已在网关生效：「规则清单」上方不应出现“有若干条规则被数据面结构性拒绝，未参与判定”的提示；出现时按列出的原因（例如规则编译失败、规则尚未同步到数据面、引用的内部系统清单不存在）修正后重新保存。在规则详情中选择「查看命中的请求」，确认试点调用确实命中了该规则。
10. 在「概览」的「网关请求」「路由策略命中」和「对话」页签中核对试点调用的路由结论、命中记录和多轮会话。
11. 试点运行一段时间后打开「会话诊断」，查看「问题」页签中识别出的「上游模型服务报错」「请求超出上下文上限」「被额度或限流拒绝」「上游换了模型」等问题；处理后选择「已处理，开始观察」并写明所做调整，系统在第 7 天自动复查是否好转。
12. 在「操作审计」中确认本次上线的配置变更记录完整（需要全链路审计模块）。
13. 根据验证结果和真实用量调整策略，再逐步增加部门、模型服务、MCP 服务和成员。

上线检查清单

- ☐ 部门、负责人和成员归属正确
- ☐ 租户管理员只授予必要人员；系统管理员、安全管理员、审计员由不同人员担任
- ☐ 「安全策略」中的会话超时、登录方式、强制两步验证和密码策略已确认
- ☐ 模型供应商验证成功，密钥未在文档或工单中泄露
- ☐ 模型服务名称、描述和模型顺序正确，「试跑」成功
- ☐ 可用范围只包含目标部门、职能组或成员
- ☐ 配额、限流保护、全局限额的周期、单位和并发已复核
- ☐ 敏感数据规则的正则表达式与类型，以及流量规则中选中的检测类型和命中处置已经复核

- ☐ 内容检测的检测范围、风险分类、命中动作、「护栏检测出错」处理方式和拒答话术已经验证
- ☐ 「验证规则」的模拟结果与真实调用的路由结论一致，规则清单没有结构性拒绝提示
- ☐ MCP 连接、工具权限、发布状态和可用范围已验证；临时授权设置了截止日期，非授权成员无法调用
- ☐ 「概览」中的网关请求、路由策略命中和对话记录可查询，「会话诊断」中没有未处理的高影响问题
- ☐ 通知渠道发送测试成功，告警规则已覆盖关键调用质量指标
- ☐ 已明确内部支持联系人和异常升级方式

26. 上线检查、稳定运行与管理交接

运维范围：本章的平台级检查由私有部署的平台管理员执行；SaaS 方式下由服务运营方执行，租户管理员只检查本租户范围内的项目。受许可模块控制的页面（“护栏”需要“模型护栏”，“流量规则”需要“动态路由”，“操作审计”需要“全链路审计”，用量与成本需要“FinOps”）只在许可包含相应模块时出现。

上线检查以真实账号和真实入口完成，重点确认成员能完成工作、无权限人员被正确拒绝、异常能够定位和恢复。

26.1 完成上线前检查

平台部署检查（平台管理员）：

1. 安装器输出的装后验收全部通过；在“部署配置 → 最近一次验收”中结果为“已通过”，没有阻断项。
2. 已保存 root 初始密码，并在首次登录时修改密码、绑定两步验证；平台管理人员已使用个人平台账号。
3. “系统更新”中当前版本、部署方式和已启用能力符合交付清单。
4. “系统运行监控”中控制面、数据面和依赖均无“异常”项；“网关 → 概览”中中心节点及远端节点在线，“入口池可服务”。
5. “网关 → 入口与证书”中业务访问地址、管理访问地址和入口证书正确，“入口自检”全部就绪；外部客户端可以通过业务入口访问，防火墙或安全组已放行业务入口与管理入口端口。
6. “部署配置”中运维通知渠道已配置；“数据依赖”中各依赖的部署方式与管理方符合方案。
7. 已执行一次备份，记录备份标识，并将归档转存到安装主机以外；已明确恢复责任人和维护窗口。

租户使用检查（租户管理员）：先逐项完成第 25 章的「上线检查清单」，再补充以下检查：

1. “授权许可”中授权用户数、到期时间和高级模块符合合同。
2. 需要单点登录时，“SSO 配置”中的身份源已启用并完成一次登录。
3. 使用试点成员账号核对首页、能力列表和可用范围，并分别验证模型服务、MCP 服务、API 服务、外部智能体和 API 密钥的正常调用。
4. 需要系统邮件时，“SMTP 配置”已通过测试邮件验证。
5. 记录未完成项、影响范围、临时处理方式和负责人，再决定是否扩大使用范围。

26.2 查看平台负载并处理容量问题

1. 选定同一时间范围，在“概览”中查看请求量、错误率和响应耗时；平台管理员同时在“系统运行监控”中查看各组成部分的状态与实例资源占用。
2. 按模型服务、API 密钥、外部智能体或网关节点缩小影响范围。网关节点的“5 分钟请求数”“5 分钟错误率”和“回传积压”可用于判断单个节点的负载。
3. 打开代表性调用，区分上游响应慢、工具执行慢、护栏处理耗时或重复调用。上游不可达时，在“出口网络”检查平台访问公网的延迟与丢包。
4. 检查模型服务的“可用范围与配额”、流量规则的“全局限额”、租户的“每月 Token 上限”和上游供应商状态。
5. 整机资源紧张导致请求被拒绝时，在“网关 → 过载保护”核对处理器与内存的高、低水位，并在“网关 → 节点”的节点详情查看“本机负载”与“过载保护”状态。
6. 一次只调整一类限制，并记录原值和恢复值。
7. 使用正常业务请求验证恢复，再继续观察同一时间窗口的错误和耗时变化。

“最多在线人数”与模型请求限制属于不同层级。登录受限时检查“账号与会话安全 → 安全策略”中的在线人数上限与会话设置；模型请求被拒绝时检查配额、限额和过载阈值。

26.3 配置和检查平台安全控制

1. 核对“SSO 配置”中的身份源，以及“安全策略”中的空闲超时、绝对超时、登录方式、强制两步验证和密码策略。
2. 平台管理员核对“管理面访问网段”是否只允许运维网段访问管理控制台，以及“SSRF 出站防护”的放行范围。
3. 核对模型服务、MCP 服务、技能和外部智能体的可用范围，以及路由策略中针对 MCP 工具的规则。
4. 核对“护栏”中的敏感数据规则、词库与问答，以及流量规则中内容检测的检测项和命中处置。
5. 核对外部连接使用的凭据、HTTPS、允许访问的出口目标和审批方式；外部智能体需经“批准”后才能配置调用方访问。
6. 使用一个允许样例和一个拒绝样例验证每项重要策略。
7. 在护栏命中记录、“操作审计”和调用记录中确认命中结果、执行结果和操作人员可以关联。

安全检查不能只使用管理员账号。至少使用普通成员、业务管理人员和平台管理员三类账号核对界面可见性与直接访问结果。

26.4 完成平台管理交接

1. 为接任人员分配与职责相符的管理角色：租户内通过“角色管理”，平台级通过“平台账号”授予“平台管理员”，并使用其账号核对可见菜单。
2. 确认组织、模型、MCP、外部智能体、护栏、告警、审计、备份恢复、升级和网关节点的负责人。
3. 轮换由原管理员保管的 API 密钥、上游账号密钥、入口证书私钥、DNS 服务商凭据和连接凭据。备份文件含主密钥，应与密钥服务的密钥目录 `/etc/zenova/secret-service/keys` 的主机级备份一起按密钥材料移交保管责任。
4. 原管理员掌握 root 密码时，在安装主机执行 `sudo zenova root-password set` 重置。该命令打印新的初始密码，清除 root 的两步验证并让 root 的全部会话失效。

- 5. 由接任人员完成一次模型或能力变更、一次异常定位和一次备份。
- 6. 核对支持联系人、维护窗口、运维通知渠道和紧急处理方式。
- 7. 确认接管完成后，通过“移除”角色成员或“撤销平台管理员”收回原管理员不再需要的权限，并核对变更记录。

26.5 提交支持请求

提交支持请求时，按第 27 章准备信息并遵守其中的保密要求。

27. 联系支持所需信息

为了快速定位问题，请提供：

- 问题发生的准确时间和时区，以及出现问题的页面入口；
- 当前租户、部门和成员账号名称，以及账号持有的角色；
- 交付版本号（平台管理员可在「系统更新」中查看）和「授权许可」中显示的已开通模块；
- 受影响的模型服务、模型、MCP 服务、API 服务或外部智能体的名称；
- 期望结果、实际提示和可复现步骤，以及问题能否稳定复现；
- 页面错误提示或经过脱敏的截图。截图应保留导航、字段名称、状态和错误上下文，只遮盖真实敏感值；
- 管理员在「概览 → 网关请求」详情中看到的「请求 ID」「Trace ID」和路由结论，或成员在「首页 → 调用日志」中看到的「Request ID」；涉及多轮会话时附上会话 ID；
- 与问题相关的规则名称，以及「验证规则」的模拟结果（如已执行）；
- 「会话诊断」中对应问题的名称与证据（如已识别）；
- 最近是否修改过模型供应商、模型服务、MCP 服务、护栏、流量规则、可用范围、配额、凭据、网络、安全策略、成员角色或平台版本（可在「操作审计」中核对），以及已经尝试的处理；
- 涉及安装、升级或备份时，附上 `sudo zenova status` 的输出、升级历史中的结果说明或备份任务的失败原因。

不要在邮件、聊天或工单中提供密码、API 密钥、访问令牌、私钥、数据库密码、备份文件、应用密钥、完整连接串或未经授权的业务数据。支持人员如需进一步信息，应通过组织认可的安全渠道收集。

界面入口与可见字段可能因账号角色、许可模块和交付版本不同而变化。截图中的邮箱、密钥、令牌、Header 值和其他敏感信息均已局部脱敏。

附录 A. 角色与入口速查

能力中心

位置	内容	说明
登录后的首页	能力中心「首页」，默认打开「使用情况」	见 1.2 首页与导航

位置	内容	说明
顶部页签	「首页」「模型广场」「MCP 广场」「API 广场」「知识库」「技能广场」「插件广场」「外部智能体广场」	「首页」登录后显示；「知识库」需要许可证包含知识库能力且部署已安装；其余目录页未登录也可打开
「首页」左侧导航	用量：「使用情况」「模型用量」「调用日志」；资源：「模型服务」「API 密钥」「凭据管理」；个人设置：「个人资料」「偏好设置」	见 第 8 章 与 3.8 节
右上角账号菜单	显示名称与邮箱、「管理控制台」「退出登录」	账号至少能看到一个管理菜单时才显示「管理控制台」

管理角色与侧边栏菜单

租户管理控制台的菜单由账号的管理范围、持有的角色和许可证包含的模块共同决定（见 1.2 首页与导航）。下表列出各角色在许可模块齐全时可见的侧边栏菜单。

角色	获得方式	侧边栏可见菜单	常用章节
成员	加入租户的账号	无管理菜单，使用能力中心	第 8 章、11.8、12.3
租户管理员	租户属主；或在「添加成员」中开启「管理员权限」，或在成员操作菜单中分配	1.2 节所列全部租户菜单，包括只对租户管理员显示的「会话诊断」「角色管理」	第 3 章至第 25 章
系统管理员	「角色管理」中添加成员	「概览」；「搭建」分组全部菜单；「成员与部门」「职能组」；「网关扩展」；「报表导出」；「系统设置」分组全部菜单	第 4 章至第 7 章、第 11 章、第 12 章、22.6、23.9
安全管理员	「角色管理」中添加成员	「SSO 配置」「账号与会话安全」「护栏」「威胁情报」「流量规则」（「全局限额」「缓存」页签无法加载与保存）	3.6、3.7、第 13 章至第 15 章
审计员	「角色管理」中添加成员	「操作审计」（需要全链路审计）	21.2、22.7
运营	「角色管理」中添加成员	「告警规则」	22.1至22.4
财务	「角色管理」中添加成员	无；由管理员提供「概览」页面地址后直接打开	17.2、第 18 章
只读审计	「角色管理」中添加成员	无；由管理员提供「概览」页面地址后直接打开	20.1、20.3

角色	获得方式	侧边栏可见菜单	常用章节
平台管理员	安装时创建的 root；或在「平台账号」中被授予「平台管理员」的平台账号	平台范围菜单：「SSO 配置」「账号与会话安全」「平台账号」「告警规则」「系统运行监控」「出口网络」「通知横幅」「系统更新」「备份与恢复」「部署配置」「SSRF 出站防护」「网关」「租户管理」	第 23 章、24.1、24.8、14.6、20.6、22.8

系统管理员、安全管理员、审计员三个角色两两互斥；运营、财务、只读审计可以与任一角色叠加（见 3.4 角色管理）。许可证未包含模型护栏、动态路由或全链路审计时，「护栏」「流量规则」「操作审计」不显示。

「概览」页签与角色

「概览」只在租户管理员和系统管理员的侧边栏中显示；其他管理角色需要时直接打开概览页面地址，页面按角色显示可用的页签。

页签或页面	可以查看的角色	说明
「网关总览」	租户管理员、系统管理员、安全管理员、审计员、运营、财务	用量、费用与耗时面板需要 FinOps；财务只看到费用相关面板；「预算」面板只对租户管理员和系统管理员显示，见 17.2
「网关请求」	租户管理员、运营、财务、只读审计	运营只看到本人发起的请求；只有财务角色的账号不能打开请求详情，见 20.1
「对话」	租户管理员、只读审计	见 20.3
「路由策略命中」	租户管理员（安全管理员可见页签，但无法加载命中数据）	规则来自「流量规则」，见 20.5
「网关告警」	租户管理员、运营	见 22.3
「会话诊断」	租户管理员	侧边栏「概览」分组下的独立菜单，见 20.4

不在侧边栏中的入口

入口	打开方式	说明
「验证规则」	「流量规则 → 规则清单」右上方的「验证规则」	见 15.3
开放 API 规范文件	向交付方索取；访问入口不提供在线文档	见 16.4

附录 B. 常用任务索引

按使用顺序列出常用任务及对应章节。标注「平台管理员」的任务只在平台范围的管理控制台中操作。

开通与组织

常用任务	参考章节
了解产品范围、许可模块与推荐配置顺序	第 1 章 • Zentrix 概览与配置流程
核对菜单的可见条件	1.2 首页与导航
了解各配置对象及其入口	第 2 章 • 平台对象和边界
导入许可证并核对已开通的模块	23.9 授权许可、通知横幅、品牌与 SMTP 配置
建立部门并添加成员	3.1 新建部门、3.2 添加成员
处理成员离职与恢复入职	3.3 维护成员与已离职成员
为系统管理员、安全管理员、审计员等角色指定成员	3.4 角色管理
建立跨部门的职能组	3.5 职能组
接入钉钉、飞书、企业微信、企业 IAM、LDAP 或目录推送	3.6 SSO 配置
设置会话时长、登录方式、密码策略与 令牌有效期	3.7 账号与会话安全
开启两步验证与通行密钥	3.8 个人账户安全：两步验证与通行密钥

接入模型与开放模型服务

常用任务	参考章节
新建模型供应商并获取可用模型	4.1 新建并验证供应商
排查供应商连接失败	4.2 可用性探查与连接排查
调整超时、重试与健康检查参数	4.4 编辑供应商与运行参数
添加上游账号	4.5 账号与上游余额
声明模型的能力用途	4.6 模型能力用途与元数据
核验上游是否提供所声明的模型	4.7 上游真实性核验
测试上游的并发容量与延迟	4.8 性能测试
设置模型售价与阶梯计价	5.1 设置模型售价、5.2 阶梯计价
参照目录价定价或批量填充售价	5.3 参照目录模型定价、5.4 按目录价批量填充

常用任务	参考章节
维护模型广场中的模型分类	5.5 模型分类
新建模型服务	6.1 新建模型服务
试跑模型服务并查看失败归因	6.5 试跑与失败归因
为内容安全、敏感信息识别、意图识别等功能指定默认模型	6.9 全局设置：默认模型与美元汇率
设置模型服务的可用范围与有效期	7.1 选择授权范围并设置有效期
设置对象级与服务级配额	7.2 配额字段说明、7.3 服务级配额
为单个 API 密钥设置额度	7.5 单个 API 密钥的额度
处理额度用尽导致的拒绝	7.6 配额拒绝排查

成员使用与业务系统接入

常用任务	参考章节
浏览模型广场并在线试用	8.1 浏览模型广场、8.2 在线试用
获取 API 密钥并发起首次调用	8.3 获取 API 密钥并发起首次调用
撤销本人的 API 密钥	8.4 管理本人的 API 密钥
查看本人用量与调用日志	8.5 查看本人用量与调用日志
为业务系统准备 API 密钥	9.2 API 密钥的创建与管理
轮换、禁用或删除 API 密钥	6.7 API 密钥管理、9.3 轮换、禁用和应急处理
设置能力授权的生效与到期时间	9.4 分发能力和设置生效时间
选择调用认证方式，启用身份源令牌调用	9.5 认证方式支持边界
按调用方计数与限额	9.7 调用身份、计数与配额
定位鉴权失败	9.8 鉴权故障定位
查找 Base URL、鉴权头与接口清单	10.1 Base URL 与鉴权、10.2 接口清单
用 Claude Code、Codex 或 Gemini CLI 调用任意对话模型	10.3 协议自动转换
处理流式输出、重试与心跳	10.4 流式输出、重试与心跳
调用嵌入、重排序、图像、语音与批量判断接口	10.5 嵌入与重排序、10.6 图像与语音、10.7 批量判断
按错误码处理调用失败	10.8 错误格式与错误码
指定会话归属，续接 Responses 对话	10.9 会话识别

接入工具与外部能力

常用任务	参考章节
接入远程 MCP 或把 HTTP 接口生成为 MCP 工具	11.1 创建 MCP 服务
配置 MCP 鉴权与请求头	11.2 配置鉴权与请求头
发布并分发 MCP 服务	11.3 发布与分发
验证 MCP 连接并排查不可用	11.4 MCP 连接验证与排查
限制单个工具的调用	11.5 设置工具调用权限
查看 MCP 调用情况、管理访问令牌	11.6 查看调用情况与可用性、11.7 管理访问令牌 与分类
成员获取 MCP 接入配置	11.8 成员在能力中心获取 MCP 接入配置
在 MCP 客户端中管理网关	11.9 通过 MCP 管理网关
以 API 服务方式开放上游接口	11.10 API 服务
接入、审批外部智能体	12.1 接入外部智能体
配置外部智能体上游凭据与调用方 Token	12.2 配置上游凭据与调用方 Token
分发外部智能体或发布为 MCP 工具	12.3 分发外部智能体与发布为 MCP 工具
查询外部智能体任务	12.4 监控外部智能体与查询任务
导入、检查并分发技能	12.5 导入与分发技能
导入或导出插件	12.6 导入与导出插件
从 Nacos 同步外部智能体与 MCP Server	12.7 同步外部注册中心

护栏与内容检测

常用任务	参考章节
查看内置敏感数据规则	13.1 敏感数据页签与内置规则
新增、导入或导出自定义敏感数据规则	13.2 新增、导入与导出自定义规则
试运行敏感数据检测	13.3 试运行
在流量规则中启用敏感数据检测并设置处置	13.4 在流量规则中启用检测与命中处置
核对脱敏结果与原文证据	13.5 脱敏结果与审计
维护词库、问答与内部系统白名单	14.1 维护词库、问答与内部系统白名单
在流量规则中添加内容检测	14.2 在流量规则中添加内容检测

常用任务	参考章节
识别规避写法，检测危险工具调用	14.3 规避写法识别与工具调用防护
设置检测失效处理、拒答话术与流式检测	14.4 失效处理、拒答话术与流式输出
验证内容检测并配置命中通知	14.5 上线验证与命中通知
开启 SSRF 出站防护与 Web 应用防护（平台管理员）	14.6 SSRF 出站防护与 Web 应用防护

路由与流量保护

常用任务	参考章节
理解规则的入口、类型与生效顺序	15.1 入口与生效顺序
创建模型路由或增强规则，配置匹配条件	15.2 创建规则与配置条件
用「验证规则」和真实调用验证规则	15.3 验证规则与真实调用
配置全局限额、并发限制与临时豁免	15.4 配置全局限额
配置处置动作、改写与异常处理	15.5 处置动作与异常处理
配置并验证熔断	15.6 配置并验证熔断
配置流量分配与故障转移	15.7 流量分配与故障转移
维护规则、处理校验失败与保存冲突	15.8 维护、校验与故障定位
按意图分流	15.9 意图集
配置语义缓存	15.10 语义缓存
转换请求和响应	16.1 请求和响应变换
安装、绑定和升级网关扩展	16.2 安装、配置和绑定网关扩展
配置插件的请求屏蔽、防重放与并发上限	16.3 配置插件网关策略
获取开放 API 文档	16.4 获取开放 API 文档
完成集成上线检查	16.5 集成上线检查

用量与成本

常用任务	参考章节
管理员查看全租户用量与费用	17.2 管理员查看用量与成本
核对 Token 消耗及其归属	18.1 分析 Token 和调用量

常用任务	参考章节
核对调用量、成功率与调用主体	18.1 核对调用量、成功率与调用主体
分析 Token 消耗趋势	18.1 分析 Token 消耗趋势
识别高消耗接入凭据	18.2 识别高消耗接入凭据
分析成本预算与费用尖刺	18.3 成本预算和尖刺
分析可靠性与耗时	18.4 分析可靠性与耗时
分析接入凭据的耗时和调用明细	18.4 分析接入凭据的耗时和调用明细
切换统计范围并导出汇报	18.5 切换统计范围与导出
处理负载与容量异常	18.6 处理负载与容量异常

观测与排障

常用任务	参考章节
按问题选择观测入口	19.1 选择入口
了解观测数据的来源、依赖与安全控制	19.2 数据来源与依赖、19.4 数据安全
把租户指标接入 Prometheus	19.3 Prometheus 指标抓取
查询单次网关请求	20.1 查询网关请求
阅读请求详情中的路由、改写与上游尝试	20.2 阅读请求详情
对比客户端请求与上游请求	20.2 对比客户端请求与上游请求
按会话查看多轮调用	20.3 查看对话与会话详情
处理会话诊断问题	20.4 处理会话诊断问题
开启会话评审，查看成员与最佳实践	20.4 使用总览、成员、任务场景与最佳实践
查看路由策略命中	20.5 查看路由策略命中
追踪请求中的治理动作	20.5 追踪请求中的治理动作
查看平台运行状态与出口网络（平台管理员）	20.6 平台运行状态与出口网络
按现象确定根因定位顺序	20.7 根因定位顺序、第 21 章 · 问题诊断与处理
定位失败调用	21.1 网关请求：定位失败调用
定位引起问题的配置变更	21.2 操作审计：定位配置变更
追溯 AI 调用与合规决策	21.2 追溯 AI 调用与合规决策

告警、导出与审计

常用任务	参考章节
创建并验证钉钉、飞书、企业微信、Webhook 或邮件渠道	22.1 创建通知渠道
启用内置告警规则，新建网关策略告警	22.2 配置告警规则
配置质量告警与护栏命中通知	22.2 配置质量告警与护栏命中通知
确认、静默、指派或解决告警事件	22.3 处置告警事件
处理通知投递失败	22.4 处理投递故障
了解调用记录与内容的留存	22.5 调用记录与内容的留存
导出调用与用量数据，配置定时导出	22.6 导出调用与用量数据
查询操作审计、校验完整性并导出证据	22.7 查询操作审计并导出证据
处理平台告警，配置平台通知出口（平台管理员）	22.8 平台管理员的告警目录与通知出口

平台运行与维护（平台管理员）

常用任务	参考章节
升级平台，启用或停用可选能力	23.1 执行系统升级
备份平台数据	23.2 备份平台数据
恢复备份并演练	23.3 恢复和演练
日常巡检与查看服务日志	23.4 日常巡检与服务日志
调整过载阈值与数据库连接上限	23.5 运行参数边界
修改部署配置、入口端口与绑定网卡	23.6 维护部署配置与入口监听
添加并维护远端网关节点	23.7 管理网关节点
创建租户，设置租户配额	23.8 管理租户
发布通知横幅，设置品牌与 SMTP	23.9 授权许可、通知横幅、品牌与 SMTP 配置
变更公开访问地址与入口证书	24.8 变更访问地址与入口证书

日常维护与变更

常用任务	参考章节
维护管理角色、平台账号与登录会话	24.1 维护管理角色、平台账号与登录会话

常用任务	参考章节
同步组织、维护部门并管理成员会话	24.1 同步组织、维护部门并管理成员会话
维护模型供应商	24.2 维护模型供应商
维护模型服务	24.3 维护模型服务
维护 MCP 服务	24.4 维护 MCP 服务
维护外部智能体服务	24.5 维护外部智能体服务
核对配置变更	24.6 核对配置变更
接入新的模型与外部能力	24.7 接入新的模型与外部能力

上线与交接

常用任务	参考章节
按试点范围分阶段上线	第 25 章 · 分阶段上线
完成上线前检查	26.1 完成上线前检查
查看平台负载并处理容量问题	26.2 查看平台负载并处理容量问题
配置和检查平台安全控制	26.3 配置和检查平台安全控制
完成平台管理交接	26.4 完成平台管理交接
提交支持请求	26.5 提交支持请求、第 27 章 · 联系支持所需信息