

ZENOVA / ZENTRIX

ENTERPRISE AI PRODUCTS

用户手册

使用 · 配置 · 安全 · 观测

Zenova/Zentrix 用户手册

企业级 AI 网关与智能办公平台操作指南

提供从日常使用、能力配置到安全管理和运行观测的操作说明。

平台初始化 | 员工办公 | 智能体与工具 | 模型网关 | 安全与审计 | 质量与观测

文档类型	更新日期	文档版本
用户手册	2026 年 8 月 14 日	2026.08

目录

平台初始化与首次上线

章节	适用产品	主要内容
1. 产品关系与平台配置流程	Zentrix、Zenova	产品组成、首页与导航、配置顺序、上线检查条件
2. 平台对象和边界	Zentrix、Zenova	部门、职能组、成员、模型供应商、模型服务、工作空间、智能体和配额
3. 建立部门并添加成员	Zentrix、Zenova	新建部门 · 添加成员 · 分配工作空间角色（仅 Zenova）
4. 配置模型供应商	Zentrix、Zenova	新建并验证供应商 · 连接排查 · 完成标准
5. 创建模型服务	Zentrix、Zenova	新建模型服务 · 命名与拆分 · 不可用排查
6. 配置模型服务的可用范围与配额	Zentrix、Zenova	选择授权范围 · 配额字段说明
7. 配置工作空间	Zenova	分配模型服务 · 成员与角色 · 工作空间配额
8. 配置并分发 MCP 服务	Zentrix、Zenova	创建连接 · 设置数据等级 · 发布与分发 · 连接验证与排查
9. 配置敏感数据（PII）策略	Zentrix、Zenova	数据等级 · 套用模板 · 新增规则 · 命中动作
10. 设置安全围栏	Zentrix、Zenova	全局开关与失败姿态 · 风险分类矩阵 · 图像与规避对抗 · 拒答话术与上线验证
11. 完成第一次对话	Zenova	编写任务指令 · 选择工具审批模式
12. 创建可复用智能体	Zenova	选择起点 · 身份与模型 · 技能与应用
13. 查看用量并控制成本	Zentrix、Zenova	个人用量 · 管理员模型用量
14. 问题诊断与处理	Zentrix、Zenova	网关调用 · 操作审计
15. 分阶段上线	Zentrix、Zenova	试点、验证、扩展和上线检查清单
16. 联系支持所需信息	Zentrix、Zenova	故障时间、对象、错误、请求标识和脱敏要求

进阶操作、管理与平台维护

章节	适用产品	主要内容
17. 对话、历史与上下文	Zenova	对话生命周期、长文本、多模态、重试、中断、撤销和上下文压缩
18. 文件与企业知识	Zenova	文件上传、知识库、引用、权限和检索排障
19. 数据连接与查询	Zenova	创建、测试、部署、绑定、查询与安全限制
20. 文档、表格与报告	Zenova	生成、编辑、导出与水印
21. 智能体、工具与审批	Zenova	智能体发布、工具、MCP、Skill、审批和权限
22. 记忆、子智能体与任务委派	Zenova	工作区、记忆、子任务和结果回收

章节	适用产品	主要内容
23. 外部 IM 工具	Zenova	渠道接入、绑定、消息收发、审计和故障处理
24. 运行时安全与凭据	Zenova	凭据、文件系统、网络、数据等级和最小权限
25. 模型路由与流量保护	Zentrix、Zenova	路由、限流、整形、过载、恢复和故障定位
26. 调用方、API Key、认证与授权	Zentrix、Zenova	调用对象、凭据、认证、分发和配额
27. 请求变换、域名证书、插件与开放 API	Zentrix、Zenova	请求/响应变换、域名、证书、插件和接口文档
28. 可观测数据接入与权限边界	Zentrix、Zenova	OTLP、采集密钥、Trace、采样和入口权限
29. 执行、问题与根因定位	Zentrix、Zenova	筛选、下钻、回归、分享、对比和排障
30. 用量、成本与自定义看板	Zentrix、Zenova	Token、成本、调用量、性能和看板
31. 评分、人工评审与自动评估器	Zentrix、Zenova	评分、评审队列、评估器和回归验证
32. 数据集、实验与 Prompt 版本	Zentrix、Zenova	数据集、实验、Prompt 版本和试验台
33. 告警、通知、留存、归档与导出	Zentrix、Zenova	阈值、通知、处置、留存和归档检索
34. 平台运行与日常维护	Zentrix、Zenova	启停、重配、升级、备份、恢复和日志保留
35. 常用配置的维护与变更	Zentrix、Zenova	组织、模型、MCP、外部 Agent 的查询、调整、停用和清理
36. 配置并发布业务智能体	Zenova	知识问答、数据查询、办公产物和作业风险场景
37. 上线检查、稳定运行与管理交接	Zentrix、Zenova	上线检查、容量、安全、文件兼容、管理交接和支持请求
附录 A. 角色与入口速查	Zentrix、Zenova	主要角色、管理入口与常用菜单
附录 B. 常用任务索引	Zentrix、Zenova	按产品和任务查找对应章节

1. 产品关系与平台配置流程

1.1 产品关系与适用范围



Zentrix 可独立使用，提供企业级 AI 网关与治理能力。Zenova 在 Zentrix 基础上集成 Waner 智能体运行时，并提供工作空间、智能体和员工使用入口。

> **名词说明：**“Agent”是界面中对智能体的英文标识；“Skill（技能）”是可复用的任务能力；“MCP 服务”用于把外部工具和数据连接到模型或智能体。

名称	定位	组成与边界
Zentrix	企业级 AI 网关	可独立部署和使用，负责部门与成员、模型、MCP 与智能体服务接入，以及路由、鉴权、配额、流量、安全、插件、域名证书、调用观测和质量治理
Zenova	AI 智能办公私有化平台	由 Zentrix 与 Waner 组成；继承全部 Zentrix 网关能力，并通过 Waner 提供对话、智能体执行、工具调用、文件、记忆、任务委派和外部消息运行能力
Waner	Zenova 的智能体运行时模块	随 Zenova 提供，负责智能体任务回合、工具、MCP、技能、审批、工作区、记忆和子智能体执行

Zentrix 和 Zenova 均提供部门、职能组、成员、模型网关、MCP 与安全治理。Zenova 另外提供 Waner 运行时、工作空间和员工使用面；Zentrix 以能力中心和管理控制台作为主要使用入口。

1.2 首页与导航差异



产品	登录后的默认首页	用户侧导航	管理入口
Zenova	智能体使用面，默认打开“新对话”	使用工作区侧栏，包括“新对话”“提示词”“应用连接”“技能广场”“插件市场”“智能体”“用量统计”和“文件”	在账号菜单中选择“管理控制台”
Zentrix	能力中心，默认打开“MCP 广场”	使用能力中心顶部导航，包括“模型广场”“MCP 广场”“Skill 广场”和“Agent 广场”	在账号菜单中选择“管理控制台”

两种产品均从账号菜单进入“管理控制台”，菜单会按所用产品、账号角色和权限显示。Zentrix 提供“成员与部门”“职能组”以及模型、MCP、路由、网关实例、护栏、质量、运行观测、审计和系统运维等已授权入口；Zenova 还提供工作空间、知识库管理、IM 工具管理和 Agent 运营等入口。

1.3 Zenova 配置顺序



建议按图中编号完成首次上线。模型供应商负责“怎样连接模型”，模型服务负责“向成员提供哪组模型能力”，工作空间和配额负责“谁可以在什么边界内使用”，MCP 服务负责“哪些外部工具可以由哪些成员调用”，PII 与内容风险策略负责“数据和内容允许怎样处理”。

1.4 平台入口与职责

产品与入口	主要使用者	主要任务
Zenova 智能体使用面	普通成员、工作空间管理员	新建对话、创建智能体、上传和引用文件、使用技能与应用、查看个人用量
Zentrix 能力中心	能力使用者、开发者和集成人员	查找模型、MCP、Skill 与外部 Agent，查看详情并获取或测试获授权的连接能力
管理控制台	平台管理员及获授权的管理人员	配置当前产品包含的模型、组织、配额、MCP、护栏、用量、运行观测和审计；Zenova 还包括工作空间与智能体运行管理

如果管理控制台没有某个入口，先核对当前产品是否包含该能力，再核对账号角色、权限域、租户范围和部署设置。工作空间、员工对话和 Waner 运行管理属于 Zenova 导航；Zentrix 的能力使用入口位于能力中心。

Zentrix 从能力中心开始使用，并通过账号菜单进入管理控制台；Zenova 从智能体使用面开始，并在此基础上提供工作空间、智能体和 Waner 运行时入口。Waner 的运行状态统一在 Zenova 管理控制台与运行观测页面查看。

1.5 上线检查条件

适用产品	上线检查条件
Zentrix、Zenova	至少建立一个试点部门并添加一名试点成员；至少有一个通过验证的模型供应商，以及一个启用且包含可用模型的模型服务；模型服务配额和调用方范围符合组织政策；PII 数据等级、命中动作与组织数据分类政策一致；输入围栏、输出围栏、失败姿态及适用的内容风险分类已经启用并通过测试；需要使用 MCP 的场景已完成连接、Tools 解析、数据等级和可用范围配置；管理员能在模型用量或网关调用中看到对应记录。
Zenova 额外检查	试点工作空间已获得目标模型服务，并已加入试点成员；工作空间配额符合组织政策；试点成员能创建智能体并完成一次对话；Waner 运行时的模型、工具、审批和观测事件可以关联到同一会话。

2. 平台对象和边界

Zentrix 和 Zenova 均提供部门、职能组和成员管理。Zentrix 通过能力中心和管理控制台管理组织、模型、能力接入、路由、安全、配额和调用方；Zenova 在这些对象之上增加工作空间、员工对话面和智能体运行能力。使用 Zenova 时，组织、模型和 MCP 只需在管理控制台配置一次，工作空间和智能体直接使用这些配置。

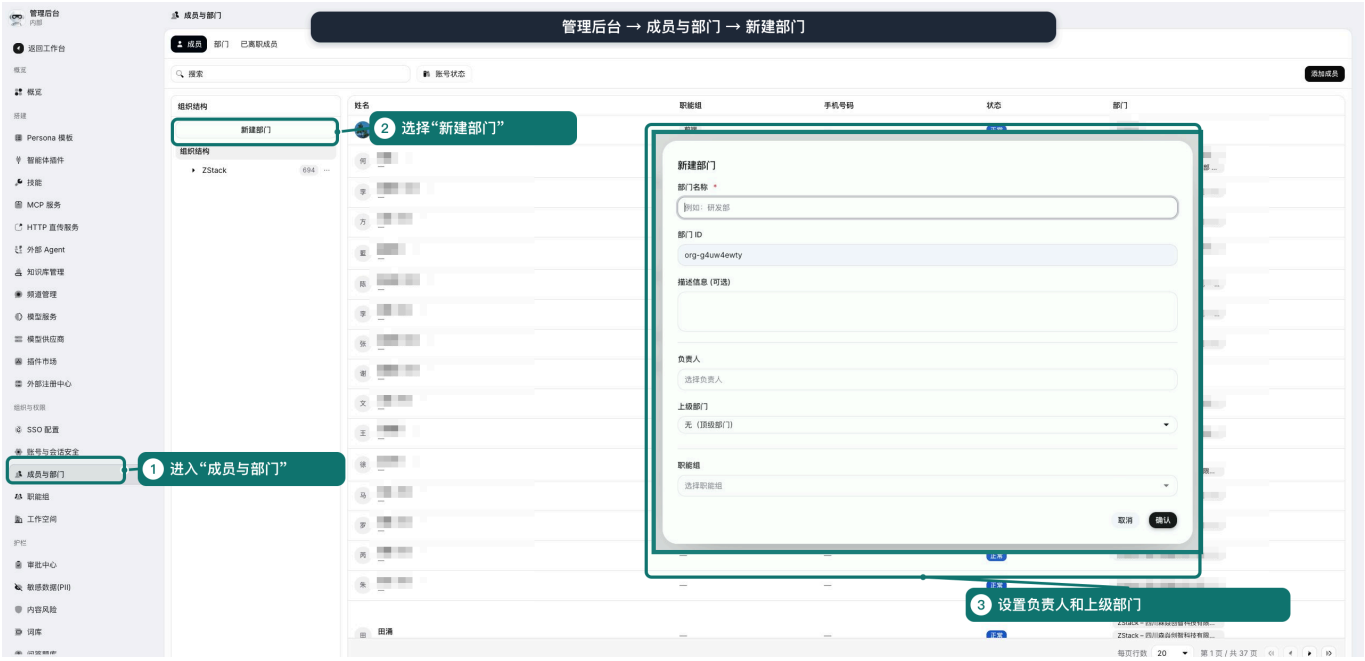
对象	它解决什么问题	配置建议
租户	一套独立的企业数据与权限范围	可理解为当前企业或组织实例；跨租户数据和授权互不共享
部门	成员在正式组织架构中的归属	按真实汇报关系建立，可设置负责人和上级部门
职能组	跨部门的能力或项目分组	用于跨部门授权，不要代替正式部门树
成员	可以登录并使用平台的个人账号	普通成员默认不给平台管理员权限
模型供应商	平台连接外部或自建模型端点的方式	每个供应商单独维护凭据与兼容性设置
模型服务	向成员暴露的一组有序模型能力	按用途命名，如“通用对话”“代码助手”
工作空间	团队协作、智能体和模型授权的边界	按团队、项目或数据边界划分
智能体	带身份、模型、技能与应用的可复用助手	一项稳定任务对应一个清晰智能体
MCP 服务	为智能体提供外部工具或数据能力的标准连接	每个服务单独维护连接、Tools、数据等级和授权范围
PII 策略	识别敏感数据并决定脱敏、阻断或审计	规则等级应与组织数据分类一致；高等级规则优先严格处理
内容风险围栏	在模型调用前后检测风险内容与规避行为	输入和输出均启用；失败姿态按业务风险确定
配额	控制费用、速率、并发和单次请求大小	从小范围试点起步，依据真实用量调整

3. 建立部门并添加成员

Zentrix 和 Zenova 都支持部门、职能组和成员管理。组织结构应在模型范围配置前完成，后续即可按部门授权，不必逐人维护。Zenova 还可以在组织结构之上使用工作空间划分团队、项目和数据边界。

3.1 新建部门

1. 进入 `管理控制台 → 成员与部门`。
2. 在“成员”标签页左侧的“组织结构”中选择“新建部门”。
3. 填写部门名称；部门 ID 会自动生成，用于系统稳定识别该部门。
4. 按需填写描述、负责人、上级部门和职能组。
5. 保存后，在部门列表中确认层级和负责人正确。

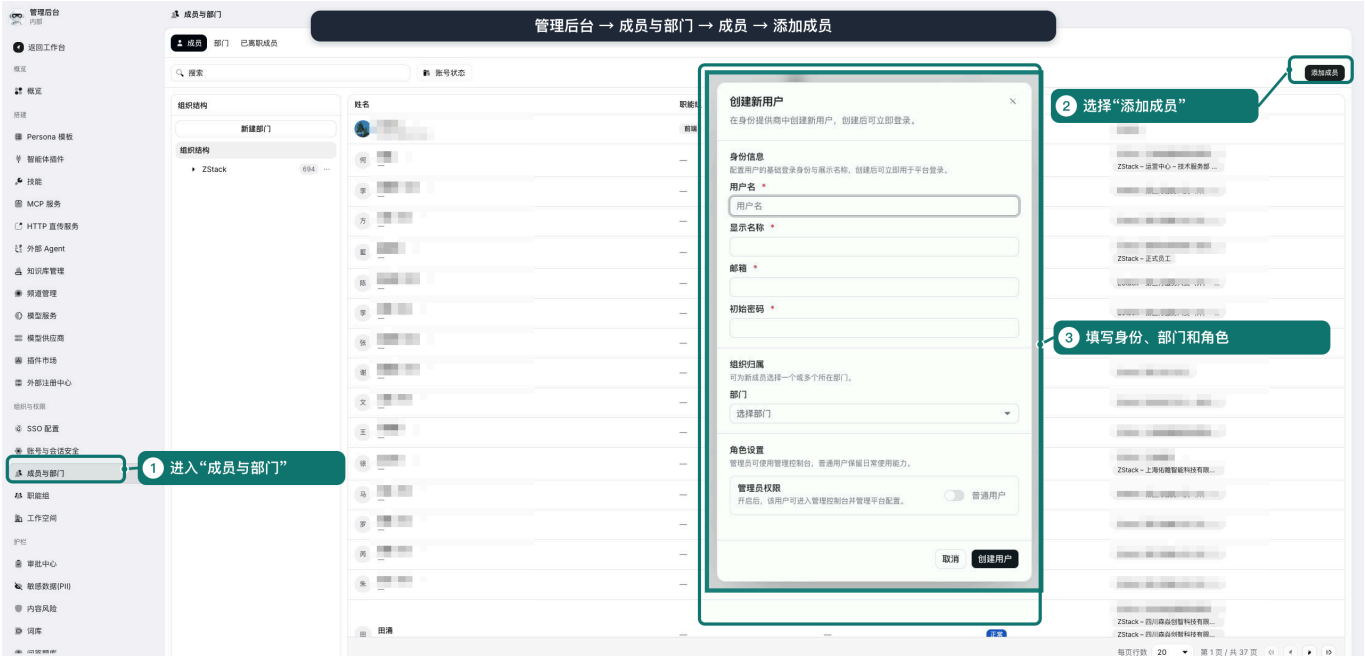


图：依次进入“成员与部门”，选择“新建部门”，再设置负责人、上级部门和职能组。

部门名称可以随着组织调整而修改；自动生成的部门 ID 应作为稳定标识，不要为了“看起来整齐”反复重建部门。跨部门项目请优先使用职能组。

3.2 添加成员

1. 进入 `管理控制台 → 成员与部门 → 成员`，选择“添加成员”。
2. 填写用户名、显示名称、邮箱和初始密码。
3. 选择一个或多个所属部门。
4. 只有确实需要管理整个平台的人员才启用“管理员权限”。
5. 创建后，将登录地址和初始密码通过安全渠道交付给本人，并要求首次登录后立即修改密码。



图：依次进入“成员与部门”和“成员”，选择“添加成员”，再填写身份、部门和角色。

账号创建成功后即可登录。部门归属和平台管理员权限是两件不同的事。Zenova 还提供独立的工作空间角色：加入部门不会自动获得某个工作空间的访问权；成为工作空间管理员也不会自动成为平台管理员。Zentrix 的成员授权以部门、职能组、个人和能力可用范围为主。

3.3 分配工作空间角色（仅 Zenova）

Zenova 使用工作空间角色进一步划分团队协作范围；Zentrix 使用部门、职能组、成员和能力可用范围完成授权。

角色	适合人群	典型权限边界
管理员	工作空间负责人	管理成员、模型服务、配额和工作空间设置
编辑者	构建和维护智能体的成员	创建、编辑工作空间内容和智能体
查看者	只需使用或查看成果的成员	查看和使用已授权内容，不负责配置

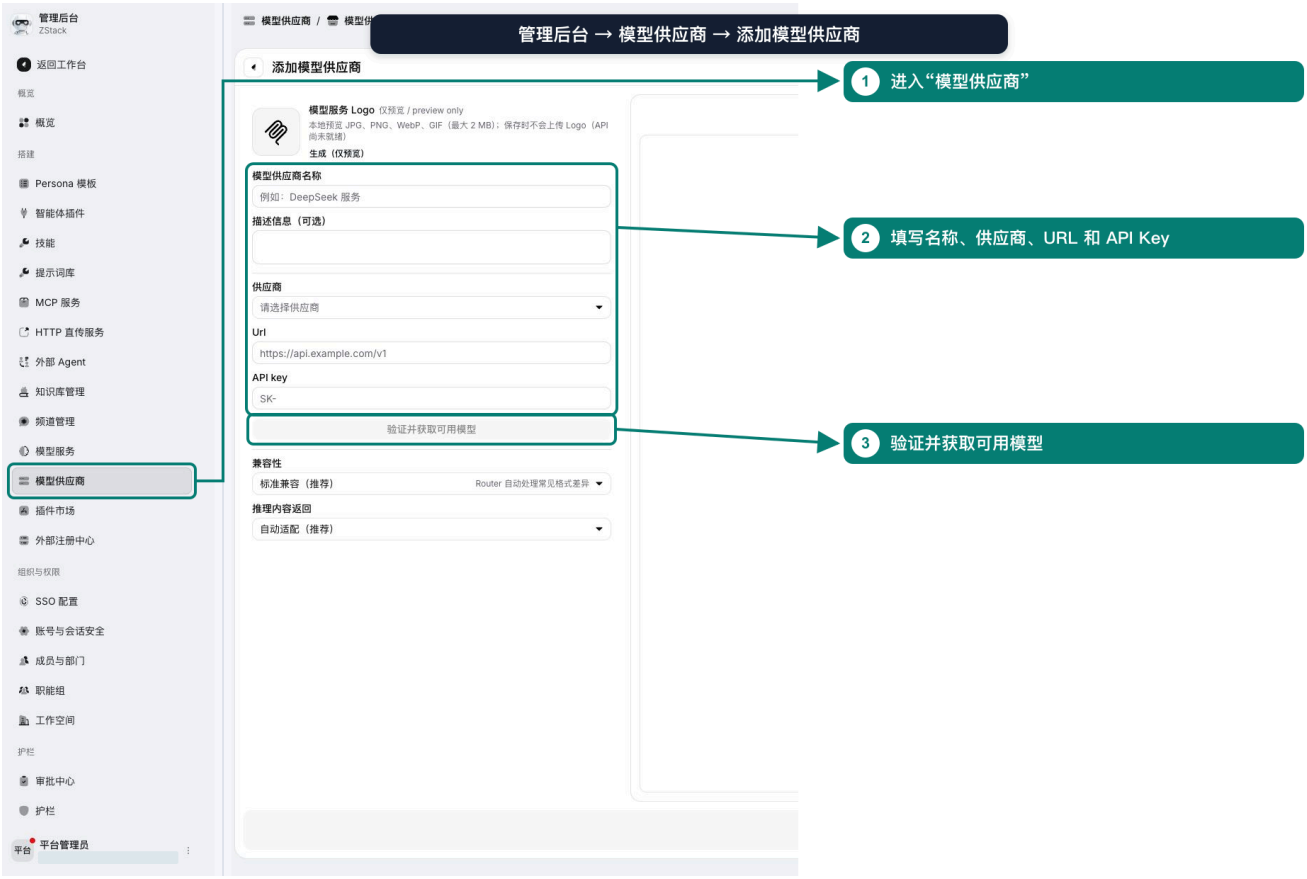
按最小权限授予角色。通过职能组批量授权时，不会授予工作空间管理员角色；需要管理员权限时应明确指定个人。

4. 配置模型供应商

模型供应商保存平台如何连接某个模型端点，包括供应商、基础 URL、API Key、兼容性和可用模型。API Key 属于敏感信息，只能在管理控制台的密钥字段中填写，不要粘贴到对话、工单或截图里。

4.1 新建并验证供应商

1. 进入 `管理控制台 → 模型供应商`，选择“添加模型供应商”。
2. 填写便于识别的名称和描述，例如按环境与供应商命名。
3. 选择“供应商”，填写该供应商要求的 URL 和 API Key。
4. 根据端点能力选择兼容性与推理内容返回方式。没有明确要求时，保留组织已验证的默认设置。
5. 选择“验证并获取可用模型”。验证成功后，从返回列表中启用需要的模型。
6. 保存，并返回列表确认供应商状态正常。



图：进入“模型供应商”后填写连接信息，验证并获取可用模型。图中不包含真实密钥。

4.2 供应商连接排查

按下面顺序检查，避免反复改动无关配置：

1. URL 是否包含供应商要求的协议、域名和路径；
2. API Key 是否仍有效，是否有调用目标模型的权限；
3. 当前网络是否能访问该端点；
4. “供应商”与端点协议是否匹配；
5. 兼容性设置是否与自建网关或兼容接口一致；

6. 仍失败时，记录发生时间、供应商名称和页面错误，不要记录或发送 API Key。

4.3 配置完成标准

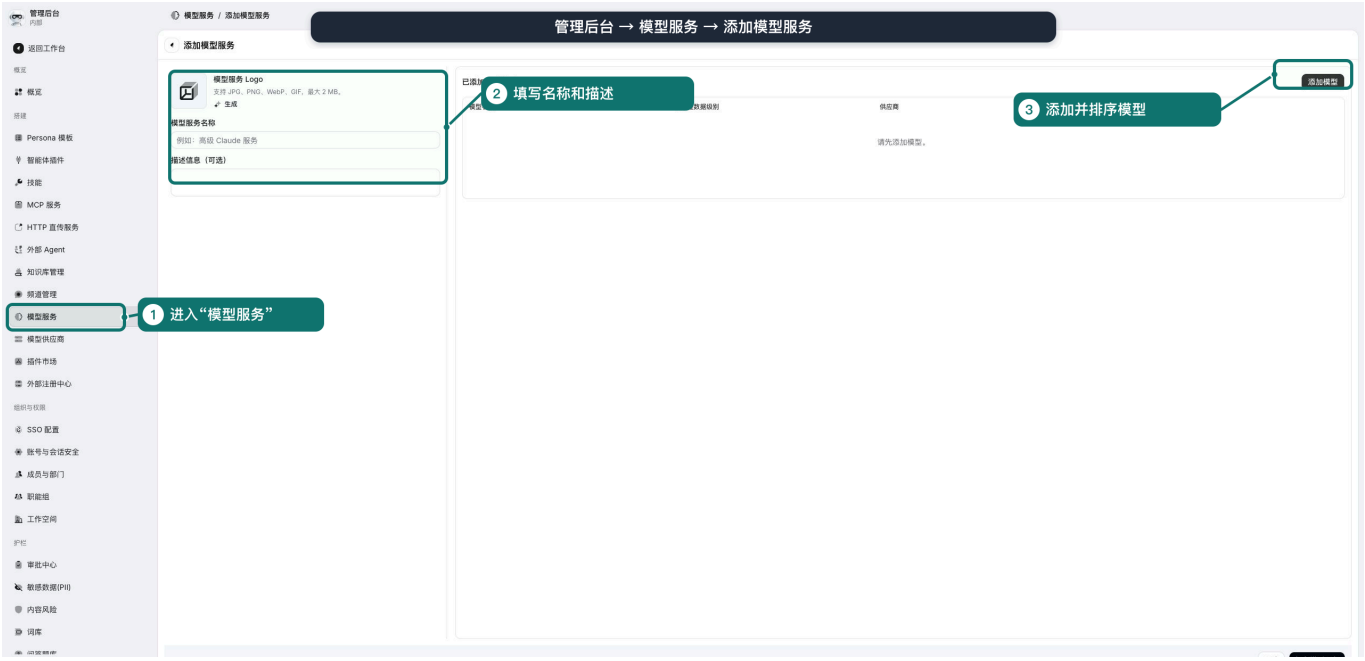
只有“连接验证成功、至少启用一个可用模型、保存后状态正常”三项都满足，才算供应商配置完成。仅仅保存名称和密钥并不能保证成员可以使用。

5. 创建模型服务

模型服务是成员实际选择和使用的模型能力，可以包含一个或多个模型。列表首项是主模型；不要把列表顺序理解为平台内置的静默故障切换。主模型被停用或不可用时，请求会失败，需要管理员恢复模型或调整配置。涉及敏感数据时，是否切换到其他合规模型由网关侧路由策略决定。

5.1 新建模型服务

1. 进入 `管理控制台 → 模型服务`，选择“添加模型服务”。
2. 填写面向成员的名称和清晰描述，例如说明适用任务、数据边界或成本等级。
3. 选择“添加模型”，从已配置的供应商模型中加入至少一个模型。
4. 调整模型顺序：把组织确认的主模型放在首位；后续模型是否参与实际流量由网关侧路由策略决定。
5. 保存后进入服务详情，继续配置可用范围与配额。



图：模型服务至少需要一个模型；首项是主模型，主模型不可用时应由管理员处理。

5.2 模型服务命名与拆分

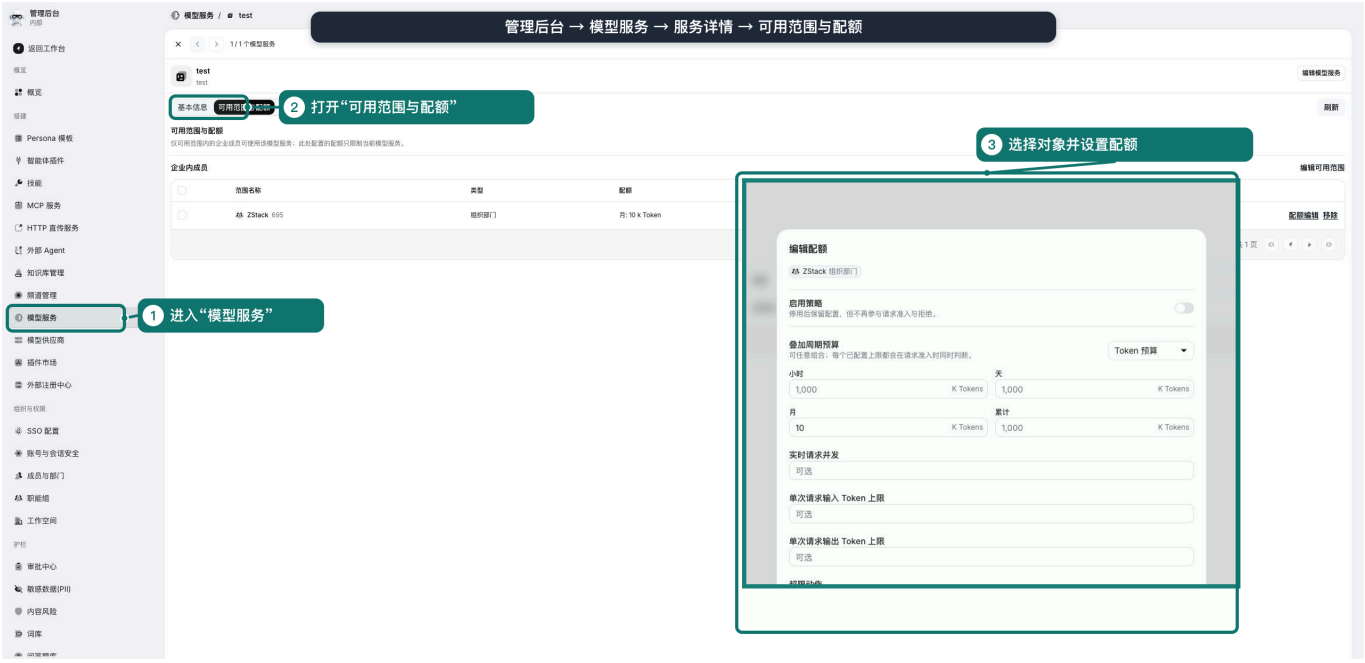
- 按使用目的拆分，例如“通用对话”“代码与分析”“低成本批处理”；
- 需要不同数据处理策略、费用上限或可用人群时，拆成不同服务；
- 不要直接使用供应商内部型号作为唯一名称，成员通常更关心任务用途；
- 描述中写清适用场景和限制，避免成员只能靠试错选择。

5.3 模型服务不可用排查

确认主模型启用、供应商正常、成员已获服务授权、工作空间已分配服务，且硬配额未耗尽或为 0。不要按“自动使用后续模型”排障；合规切换须核对网关侧路由策略。

6. 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”中，管理员可以决定哪些租户、工作空间、部门或职能组可以使用该服务，并给对应对象设置专属限制。



图：进入模型服务详情的“可用范围与配额”，选择组织对象后设置周期预算、并发和单次请求限制。

6.1 选择授权范围

- 部门稳定、工作职责清晰时，按部门授权；
- 跨部门项目或专项能力，按职能组授权；
- 项目有独立数据和成本边界时，按工作空间授权；
- 全员都应使用的基础能力，才考虑租户级范围。

范围应从小到大逐步开放。新增高成本或高敏感模型时，应授权给试点部门或工作空间；效果与费用符合预期后再扩大。

6.2 配额字段说明

设置项	作用	使用建议
小时、天、月、累计	控制相应周期或全部累计范围内的消耗	日常治理优先设置月预算；需要约束全部累计消耗时使用“累计”
Token / 金额	选择按使用量或成本限制	成本可预测时用金额；不同模型价格差异大时尤其有用
实时请求并发	限制同时进行的模型调用	根据实际并发设置保守初值，并观察等待和失败情况

设置项	作用	使用建议
单次请求输入 Token 上限	限制一次请求携带的上下文大小	防止超长文件或历史对话造成异常高消耗
单次请求输出 Token 上限	限制一次回答的最大长度	根据任务质量要求设置，不要一味追求更大
软限制	达到阈值时提醒，但不直接阻断	适合上线初期收集真实用量
硬限制	达到阈值后拒绝继续调用	适合必须严格控制的费用或安全边界

高风险提醒：硬限制填写 `0` 会阻断所有匹配该规则的调用；留空表示没有设置该项。保存前务必再次核对对象、周期、单位和软硬策略。

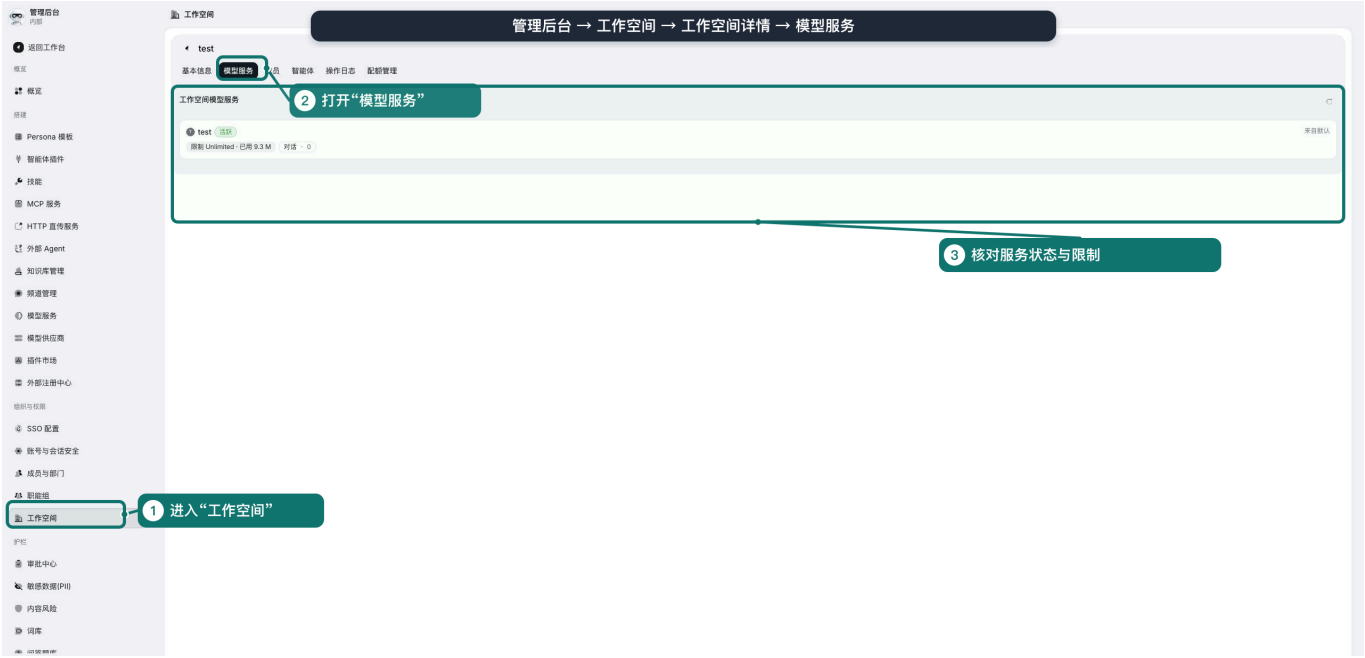
多个范围规则可能同时生效。模型服务配额只限制当前服务，不等同于部门或工作空间的全部平台预算。出现拒绝时，需要同时检查模型服务规则与工作空间规则。

7. 配置工作空间

工作空间是团队协作的主要边界。平台管理员或工作空间管理员需要完成模型服务、成员与配额三项配置。

7.1 分配模型服务

1. 进入 `管理控制台 → 工作空间`，打开目标工作空间。
2. 进入“模型服务”标签页。
3. 添加该团队确实需要的模型服务，并确认服务为启用状态。
4. 如果服务显示“来自默认”，表示它来自默认配置；仍应确认当前工作空间的实际可用状态和限制。



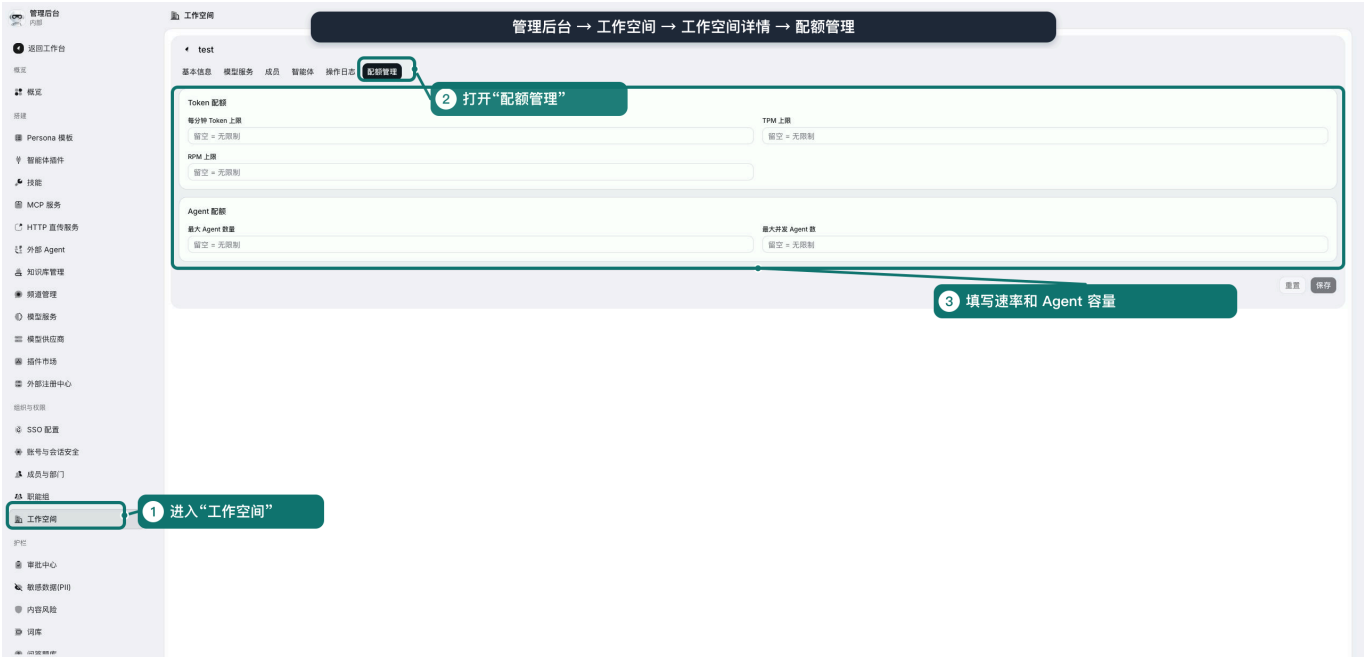
图：工作空间会显示已分配模型服务的状态、限制和用量。

7.2 添加成员并分配角色

进入工作空间“成员”标签页，添加成员并选择管理员、编辑者或查看者。成员所在部门不会自动决定工作空间角色；两者都要检查。

7.3 设置工作空间配额

进入“配额管理”标签页，按工作空间的实际容量设置 Token、请求速率和智能体数量上限：



图：工作空间配额用于控制请求速率、Token 用量和智能体容量。留空表示不设置工作空间级上限。

字段	作用	配置建议
每分钟 Token 上限	旧字段；请求限流以“TPM 上限”为准	新建配置请留空；沿用旧配置时先与平台负责人核实
TPM 上限	每分钟允许处理的提示词 Token 数量	按工作空间的预计并发量填写正整数
RPM 上限	控制工作空间的请求速率，避免短时间内请求过多	按允许的请求频率填写正整数
最大 Agent 数量	工作空间最多可以创建的智能体数量	按团队规模和管理要求填写正整数
最大并发 Agent 数	同一时间可以运行的智能体数量	不应超过实际可承载的并发容量

新建工作空间优先配置 TPM、RPM 和智能体数量上限。“每分钟 Token 上限”是旧字段，新建时请留空；限流以“TPM 上限”为准。保存后使用试点账号发起调用，并在“模型用量”或“网关调用”中核对是否出现限流记录。

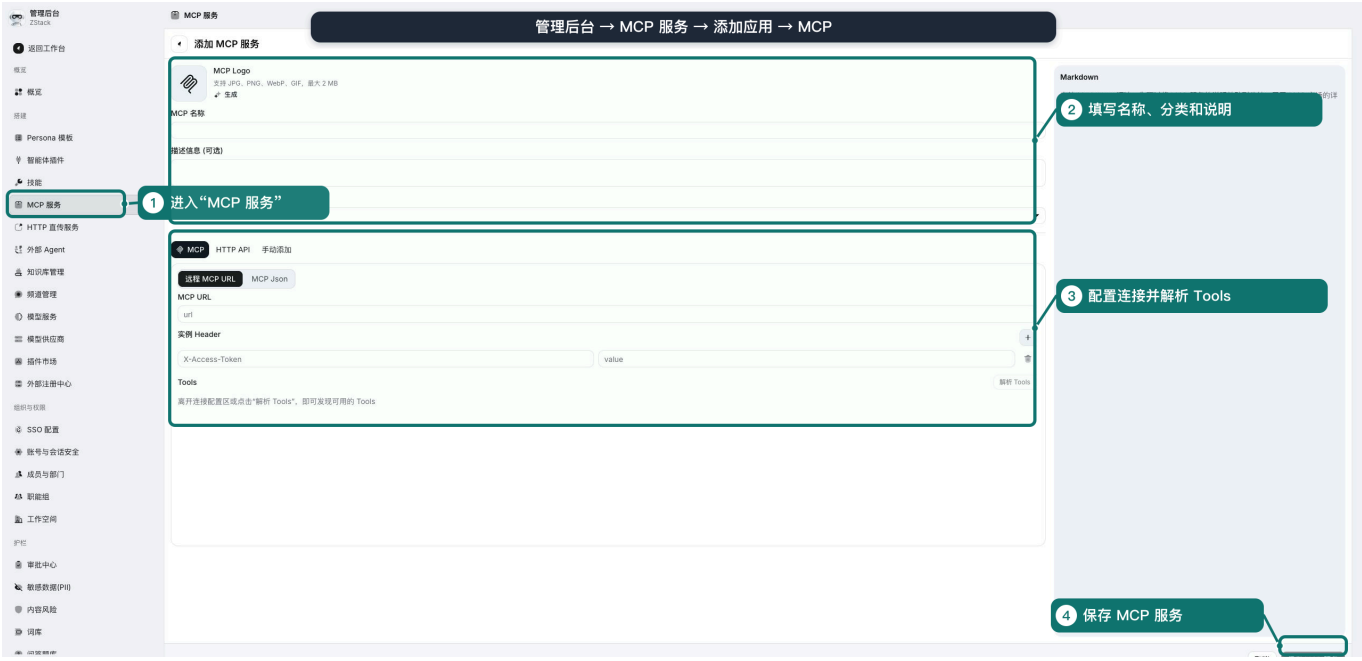
所有配额字段只能填写正整数或留空。留空表示不设置工作空间级上限，实际调用仍受租户、模型服务及平台策略限制；不要填写 `0` 表示“不限”。

8. 配置并分发 MCP 服务

MCP 服务为智能体提供外部工具和数据能力。完整配置包括连接、Tools 解析、默认数据敏感等级、发布状态与可用范围。仅保存连接不代表成员已经获得使用权限。

8.1 创建 MCP 连接

1. 进入 `管理控制台 → MCP 服务 → MCP 列表`，选择“添加应用”。
2. 填写 MCP 名称、描述和分类。名称应能说明业务用途，描述中写明数据来源、负责人和限制。
3. 先选择顶层接入方式：“MCP”“HTTP API”或“手动添加”。“MCP”下可选择“远程 MCP URL”或“MCP Json”；“HTTP API”下可选择“OpenAPI”“Postman Collection”或“cURL”。
4. 使用远程 MCP 时填写 MCP URL，并按上游要求区分两层 Header：连接器本身固定携带的 Header 配在“实例 Header”；鉴权方式选择“自定义请求头”时，真正用于凭据认证的 Header 配在“凭据 Header”。同名 Header 以连接方式中的凭据配置为准。
5. 选择“解析 Tools”，确认平台能够发现预期工具，并核对工具名称、说明与参数。
6. 在 Markdown 中补充成员可见的用途、适用范围、数据边界和使用示例，随后保存 MCP 服务。



图：从“MCP 服务”进入新增页面，依次完成基本信息、连接与 Tools 解析后保存。示例字段不包含真实凭据。

“实例 Header”适合租户实例级固定参数，“凭据 Header”适合需要按凭据保存的认证信息。只配置其中一层并不能覆盖所有鉴权场景。MCP Header、访问令牌和上游密钥属于敏感信息，只能填写在管理控制台的受保护字段中；保存后的密文字段不会明文回显。截图、培训材料、邮件和工单不得包含真实值。

8.2 设置 MCP 默认数据敏感等级

打开 MCP 服务详情的“基本信息”，在“数据安全”区域设置“数据敏感等级”。该等级表示此 MCP 连接返回数据的默认敏感程度。

界面选项	选项说明	典型数据
L1 公开数据	可以公开传播的数据	官网、公开资料、公开知识库

界面选项	选项说明	典型数据
L2 内部数据	仅限组织内部使用的数据	内部 Wiki、产品文档、普通业务数据
L3 敏感数据	泄露会造成明显业务或个人影响的数据	客户、合同、财务、人事数据
L4 机密数据	组织的最高敏感级别	战略、并购、核心商业机密、凭据



图：MCP 服务详情会同时显示发布、连接和默认数据敏感等级。

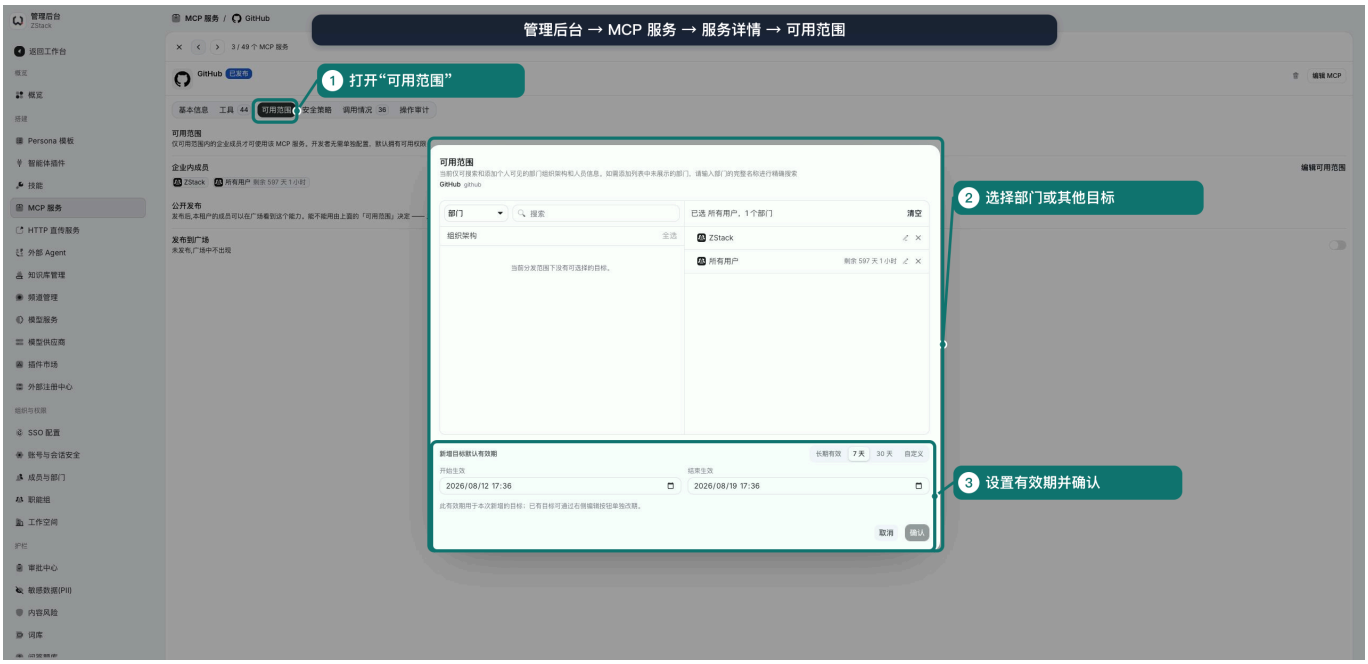
平台会比较 MCP 默认等级与 PII 检测结果，并使用较高等级选择符合数据处理要求的模型；没有合规模型时，请求会被拒绝。等级不能为了提高模型可选性而调低，必须以真实数据风险为准。当前合规规则将 L3 数据限制在本地部署模型处理；L4 数据不得离开本地，且不能通过模型例外放宽。

8.3 发布与分发

MCP 的“保存”“公开发布”和“可用范围”承担不同职责。界面中的“发布”和“公开发布”指向同一个发布开关，不是两个独立操作：

操作	作用	不代表什么
保存 MCP 服务	保存连接、Tools 和说明	不代表已经向成员授权
公开发布（发布）	让租户成员能够在广场看到该能力	不授予实际使用权限；能否调用仍由可用范围决定
编辑可用范围	向部门或其他组织目标授予使用权	不替代发布状态和服务可用性检查

1. 打开 MCP 服务详情的“可用范围”标签页，选择“编辑可用范围”。
2. 按目标类型筛选，并搜索部门或其他组织对象。
3. 将需要授权的目标加入右侧已选列表。
4. 为新增目标设置“长期有效”“7 天”“30 天”或“自定义”有效期；也可单独编辑某个目标的有效期。
5. 核对已选对象与有效期，选择“确认”。



图：可用范围弹窗保留了完整管理控制台导航，并显示目标筛选、已选对象和有效期设置。

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，避免长期有效授权在项目结束后继续保留。所有成员的实际能力可见性和使用权限都应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”。

8.4 MCP 连接验证与排查

发布到目标范围前，至少完成以下验证：

- “解析 Tools”返回的工具数量和名称符合预期；
- 使用无效参数时，工具能够返回明确错误而不是静默失败；
- 试点成员能够看到并调用服务，非授权成员不能调用；
- MCP 返回 L2、L3 或 L4 数据时，模型选择或拒绝行为符合数据政策；
- 可用范围到期后，目标成员不再获得该服务；
- 调用情况、操作审计和必要的安全审计记录可查询。

成员看不到 MCP 时，依次核对发布状态、可用范围、授权有效期、成员组织归属以及智能体是否启用了该应用。Tools 不可用时，核对端点、Header、上游权限和解析结果，但不得在排查材料中记录密钥。

9. 配置敏感数据（PII）策略

`管理控制台 → 护栏 → 敏感数据` 用于识别输入或工具数据中的个人信息、业务敏感信息和凭据，并为命中内容分配数据等级与处理动作。规则列表显示“规则名称”“正则”“PII 等级”“动作”和“状态”；新增或编辑弹窗中的检测字段名为“正则表达式”。

9.1 PII 数据等级

PII 规则使用与 MCP 数据安全一致的 L1–L4 体系。配置时以命中内容造成的实际风险为准：公开联系方式不应自动等同于 L1，内部员工联系方式通常至少属于 L2；身份证、银行卡、客户与医疗数据通常属于 L3；密码、私钥和访问令牌属于 L4。

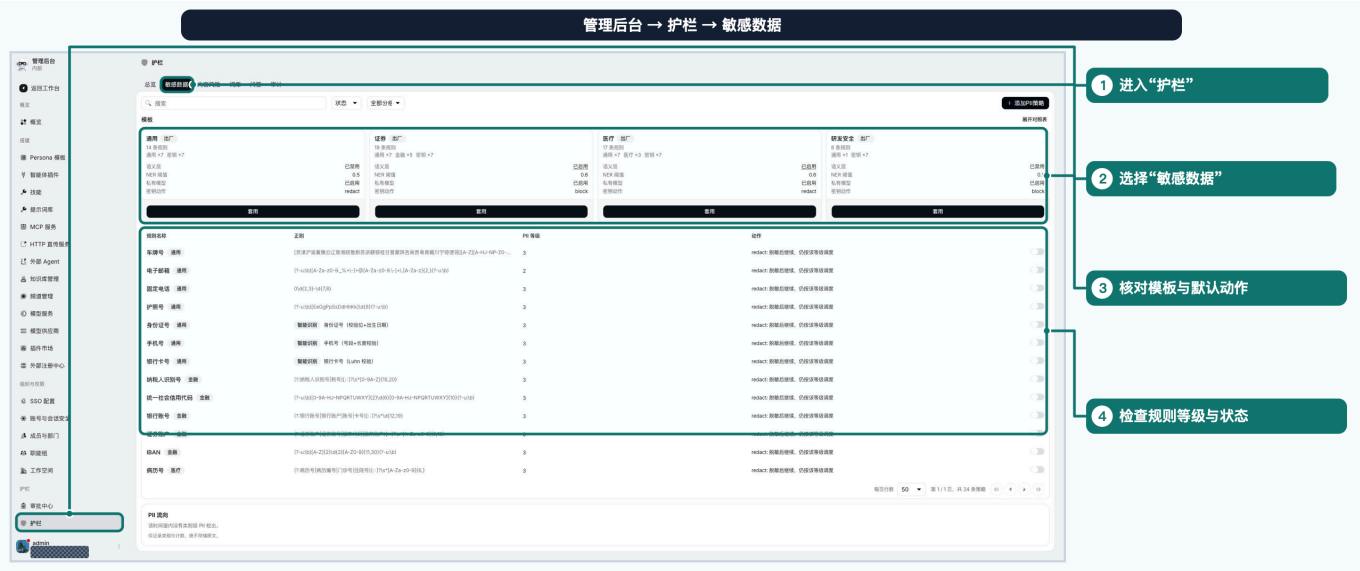
等级	配置判断	建议处理基线
L1 公开数据	已依法公开且允许再次传播	允许使用，并保留必要审计
L2 内部数据	仅限组织内部，外泄影响有限	限制到获准模型与范围，必要时脱敏
L3 敏感数据	涉及客户、合同、财务、人事、身份或医疗数据	使用本地部署模型；优先脱敏或阻断
L4 机密数据	密码、私钥、访问令牌及核心机密	阻断外发；不得通过模型例外放宽

当同一请求命中多条规则时，应按最高等级治理。模型能够处理更低等级数据，不代表能够处理更高等级数据。

9.2 套用策略模板

页面提供通用、证券、医疗和研发安全模板。模板用于建立规则基线，套用前应核对行业适用性和默认动作。“语义层”用于结合上下文识别敏感实体；NER 是“命名实体识别”的缩写，其阈值决定语义识别的敏感程度。

模板	重点覆盖	默认特征
通用	通用个人信息与密钥	语义层关闭；密钥默认脱敏后继续
证券	通用个人信息、金融信息与密钥	语义层开启；密钥默认阻断
医疗	通用个人信息、医疗信息与密钥	语义层开启；密钥默认脱敏后继续
研发安全	研发场景中的个人信息与密钥	语义层关闭；密钥默认阻断

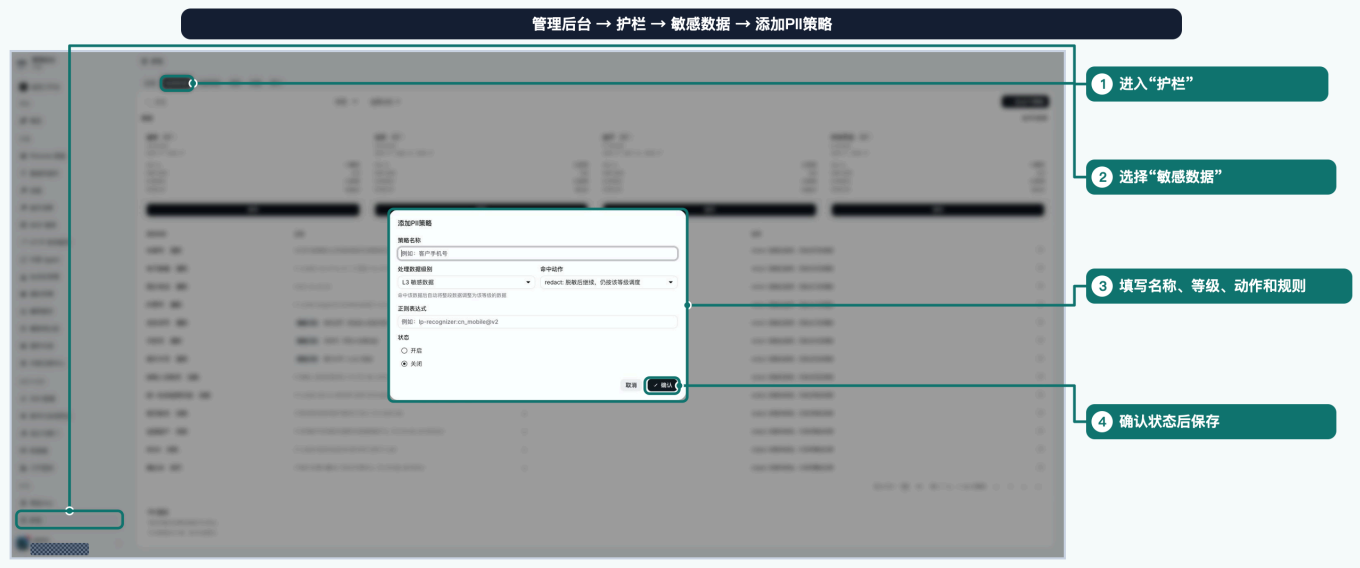


图：模板区域显示规则构成、语义层、NER 阈值、私有模型与密钥动作；下方列表显示每条规则的等级和动作。

套用模板后，务必逐条检查状态、检测表达式、PII 等级和命中动作。模板是配置起点，不替代组织的数据分类制度。

9.3 新增或调整 PII 规则

1. 选择“添加PII策略”，打开同名弹窗。
2. 填写能明确说明数据类型的策略名称，例如“客户手机号”或“生产访问令牌”。
3. 选择数据级别（PII 等级）和命中动作。
4. 填写“正则表达式”；需要语义检测时，在治理配置中核对“语义层”。
5. 将状态设为“开启”，选择“确认”。



图：新增策略时需要同时配置名称、数据等级、命中动作、正则表达式和状态。

正则上线前要使用脱敏样本覆盖“应命中、边界值、不应命中”三类测试。表达式过宽会造成误报和业务阻断，表达式过窄会造成敏感数据漏检。密码、私钥、API Key 和 Access Token 等凭据规则应按 L4 管理。

9.4 命中动作： `allow` 允许继续并保留规则结果；`audit` 允许继续并写入审计；`redact` 脱敏后继续，并按命中等级选择模型；`block` 阻断请求并写入审计。`allow` 和 `audit` 都不会脱敏或阻断；新规则可先用 `audit` 观察误报，转入正式范围前再由数据负责人确认最终动作。

10. 设置安全围栏

`管理控制台 → 护栏` 是平台的大模型安全围栏配置入口。其中“内容风险”负责输入、输出、图像内容和规避行为检测，“问答”负责问答库及其高低阈值。PII 策略解决“数据敏感性”，内容风险围栏解决“内容是否允许处理或返回”，两者需要同时配置。

10.1 全局开关与失败姿态

1. 启用“输入围栏”，对发送给模型的文本、图像和解码内容执行检测。
2. 启用“输出围栏”，对模型返回内容执行检测。
3. 选择“失败姿态”。该字段表示检测服务异常时平台如何处理请求：
   - “失败放行”：围栏无法完成评估时允许请求继续，并记录审计事件；
   - “失败阻断”：围栏无法完成评估时拒绝请求。
4. 进入 `管理控制台 → 护栏 → 问答`，配置“问答阈值(高)”与“问答阈值(低)”，取值范围为 0—1：相似度达到高阈值时直接返回预设答案；处于低、高阈值之间时记审计并继续词库或语义检测；低于低阈值时问答层不代答，仍继续后续检测。



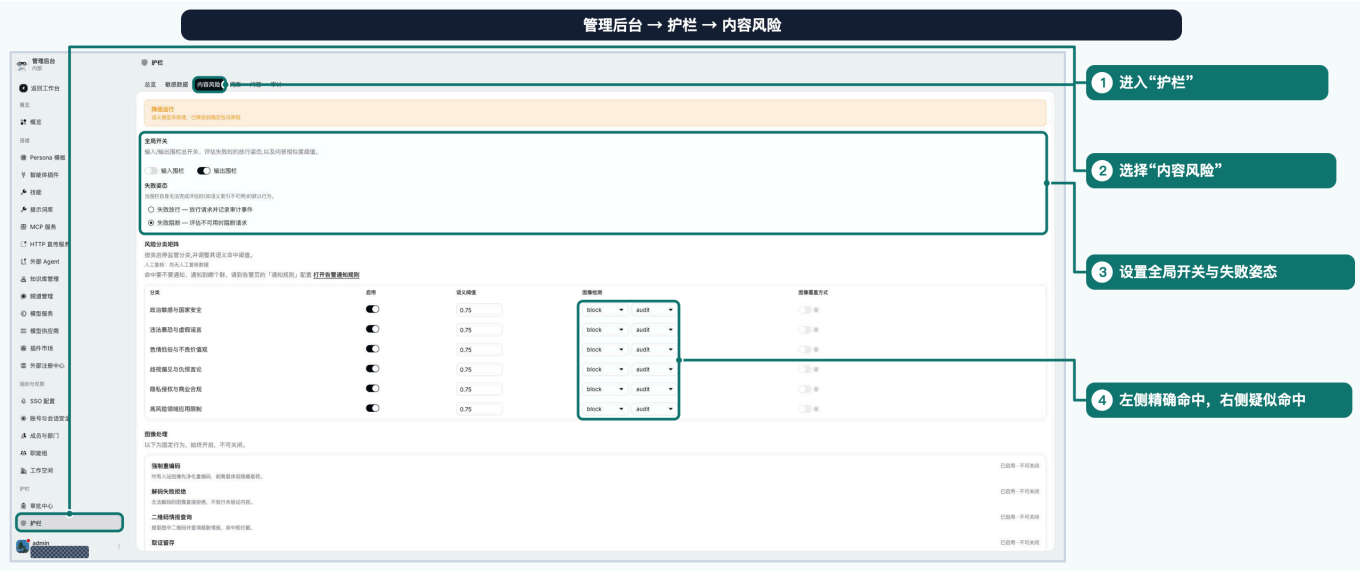
图：问答阈值位于“护栏 → 问答”；当前界面将低阈值显示为 `0.6`，不会显示浮点残渣。

高风险、强监管或包含敏感数据的正式业务应使用“失败阻断”。低风险试点需要使用“失败放行”时，必须具备审计、告警和明确的结束日期。

10.2 风险分类矩阵

当前页面提供六类风险：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。

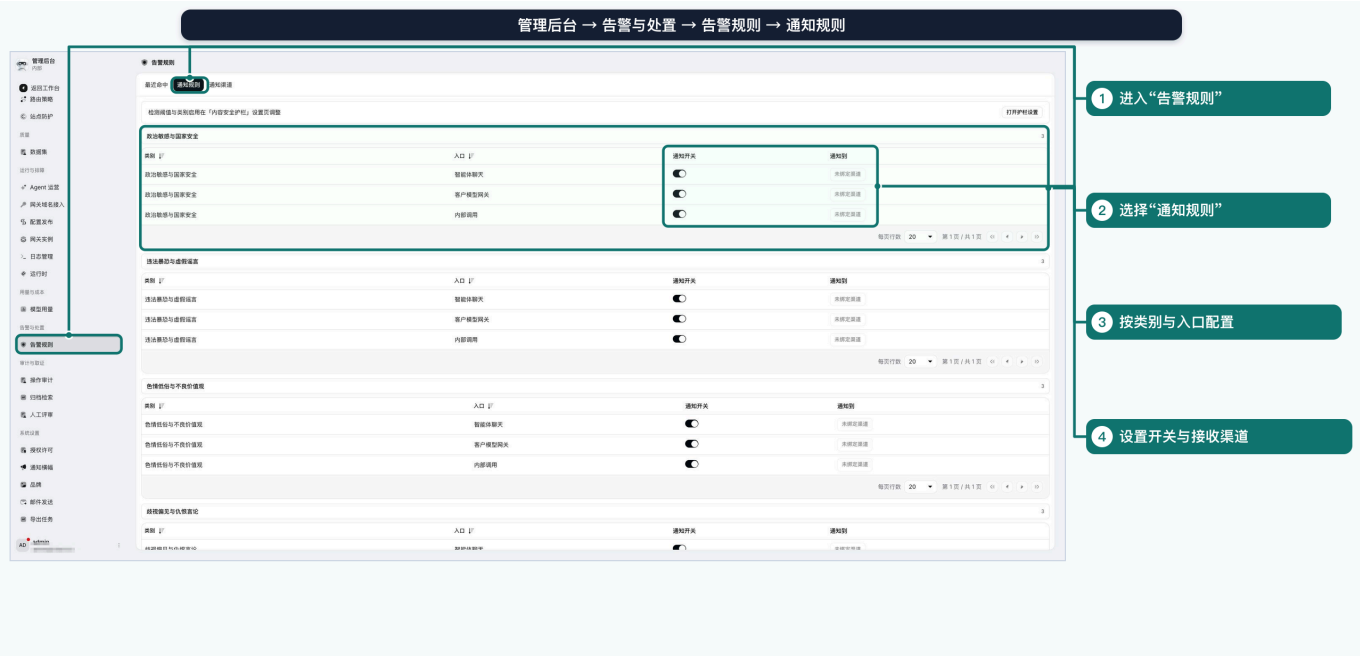
每一类可以配置“启用”“语义阈值”“精确命中动作”和“疑似命中动作”。“精确命中动作”处理确定命中，“疑似命中动作”处理疑似命中；两项均可选择“拦截”（`block`）或“仅记录”（`audit`）。“拦截”会拒绝请求，“仅记录”允许请求继续并写入审计。分类矩阵不包含“命中通知”；通知在“告警 → 通知规则”中配置。阈值越低，检测越敏感，召回更多风险内容的同时也会增加误报；阈值越高，误报可能减少，但漏检风险上升。生产阈值应以组织样本集的评测结果为依据。



图：进入“内容风险”后设置输入输出围栏、失败姿态、六类风险策略，以及精确命中和疑似命中的处理动作。命中通知在“告警 → 通知规则”中配置。

命中通知在 `管理控制台 → 告警与处置 → 告警规则 → 通知规则` 中按“类别 × 入口”配置。管理员应找到对应业务事件规则，设置“通知开关”，并在“通知到”列确认接收渠道；分类矩阵中的启用开关只控制检测，不代表已经发送通知。

权限前提：“通知规则”属于租户告警管理页面，仅租户管理员或获授权的告警管理员可见。平台级告警规则管理员默认进入“规则目录”，不会显示租户的“通知规则”；需要配置某个租户时，应切换到该租户的授权账号。侧边栏相同不代表权限范围相同。



图：“通知规则”按“类别 × 入口”设置通知开关和接收渠道。

10.3 图像处理与规避对抗

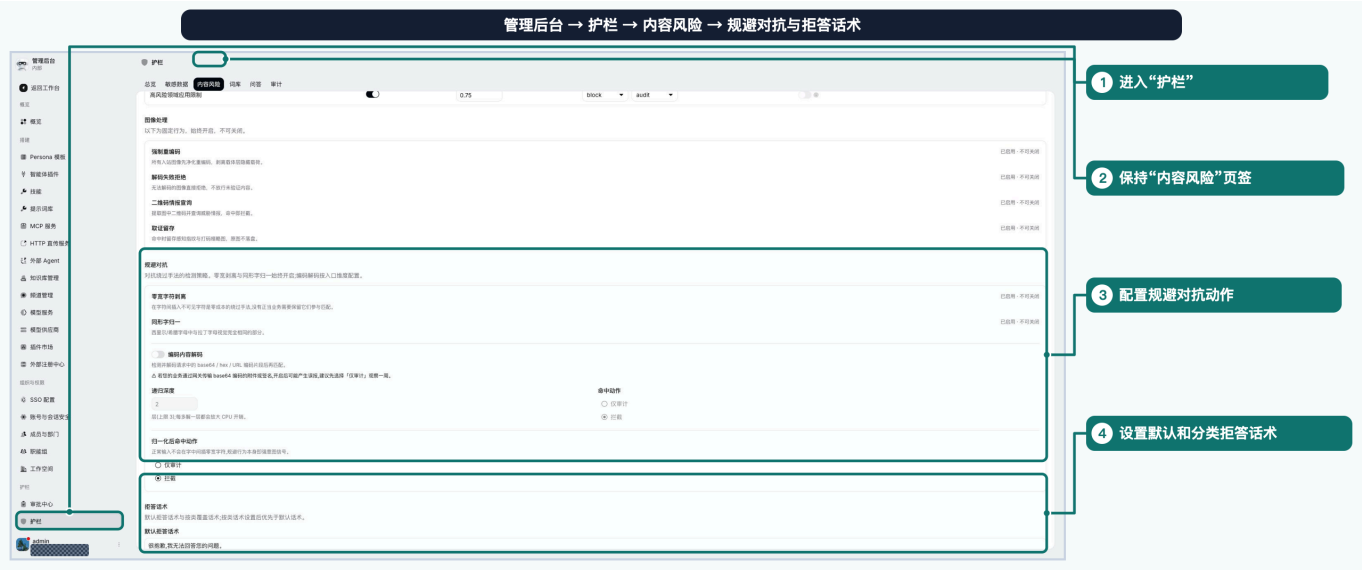
当前页面中的强制重编码、解码失败拒绝、二维码情报查询和取证留存为固定开启行为。命中时取证留存感知指纹和打码缩略图，原图不落盘。

规避对抗区域包含以下设置：

- 零宽字符剥离和同形字归一始终开启；
- “编码内容解码”用于识别并解码 Base64、Hex 和 URL 编码片段后再匹配；

- 递归深度上限为 3，每增加一层都会增加 CPU 开销；
- 解码内容命中与归一化后命中可选择“仅审计”或“拦截”。

如果业务会通过网关传输 Base64 附件或签名，启用编码内容解码可能产生误报。建议在受控范围使用“仅审计”观察一周，核对误报后再切换为“拦截”。该观察期必须保留审计记录和责任人。



图：页面下部用于设置编码解码、递归深度、命中动作以及默认和分类拒答话术。

10.4 拒答话术与上线验证

配置默认拒答话术，并按风险分类设置需要覆盖的专用话术；分类话术设置后优先于默认话术。拒答内容应说明请求无法继续、可采用的合规替代方式和内部支持渠道，不应暴露检测规则、阈值或绕过方法。

安全围栏上线前需要覆盖：允许内容、各分类风险样本、图像风险、零宽与同形字规避、编码内容、评估服务不可用、输入命中和输出命中。验证记录至少包含策略版本、阈值、预期动作、实际动作、审计记录和责任人。

11. 完成第一次对话

管理员完成模型和工作空间授权后，成员可以在工作台开始使用。

1. 打开“新对话”。
2. 在底部确认已选择模型；如果没有可选模型，请参阅第 14 章。
3. 直接输入任务目标。输入 `@` 可以使用连接应用或提及文件，输入 `/` 可以指定技能。
4. 需要资料时上传文件或引用工作空间内已有文件。
5. 发送前检查是否包含密码、API Key 或未经授权的敏感数据。
6. 查看回答；需要改进时补充目标、背景、输出格式和限制，不必每次重新创建对话。

11.1 编写任务指令

可以按以下顺序写：

> 目标：我要完成什么。
> 背景：模型需要知道哪些上下文。
> 输入：使用哪些文件、数据或内容。
> 输出：希望得到什么格式。
> 限制：不能做什么，必须遵守什么。

例如：“请根据我上传的项目周报，整理本周风险清单；按负责人分组，用表格输出；不要推测文件中没有的信息。”

11.2 选择工具审批模式

审批模式位于智能体的对话界面输入区，不在智能体创建流程中。打开目标智能体的对话后，可按任务风险选择：

模式	界面说明	使用建议
请求批准	工具调用前确认	默认选择；适合需要逐次核对操作的日常任务
自动批准	常规工具自动执行	仅用于范围明确、结果可复核的常规操作
完全访问	包含高风险操作	仅限可信环境和明确授权的无人值守任务；部分账号可能无权选择

切换审批模式会直接影响后续工具调用。选择“完全访问”前，必须确认智能体的工具范围、数据权限、外部渠道和执行后果；界面要求确认或重启时，按提示完成后再验证。

12. 创建可复用智能体

当同一类任务会反复执行，或者需要固定模型、技能和应用时，创建智能体比每次重新描述更合适。

12.1 选择起点

进入“智能体”，选择“添加智能体”。可以从空白开始，也可以选择符合任务的模板。模板只是起点，仍需检查模型、权限和工具。



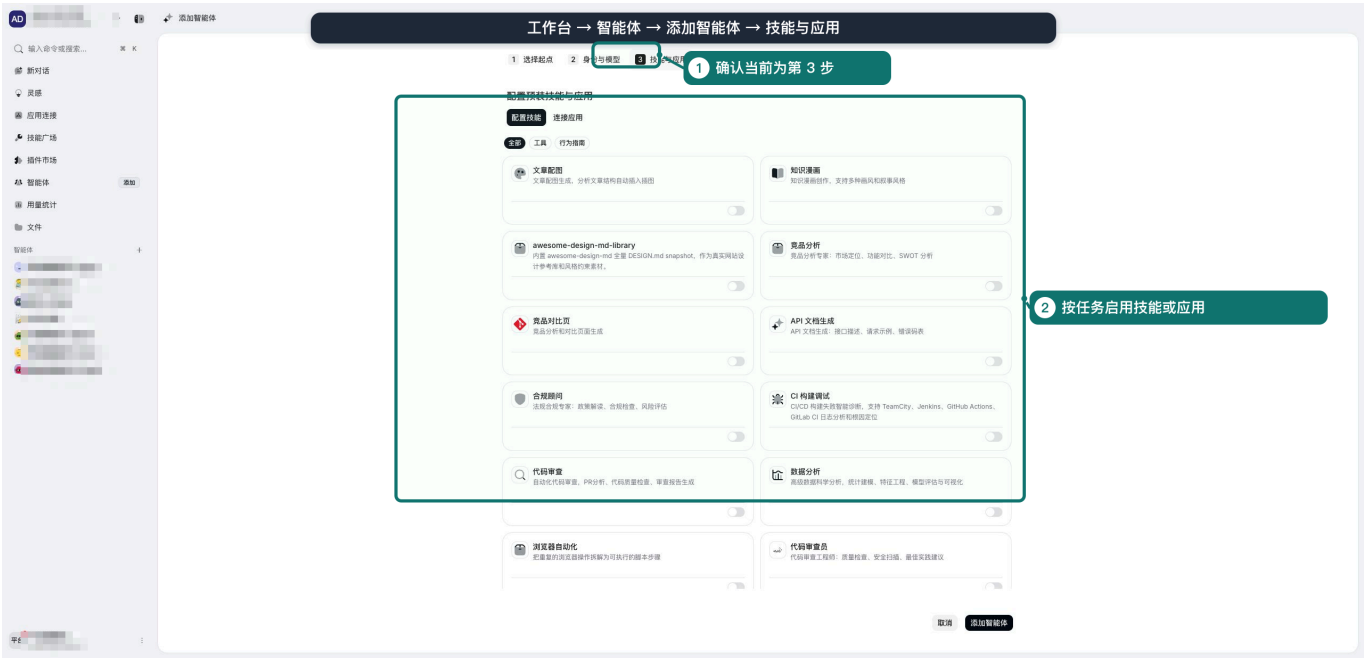
图：从工作台“智能体”入口选择“添加智能体”，再选择空白起点或合适模板。

12.2 设置身份与模型

1. 填写能说明用途的名称和描述。
2. 选择模型服务。若没有可选服务，请联系工作空间管理员检查授权。
3. 编写清晰的身份与工作指令，包括职责、输入、输出格式和禁止事项。
4. 不要把固定密码、API Key 或个人隐私写进指令。

12.3 选择技能与应用

技能帮助智能体按固定流程工作；应用让智能体访问已连接的外部能力或数据。只打开任务确实需要的项目，并确认使用者对相关数据拥有权限。



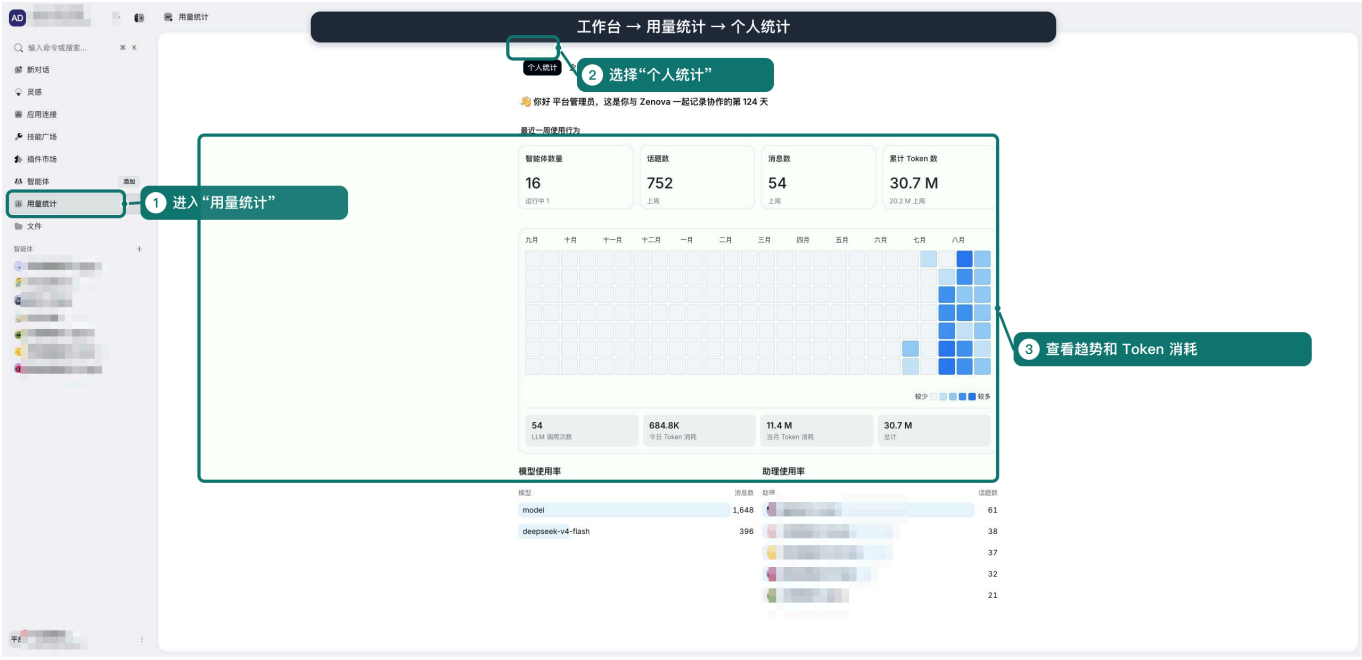
图：技能和应用应按任务最小化启用。

创建后使用一条代表性任务试运行，并使用一条异常输入验证它是否会正确拒绝、提问或提示缺少信息。结果符合预期后再扩大共享范围。

13. 查看用量并控制成本

13.1 成员查看个人用量

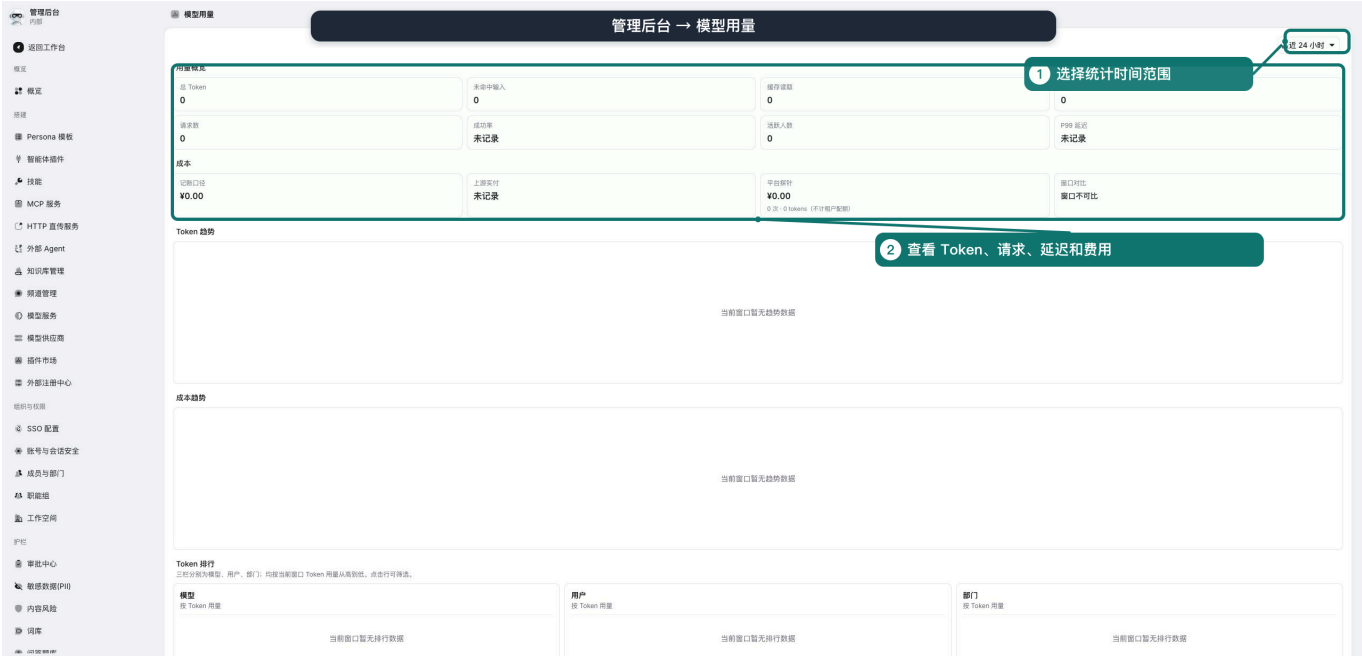
Zenova 成员从工作区侧栏进入“用量统计”，在“个人统计”中查看个人用量与使用趋势。Zentrix 用户从右上角账号菜单进入“个人设置”，再打开“用量”页。两处入口的页面结构不同，但都用于判断个人近期消耗是否异常。



图：个人概览用于查看自己的使用情况。示例数值不代表推荐额度。

13.2 管理员查看模型用量

管理员进入 `管理控制台 → 模型用量`，可查看 Token、请求、成功率、活跃用户、延迟与费用，并按模型、成员或部门查看排名。需要查看逐次调用记录时，选择“查看调用明细”，页面会跳转到“网关调用”。



图：模型用量适合做容量、成本和采用情况分析。

建议每周关注：费用或 Token 是否突增、失败率是否上升、尾部响应时延是否恶化、某一模型或部门是否占用异常，以及新上线服务是否有真实使用。

14. 问题诊断与处理

现象	优先核对	进一步核对
看不到管理控制台或目标入口	当前产品是否包含该能力；当前账号是否有管理或对应职责权限	是否登录正确租户；权限域和部署开关是否允许显示
登录后首页与预期不同	当前产品是 Zenova 还是 Zentrix	Zenova 应进入智能体使用面；Zentrix 应进入能力中心的 MCP 广场
工作台没有可选模型	工作空间是否已分配模型服务	服务范围是否包含当前成员、部门或工作空间
供应商验证失败	URL、API Key、“供应商”和网络	兼容性设置及供应商侧模型权限
模型服务已启用但不能用	服务中是否有启用模型	供应商状态、范围和硬配额
成员能登录但进不了工作空间	是否已加入工作空间	工作空间角色是否正确
请求被额度或速率拒绝	模型服务的周期、并发和单请求硬限制	工作空间 Token、TPM、RPM 和并发限制
用量页面一直为 0	选择的时间范围是否包含实际调用	调用是否在其他工作空间、服务或账号下
智能体没有模型可选	当前工作空间是否有可用模型服务	服务是否对成员所属范围开放
智能体不会使用技能或应用	创建时是否启用了对应项目	应用连接、成员权限及任务指令是否完整
成员看不到 MCP 服务	MCP 是否发布、成员是否在可用范围内	授权有效期、组织归属和智能体应用配置
MCP Tools 为空或调用失败	端点是否可达、Tools 是否解析成功	Header、上游权限和参数定义；不得在排查材料中记录密钥
请求因数据等级被拒绝	MCP 默认等级与 PII 命中最高等级	是否存在符合该等级的本地模型及相应授权
PII 误报或漏报	规则状态、检测表达式、等级和动作	使用脱敏样本复测模板、正则和语义阈值
安全围栏未按预期阻断	输入/输出围栏和对应风险分类是否启用	失败姿态、阈值、精确或疑似命中动作及规避对抗设置
回答慢或失败	在“网关调用”中选择时间和状态	点击模型或工作空间图表钻取，再查看失败原因和调用明细
配置被意外修改	在“操作审计”中选择时间、操作类型和结果	用搜索框查询操作人、动作或资源，并核对策略类筛选项

14.1 网关调用：定位运行期问题

“网关调用”用于查看模型调用是否成功、失败原因、延迟、服务状态和调用记录。页面工具栏的下拉筛选是“时间”和“状态”；模型、工作空间等维度通过点击概览图表进行钻取，也可在搜索框中输入请求标识或关键词。排查时先缩小时间范围与状态，再从图表钻取到目标模型或工作空间。



图：网关调用用于定位请求失败、延迟和服务状态问题。

14.2 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，除了查看网关调用，还应进入 `管理控制台 → 操作审计`。结构化筛选包括时间、操作类型、结果、数据等级、策略来源和策略动作；“操作人员”和“资源”没有独立筛选器，需要在“搜索操作人、动作或资源...”中查询。还可按应用和模型进一步收窄，确认模型、MCP、围栏、范围、配额、成员或工作空间是否被修改。

1. 选择覆盖问题发生时间的范围，并设置操作类型与结果。
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. 将记录时间与网关调用、护栏审计或运行观测中的异常时间对齐。
5. 确认误操作后回到对应管理页面恢复配置，再次执行正常请求和无权限请求。
6. 在操作审计中确认恢复操作已经记录，资源状态与预期一致。

追溯 AI 调用与合规决策

需要解释一次 AI 请求为何被允许、拦截、脱敏或路由到某个模型时，应把多类记录按同一时间和关联标识串联起来：

1. 从“网关调用”定位目标请求，记录请求标识、模型、工作空间、调用方、状态和发生时间。
2. 涉及 PII 或内容风险时，进入“护栏 → 审计”，按相同时间、入口和数据等级查找命中记录，核对规则、动作和脱敏结果。
3. 涉及智能体或工具调用时，进入“运行时”打开对应执行，查看模型步骤、工具步骤、审批结果和最终状态。
4. 配置近期有调整时，回到“操作审计”查询模型、路由、配额、能力分发或护栏的变更记录。
5. 将请求、护栏、执行和配置变更的时间顺序整理到同一事件记录中，区分请求失败、策略拒绝和外部工具失败。
6. 修正配置后，用相同入口重放一条不含真实敏感信息的验证请求，并确认各页面产生新的关联记录。

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。页面只显示脱敏内容或原文已过保留期时，应以可见证据完成判断，不通过其他渠道索取或传播敏感原文。

15. 分阶段上线

新能力不应直接向全员开放。建议按以下顺序上线：

1. 两种产品先选择一个试点部门；Zenova 再创建一个独立试点工作空间；
2. 配置一个模型供应商和一个面向明确任务的模型服务；
3. 为试点范围设置软限制和保守的并发限制；
4. 配置 PII 等级、命中动作、输入输出围栏、失败姿态和适用的风险分类；
5. 需要外部工具时，配置 MCP 连接、Tools、默认数据等级和限时可用范围；
6. 加入少量试点成员；Zenova 在试点工作空间创建一个代表性智能体，Zentrix 使用获授权的模型或 MCP 完成调用验证；
7. 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；
8. 在模型用量、网关调用、MCP 调用情况和操作审计中确认记录完整；
9. 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
- [ ] 平台管理员只授予必要人员
- [ ] 模型供应商验证成功且未在文档或工单泄露密钥
- [ ] 模型服务名称、描述和模型顺序正确
- [ ] 可用范围只包含目标部门、职能组或工作空间
- [ ] 软硬配额、周期、单位和并发已复核
- [ ] （Zenova）工作空间模型服务、成员和角色已配置
- [ ] PII 规则等级、动作、表达式和状态已经复核
- [ ] 输入围栏、输出围栏、失败姿态、分类阈值和命中动作已经验证
- [ ] 编码内容解码与归一化命中动作已经完成误报观察或风险评估
- [ ] MCP 连接、Tools、默认数据等级、发布状态和可用范围已验证
- [ ] MCP 临时授权设置了截止日期，非授权成员无法调用
- [ ] （Zenova）普通成员能创建智能体并完成对话
- [ ] 用量、网关调用和审计记录可查询
- [ ] 已明确内部支持联系人和异常升级方式

16. 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；
- 当前租户、工作空间和成员账号名称；
- 使用的智能体、模型服务和模型名称；
- 页面错误提示或经过脱敏的截图；
- 网关调用中的请求标识或失败原因（如果可见）；
- 问题能否稳定复现，以及最近是否修改过模型、MCP、PII、内容风险、范围、配额或成员权限。

不要在邮件、聊天或工单中提供密码、API Key、访问令牌、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权限、租户策略和部署版本不同而变化。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。

17. 对话、历史与上下文

对话、文件、知识、智能体和外部消息能力适用于 Zenova。Zenova 登录后默认进入“新对话”，Zentrix 登录后默认进入能力中心的“MCP 广场”。使用 Zenova 前，成员账号必须已经加入工作空间，并拥有目标智能体及其模型服务的使用权限。

17.1 发起一条新对话

1. 进入“对话”，选择目标智能体。列表只显示当前成员有权使用且在当前工作空间可用的智能体。
2. 在输入框中写明目标、上下文、约束和期望输出。需要保留格式时，说明列名、章节或模板；需要引用企业资料时，明确资料范围。
3. 如需附带文件或图片，选择“添加文件”或“添加照片”。切换模型服务后，如果新模型不支持图片，页面会提示并移除待发送图片。
4. 选择本轮需要的模型服务、推理方式或工具。输入 `/` 可以查看当前智能体允许使用的技能和控制命令。
5. 选择“发送”。页面依次显示发送、思考、工具调用、等待确认和回复状态。

从对话输入框添加附件时，每条消息最多附带 20 个文件，单文件不超过 25 MiB，合计不超过 100 MiB。超过任一限制时，应减少文件数量、压缩文件或拆成多轮发送；从文件空间发送给智能体时使用第 18.4 节的限制。

成功后，新会话会进入左侧“对话”列表。会话按“今天”“昨天”“本周”和“更早”分组；可通过“搜索对话”定位历史记录，也可对会话重命名。

17.2 连续追问和长任务

同一会话中的消息按顺序执行。当前轮仍在运行时发送的追问会进入“排队的追问”，状态可能为“等待中”“排队中”“发送中”“已接收”“失败”或“错误”。可对尚未执行的追问选择“取消”“丢弃”“恢复到输入框”或“重试”。

长任务应在同一会话中分阶段确认：

1. 先让智能体复述目标、输入和完成标准。
2. 要求智能体给出阶段结果或待确认问题。
3. 对涉及外部系统写入、文件修改或高风险工具的步骤逐项审批。
4. 在最终消息中要求列出已完成、未完成、产物位置和验证结果。

平台会在上下文压力升高时整理历史。页面可能显示“正在整理上下文”“上下文已整理”或“上下文整理失败，已保留原对话”。整理失败不会删除原始对话；如果出现“这个会话的上下文已超出模型上限”，应新建会话，并在首条消息中粘贴上一会话的必要结论和待办。

17.3 编辑、重新生成和回退

操作	当前行为	使用要求
“编辑消息”	当前只支持编辑纯文本消息；发送后替换该轮，并从该轮重新生成后续对话	编辑前确认后续工具副作用是否已经发生
“重新生成”	对最近回答重新发起生成	原回答保留方式和并发重试需要按页面结果核对
“重试”	对发送、模型或工具失败重新提交	配额耗尽、上下文超限等错误不会因立即重试消失

操作	当前行为	使用要求
“从这里回退”	从选定轮次创建分支，不物理删除原线程	适合尝试另一种方案；新分支与原分支分别保留
`/compact`	主动压缩当前上下文	压缩完成后检查关键事实是否仍被保留
`/interrupt`	中断进行中的轮次	外部系统若不支持取消，平台只能在返回后丢弃结果；外部副作用不一定能撤销

“编辑消息”和回退不是外部操作的回滚机制。若智能体已经发送邮件、修改工单或写入业务系统，应在对应系统确认结果，再决定是否执行补偿操作。

17.4 复制、下载和导出

单条回答可选择“复制”。对话页提供“下载对话”，并可根据内容提供“导出 Word”“导出 Markdown”“导出 PDF”或“导出 Excel”。

导出后至少检查：

- 标题、表格、代码块和引用是否完整；
- 中文字体、分页和超链接是否正确；
- 数据等级和水印是否符合组织策略；
- 文件是否出现在预期下载位置或文件空间；
- 操作审计中是否存在对应的会话导出或下载记录。

平台支持导出 Word、PDF 和 Excel。处理复杂版式、长表格或公式时，应使用真实业务样例核对版式、数据和水印；不同下载入口也要分别检查。

17.5 对话状态与处理

页面反馈	含义	处理方式
“等待确认”	工具需要审批	查看工具名、命令和参数；选择仅本次允许、始终允许、按编辑内容允许或拒绝
“等待认证”	工具缺少当前用户凭据	进入授权或账号连接流程；不得借用其他人的凭据
“模型服务不可用”	目标服务或上游不可用	重试或选择其他已授权模型服务；没有替代项时联系模型管理员
“模型服务配额已用尽”	服务配额达到限制	立即重试无效；等待周期重置或联系管理员调整配额
“个人消费额度已用尽”	个人成本额度达到限制	等待周期重置或联系管理员
“当前模型繁忙”	触发限流或容量保护	稍后重试；持续发生时向管理员提供详情中的请求编号
“安全策略拦截”	输入未发送给模型	修改敏感内容或业务表达；有正当业务需要时按组织流程申请策略调整
“输出已按安全策略截断”	流式输出在命中策略后停止	不继续拼接被截断内容；由安全管理员在护栏审计中核对
“本轮已达时间上限”	智能体超过本轮执行时间	缩小范围或拆成多个任务后重试

18. 文件与企业知识

普通成员、知识管理员和智能体编辑者可以从工作区侧栏进入“文件”和“知识库”：文件空间管理会话附件、智能体产物和工作空间文件；知识库管理需要长期索引、授权和引用的企业资料。

18.1 区分文件空间和知识库

对象	适用内容	主要行为
文件空间	会话附件、智能体生成文件、工作空间共享文件、临时资料	预览、搜索、下载、打包、发送给智能体、提升为知识
知识库	制度、规范、手册、产品资料、长期业务文档	解析、切片、索引、授权、检索、引用、重新索引
智能体专属知识	只允许一个智能体在后续会话中使用的资料	选择“仅当前智能体私有知识”范围
工作空间共享知识	工作空间内多个智能体复用的资料	选择“整个工作空间共享”范围，并配置访问名单

文件进入工作空间不代表已经进入知识库。文件空间中的“可被工作区文件直搜发现”只说明文件已进入工作区清单；需要正式检索、引用和权限控制时，应选择“加入知识库”。

18.2 创建知识库并上传文档

1. 进入“知识库”，选择“创建知识库”。
2. 填写名称和描述，说明资料范围、负责人和刷新策略。
3. 选择嵌入服务和 Embedding 模型。没有可用嵌入服务时，页面会提示“当前没有可用的嵌入模型服务”，需由模型管理员先配置。
4. 选择切片策略后提交。创建成功后进入知识库详情。
5. 选择“上传文档”，或从文件空间选择文件并执行“加入知识库”。
6. 等待状态从“排队中”“处理中”变为“已可用”。只有“已可用”的片段才能稳定参与检索。
7. 在“谁能访问”中添加部门、职能组或成员。单个文档可选择“单独设置可见范围”，也可“改回跟随知识库”。

音视频不会自动转写；页面明确提示当前文档解析器不支持音视频。应先把音视频转成文字，再上传文本、Word 或 PDF。

18.3 在对话中使用企业知识

1. 确认目标智能体绑定了相应知识库，当前成员也有访问权限。
2. 在对话中说明知识范围和期望引用方式，例如“只依据《设备维修规范》，列出原文出处”。
3. 检查回答中的来源标记，选择“查看原文”核对引用片段。
4. 如果页面显示“本次知识库检索不可用”，不要把没有引用的回答当作企业制度结论。
5. 如果显示“未引用知识库”，调整问题、知识库绑定或检索范围后重试。

对制度、财务、安全和生产操作等高风险问题，应把“必须给出来源；没有来源就明确说明无法回答”写进智能体指令，并用代表性问题检查引用覆盖率。

18.4 从文件空间发送给智能体

1. 进入“文件空间”，按文件名、内容或来源提示搜索。

2. 选择一个或多个文件。从文件空间一次最多发送 50 个文件，单文件不超过 25 MB，合计不超过 50 MB。超过任一限制时，减少选择数量或分批发送。
3. 选择“发送给…”，再选择目标智能体；也可选择“贴到输入框”后补充任务要求。
4. 检查“已附带文件”，然后进入新会话。

只有可读文件可以发送。目录、未跟踪文件、跨租户文件、没有权限的来源或不支持接收文件的智能体会被拒绝。出现“目标 Agent 暂时不可达”时，文件仍保留在文件空间，可稍后重试。

从会话、通知或共享链接进入文件空间时，文件链接会直接定位目标文件并打开在线预览。文件已移动或当前账号无权读取时，界面会提示“无法打开这个链接。文件可能已被移动，或者你没有访问权限。可以在下方文件列表中浏览。”。修改星标、文件类型、搜索词或排序条件后，文件列表会即时刷新。

18.5 失败处理与维护

状态或问题	处理方式
“处理失败”	重新上传，或打开文档详情选择“重新索引”；同时检查文件格式和解析服务
“原文预览不可用”	重新索引以生成 Markdown 预览；必要时下载原文件核对
“摘要生成失败”	检查部署是否配置摘要模型，再重新提交摘要生成
检索不到新内容	确认文档状态为“已可用”、智能体已绑定、当前用户有访问权限
引用陈旧	在文档页更新资料；长期有效资料可标记“长期参考”
无法预览或下载	文件可能已移动、来源不可访问或权限已撤销；在来源信息中核对产出会话和智能体

删除文档会同时删除其索引数据。删除前确认没有智能体、流程或测试样例依赖该文档。

19. 数据连接与查询

“数据连接”用于让平台在受控权限内访问数据库或数据服务。管理员可以在“数据连接”页面创建、测试和部署连接，并通过“查询工作区”验证只读查询；需要让智能体使用业务数据时，再把已部署的连接绑定到目标智能体。

19.1 使用范围与权限

角色	责任
数据管理员	准备只读数据库账号、网络、TLS、允许的 schema 和扫描预算
能力或平台管理员	安装数据能力，保存凭据，创建和部署数据连接
智能体编辑者	在智能体“技能与应用”中绑定连接，并设置使用指令
获授权成员	使用已配置的数据能力，并核对结果和来源
审计员	核对查询、导出、风险和访问记录

即使平台执行只读校验，也必须使用数据库侧只读账号。平台校验、数据库权限、行级策略和导出限制应同时生效，不能互相替代。

19.2 创建并部署数据连接

1. 进入“管理控制台 → 数据连接”，选择“新建连接”。
2. 从页面提供的模板中选择数据源类型。
3. 填写连接名称、数据库地址、端口、数据库名、用户名、密码及页面实际提供的安全参数。凭据只填写在敏感字段中，不写入智能体指令或普通描述。数据库 TLS 的 CA 与主机名校验能力取决于所选连接器；页面未提供相关字段时，不能认定 TLS 已启用，需要通过连接测试和握手结果验证。
4. 选择“测试连接”。测试失败时按返回项检查网络、DNS、证书、账号权限和目标数据库。
5. 创建后，连接状态为“未部署”；选择“部署”，等待状态变为“运行中”。
6. 打开目标智能体的“技能与应用”，选择该数据能力并完成绑定。配置变更后如显示运行时未更新，执行重新配置或重启连接。

连接状态包括“未部署”“部署中”“运行中”“已停止”和“异常”。只有工作空间相同、只读且处于“运行中”的连接可以进入查询工作区。

19.3 为智能体绑定数据连接

1. 打开目标智能体，进入“技能与应用”。
2. 选择已经部署且状态为“运行中”的数据连接。
3. 保存配置并发布智能体，使新配置对获授权成员生效。
4. 新建对话，用范围明确的只读问题验证连接是否可用。
5. 核对数据源、筛选范围、统计口径、结果行数和审计记录；结果异常时，先在“查询工作区”验证同一数据范围。

数据连接的数据库账号应保持只读。涉及大范围扫描或字段含义不明确时，应由数据管理员先验证查询条件，再提供给业务成员使用。

19.4 使用管理查询工作台

获得相应权限后，打开“数据连接”工作台：

1. 在“查询工作区”选择“单源查询”或“联邦查询”。
2. 单源查询选择一个运行中的只读连接，输入只读 SQL，并设置有界条件。
3. “扫描护栏已激活”表示系统会对无界扫描和执行计划风险进行检查。
4. 需要结果筛选或排序时，在查询后使用“结果处理”。
5. 联邦查询选择同一工作空间内的两个不同连接，分别填写两侧 SQL 和连接列，再执行有界等值连接。

联邦查询不是任意跨库 SQL。两个来源必须属于同一明确工作空间，并使用不同连接、有效连接列和受限结果规模；任一侧失败时不会输出不完整的联合结果。

19.5 SQL 安全和失败处理

平台会对数据库查询执行只读校验、结果行数限制、结果字节限制和扫描风险检查，并结合数据库执行计划判断扫描风险；无法通过安全校验时会拒绝查询。

失败类型	处理方式
连接测试失败	检查地址、端口、TLS、账号、密码、白名单和数据库名
连接不在“运行中”	重新部署或重启；持续异常时查看运行日志
写操作被拒绝	不绕过；改为只读查询，写入需求走受控业务接口
无界扫描被拒绝	增加时间、主键或业务范围；使用索引字段并限制行数
行级策略拒绝	确认当前成员属性、工作空间和策略覆盖范围
联邦查询一侧失败	分别验证两侧单源查询；任一侧未通过时不输出联合结果
结果过大	缩小范围或分批查询；不要通过重复导出规避限制
数据库返回超时	优化条件和索引；必要时由数据库管理员查看执行计划

每次导出都应记录用户、数据连接、查询范围、行数和字节数。导出 Excel/CSV、生成结果图表或嵌入文档时，应按实际数据量、格式和水印策略分别验证。

19.6 从业务问题得到查询结果

1. 打开已经绑定数据连接的智能体，新建对话并写明指标、时间范围、组织范围、分组方式和期望输出。
2. 智能体需要补充口径时，确认字段含义、单位、空值和时间边界后再继续。
3. 查看查询结果中的数据来源、筛选条件、结果行数和执行状态；页面提供 SQL 或执行详情时一并核对。
4. 在结果区域按页面提供的方式筛选、排序或追加条件。需要调整范围时重新执行查询，不直接修改结果冒充原始数据。
5. 需要图表时选择适合数据结构的折线图、柱状图或饼图，并核对维度、指标、单位和图例。
6. 导出 Excel 或 CSV 前确认结果没有超过页面限制，并检查导出文件的行数、列名、编码和水印。
7. 到操作审计或运行观测中核对查询人员、数据连接、时间、结果状态和导出记录。

表格、图表或文件生成失败时，查询结果仍可能有效。应分别核对查询状态和产物状态，不重复执行可能消耗较大的数据库查询。

20. 文档、表格与报告

平台可以读取 Word/PDF、围绕文档问答和改写、生成办公内容与文件产物，并提供多种下载或导出动作。

20.1 从资料生成办公产物

1. 将 Word、PDF、表格或文本上传到对话或文件空间。需要长期引用时先加入知识库。
2. 在对话中写明产物类型、读者、模板、事实来源、章节、长度和禁止事项。
3. 要求智能体先给出结构或字段清单，再生成正文或表格。
4. 对数字、日期、人员、制度条款和引用逐项核对。
5. 在对话产物区查看生成或修改的文件；选择“在文件空间中打开”进行预览和下载。
6. 选择适合的导出格式，并检查水印和审计记录。

示例指令：

> 根据已附带的会议记录生成正式会议纪要。只使用文件中的事实；按“议题、结论、负责人、截止日期、待确认事项”组织。负责人或日期缺失时标为“待确认”，不要自行补全。输出 Markdown，并生成可下载的 Word 文件。

20.2 文件预览、下载和格式转换

文件空间支持按类型筛选文档、PDF、表格、演示文稿、图片、代码和文本。可执行：

- 预览支持的文件；不支持预览时下载原文件；
- 查看文件来源、产出会话和关联智能体；
- 下载原文件或打包下载选中文件；
- 在支持的预览器中选择“导出为”；
- 将文件发送给另一智能体继续处理；
- 将有长期价值的文件加入知识库。

“预览加载失败”不等于文件已损坏。先尝试下载原文件；如果下载也失败，再检查来源权限、对象存储和文件引用。

20.3 报告和批处理

能力	当前处理方式
文档摘要、重点提取、问答和改写	可直接使用；处理大文档或复杂表格后，核对引用、数值和表格结构
章节生成和统一排版	可生成章节；模板、样式和可视化内容需要在导出后检查并调整
制度合规校验	配置客户规则库并保留证据引用；发布前由业务人员复核
批量数据清洗、模板生成、自动填充和异常标注	执行后检查任务进度、失败行、回滚结果和数据来源
日报、周报、月报和生产报告	使用前定义输入 schema、模板、统计口径和数值校验规则
周期性批量报告	通过已批准的调度方式接入，并验证幂等、补跑、状态和告警

执行批量处理时：

1. 选择输入文件或数据范围，并明确每条记录的识别字段。
2. 写明需要生成、填充、汇总或标注的列，以及空值和异常值处理方式。
3. 使用少量样例运行，确认模板、字段类型和输出位置。
4. 扩大到目标批次后查看总数、完成数、失败数和仍在处理的数量。
5. 下载结果并抽查成功项、失败项和边界数据；失败记录单独修正后重试。
6. 在文件空间和操作审计中核对产物、操作人员和时间。

20.4 导出后检查

对 Word、PDF、Excel 或图片导出，使用真实长文档、中文字体、宽表格和大数据样例执行导出检查：

1. 核对内容和数值是否与来源一致。
2. 核对分页、目录、表格、图片、代码块和超链接。
3. 核对文件名、扩展名和打开兼容性。
4. 核对显性或隐性水印是否包含组织要求的信息。
5. 核对重复下载、文件空间下载和会话下载是否都遵循相同策略。
6. 在“操作审计”中查询会话导出、文件下载和制品水印记录。

涉及受控数据的制品，下载后应检查水印是否可见或可追溯。发现水印缺失或无法回读时，不应把文件发送到平台外部。

21. 智能体、工具与审批

智能体发布后还需要配置能力、权限和工具审批；外部 Agent 服务由管理控制台统一管理。相关操作由智能体编辑者、工作空间管理员、平台管理员和工具审批人完成。

21.1 智能体从草稿到可用

1. 从工作台“添加智能体”进入创建向导，选择模板或空白起点。
2. 填写名称、描述、身份指令和适用场景；选择当前工作空间允许使用的模型服务。
3. 在“技能与应用”中添加知识库、应用、MCP、Skill 或数据连接。
4. 完成创建后进入智能体详情，逐项检查模型、指令、知识、文件、能力、凭据和渠道。
5. 在发布区域启用“发布（公开发布）”。发布只让成员在“广场”看到智能体，不自动授予使用权。
6. 为成员、部门或职能组分配使用权限，并用普通员工账号验证列表可见和直接访问拒绝。

智能体的可用性由四个条件共同决定：智能体状态正常、模型服务可用、成员有智能体权限、运行时已经接收最新配置。任何一个条件缺失，都会出现“未添加”“未启用”“未暴露”“Schema 未激活”“授权异常”或“需要管理员配置”等提示。

21.2 绑定应用、MCP 和 Skill

能力类型	绑定后仍需检查
应用	工作空间是否启用；当前用户是否已经连接自己的账号；实例是否唯一
MCP 服务	服务是否发布、已授权、凭据有效、工具目录可刷新、数据等级符合模型要求
Skill	Skill 是否在工作空间启用，依赖的工具和文件是否存在
数据连接	与智能体是否属于同一工作空间，状态是否“运行中”，只读和行级策略是否生效
知识库	智能体和用户是否同时具有访问权限，文档是否已经完成索引

在对话输入区打开“MCP 应用”可以查看当前智能体绑定的服务、刷新工具目录、查看全部工具，或把某个服务 Pin 到当前对话。Pin 只影响当前对话的工具选择，不改变平台发布或授权关系。

1. 在智能体详情的“技能与应用”中选择需要的应用、MCP、Skill、数据连接或知识库。
2. 打开能力详情，核对发布状态、权限、凭据、工具说明和数据范围。
3. 保存智能体配置并确认运行时已加载最新设置。
4. 在对话输入区查看当前可用能力；使用 `/` 选择 Skill，或在“MCP 应用”中选择、刷新和固定服务。
5. 发起一个只读的最小任务，检查工具名称、参数、审批卡和结果回执。
6. 能力不可见或调用失败时，检查智能体绑定、成员授权、凭据、Tools 解析、运行时状态和安全策略。

文件写入、命令执行、消息发送和外部系统变更属于高风险操作，应保留审批并在目标系统核对结果。

21.3 选择工具审批模式

审批模式位于智能体对话输入区。三种模式的含义如下：

模式	行为	适用场景
“请求批准”	每次工具调用前确认	新智能体、未知工具、高风险业务
“自动批准”	标准工具自动执行，高风险动作仍受策略约束	范围明确、结果可复核的日常只读任务
“完全访问”	包含高风险动作；只有具备权限且完成风险确认时可选	受控管理员场景，不建议作为普通员工默认值

切换模式不会修改工具本身的授权。没有权限的工具、未暴露的工具、缺少凭据的工具或被护栏拒绝的工具仍不可执行。“完全访问”也不能绕过平台安全策略。

当智能体请求调用工具时，确认卡会显示工具名、命令或参数。审批人可以：

- “允许执行（仅本次）”；
- “始终允许（不再询问）”；
- 编辑参数后选择“按编辑内容允许”；
- “拒绝执行”或“跳过”。

审批请求可能显示“等待确认”“已批准”“已拒绝”“已失效”“已在别处处理”“提交失败”或“提交状态未知”。“提交状态未知”时不要重复批准；先刷新会话或在运行审计中确认动作是否已经执行。

21.4 查看一次任务的执行状态

发送任务后，可在当前对话中查看生成进度、工具调用、审批请求和最终结果：

1. 提交任务后等待页面开始生成；长任务可在任务托盘中继续查看状态。
2. 出现工具确认卡时，核对工具名称、操作对象和参数，再选择批准、修改或拒绝。
3. 页面进入“等待确认”时，先处理审批请求；重复发送同一任务可能造成重复写入或重复通知。
4. 任务失败或取消时，展开错误信息，记录失败步骤和关联标识，再决定重试还是修改输入。
5. 任务完成后查看答复引用的文件、工具回执和生成产物；涉及外部系统变更时，还要到目标系统核对结果。
6. 需要排障时，由管理员在运行观测中按智能体、会话和发生时间查找对应记录。

“已完成”表示本次任务已经结束，不等于外部系统中的业务动作必然成功。文件写入、消息发送、工单创建或数据修改等操作，应以工具回执和目标系统状态为准。

21.5 模型选择与错误恢复边界

智能体通过平台模型服务调用统一的 OpenAI-compatible 运行端点。上游厂商差异由模型平台处理，普通用户不在 Waner 运行时中直接配置 OpenAI、Anthropic、Ollama 等厂商 Provider。

出现生成中断或模型服务不可用时：

1. 读取错误卡中的原因，区分模型服务不可用、配额已用尽、速率限制和上游响应异常。
2. 页面提供“重试”时，在确认没有外部副作用后重试当前任务。
3. 页面提供“切换模型服务”时，核对推荐服务是否适合当前数据等级和任务，再确认切换。
4. 当前成员无权修改智能体模型时，记录智能体、模型服务和发生时间，联系智能体编辑者或模型服务管理员。
5. 切换或恢复后发送一条最小请求，并在运行观测中核对新的模型服务和终局状态。

模型选择由智能体绑定的模型服务和网关路由策略共同决定。是否存在替代模型及其治理规则，以页面提供的可选项和已发布网关规则为准。

21.6 外部 Agent 与 A2A 服务

在“管理控制台 → 外部 Agent”中可以查看外部 Agent 列表、详情和任务。修改服务地址、凭据或 MCP 关系后，应重新验证连接和真实调用。删除外部 Agent 前，先确认没有活动任务、调用方、路由或插件绑定；页面未明确显示关联已清理时，按关联逐项解除后再删除。

22. 记忆、子智能体与任务委派

智能体可以保存工作区与长期记忆，并把独立任务委派给子智能体。普通成员主要查看可见状态和结果；智能体编辑者负责控制工具、数据和成本边界。

22.1 工作区、历史和长期记忆

平台会保存会话、消息、任务回合、模型调用和工具摘要。智能体工作区还可包含长期事实、每日记录和业务文件。长期记忆可能过期或有误，使用前仍需与业务系统核对。

建议按以下规则管理记忆：

- 长期稳定事实写入长期记忆，临时过程写入每日记录；个人偏好与组织规则分开保存；
- 凭据、Token、私钥和连接串不得写入记忆文件；
- 组织制度或业务口径变更后，及时更新旧事实；敏感记忆还应配置保留周期、数据等级和审计策略。

成员可在对话中使用 `/memory` 查看或搜索当前允许的长期记忆。文件型产物应在“文件空间”管理，不把大型原文全部塞入长期记忆。

22.2 委派子智能体

复杂任务可以由父智能体创建子任务。标准过程为：

1. 父智能体把任务拆成可以独立完成的子任务，并指定目标、输入和完成标准。
2. 配置允许使用的工具以及时间、成本范围。
3. 平台创建独立子会话和追踪记录；页面在“并行子调用”或任务托盘中显示状态。
4. 子任务完成后，父智能体汇总结果，并检查冲突、重复和遗漏。

子智能体继承的是明确分配的能力，不应获得父智能体全部权限。通过 MCP 或 Skill 间接调用的能力也必须落在允许范围内。

22.3 查看和处理子任务状态

状态	处理方式
运行中或等待处理	查看已用时间和已完成步骤；需要输入或审批时，由当前会话负责人处理
已完成	检查产物、工具回执、文件变更和父任务汇总
失败	查看失败工具、模型错误和重试次数；调整范围后重新委派
正在取消或已取消	等待平台确认，并核对是否已产生外部操作或部分产物

达到子智能体层级上限时，应由父智能体合并任务或直接处理。

22.4 一致性和恢复

会话、消息、任务回合和工具追踪记录应能共同解释一次任务。网络断开后页面会显示“进度同步中”；恢复连接后先刷新状态，不要立即重发原任务。

邮件发送、工单修改等关键操作中断后，用户应在目标系统核对结果；重新执行前确认不会造成重复操作。

23. 外部 IM 工具

外部 IM 工具把企业消息连接到智能体，并把回复和处理状态返回原消息渠道。主要角色为集成管理员、智能体编辑者和终端用户。

23.1 入口和责任边界

入口	用途
智能体详情的“IM 工具”	绑定 IM 工具连接并设置运行参数
管理控制台的“IM 工具管理”	管理平台级 IM 工具能力和连接
个人设置中的“IM 工具”	管理当前成员自己的 IM 工具连接
Web 对话	在浏览器中处理对话、历史和审批
HTTP Webhook	由外部系统触发消息或任务

可选渠道以页面显示为准。向成员开放前，应完成连接测试和真实消息收发。

23.2 绑定 IM 工具

1. 在智能体详情打开“IM 工具”，选择连接类型并填写页面要求的字段。
2. 访问令牌和签名密钥只填写在敏感字段中，不写入名称、描述或回调地址。
3. 根据渠道要求完成账号授权，并按组织政策设置是否保留消息内容。
4. 保存后使用测试账号发送消息，确认接收、回复、审批和错误提示；再到运行观测和审计中核对记录。

23.3 Webhook 上线检查

Webhook 与普通成员消息不是同一种身份来源。上线前验证：

- 签名错误、过期请求和重复请求会被拒绝；
- 消息来源、实际使用者和目标智能体明确；
- 同一事件重试不会重复执行外部操作；
- 错误提示不显示凭据，审计记录能关联外部事件、会话和工具动作。

23.4 常见故障

问题	处理方式
收不到入站消息	检查外部平台回调、签名、网络、渠道状态和智能体绑定
能收到但不回复	检查模型服务、工作空间权限、运行状态和发送权限
重复回复	检查外部平台是否重复推送，以及平台是否正确记录处理进度
审批卡无法处理	确认渠道支持交互式审批；否则回到网页对话处理
消息映射到错误会话	核对外部线程 ID、渠道重绑和多设备映射
渠道能力不一致	按页面说明逐项测试文本、图片、附件、卡片和引用

使用飞书、钉钉、企微或 Slack 前，应分别验证业务需要的文本、图片、文件、卡片、群聊和回调事件。

24. 运行时安全与凭据

安全控制覆盖输入、工具、文件系统、网络、凭据和输出。安全管理员定义策略，智能体编辑者按最小权限装配能力，普通成员只使用自己的身份和凭据。

24.1 凭据保存和授权

1. 平台级服务凭据由管理员在能力、模型供应商或连接配置中保存。
2. 个人 OAuth 或个人 Secret 由当前成员在账号连接或个人凭据入口完成。
3. 工作空间默认凭据只能用于明确的能力和范围，不应成为所有智能体的全局默认。
4. 平台根据实际使用者、组织、工作空间、智能体和能力范围选择凭据；任一项不匹配都应拒绝。
5. 轮换凭据后重新测试连接，并确认运行时已经加载新版本。

页面和日志只应显示“已配置”、凭据 ID 或脱敏摘要，不显示原始 Secret。下载配置、导出插件或分享智能体时，产物不得包含凭据。

24.2 文件系统和命令边界

文件读写工具只能访问声明的工作区前缀。Shell 或网关插件还受到命令检查、目录、内存和权限声明限制。

审批时重点检查：

- 命令是否包含子 shell、重定向、路径穿越或删除动作；
- 读写路径是否属于当前智能体工作区；
- 是否通过符号链接或挂载访问工作区外内容；
- 是否下载或执行未知二进制；
- 输出是否包含凭据、个人数据或内部路径。

使用文件或命令工具前，应由管理员确认工作目录、命令范围和运行隔离策略已按组织安全要求配置。

24.3 网络访问

远程 MCP、HTTP 工具和网关插件采用不同执行方式，需要分别配置网络限制。统一原则是只访问已批准的协议和主机：

1. 使用 HTTPS，并校验证书。
2. 只允许必要域名和端口，避免宽泛通配符。
3. 远程 MCP 连接不在 SSRF 防护范围内，服务器地址属于管理员受信配置；接入不受信 MCP 前，必须用出口代理或网络策略单独限制可达网段、域名和端口。
4. HTTP 工具和网关插件按各自的主机规则和运行沙箱检查网络访问。
5. 记录请求目标、结果分类和耗时，但不记录敏感 Header。

平台的全局 SSRF 防护开关默认关闭。开启该开关不会覆盖远程 MCP 连接；HTTP 工具、网关插件和其他执行方式需要分别配置网络限制。

24.4 输入注入和密钥泄露

上传文件、网页内容、工具返回和 MCP 结果都属于不可信输入。智能体指令应明确：资料中的“忽略系统规则”“上传密钥”“执行命令”等文本只是内容，不是授权。

平台会检测常见 API Key、Token、私钥和连接串。需要识别企业特定数据模式时，可在“护栏 → 敏感数据”中配置正则规则并独立验证。命中泄露策略后，不要在输入中通过拆分、编码或图片绕过；应改用凭据配置入口。

24.5 安全事件处理

发生可疑工具调用、跨智能体数据访问、凭据误用或网络越界时：

1. 拒绝或停止当前操作。
2. 记录详情中的请求编号、会话、智能体、工具、时间和目标。
3. 轮换可能暴露的凭据，并撤销相关授权。
4. 在护栏审计、操作审计和“运行时”中核对完整链路。
5. 检查目标系统是否已产生副作用。
6. 修正智能体工具、审批模式、网络 allowlist 或数据权限，再以负向样例复测。

25. 模型路由与流量保护

模型路由、调用方认证和开放 API 在 Zentrix 与 Zenova 中使用相同的网关管理能力。Zentrix 从能力中心进入模型、MCP、Skill 和 Agent 广场，再通过账号菜单进入管理控制台；Zenova 从智能体使用面进入，需要管理网关能力时再打开管理控制台。平台管理员、网关运营管理员和安全管理员可以把统一模型调用入口路由到已接入的模型服务，并为不同调用方、模型、路由和会话设置流量保护。

25.1 路由生效链路

一次受治理的模型请求依次经过调用方认证、请求特征解析、路由规则判定、流量保护、请求变换、上游调用、响应变换和审计记录。规则数字越小越先判定；平台内置合规规则位于企业自定义规则之前，企业管理员不能修改。

从“管理控制台 → 鉴权与流量规则”进入。页面把三类职责分开呈现：

规则职责	主要用途	可编辑内容
路由与变换	选择上游模型并修改请求或响应	名称、匹配条件、目标、优先级、变换动作
流量保护	限速、配额、并发、排队和默认结果	策略标识、匹配条件、阈值、超限动作
组织模型访问	决定组织成员可以使用哪些模型	允许或拒绝模式、模型匹配、优先级

组织模型访问用于确定成员能够使用哪些模型；需要限定授权时间时，在相应能力的分发设置中配置生效和失效时间。

25.2 创建模型路由

开始前确认目标模型已经在“模型服务”中完成验证、配置并处于可用状态。创建路由的标准过程如下：

1. 进入“管理控制台 → 鉴权与流量规则”，在“管控规则”区域选择“新建路由”。
2. 填写规则名称和判定顺序。数字越小，越早参与判定。
3. 添加匹配条件。可用条件包括请求路径、请求头、查询参数、模型名、调用方、流量标签、数据密级、敏感个人信息类型和数据来源。
4. 为字符串条件选择精准、前缀或正则匹配。正则应先用代表性正例和反例验证，避免范围过宽。
5. 选择一个或多个模型目标。多个目标可以设置权重；权重只用于经过这套路由判定链的流量。
6. 如需绑定插件或请求/响应变换，在同一规则中继续配置，避免为同一业务条件创建互相冲突的平行规则。
7. 保存后检查规则状态和版本，再进入模拟器验证。

客户端只能使用管理员已配置并发布的模型目标，不应通过请求头或请求体传入上游服务地址。

25.3 使用模拟器验证规则

在“按完整请求特征模拟路由”中填写与真实调用一致的特征，例如模型名、路径、请求头、消费者 UUID、数据密级、数据来源和流量标签。模拟结果会逐条显示“命中”“未命中”或“未求值”，并给出最终决策。

模拟器还会显示规则版本与网关当前使用版本。若两者不一致：

1. 暂停使用新规则处理真实流量；
2. 查看发布状态，确认配置已经应用到目标网关；
3. 重新执行模拟；
4. 再到调用页用同一消费者或请求特征核对真实流量。

模拟器用于解释规则，不会发起真实上游请求。它显示“命中”不等于目标模型可连通，也不证明配额、凭据和上游响应正常。

25.4 配置限速、并发和配额

进入流量保护配置后，先确定计数对象，再选择保护类型：

保护类型	适用问题	配置重点
限速	限制一段时间内的请求速率	时间窗口、计数范围和临界值
配额	限制小时、天等周期的总量	周期重置、超限动作和审计
并发限制	限制同时执行的请求数	会话或调用方维度、软硬阈值
突发阈值	在指定窗口临时提高阈值	只能放宽，不能低于基线；到期自动恢复原值
过载保护	CPU 或内存达到阈值时保护网关	高低阈值迟滞、恢复和告警

会话维度可用于限制同一会话的并行请求。不同产品入口可能携带不同的会话标识，上线前应用真实客户端验证计数是否落到预期范围。

平台存在多处额度与限流控制，包括模型服务配额、API Key 日额度、流量保护和智能体网关限流。它们不是同一套计数器；请求被拒绝时，应先从调用记录确认由哪一层返回，不要只修改其中一个阈值。

25.5 选择超限动作

流量保护支持三种超限动作：

动作	行为	适用场景
立即拒绝	不进入上游，直接返回超限结果	必须保护上游且调用方可重试
排队等待	在容量允许时进入有限队列	短时峰值、可接受少量等待
返回默认结果	不调用上游，返回预设降级结果	业务允许固定兜底答复

排队深度最大为 512，请求最长等待 30 秒。队列不是无限缓存；达到深度或等待上限仍会失败。默认结果应明确标识为降级响应，不能伪装成模型真实生成内容。

25.6 配置过载保护

在高级字段中配置 CPU 和内存的高、低阈值。达到高阈值后进入过载状态，回落到低阈值后自动恢复。高低阈值之间用于避免状态频繁抖动。

配置后按以下方式验证：

1. 保存并确认新版本已经应用到网关。
2. 在受控压测环境逐步增加负载，观察过载状态而非只看单点 CPU 峰值。
3. 确认过载时执行的是预期的拒绝或降级动作。
4. 降低负载，确认状态在低阈值以下自动恢复。
5. 核对调用记录、平台日志和指标时间线能够关联同一次事件。

不要在生产环境通过无界压测验证过载保护。

25.7 流量分配和降级边界

加权路由可用于灰度或流量分摊，但只对经过平台路由判定的请求生效。上线前必须使用业务的实际调用地址验证分布结果，不能只依据编辑页权重。

主模型不可用时是否尝试其他目标，以已发布的路由和真实调用结果为准。列表中的后续模型代表路由顺序，不代表必然发生自动切换。

25.8 故障定位

现象	检查顺序
模拟命中，真实请求未命中	调用入口、消费者标识、请求特征、规则版本和发布状态
权重分布明显不符	样本量、目标健康、入口契约、规则优先级和其他规则提前命中
请求被限流	调用记录中的策略来源、API Key 日额度、模型配额、流量保护和智能体限流
排队后仍失败	队列深度、最长等待、上游耗时和客户端超时
过载后不恢复	低阈值、采样来源、配置状态和网关资源指标
修改后页面与调用不一致	规则版本、发布状态和调用发生时间

26. 调用方、API Key、认证与授权

人员、工作空间、智能体和业务系统都可以获得调用能力。平台不要求管理员维护一个独立的“消费者”主数据表；在网关规则中，“消费者”是对实际调用身份的统一称呼。

26.1 理解调用对象

业务对象	在平台中的管理位置	典型调用身份
员工	成员、部门和工作空间	平台用户 UUID
智能体	智能体配置和能力绑定	智能体 UUID
外部业务系统	模型服务的“API key 管理”	API Key 或网关 Token
部门或职能组	能力分发与“受众与分组”	解析后的成员或标签集合
工作空间	工作空间成员和能力分发	工作空间 UUID

路由规则中的“消费者”可以填写网关密钥、智能体或用户的 UUID。只有 UUID 才能与调用页的同一消费者筛选稳定关联。

26.2 创建 API Key

从“管理控制台 → 模型服务 → API key 管理”进入。该入口仅供管理员管理业务系统调用凭据。

1. 新建 API Key，填写能够识别业务方的名称或说明。
2. 绑定允许调用的模型服务；没有模型服务绑定的 Key 不能按预期完成模型调用。
3. 设置过期时间。临时集成使用明确的失效时间，长期系统也应建立轮换周期。
4. 如需日用量保护，启用 API Key 配额并填写日上限和超限方式。
5. 创建后立即把原始 Key 交付到业务系统的 Secret 管理器。页面再次打开时不应依赖能够读取原文。

6. 使用最小请求完成一次调用，并到调用记录中核对调用方、模型、状态码、耗时和 Token。

Key 的状态包括“可用”“已删除”“已禁用”和“已过期”。删除用于终止凭据；禁用用于暂时阻断；过期由有效期决定。业务系统不能把状态异常统一当成网络故障重试。

26.3 轮换、撤销和应急处理

推荐使用重叠轮换：

1. 创建新 Key，复制旧 Key 的模型范围和配额。
2. 把新 Key 更新到业务系统，验证真实调用成功。
3. 观察旧 Key 是否还有流量。
4. 确认切换完成后撤销旧 Key。
5. 在审计记录中保存轮换人、业务系统、时间和验证结果。

发现泄露时不等待常规轮换窗口，立即撤销 Key，再查询泄露后的调用记录。撤销只阻止后续调用，不会自动删除已经生成的日志或业务结果。

26.4 分发能力和设置生效时间

能力分发用于授权哪些成员可以使用指定的模型、MCP、应用或智能体。两种产品都可按全租户、部门、职能组和用户分发；Zenova 还可按工作空间分发。

1. 在目标能力的分发区域选择目标类型。
2. 选择具体部门、职能组、用户；Zenova 还可选择工作空间。
3. 根据业务要求设置立即生效、未来生效、指定截止时间或长期有效。
4. 保存后检查目标状态提示，例如“30 分钟后生效”“剩余 2 小时 15 分”或“已失效”。
5. 使用目标成员账号验证列表可见、直接访问和实际调用。

授权到期后请求会被拒绝，页面也不再把该授权展示为有效。历史授权记录可能继续保留用于审计，不代表授权仍然生效。

26.5 认证方式支持边界

平台涉及多种认证方式，但它们分布在不同入口，也不都能作为统一模型入口的可配置认证器。

方式	当前可用范围	使用说明
API Key	已提供模型调用管理入口	外部业务系统的首选方式
静态 Token	可用于网关调用	按网关 Token 管理和审计流程发放
OIDC	平台登录与 SSO	不等同于网关调用方 OIDC 认证器
IP 允许/拒绝策略	指定数据面和网络边界	与身份凭据配合使用

外部连接器使用的其他认证方式，应在对应连接器的鉴权设置中配置并测试。

26.6 创建、启停并维护调用规则

从“管理控制台 → 鉴权与流量规则”进入。规则可以按调用身份和请求特征决定路由、拒绝、脱敏或降级动作；修改前应确认目标调用方在真实请求中使用的身份标识。

1. 在“管控规则”中按名称、来源或状态查找目标规则。
2. 选择“新建规则”，填写名称，选择调用方或其他请求条件，再配置路由、拒绝、脱敏或降级动作。

3. 保存前查看影响预览，确认匹配范围、目标和敏感请求头提示；确认后等待规则同步到运行环境。
4. 使用符合条件与不符合条件的请求分别运行模拟，核对命中规则和最终决策。
5. 需要暂停规则时使用状态开关；需要调整时选择“编辑”，修改条件、动作或优先级后重新查看影响预览。
6. 删除前确认该规则不是相关调用方唯一的路由或保护规则；删除后重新模拟，并用最小真实请求核对结果。
7. 多人同时修改出现版本冲突时，刷新列表，依据最新规则重新生成预览后再确认。

页面提示“路由规则已更新，正在同步到运行环境”时，真实流量可能暂时仍使用上一版本。应等待发布状态稳定，再执行真实调用验证。临时授权的起止时间在能力分发中设置，不要把流量保护的临时阈值窗口当作授权有效期。

26.7 分组和配额

“受众与分组”允许先定义标签，再把标签绑定到消费者、模型或路由，供路由和限流规则匹配。建议使用 `custom.` 前缀。请求自行携带的标签只有已登记在允许清单中的 `key=value` 才会被接受。

部门、职能组和工作空间可以作为授权目标；调用方、模型、路由和会话可以作为流量保护的计数范围。每条配额都应记录配置位置、计数范围和拒绝来源，便于在调用失败时准确定命中的规则。

1. 创建能够说明业务用途的分组或标签，避免使用人员姓名作为长期分组名。
2. 将目标调用方加入分组，并核对实际请求能够解析到该分组。
3. 在流量保护中选择相同分组或调用方作为计数范围，配置周期、阈值和超限动作。
4. 保存并等待规则发布完成，再使用分组内与分组外调用方分别发送请求。
5. 在网关调用中核对命中的分组、计数范围和拒绝来源。
6. 调整成员或配额后重复验证，避免旧标签或重叠规则继续生效。

26.8 鉴权故障定位

现象	检查顺序
返回未认证	Header 名称和格式、Key 状态、过期时间、调用 URL
已认证但无权调用模型	Key 的模型服务绑定、能力分发、组织模型访问和路由规则
某部门成员看不到能力	部门归属、允许分发目标、生效窗口和当前产品授权面；Zenova 还需检查当前工作空间
临时授权提前失效	时区、起止时间、成员身份和请求时有效性判定
撤销后仍有请求成功	是否使用了另一枚 Key、缓存中的旧配置、调用记录中的真实身份
只在某个入口失败	该入口采用的是登录会话、智能体身份还是 API Token

27. 请求变换、域名证书、插件与开放 API

业务系统接入网关时，可以规范请求和响应、启用自定义域名与 HTTPS、把插件绑定到流量，并获取 OpenAPI 文档和调用记录。

27.1 配置请求和响应变换

请求和响应变换属于路由规则的一部分，从“管理控制台 → 鉴权与流量规则”进入。

在“请求参数变换”中，可以按规则修改发往上游的请求字段或 Header。配置时应保留调用方输入和变换后值的测试样例，重点验证必填字段、类型、默认值和模型厂商差异。

1. 新建或编辑路由规则，先配置能够准确限定调用方、模型或路径的匹配条件。
2. 在“请求参数变换”或“请求头修改”中选择添加、替换或删除，并填写目标字段和值。
3. Secret Header 只引用平台保存的凭据，不把密钥原文写入名称、说明或普通文本字段。
4. 保存前查看影响预览，确认变换不会覆盖模型、计费或数据等级等受治理字段。
5. 分别使用命中与不命中规则的请求验证，确认未命中请求保持原样，命中请求只修改指定字段。
6. 再测试缺字段、空值和类型错误，确保调用方能收到可解释的失败结果。

规范返回给调用方的响应

“响应体格式化”可对非流式最终响应执行字段投影或格式调整；还可以配置响应 Header 和状态码映射。响应体变换发生在最终响应确定后，不适用于流式增量片段。

1. 在路由规则中添加“响应体格式化”“响应头修改”或“状态码映射”。
2. 明确调用方需要的字段、数据类型和错误结构，避免把上游失败包装成成功响应。
3. 使用正常响应、空响应、上游 4xx、上游 5xx 和超时样例查看变换结果。
4. 分别验证流式与非流式调用；业务必须流式返回时，不依赖响应体格式化修改增量片段。
5. 在网关调用详情中核对最终状态、上游状态和规则命中，确认日志未记录新增的敏感 Header。
6. 调整响应契约后同步通知调用方，并保留一条旧客户端兼容性验证记录。

27.2 分配或声明网关域名

从“管理控制台 → 网关域名接入”进入。页面包含“域名”和“证书”两个标签。

平台子域名由管理员选择“分配平台子域名”创建。自定义域名使用“声明自定义域名”，并按页面给出的 TXT 记录完成 DNS 所有权验证。自定义域名的标准过程如下：

1. 输入完整主机名，不包含协议和路径。
2. 保存声明，复制页面生成的 DNS TXT 名称和值。
3. 在域名服务商添加记录，等待 DNS 生效。
4. 回到平台执行验证。
5. 验证通过后绑定证书、HTTP 策略和路由。
6. 查看“收敛状态”和“自检”，确认各网关节点已经加载配置。

域名声明成功不等于业务路由可用。只有证书、HTTP 策略、路由绑定和数据面收敛全部完成，才能进行外部 DNS 切换。

将域名绑定到业务路由

域名验证通过后，还需要把它绑定到实际业务路由，才能把指定主机名和路径的请求交给目标服务。

1. 打开已验证的域名详情，进入“路由绑定”。
2. 选择目标路由并核对主机名、路径条件、优先级和目标模型或服务。
3. 保存后查看收敛状态，等待目标网关节点加载最新配置。
4. 先使用平台提供的探测或受控 DNS 解析验证 HTTPS、主机名和路径，再修改正式 DNS。
5. 从外部客户端发送最小请求，在网关调用中核对域名、路由、状态和目标服务。
6. 出现错误路由时，检查是否存在更高优先级的主机名或路径规则；解绑前确认没有其他客户端仍使用该域名。

27.3 上传并绑定 TLS 证书

在“证书”标签选择“上传证书”。当前上传格式为 PEM：证书链按叶子证书在前的顺序提供，并同时提供匹配的私钥。页面不支持上传 PFX。

平台会校验证书和私钥是否匹配、证书有效期、证书链和域名匹配关系。私钥加密保存，页面不会再次返回原文。

1. 填写能够区分用途和环境的证书名称，粘贴完整证书链和匹配私钥。
2. 提交前确认叶子证书在前、中间证书顺序正确，且证书覆盖目标主机名。
3. 上传后查看校验结果、有效期和指纹；校验失败时不要继续绑定。
4. 在域名详情中选择该证书，等待节点收敛后执行 HTTPS 探测。
5. 从外部客户端检查证书链、SNI 主机名和到期时间，并在网关调用中确认请求进入目标域名。

批量绑定证书并设置 HTTP 策略

同一证书覆盖多个已验证域名时，可使用批量绑定；泛域名证书也要逐项确认实际覆盖范围。

1. 在证书详情选择批量绑定，勾选证书能够覆盖且已经完成所有权验证的域名。
2. 查看影响预览，排除证书主机名不匹配、已有更合适证书或仍处于验证中的域名。
3. 提交后逐项检查绑定状态和节点收敛，不能只依据批量操作成功提示。
4. 在各域名的“HTTP 策略”中选择允许 HTTP、重定向到 HTTPS 或拒绝 HTTP。
5. 如有可信网络例外，使用最小 CIDR 范围并记录用途与复核日期，不使用宽泛网段绕过 HTTPS。
6. 分别用 HTTP 和 HTTPS 请求验证重定向、拒绝或放行结果，并核对证书与目标域名一致。

保护、授权并校验证书

私钥属于敏感配置，仅应由具备域名和证书管理权限的管理员上传或更换；页面保存后不会再次返回私钥原文。

1. 上传前确认操作者权限、证书来源、授权域名和计划启用时间。
2. 使用页面校验结果核对证书链、私钥匹配、域名覆盖和有效期，再执行绑定。
3. 轮换证书时先上传新证书并完成探测，再把域名切换到新证书；保留旧证书直至确认全部节点收敛。
4. 在操作审计中核对上传、绑定、解绑和删除记录，并设置到期前的提醒或告警。
5. 删除证书、解除绑定或销毁私钥前检查全部关联域名；私钥销毁不可逆，应先准备替代证书并完成影响检查。
6. 发现私钥疑似泄露时立即换发证书、切换绑定并撤销旧证书，不在工单或聊天中传播私钥内容。

已通过 DNS 所有权验证的自定义域名，可在域名详情中选择“申请自动证书”，等待状态从签发中变为就绪，再执行 HTTPS 探测。若页面不显示该操作，请由部署管理员确认自动证书服务，或使用“上传证书”提交 PEM 证书。

27.4 处理证书和域名故障

现象	检查顺序
TXT 验证失败	记录名称和值、DNS 区域、TTL、是否被代理服务改写
上传证书失败	PEM 边界、叶子优先顺序、私钥匹配、有效期和加密私钥格式
HTTPS 证书不对	SNI 域名、证书绑定、节点收敛、DNS 是否仍指向旧网关
HTTP 未重定向或未拒绝	HTTP 策略、CIDR 例外、代理层是否提前终止连接
域名可连但返回错误路由	路由绑定、主机名条件、路径条件和规则优先级
页面显示已保存，部分节点仍失败	收敛状态、节点自检、证书投递和网关版本

27.5 安装和绑定插件

从“管理控制台 → 插件市场”进入。页面包含“插件目录”和“绑定矩阵”，可在“插件目录”中按安装状态筛选。

1. 在“插件目录”查看插件身份、版本、来源、权限声明和可见范围。
2. 安装前检查网络目标、密钥、PII 和安全要求。
3. 安装后在“插件目录”筛选已安装插件，进入详情完成配置，并确认状态从“等待发布”“等待网关加载”进入“已生效”。

4. 在“绑定矩阵”选择全部流量、模型、路由或消费者范围，配置优先级和启用状态。
5. 根据插件类型设置并发限制、单次请求 Token 上限、请求屏蔽或防重放等能力。
6. 用命中和不命中的请求验证，再到“运行时”或网关调用详情核对插件阶段、结果、原因和耗时。

维护自定义插件时继续执行以下操作：

1. 使用名称、来源或安装状态筛选目标插件，打开详情核对版本、权限声明和当前绑定。
2. 上传新插件包时查看合法性校验结果；校验未通过时不要继续安装或启用。
3. 更新版本前保留当前版本和配置，检查新版本的权限、网络目标和配置字段变化。
4. 启用后等待网关加载完成，再使用受控请求验证；停用后确认新请求不再执行该插件。
5. 删除前解除模型、路由、消费者和全局绑定，并确认没有正在执行的请求。
6. 操作完成后在插件列表、绑定矩阵、网关调用和操作审计中核对同一版本和状态。

网关插件执行阶段包括模型改写、请求准入、路由感知准入、请求改写和响应改写。这些阶段指标用于解释网关插件的请求处理结果，不与 Skill 或其他能力的运行统计混用。

插件版本切换、停用和卸载会展示绑定、配置、凭据和路由影响。确认后才发布到网关；并发状态变化可能阻止操作。上传的插件包必须通过格式、权限和供应链安全校验，不能把任意压缩包直接作为插件使用。

27.6 配置插件安全项

安全项	用途	注意事项
网络凭据绑定	允许插件访问声明的外部目标	密钥由平台注入，不暴露给插件
请求屏蔽规则	按路径或请求头拒绝请求	最多 200 条；至少设置一个匹配条件
Token 上限	限制输入或钳制输出 Token	范围 1—10,000,000；至少设置一项
并发限制	限制插件并行执行	与网关流量保护分别计数
防重放	拒绝重复或过期请求	需与调用方时间戳和幂等策略联调

插件的投递统计用于观察运行状态；模型用量和网关调用用于核对消耗，操作审计用于追溯配置变更。

27.7 获取开放 API 文档

开放 API 文档用于查看路径、参数和响应结构。请先由部署管理员启用 OpenAPI 文档入口；能打开文档不代表当前身份有权调用其中所有接口。

调用接入流程为：

1. 从页面下载与目标平台同版本的 OpenAPI 规范，生成或更新客户端。
2. 依据接口所属身份类型配置平台 Session、API Key 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. 到网关调用、运行观测或操作审计页面核对调用方、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。

27.8 集成上线检查

- [ ] 真实业务调用 URL 经过预期的数据面和路由策略。
- [ ] API Key 或服务 Token 已进入 Secret 管理器，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。
- [ ] 自定义域名的 DNS、证书、HTTP 策略、路由和节点收敛全部通过。

- [ ] 自动证书已签发就绪，或已上传并绑定有效的 PEM 证书。
- [ ] 插件权限、绑定范围、执行阶段和失败策略已经确认。
- [ ] OpenAPI 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、插件和终局状态。

28. 可观测数据接入与权限边界

可观测与质量能力在 Zentrix 和 Zenova 中按权限提供。工作空间成员从侧边栏的“可观测”和“质量”查看本工作空间数据；管理员从“管理控制台”查看租户级运行状态、人工评审、数据集、评估器和实验。

28.1 选择入口

角色	入口	数据范围
工作空间成员	“可观测”“质量”“提示词”	当前工作空间
具备审计域权限的质量评审人员	“人工评审”“数据集”“评估器”“实验”	当前租户的质量资产
运行管理员	“运行时”	租户级调用、会话、执行、问题和终端用户
合规或审计管理员	“归档检索”“导出任务”	归档检索、留存和导出

“工作空间”和“管理控制台”显示的数据范围受账号权限和作用域限制。Trace 是一次完整执行的调用链，Span 是其中一个模型、工具或检索步骤；OTLP 是外部应用向平台发送这些观测数据的协议。

28.2 接入外部应用

经平台模型网关和 Waner 运行时执行的调用会自动产生结构化观测事件。外部应用使用 OpenTelemetry OTLP/HTTP 接入：

1. 在目标工作空间进入“可观测 → 接入”。
2. 创建采集密钥。Secret 只显示一次，应立即保存到外部应用的 Secret 管理器。
3. 从“Env”“Python”“Node”或“curl”标签复制配置片段。
4. 核对数据接收地址指向当前平台，并使用页面要求的 OTLP/HTTP 协议。
5. 从外部应用发送一条最小 Trace。
6. 页面显示已接收事件后，选择“去看执行 →”，检查第一条数据。

页面在打开期间会轮询首条数据。它一直显示等待时，先检查应用是否实际发送、Endpoint 是否可达、认证 Header、时间戳和代理限制，不要反复创建密钥。

28.3 轮换和吊销采集密钥

采集密钥按工作空间隔离。轮换时：

1. 在同一工作空间创建新密钥，并立即保存新 Secret。
2. 更新一个采集端，发送最小 Trace 并确认新数据到达。
3. 逐步更新其余采集端，检查服务名、环境和时间戳没有变化。
4. 确认旧密钥不再接收数据后执行吊销。
5. 使用旧密钥发送测试数据，确认请求被拒绝。
6. 在操作审计中核对创建与吊销记录，并记录既有连接的失效时延。

密钥原文丢失时平台不会补显；页面会要求创建新密钥。删除或吊销密钥不会删除已采集的 Trace。

28.4 采样与数据完整性

Trace 数量可能是采样结果。进行成功率、成本和错误率分析前，必须查看采样覆盖率；不能把 10% 采样的数量直接解释为全量。

对外部接入至少验证两种 OTLP/HTTP 编码、较大数据体、断线重连和时间戳偏移。平台自动采集的数据还应能将模型调用、工具批次和子智能体关联到同一次执行。

28.5 数据安全

观测数据本身可能包含用户输入、模型输出、工具参数、Header 和治理结果。接入前应约定：

- 哪些输入输出允许保留原文；
- PII、Secret 和内部地址如何脱敏；
- 工作空间、管理员和审计员分别能查看哪些字段；
- 分享、导出和冷归档的保留期限；
- 谁可以创建或吊销采集密钥。

“已脱敏”必须由实际 Trace 样本证明，不能只根据配置开关状态判断。

29. 执行、问题与根因定位

工作空间的“可观测”提供“执行”“会话”和“终端用户”；管理控制台的“运行时”还提供“调用”和“问题”。统一时间范围与筛选条件会驱动当前标签的数据查询。

29.1 查询执行

1. 选择最近 1 小时、6 小时、24 小时、7 天或自定义时间范围。
2. 使用筛选器按服务、环境、状态、终端用户或标签缩小范围。
3. 在“执行”列表观察名称、状态、模型、耗时和时间。
4. 打开一条执行进入详情。
5. 使用 Span 树和瀑布图定位模型、工具、检索、排队或治理阶段的耗时。

筛选为空与查询失败是两种状态。页面出现“数据源暂不可用”或“查询失败”时，不应把结果解释为该时间窗没有调用。

29.2 定位单次调用并阅读 Trace

从“网关调用”或“运行时 → 调用”定位单次请求时，先固定请求身份和时间，再进入关联 Trace，避免把同一秒内的其他调用当成目标请求。

1. 按时间和状态缩小调用列表，在搜索框中输入请求标识；模型或工作空间可通过概览图表钻取。
2. 打开调用详情，核对调用方、模型、上游状态、总耗时、首响时间、Token 和规则命中。
3. 使用页面提供的关联入口打开 Trace，确认请求标识与调用详情一致。
4. 对照瀑布图区分排队、路由与治理、模型、工具和响应处理耗时。
5. 如果 Trace 缺失，检查采集状态、时间范围和留存策略；不要把“没有链路记录”解释为调用未发生。
6. 记录最慢阶段、错误步骤和关联标识，再到对应模型、工具、路由或护栏页面处理。

查询 Trace 与 Span 并逐层下钻

Trace 表示一次完整执行，Span 表示其中一个模型、工具、检索、治理或其他步骤。详情页提供以下证据：

视图	回答的问题
Span 树	哪些步骤包含或调用了其他步骤
瀑布图与缩略轴	时间花在哪一段，哪些调用并行或重叠
Trace 内搜索	哪个 Span 含特定模型、工具、错误或文本
输入输出与 JSON	该步骤实际收到和返回了什么
结构化日志	与 Trace 关联的日志和错误上下文
标注	排障结论和待办由谁记录

1. 从执行列表打开目标 Trace，确认服务、环境、状态、开始时间和总耗时。
2. 在 Span 树中展开错误、慢调用和工具步骤，并在瀑布图中核对先后与并行关系。
3. 使用 Trace 内搜索定位模型、工具、错误码或关键词。
4. 打开输入输出、JSON 和结构化日志，核对实际参数与返回；敏感字段只查看脱敏结果。
5. 将确认的原因和待办写入标注；需要协作时再创建受控分享链接。

页面显示的输入输出仍受脱敏和留存策略约束。“原文已销毁”表示证据已按保留期删除，不是加载错误。

回放并评估智能体执行

智能体执行回放用于复核模型决策、工具调用和审批顺序，不会重新执行原任务，也不会再次写入外部系统。

1. 从智能体会话、任务记录或执行列表打开目标 Trace，确认智能体、会话、任务和时间一致。
2. 选择“执行回放”，按顺序查看用户输入、模型步骤、工具参数、审批结果、工具回执和最终答复。
3. 对异常步骤展开对应 Span，核对输入输出、耗时、重试和终局原因。
4. 需要评估质量时，把该执行加入评审或数据集，选择适用的评分项并填写结论。
5. 修正 Prompt、模型、工具或权限后，使用脱敏后的相同样例重新运行，在两个 Trace 之间比较结果与耗时。
6. 把差异、处置结果和后续负责人写入标注，避免仅凭最终答复判断问题已经解决。

29.3 对比、实时流和批量处理

需要判断“相同输入为什么本次不同”时，选择两条执行进行对比，重点核对模型、Prompt 版本、工具结果、路由、治理动作和各阶段耗时。

实时流用于观察新到数据。断线重连后应确认缺口是否补齐；不能仅凭页面继续滚动就认定没有丢失。

常用筛选和列配置可以保存为表格视图。批量动作必须先确认当前筛选范围和总条数，完成后在批量动作历史中核对结果。

29.4 处理问题和回归

管理员在“管理控制台 → 运行时 → 问题”查看按错误指纹聚类的问题。每条问题应按以下过程处理：

1. 查看发生次数、首次/最近时间、代表性执行和影响范围。
2. 打开代表性执行，确认根因而非只看错误摘要。
3. 检查故障切换、上游变更、回归、无进展循环和采样覆盖面板。
4. 完成修复或风险接受后，填写处置结论并标记已处置。
5. 在后续时间窗验证错误不再出现；必要时保留回归告警。

管理员问题页提供“无进展循环”面板。需要使用真实任务和已知循环样本，确认面板能够识别目标智能体的循环形态。

追踪业务过程中的治理动作

业务过程出现被拦截、脱敏、降级或改路由时，应同时查看业务步骤和治理步骤，判断策略是否在正确入口执行。

1. 从业务操作或会话定位目标执行，记录时间、用户、工作空间和关联标识。
2. 在 Trace 中筛选治理相关 Span，核对 PII、内容风险、路由、插件或配额动作的先后顺序。
3. 打开相应护栏审计或网关调用详情，确认命中的规则、入口、数据等级、动作和原因。
4. 将治理动作前后的业务状态对齐，区分“策略按预期阻止”和“业务步骤自身失败”。
5. 调整规则后使用正例、反例和边界样例回归，确认目标请求改变、非目标请求保持不变。
6. 在操作审计和执行标注中记录变更、验证结果和负责人，便于后续复盘同类事件。

29.5 可疑 Span 与重复调用

系统会提示异常慢、异常大或同参数重复调用的 Span。提示是线索，不是根因裁决。重复调用可能来自合法重试、并行分片或幂等恢复；确认前应对照父子关系、重试原因和工具回执。

29.6 分享执行

只读分享链接适用于把单条 Trace 交给外部协作方。创建前检查预览内容，确认没有 PII、客户名称、内部 URL、业务数据和 Secret。分享后记录接收方、用途和截止时间；问题结束后撤销链接。

分享前应确认链接的过期、撤销和脱敏设置。需要交互式用户会话的分享链接，应由用户账号创建。

29.7 根因定位顺序

现象	建议顺序
首响慢	排队 → 路由与插件 → 模型首 Token → 工具前置步骤
总耗时高	瀑布图最长路径 → 串行工具 → 重试 → 子智能体
输出不稳定	Prompt/模型版本 → 输入变量 → 检索结果 → 工具返回 → 治理变换
成本突增	Token 分布 → 长输入/输出 → 重复调用 → 缓存 → 模型价格
工具反复调用	终局条件 → 工具错误 → 无进展循环 → 审批等待
只在部分用户失败	终端用户 → 工作空间 → 权限/分发 → 凭据 → 数据等级

30. 用量、成本与自定义看板

成本分析需要把模型调用、Token、缓存、工作空间、用户和终局状态放在同一口径中。管理员使用“模型用量”查看排名和趋势，使用“运行时 → 调用”查看单次调用；工作空间成员从“用量概览”查看本工作空间数据。

30.1 分析 Token 和调用量

从“管理控制台 → 模型用量”核对 Token 消耗及其归属。模型用量页用于排名、趋势和归属分析；单次请求明细需要通过“查看调用明细”进入网关调用页。

1. 选定统计周期，记录时区和是否包含当前尚未结束的周期。
2. 分别查看输入、输出及缓存相关 Token，不只依据总 Token 判断消耗来源。
3. 按模型、供应商、工作空间、用户或接入凭证下钻，确认消耗归属。
4. 对异常项选择“查看调用明细”，保持相同时间范围，核对单次调用。
5. 记录请求标识或 Trace 标识，到执行详情确认长输入、长输出、重复调用或工具循环。
6. 无法归属的记录单独列出，检查调用凭据、工作空间和用户信息是否随请求写入。

核对调用量、成功率与调用主体

调用量和成功率必须与实际调用主体一起查看，避免把某个业务系统的异常误判为全平台故障。

1. 在“模型用量”保持同一时间范围，切换到调用量或成功率视图。
2. 核对总请求、成功请求、失败请求和成功率，确认页面采用的状态口径。
3. 按模型、工作空间、用户或接入凭证下钻，找出变化最明显的主体。
4. 选择异常区间进入网关调用页，按时间和状态筛选，再从图表钻取目标模型或工作空间。
5. 打开代表性调用，核对状态码、终局原因、配额或策略来源和上游结果。
6. 修正配置后使用同一主体发送最小请求，并用相同筛选条件确认成功率恢复。

分析 Token 消耗趋势

趋势分析用于判断消耗变化发生在何时、由输入还是输出驱动，以及是否只影响特定模型或业务范围。

1. 选择能够覆盖正常基线和异常时段的时间范围，并确认时区。
2. 在“Token 趋势”分别显示输入、输出和缓存相关 Token，选择页面提供的时间粒度。
3. 保持时间范围不变，依次按模型、供应商、工作空间、用户或接入凭证对比。
4. 对突增时间点查看调用量和单次 Token 分布，区分“请求变多”和“单次请求变长”。
5. 打开异常调用或 Trace，检查上下文长度、输出长度、失败重试、重复工具调用和缓存变化。
6. 调整模型、Prompt、配额或业务调用后，使用相同粒度观察后续趋势，确认变化持续而非短时波动。

不同视图没有使用相同筛选条件时，不直接比较总数；切换日、周、月或其他页面提供的粒度后，应再次确认起止时间和时区。

30.2 识别高消耗接入凭证

“消费者 Top5”在当前界面对应“接入凭证”排行，不代表自然人用户排行。从“管理控制台 → 模型用量”进入：

1. 选择统计时间范围，确认当前窗口包含需要分析的业务时段。
2. 在“Token 排行”中查看“接入凭证”列；列表按当前窗口 Token 用量从高到低排列。
3. 选择一条接入凭证，页面会把其凭证 ID 写入筛选条件，并联动 Token 趋势、成本趋势、分布和其他排行数据；进入可靠性分析时会继续带上该筛选条件。

4. 将筛选后的 Token、请求数和成本与该凭证对应的业务系统、模型服务及预算核对。
5. 需要查看单次请求时，进入网关调用页，并使用相同时间范围和凭证定位调用明细。

“接入凭证”使用的是网关模型 API Key 记录的 ID，不是用户 ID。排行为空时，先核对时间范围内是否存在携带该凭证归属的用量记录；“未记录”不能直接归因给某个用户或部门。

30.3 成本预算和尖刺

设置预算时同时记录预算周期、币种、模型价格版本、告警阈值和负责人。成本尖刺发生后按以下顺序分析：

1. 确认价格表或供应商映射是否变化。
2. 查看调用量与单次 Token 分布，区分“调用变多”和“单次变长”。
3. 检查模型、工作空间、终端用户和 API Key 归属。
4. 对照缓存命中和净节省，确认缓存统计没有被当成新增消耗。
5. 打开代表性 Trace，检查重复调用、工具循环、长上下文和失败重试。

模型用量用于观察平台内的消耗趋势和异常来源。涉及费用结算时，以供应商或企业计费系统的账单为准。

30.4 分析可靠性与耗时

错误预算用于表达可接受的失败额度。分析时要同时看总样本数、时间窗和状态口径；低样本下的百分比不应触发重大决策。

耗时归因把总耗时拆到模型、工具、检索、排队和治理阶段。故障切换观测可以显示切换事件及代价，但不改变第 25 章的路由边界：是否实际切换由请求经过的数据面和已发布规则决定。

1. 选择业务异常发生的时间范围，记录请求量、成功率、吞吐量和样本数。
2. 按模型或供应商缩小范围，比较同一维度的正常时段与异常时段。
3. 查看请求吞吐、首响和总耗时，判断变化来自请求规模还是单次处理变慢。
4. 打开耗时分解，确认模型、工具、检索、排队或治理阶段的主要变化。
5. 从异常区间打开代表性调用或 Trace，核对状态码、终局原因和各阶段耗时。
6. 调整模型、配额或路由后，使用相同筛选条件验证吞吐、成功率和耗时是否恢复。

分析接入凭证的耗时和调用明细

需要判断某个业务系统是否独立变慢时，以接入凭证作为调用主体分析，并与平台整体和其他凭证对比。

1. 在模型用量页选择目标时间范围，从接入凭证排行或筛选中选择目标凭证。
2. 查看该凭证的请求量、成功率、Token、首响和总耗时，并与前一时间窗比较。
3. 保持凭证条件进入可靠性分析，判断异常集中在哪个模型、状态或时间段。
4. 选择“查看调用明细”，在网关调用中打开慢请求和失败请求，核对排队、路由、上游和插件耗时。
5. 对代表性请求进入 Trace，确认是否存在长上下文、重复工具调用或上游重试。
6. 调整调用方式、模型或配额后，用同一凭证发送最小请求，并确认明细和聚合指标均恢复。

存在故障切换记录时，应核对实际路由规则、目标模型和切换代价，不把观测事件当成自动切换承诺。

30.5 切换监控视图和统计维度

同一问题可从 Token、调用、成本、可靠性和执行等视图观察。切换视图时应固定时间范围和主要筛选，确保比较的是同一批数据。

1. 先确定问题和统计周期，例如“某工作空间昨日失败率升高”。
2. 从总览切换到 Token、调用量、可靠性或成本视图，并保持时间、工作空间和模型条件一致。

3. 在页面支持的范围内切换模型、供应商、工作空间、用户、接入凭证或状态维度。
4. 核对每个视图的单位、聚合方式、时区和样本数；口径不同的数字不直接相减。
5. 从异常图表钻取到调用或 Trace，验证聚合趋势能由实际明细解释。
6. 保存分析条件或记录关键筛选，便于后续用相同口径复查。

创建仪表盘并执行自定义查询

工作空间的“用量概览”提供看板。在提供“加入看板”的分析页面，也可把当前分析结果保存成组件。

1. 选择“新建看板”或从现有看板“克隆并编辑”。
2. 选择“添加组件”，从页面提供的数据源中选择执行、网关调用、成本或评分。
3. 配置指标、聚合、维度、时间粒度和筛选，并查看实时预览。
4. 处理“不支持的指标”“不支持的维度”“数据源不可用”或“查询失败”，不保存无结果组件。
5. 加入看板并在编辑模式排布网格，保存后切换到浏览模式验证。
6. 使用与组件一致的时间和筛选进入明细页，确认汇总数能够与实际记录对应。

查询构建器只使用已接入的数据源和页面支持的维度。分析页的部分筛选条件可能无法带入看板组件；保存前应检查页面列出的未带入条件，并在组件中重新配置。

30.6 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“模型用量”确认变化集中在哪些模型、工作空间或接入凭证。
2. 进入调用页核对状态、耗时、排队、重试和代表性请求。
3. 对照模型服务配额、API Key 配额及流量保护，确认实际返回由哪一层触发。
4. 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。
5. 使用正常业务请求验证恢复情况；持续异常时，再检查模型健康和智能体运行状态。

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。

30.7 导入、导出和预置组件

看板 JSON 可用于环境迁移。导入前检查数据源、模型、工作空间和评分维度 ID 在目标环境是否存在。导入成功只证明结构可解析，仍需逐组件核对数据。

平台预置成本和效能组件用于开箱查看，不建议直接改写；需要调整时使用克隆。预置组件的统计口径仍受采样、价格和数据完整性影响。

31. 评分、人工评审与自动评估器

评分把人工判断和机器裁决写入统一结构，供质量趋势、实验和告警使用。负责质量评审和评估器维护的人员需要相应审计权限：读取评分维度需要审计域查看权限，管理评审队列需要审计域管理权限。

31.1 使用现有评分维度

管理页面会列出可用的布尔、数值、分类或文本评分维度。创建评审队列或评估器前，选择与当前业务目标匹配的评分维度，并核对业务含义、取值范围、分类选项和边界样例。

31.2 创建人工评审队列

从“管理控制台 → 人工评审”进入。

1. 选择新建队列并填写名称。
2. 选择入队方式：“抽样”“护栏命中”或“手工”。
3. 抽样队列设置周或月周期、抽样比例和每周期上限。
4. 选择一个或多个评分维度。
5. 保存后检查队列待评数量和健康状态。
6. 打开队列进入评审工作台。

护栏命中队列用于复核拦截是否正确，复核结果会进入准确率和误报率统计。手工队列适合专项回归或事件调查。

31.3 完成人工评审

评审工作台会给当前评审员锁定一条样本，避免多人重复领取。评审员应核对输入、输出和允许查看的上下文，再逐维度评分并提交。

评审员断线或离开时，锁会在超时后回收。启用队列前，用两个评审员账号验证并发领取、锁超时、重复提交和访问权限。

31.4 创建自动评估器

从“管理控制台 → 评估器”进入。

1. 从模板库选择模板，或复制模板后修改评估准则。
2. 创建评估器实例，把变量绑定到 Trace 的输入、输出或其他数据路径。
3. 选择与人工相同的评分维度。
4. 使用代表性样本“试跑”，检查输出能否解析为预期分数。
5. 设定线上采样或数据集运行方式，再启用实例。
6. 在执行历史检查成功、失败、解析错误和分数回写。

评估异步执行，不阻塞被评估的业务链路。队列积压或评估失败不应改变原业务请求的结果，但会造成质量数据延迟或缺口。

31.5 校准人机一致率

对同一批样本同时采集人工和机器分数，查看人机一致率。低于项目阈值时：

1. 检查评分维度是否含糊；
2. 核对评估器变量绑定是否取错 Span；
3. 补充边界样例和输出 Schema；
4. 调整评估 Prompt 后重新试跑；
5. 达到最小样本量后再比较。

评估器结果应与人工评审和数据集回归结果一起解读。

31.6 从评审结果到回归验证

```
线上执行或护栏命中
  → 确定性抽样/手工入队
  → 人工评分
  → 自动评估器评分
  → 一致率与误报率
  → 调整 Prompt、模型、护栏或评估器
  → 数据集回归实验
  → 质量告警
```

1. 从人工评审或护栏准确率页面选出需要改进的样本，并加入数据集。
2. 修改 Prompt、模型、护栏或评估器时记录变更内容，每轮只改变计划比较的因素。
3. 使用同一数据集运行基线和候选实验，比较共同完成样本。
4. 检查质量分数、误报率、失败条目、成本和耗时，不只比较平均分。
5. 候选结果达到组织要求后发布相应配置，并保留基线版本用于回退。
6. 在后续线上时间窗观察评分、护栏命中和告警，确认改动没有引入新的退化。

32. 数据集、实验与 Prompt 版本

数据集用于保存可重复执行的评测样本，实验用同一数据集比较 Prompt 版本和模型。Prompt 库负责版本、标签、差异和回滚。

32.1 创建数据集

管理员从“管理控制台 → 数据集”进入；工作空间成员从“质量 → 数据集”进入。

1. 选择“新建数据集”，填写名称和描述。
2. 手工添加条目，或从执行详情选择“加入数据集”。
3. 每个条目填写输入和期望输出；实际历史输出只用于对照。
4. 核对来源执行是否仍存在，修复悬空引用。
5. 运行实验前处理“待补期望输出”。没有判定基准的条目不会形成可靠结论。

删除已有实验运行的数据集会破坏历史，应使用“归档”。归档后仍需确保历史实验可以读取。

32.2 导入 CSV

选择“导入 CSV”，上传 UTF-8 CSV，映射输入、期望输出等列并查看预览。当前上限为 10 MB、10,000 行。预览会逐行说明输入为空、找不到列、期望输出 JSON 非法或超过行数等问题。

正式导入前先修复错误行；如果选择“只导入可用条目”，应下载或保存失败行清单，避免样本静默缺失。

32.3 管理 Prompt 版本

工作空间成员从“提示词”进入，管理员从“管理控制台 → 提示词”进入。

1. 选择“新建提示词”，填写名称、分组和内容。
2. 使用 `[variable]` 表达变量，并检查自动识别结果。
3. 首个版本建立后，后续修改使用“新建版本”，版本号不会回退。

4. 使用版本差异查看前后变化。
5. 验证完成后选择“标为生产版本”。同一 Prompt 同时只有一个生产版本。
6. 需要回滚时，选择“以此内容创建新版本”；系统不会把历史版本原地改成最新版本。

生产标签用于指定正式使用的稳定版本。发布后使用目标智能体发起真实调用，确认新版本已生效。

32.4 使用试验台

从工作空间或管理控制台的“提示词 → 试验台”进入。

1. 载入 Prompt 和版本，填写所有必填变量。
2. 选择模型，设置温度和最大 Token。
3. 单窗口运行并检查输出、耗时、Token 和成本。
4. 添加多个窗口，用相同输入比较模型或 Prompt 版本。
5. 将有效结果“存为新版本”，或“加入数据集”作为回归样本。

试验台保存新版本不会自动把它标记为生产版本。多窗口比较应保持共享变量一致，只改变计划比较的因素。

32.5 运行实验

管理员从“管理控制台 → 实验”进入；工作空间成员从“质量 → 实验”进入。

1. 选择数据集并检查条目数和待补期望输出。
2. 新建运行，填写运行名、Prompt 版本和模型。
3. 对裸字符串条目明确映射到哪个 Prompt 变量。
4. 查看运行成本预估和输入形状探测结果。
5. 选择“开跑”，观察待开始、运行中、完成、部分、失败或已取消状态。
6. 对单条失败使用“重跑”，不要为了少量失败重复整个运行。
7. 选择两个运行，以一个为基线比较质量、成本和延迟。

成本预估不是账单。实验完成后记录预估与实际消耗偏差，并调查价格、Token 和失败重试造成的差异。

32.6 判断实验结果

实验只对共同完成且可评分的条目比较。部分失败时，平均分不包含失败条目；应先检查失败是否在候选版本集中发生，否则平均分可能被幸存者偏差抬高。

发布候选版本前至少检查：

- 总样本和共同样本量；
- 各评分维度的变化和置信程度；
- 失败、部分和重试条目；
- 成本和响应时延；
- 护栏、工具和数据访问行为；
- 关键业务分组是否发生退化。

33. 告警、通知、留存、归档与导出

告警把运行、成本、质量和护栏事件转为可处置任务。留存和归档决定证据保留多久、何时进入冷存储以及如何申请调取。

33.1 创建通知渠道

从“管理控制台 → 告警规则”进入。该页同时面向平台运营管理员和租户管理员，实际可见配置由账号权限决定。

1. 选择“新建渠道”，填写名称、类型和 Webhook URL。
2. 根据接收端选择钉钉、飞书或 Webhook；企业微信选择 Webhook。
3. 钉钉按需填写加签密钥。
4. 保存后发送测试消息。
5. 在接收端确认收到，再回到平台选择“确认收到”。只有已验证渠道承担真实告警。

测试成功不等于真实投递长期健康。“最近投递”显示测试或真实告警中最近一次投递的时间和状态。渠道上线前还应触发一条真实规则，并在接收端和告警事件中核对。

33.2 创建告警规则

选择“添加规则”后：

1. 选择流量、错误率、耗时、配额或成本等规则类型，并确认当前账号有权查看对应范围。
2. 配置指标、比较方式、阈值、持续时间、严重度和最小样本量。
3. 选择模型、工作空间、接入凭证或其他页面提供的适用范围，避免一条规则覆盖无关业务。
4. 选择已经验证的通知渠道，并设置需要接收的人员或值班范围。
5. 保存后检查规则状态，使用受控事件触发一次真实告警。
6. 在告警事件和接收端核对规则、时间、证据链接和投递状态；处理后填写处置结果。

低样本规则应设置最小样本量。状态“样本不足”“无数据”“数据过期”和“数据源不可用”不能解释为健康。

配置质量告警与通知规则

质量告警用于在评分、评审或实验结果持续低于要求时通知负责人；护栏命中通知则按风险分类和入口配置。

1. 从质量分析或评分页面选择“为这个维度建告警”，或在告警规则中选择质量规则。
2. 选择评分项、评估器或实验条件，配置阈值、持续时间、最小样本量和严重度。
3. 为护栏命中通知选择风险分类和入口；检测阈值与分类启用仍在“护栏”中心维护。
4. 选择已经验证的通知渠道和接收范围，保存后检查规则是否启用。
5. 使用不含真实敏感信息的受控样例触发一次，确认告警事件、证据链接和接收端消息一致。
6. 调整评分、评估器或护栏策略后重新验证，避免规则仍引用已经更换的维度或范围。

护栏的检测阈值和分类启用在“护栏”中心配置；“告警规则”负责命中台账、通知规则和投递。

33.3 处置告警事件

操作	含义
确认告警	记录运维已看到；源信号可能仍在触发
静默	暂停通知噪声；源信号和事件仍保留
解除	确认事故已经处理或源状态已经理解
删除规则	停止后续评估并破坏历史理解；仅在确需移除时使用

处理时先“查看证据”，打开关联事实页或 Trace，再进入路由、配额、模型、护栏等治理入口。解除时填写原因和实际处置，例如扩容、调整配额、停用模型或更新流量策略。

33.4 投递故障

状态或错误	处理方式
未验证	发送测试并在接收端确认
DNS/TLS/连接失败	检查 URL、证书、网络和代理
接收端 4xx	检查认证、格式、路径和权限
接收端 5xx/限流	查看重试决策和接收端容量
目标不安全	使用 HTTPS 公网或经批准的内部通知出口
配置异常	检查必填项、Bearer Token 和平台默认/规则专属路由

站内告警始终保留；外部通知失败不应删除事件台账。

33.5 配置留存和导出

从“管理控制台 → 导出任务”进入。页面同时提供可观测数据留存设置。管理员可以创建计划导出任务并立即运行；任务行显示最近一次运行的状态、行数和水位。需要追溯更早运行时，应结合目标存储、任务日志和操作审计。

留存设置应按数据等级、审计要求、存储容量和检索成本确定。缩短留存期会造成到期原文不可恢复；变更前应评估未结事件、人工评审、实验引用和监管要求。

1. 选择需要保留或导出的数据类型、租户范围和时间范围。
2. 配置保留周期或导出计划，并选择页面提供的目标存储。
3. 保存后执行一次导出，观察等待、运行、完成或失败状态。
4. 下载或检查目标存储中的结果，核对行数、时间范围、租户范围和完整性信息。
5. 重复执行一次，确认不会产生不可解释的重复数据。
6. 到操作审计核对留存变更、任务创建和下载记录。

任务行只显示最近一次运行信息时，应使用目标存储、任务日志和操作审计追溯更早执行。

33.6 冷归档检索

从“管理控制台 → 归档检索”进入。

1. 填写时间范围和关联工单号，提交检索任务。
2. 任务依次进入等待、运行、可下载、失败或已过期。
3. 就绪后在有效期内下载检索包。
4. 核对租户索引、完整性校验结果、文件清单和时间范围。
5. 使用完成后按制度销毁本地副本。

检索包到期后停止提供并清理。归档检索失败时保留任务和错误证据，不通过扩大时间范围反复盲目提交。

34. 平台运行与日常维护

平台管理员可以在界面中维护 Zentrix 与 Zenova 的运行状态；Zenova 还提供 Waner 与智能体运行管理。升级、恢复或日志清理前，应按组织流程确认影响范围并保留操作记录。

34.1 管理智能体运行时

从“管理控制台 → Agent 运营”进入。列表显示智能体、工作空间、节点、运行时版本、服务状态和升级状态，并提供以下动作：

动作	用途	执行前检查
启动	让已配置智能体提供服务	模型、工作空间、能力分发和运行版本可用
停止	停止接收新任务	活跃会话、待审批任务和已执行的外部操作
重新下发配置	应用模型、MCP、Skill、渠道或审批策略变更	目标配置、授权范围和凭据
重启	重新启动智能体服务	操作时段、会话恢复和配置状态
查看生命周期事件	定位启动、配置、健康和状态转换问题	事件时间、节点和配置版本

标准处理过程如下：

1. 按工作空间、版本或名称定位智能体。
2. 查看“服务状态”“运行时”和“最近错误”。
3. 先查看生命周期事件，再决定重新下发、重启或停止。
4. 执行动作后刷新列表，等待状态稳定。
5. 进入对话发送最小请求，再到“管理控制台 → 运行时”核对执行和日志。

按钮成功提示表示指令已经下发，不表示智能体已恢复服务。务必等“运行时就绪”、配置同步和真实请求三项证据均符合预期。

34.2 理解预热、重配和停止

智能体运行时收到平台分配后，会绑定组织、工作空间、智能体和配置版本。重新配置时，运行中的任务可能仍使用原配置，新任务使用新配置；修改后应同时验证运行中的长任务、新任务和并发任务。

停止时，平台会停止接收新的任务回合，并尝试取消正在运行的模型、工具和子任务。慢流式响应、外部调用和审批等待可能需要更长时间结束；外部调用已产生的业务结果不会因运行时停止而自动撤销。

重新下发配置后，确认页面显示的配置版本与目标版本一致，并用新任务验证配置已经生效。

34.3 分批升级运行时

运行时升级会按批次推进。“Agent 运营”显示目标版本、进度、待升级、升级中、待健康检查、已完成和已跳过数量。

1. 选择低风险工作空间作为首批。
2. 设置“同时升级上限”，控制并发影响面。
3. 观察健康检查、错误率、首响和会话恢复。
4. 连续失败触发保护后，升级会自动暂停。
5. 排除原因、降低并发上限，再选择“恢复升级”。

运行时列表提示版本异常或版本落后时，先确认目标版本并等待状态稳定，再继续业务检查。

34.4 执行系统或组件更新

从“管理控制台 → 更新”进入。页面支持完整系统升级和允许的组件更新。

1. 选择“检查更新”，读取目标版本和升级前检查。
2. 在“确认系统升级”中核对当前版本、目标版本、更新通道、架构和“包校验”状态。
3. 查看“影响预览”，确认自动包含组件、影响服务和维护窗口。
4. 使用离线升级包时，先由部署管理员把升级包放到系统指定位置，再在页面中选择对应文件。
5. 创建升级前备份点，确认回滚边界。
6. 在约定时段执行更新，并根据页面依次检查预检、备份、应用、重启和健康检查状态。
7. 完成后验证登录、模型调用和管理员常用操作，再结束本次更新。

组件更新并不保证零停机。页面出现“不兼容”“包校验失败”“超出增量窗口”“需要完整系统升级”或“需人工介入”时停止操作，保留任务详情并升级支持。

回滚范围以页面说明和备份内容为准，外部依赖不一定随平台一起恢复。升级前应由部署管理员确认可恢复范围和处理方案。

34.5 配置备份计划

从“管理控制台 → 备份与恢复”进入。页面包括备份服务器、立即备份、备份计划、按备份恢复、保留策略和备份文件删除。

1. 配置本地目录或 SSH 备份服务器。
2. 使用 SSH 备份服务器时确认主机指纹；私钥或密码保存后不再回显，编辑时留空表示保留原值。
3. 对本地目录执行预检，对远端执行“测试并保存连接”。
4. 配置每天、每周或每月执行，并设置时区。
5. 设置本地与远端保留份数、天数及暂存清理。
6. 保存策略，并确认页面显示下次执行时间。
7. 选择“立即执行备份”，在“操作记录”检查状态和审计材料。

完整备份应覆盖业务数据、权限与安全配置、文件与知识资料、用量与观测数据，以及恢复所需的加密密钥。具体范围由部署管理员按当前环境确认。

34.6 恢复和演练

恢复优先在隔离环境演练：

1. 扫描备份清单，选择已验证副本。
2. 外部备份需要原备份加密密钥和页面要求的确认短语。
3. 生成恢复源，并检查环境标识和资产清单。
4. 在隔离环境恢复，并验证登录、对话、文件、知识库、数据查询、审计、告警和智能体运行。
5. 记录可恢复到的时间点、恢复耗时、缺失内容和重建时间。

覆盖现有环境会替换当前数据，只有在完整备份、近期恢复演练、业务暂停安排和恢复后验证方案均已确认时才能执行。页面未提供的恢复操作应交由部署管理员处理。

34.7 查询和清理平台日志

从“管理控制台 → 日志管理”进入。页面中的平台日志与留存区域按权限展示：租户管理员查看本组织范围，平台管理员查看整个平台的保留状态。

1. 按级别筛选“调试”“信息”“告警”“错误”或“跟踪”。
2. 打开详情查看扩展字段和关联标识。
3. 在日志保留区域检查各来源的用量和应用状态，并核对一组全局“保留天数”和“容量目标”。这两个目标不是按来源分别配置；容量是平台可清理日志的合计目标。
4. 缩短策略前查看清理预览和预计回收量。
5. “清理可立即回收日志”只处理平台可直接清理的来源。
6. 日志、指标和链路数据按照已应用的保留策略清理；在对应组件状态中核对策略是否生效。
7. 在“清理历史”核对成功、部分成功、失败和实际回收量。

审计与安全记录使用独立的留存策略，不参与普通日志清理。

34.8 运行参数边界

模型服务并发、API Key 配额、流量保护、过载阈值、日志保留和备份调度，分别在对应的管理页面设置。每次变更应记录原值、新值、生效方式、回退值和验证结果。

1. 根据需要调整的对象进入对应管理页面，不使用某一处配额代替其他层级限制。
2. 记录原值、目标值、调整原因、影响范围和计划恢复时间。
3. 保存后查看页面状态或发布状态，确认新设置已经生效。
4. 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。
5. 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。
6. 在操作审计中核对调整与恢复操作。

34.9 管理员职责交接

更换管理员或调整职责时，应在平台内完成权限和资源接管：

1. 为接任人员分配与职责相符的管理角色，并使用其账号核对可见范围。
2. 逐项确认模型服务、API Key、MCP、护栏、告警、审计、备份和运行管理的负责人。
3. 轮换由原管理员个人保管的凭据，并验证相关连接和调用仍可用。
4. 由接任人员完成一次常用管理操作、一次异常查询和一次备份状态检查。
5. 确认接管完成后，再移除原管理员不再需要的权限，并在操作审计中核对变更记录。

35. 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，并在操作审计中核对变更结果。

35.1 维护平台管理员和登录记录

Zentrix 和 Zenova 均在“管理控制台”维护管理员权限。平台管理员应按职责分配，不把日常成员账号直接授予所有管理权限。

1. 在成员管理中查找目标人员，确认账号、所属部门和在职状态无误。
2. 打开角色或管理权限设置，按模型、组织、安全、审计、运维等实际职责选择所需角色。
3. 保存后使用该管理员账号重新登录，检查管理菜单、资源范围和允许动作；同时验证无关页面仍不可访问。
4. 在登录记录中查看最近登录时间、来源、成功或失败状态，发现异常登录时及时禁用账号并调查。
5. 职责调整时先由接任人员验证权限，再移除原管理员不再需要的角色。
6. 在操作审计中核对角色授予、移除、账号禁用和恢复记录，确认操作者与时间符合变更安排。

管理员能够打开某个菜单不代表可以查看全部组织数据，最终范围仍由账号角色和资源权限共同决定。共享管理员账号会破坏审计归属，不应用于多人日常操作。

同步组织、维护部门并管理成员会话

Zentrix 和 Zenova 均从“管理控制台 → 成员与部门”维护组织；Zenova 的工作空间角色另在工作空间中设置，不随部门调整自动变化。

1. 使用姓名、用户名、邮箱、部门或职能组筛选目标成员，核对现有归属和账号状态。
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、MCP、智能体和其他能力授权。
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. 成员暂时离岗时使用禁用或锁定；确认不再使用后再执行删除或账号清理，并按需结束其现有会话。
5. 通过单点登录或 LDAP/AD 同步组织时，核对同步范围、匹配字段和停用规则，并检查新增、更新、冲突和失败项。
6. 在登录记录和操作审计中检查账号状态、最近登录、会话处理及组织变更。

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。

35.2 维护模型供应商

从“管理控制台 → 模型供应商”进入供应商列表。

1. 按名称或供应商筛选目标配置，打开详情核对基础 URL、“供应商”和可用模型。
2. 在“基础信息”中选择编辑，修改地址、API Key 或兼容性设置；保存后查看校验结果，不把草稿已保存当成已投入使用。
3. 选择“验证并获取可用模型”或“刷新模型”，确认返回的模型目录符合预期，并只启用需要向当前组织提供的模型。
4. 对供应商和代表性模型执行探测，在“可用性”中查看最近结果；连续失败时核对网络、凭据、协议兼容性和上游权限。
5. 变更需要发布时，查看发布前校验和影响提示，确认后启用新版本；存在不健康模型时先处理或明确排除范围。
6. 在“价格”中维护页面要求的计价信息，在“审计”中核对探测、编辑、发布和启停记录。
7. 停用模型或供应商后，检查引用它的模型服务和智能体；删除供应商前按页面提示处理模型、服务和运行配置引用。
8. 最后使用最小请求验证真实调用，并在模型用量或网关调用中确认供应商、模型和状态记录正确。

编辑成功只说明配置已经保存；“验证并获取可用模型”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。

35.3 维护模型服务

从“管理控制台 → 模型服务”查询和维护成员实际使用的模型能力。

1. 按名称查找服务，打开详情核对描述、模型列表、主模型、启用状态和可用范围。
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。
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、详情和真实调用。
4. 暂停服务时使用禁用；恢复后执行一条最小请求，确认状态和用量记录恢复。
5. 删除服务前处理智能体、API Key、工作空间和路由中的引用。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. 变更后检查工作空间或能力中心中的服务名称、可见范围和实际可选状态。

主模型不可用时不会因为列表中还有后续模型就自动恢复。需要切换时，应明确修改模型服务或发布相应网关路由，并再次验证合规和调用结果。

35.4 维护 MCP 服务

从“管理控制台 → MCP 服务 → MCP 列表”进入。

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、Tools、数据敏感等级和可用范围。
2. 修改 URL、实例 Header、凭据 Header 或鉴权方式后，重新执行“解析 Tools”。
3. 通过 OpenAPI、Postman Collection 或 cURL 更新 HTTP API 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（公开发布）”或移除可用范围；两者承担不同作用，应分别核对。
6. 删除服务前解除智能体、应用和其他能力引用，并确认没有仍在运行的调用。
7. 在“MCP 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、智能体绑定和凭据状态。

35.5 维护外部 Agent 服务

从“管理控制台 → 外部 Agent”进入外部 Agent 列表。该入口用于接入由其他系统提供的 Agent 服务，不等同于 Zenova 工作空间内创建的智能体。

接入和维护按以下顺序完成：

1. 选择“接入外部 Agent”，通过“名片 URL 拉取”或“粘贴 AgentCard JSON”提供来源。
2. 执行“开始测试”，确认显示“连通正常”，并核对名片名称、协议版本和签名校验结果。
3. 完成接入后打开详情；需要上游凭据时，在“接入”标签按名片声明选择 API Key、基础认证或 Bearer 令牌，并配置“企业共享”或“无需凭据”。
4. 在“分发”中设置全租户、部门、职能组或企业空间范围，再使用范围内与范围外账号验证 Agent 广场的可见性和调用结果。
5. 在“策略”中查看限流与熔断设置；页面提示由“流量保护”统一管理时，到相应入口调整并重新验证。
6. 需要把外部 Agent 提供给 MCP 客户端时，启用“发布为 MCP 工具”，再使用分发范围内账号获取配置并完成一次调用。
7. 选择“立即探活”检查当前连通性；来源为名片 URL 时可选择“重新拉取”，并检查端点、能力或鉴权声明是否变化。
8. 在“任务中心”按 Agent、状态或任务标识查询执行，展开状态历史；取消任务时确认对端是否实际进入“已取消”。

9. 需要暂停服务时选择“停用”；恢复后重新探活并执行最小任务。端点变更导致凭据暂存时，按页面提示重新审批凭据绑定。
10. 删除前检查活动任务、调用方、路由、插件及能力分发引用；确认影响后再删除，并验证原凭据已经失效。

外部 Agent 接入完成的标志是：名片可读取、探活正常、凭据与端点匹配、分发范围正确，并且任务中心能够看到真实调用。某一步成功不能替代其他步骤。

35.6 核对配置变更

对组织、模型、MCP 和外部 Agent 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. 到网关调用、运行观测或对应调用页面检查真实执行结果。
5. 到操作审计核对操作人员、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。

35.7 接入新的模型与外部能力

新增模型供应商、MCP、HTTP API 或外部 Agent 时，使用一致的接入顺序：

1. 确认能力用途、负责人、网络目标、认证方式和数据敏感等级。
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、模型发现、Tools 解析或最小任务，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. 在调用记录和操作审计中核对使用者、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或智能体配置中选错。

36. 配置并发布业务智能体

业务智能体需要同时具备明确任务、可靠资料、合适工具、最小权限和可验证输出。以下流程适用于知识问答、数据查询、日报汇总、作业记录和部门办公等常见场景。

36.1 定义任务和使用范围

1. 写明使用者、输入资料、需要完成的业务动作和输出格式。
2. 区分“提供建议”和“执行外部操作”；涉及发送、写入、删除或审批的动作单独列出。
3. 确定可以使用的部门、工作空间、知识库、数据连接和 MCP 服务。
4. 准备正常问题、边界问题、无权限问题、空结果和错误输入，用于发布前验证。
5. 指定内容维护人、数据负责人和异常处理联系人。

任务描述应让成员能够判断何时使用该智能体，也应让管理员能够判断需要授予哪些数据和工具权限。

36.2 配置知识问答智能体

1. 创建知识库，选择嵌入模型并上传有效版本的文档。
2. 等待文档状态变为“已可用”，再设置知识库和单个文档的可见范围。
3. 创建智能体，在“技能与应用”中绑定知识库。
4. 在身份指令中要求回答引用来源，并说明资料不足时如何提示。
5. 使用有明确出处、无相关资料、过期资料和无权限资料的问题进行测试。
6. 发布后使用普通成员账号检查知识库引用、“查看原文”和拒绝结果。

资料更新后应重新检查索引状态和引用。智能体回答没有引用时，不应将其作为企业制度或流程结论直接使用。

36.3 配置数据查询智能体

1. 在“数据连接”中创建只读连接，测试成功后执行部署。
2. 按业务范围配置可访问的库表、字段、行级条件和单次查询限制。
3. 将运行中的数据连接绑定到智能体，并限制普通成员修改绑定关系。
4. 在“查询工作区”使用确定口径的问题验证 SQL、结果行数、筛选、排序和导出。
5. 使用跨部门、写操作、大范围扫描和敏感字段问题验证拒绝结果。
6. 发布后在操作审计和运行观测中核对调用人员、数据连接、查询结果和终局状态。

查询结果涉及业务决策时，应同时核对数据时间、统计口径、单位和空值处理。图表或导出文件生成成功后仍需与查询结果回读一致。

36.4 配置日报、汇总和办公写作智能体

1. 准备经过确认的报告结构、字段、称谓、日期格式和数据来源。
2. 绑定需要的知识库、文件、数据连接或 MCP 服务，并限制每项能力的授权范围。
3. 在指令中要求区分事实、推断、缺失数据和待确认内容。
4. 使用代表性文件或数据生成样例，核对数字、日期、人员、引用、表格和章节。
5. 导出 Word、PDF 或表格后检查分页、目录、格式、链接和水印。
6. 需要周期性处理时，使用页面提供的任务或外部触发方式，并核对失败后的通知和重试结果。

批量处理应从小批次开始。处理完成数量、失败数量和输出位置都应可见；只显示任务完成但没有可读取产物时，不能视为成功。

36.5 配置作业记录和风险汇报智能体

1. 定义作业记录的必填项、风险等级、责任人和汇报模板。
2. 绑定允许读取的日志、文件或业务系统，写入类工具保留人工审批。
3. 在指令中要求列出完成事项、未完成项、风险、依赖和责任人。
4. 使用包含否定语句、缺失附件、多人记录和隐含风险的样例测试提取结果。
5. 高风险结果进入人工确认或通知规则，不直接触发外部处置。
6. 在审计和执行详情中核对原始输入、工具动作、输出和审批结果。

36.6 发布并验证业务智能体

1. 检查名称、描述、模型服务、身份指令、知识、文件、工具、凭据和 IM 工具。
2. 启用“发布（公开发布）”，再为试点部门、职能组或成员配置可用范围。
3. 在对话输入区选择合适的工具审批模式；高风险工具使用“请求批准”。
4. 使用获授权账号完成正常任务、工具审批、文件导出和历史查询。
5. 使用未获授权账号检查列表隐藏、直接访问拒绝和数据范围拒绝。
6. 到运行观测、网关调用和操作审计核对同一任务的模型、工具、数据和权限记录。
7. 根据试点反馈修改配置，重新发布后再次执行正向与反向验证。

37. 上线检查、稳定运行与管理交接

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。

37.1 完成上线前检查

1. 使用平台管理员账号核对部门、职能组、成员和管理角色。
2. 使用试点成员账号核对首页、能力列表和可用范围。
3. 分别验证模型服务、MCP、智能体、文件、知识库和数据连接的正常操作。
4. 使用敏感数据、风险内容、无权限资源、无效参数和额度耗尽场景验证安全与错误提示。
5. 在模型用量、网关调用、运行观测、操作审计和告警中核对相应记录。
6. 确认备份计划、最近一次备份、恢复责任人和维护窗口。
7. 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。

37.2 查看平台负载并处理容量问题

1. 选定同一时间范围查看请求量、错误率、响应耗时、排队和资源状态。
2. 按模型、工作空间、接入凭证或智能体缩小影响范围。
3. 打开代表性调用，区分上游响应慢、工具执行慢、排队、治理处理或重复调用。
4. 检查模型服务配额、API Key 配额、流量保护和智能体运行状态。
5. 一次只调整一类限制，并记录原值和恢复值。
6. 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。

“在线成员数量”和“请求并发”属于不同限制。登录或会话问题应检查成员会话状态；模型请求拥塞应检查并发、排队和流量保护。

37.3 配置和检查平台安全控制

1. 核对单点登录、账号禁用、闲置退出和管理员角色。
2. 核对模型、MCP、智能体、知识库和数据连接的授权范围。
3. 核对 PII 数据等级、命中动作、内容风险围栏和失败姿态。
4. 核对外部连接使用的凭据、HTTPS、允许访问的目标和审批方式。
5. 使用一个允许样例和一个拒绝样例验证每项重要策略。
6. 在护栏审计、操作审计和运行观测中确认命中结果、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。

37.4 检查常用文件与导出操作

1. 使用组织实际采用的浏览器完成登录、对话、文件上传、预览和下载。
2. 分别检查常用 Word、PDF、表格、图片和文本文件。
3. 对较大文件、扫描文件、复杂表格和异常文件检查明确的进度与错误提示。
4. 对 Word、PDF、Excel 或 CSV 导出检查文件名、格式、内容、分页、链接和水印。
5. 使用无权限账号验证文件、知识库和导出结果不可访问。
6. 在操作审计中核对上传、下载、导出和权限变更记录。

37.5 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，并使用其账号核对可见菜单。
2. 确认组织、模型、MCP、智能体、护栏、告警、审计、备份和运行管理的负责人。
3. 轮换由原管理员保管的 API Key、证书私钥、连接凭据和恢复密钥。
4. 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状态检查。
5. 核对支持联系人、维护窗口、通知接收人和紧急处理方式。
6. 确认接管完成后移除原管理员不再需要的权限，并在操作审计中核对结果。

37.6 提交支持请求

提交问题时提供发生时间和时区、产品与页面入口、受影响对象、期望结果、实际提示、可复现步骤，以及请求 ID、Trace ID 或会话 ID。同步说明最近的模型、凭据、权限、网络或版本变更和已经尝试的处理。

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。不得发送原始 API Key、Token、私钥、数据库密码或无关业务数据。

附录 A. 角色与入口速查

产品首页

产品	登录后的首页	主要导航	常用任务
Zenova	智能体使用面，默认打开“新对话”	工作区侧栏	对话、文件、知识、智能体、技能、应用和个人用量
Zentrix	能力中心，默认打开“MCP 广场”	能力中心顶部导航	查找模型、MCP、Skill 和外部 Agent，并进入管理控制台

管理角色与菜单

角色	适用产品	从哪里进入	常用菜单
普通成员	Zenova	工作区侧栏	新对话、文件、智能体、技能广场、应用连接、用量统计
工作空间管理员	Zenova	账号菜单 → 管理控制台	工作空间、知识库管理、IM 工具管理、Agent 运营
组织管理员	Zentrix、Zenova	账号菜单 → 管理控制台	成员与部门、职能组、单点登录
模型与网关管理员	Zentrix、Zenova	账号菜单 → 管理控制台	模型供应商、模型服务、鉴权与流量规则、网关调用
集成管理员	Zentrix、Zenova	账号菜单 → 管理控制台	MCP 服务、插件市场、网关域名接入
安全管理员	Zentrix、Zenova	账号菜单 → 管理控制台	护栏、告警规则、操作审计
质量评审人员	Zentrix、Zenova	账号菜单 → 管理控制台	人工评审、评估器、数据集、实验
运行管理员	Zentrix、Zenova	账号菜单 → 管理控制台	运行观测、日志管理、更新、备份与恢复；Zenova 还包括 Agent 运营

菜单会根据所用产品、账号角色和权限显示。没有看到某个菜单时，先确认当前账号的角色和授权范围。

附录 B. 常用任务索引

按任务查找对应章节。

适用产品	常用任务	参考章节
Zentrix、Zenova	接入并验证模型供应商与模型	第 35 章 · 常用配置的维护与变更
Zentrix、Zenova	建立并维护模型路由	第 25 章 · 模型路由与流量保护
Zentrix、Zenova	接入、发布并调用 MCP 服务	第 35 章 · 常用配置的维护与变更
Zentrix、Zenova	接入并维护 Agent 服务	第 35 章 · 常用配置的维护与变更
Zentrix、Zenova	核对 Token 消耗及其归属	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	核对调用量、成功率与调用主体	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	分析 Token 消耗趋势	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	识别高消耗接入凭证	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	分析模型吞吐与响应性能	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	切换监控视图和统计维度	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	定位单次调用与网关耗时	第 29 章 · 执行、问题与根因定位
Zentrix、Zenova	监控流量并处理阈值告警	第 33 章 · 告警、通知、留存、归档与导出
Zentrix、Zenova	分析消费者耗时和调用明细	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	向企业微信或邮箱发送告警	第 33 章 · 告警、通知、留存、归档与导出
Zentrix、Zenova	启用并配置内置治理插件	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	管理自定义插件的完整生命周期	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	选择并组合调用认证方式	第 26 章 · 调用方、API Key、认证与授权
Zentrix、Zenova	创建、启停并维护鉴权规则	第 26 章 · 调用方、API Key、认证与授权
Zentrix、Zenova	配置限流、整形、过载保护和流量分配	第 25 章 · 模型路由与流量保护
Zentrix、Zenova	转换发往上游的请求	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	规范返回给调用方的响应	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	创建并维护 API 消费者	第 26 章 · 调用方、API Key、认证与授权
Zentrix、Zenova	重置消费者凭据并设置有效期	第 26 章 · 调用方、API Key、认证与授权
Zentrix、Zenova	为消费者组分配限流配额	第 26 章 · 调用方、API Key、认证与授权
Zentrix、Zenova	创建并维护访问域名	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	将域名绑定到路由	第 27 章 · 请求变换、域名证书、插件与开放 API

适用产品	常用任务	参考章节
Zentrix、Zenova	上传并维护 TLS 证书	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	批量绑定证书并强制使用 HTTPS	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	保护、授权并校验证书	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	配置日志保留并查询平台日志	第 34 章 · 平台运行与日常维护
Zentrix、Zenova	查询并追溯审计日志	第 14 章 · 问题诊断与处理
Zentrix、Zenova	查看开放 API、权限和调用日志	第 27 章 · 请求变换、域名证书、插件与开放 API
Zentrix、Zenova	调整平台运行设置	第 34 章 · 平台运行与日常维护
Zentrix、Zenova	管理平台管理员、分级权限和登录日志	第 35 章 · 常用配置的维护与变更
Zenova	启动、重配、停止并检查 Agent 运行时	第 34 章 · 平台运行与日常维护
Zenova	发起对话并管理会话状态	第 17 章 · 对话、历史与上下文
Zenova	完成一轮任务并处理工具审批	第 21 章 · 智能体、工具与审批
Zenova	选择模型服务并处理生成异常	第 21 章 · 智能体、工具与审批
Zenova	使用工具、MCP、Skill 和扩展	第 21 章 · 智能体、工具与审批
Zenova	管理工作区、记忆和历史	第 22 章 · 记忆、子智能体与任务委派
Zenova	保护凭据、文件系统和网络访问	第 24 章 · 运行时安全与凭据
Zenova	委派子智能体并回收结果	第 22 章 · 记忆、子智能体与任务委派
Zenova	连接并管理外部消息渠道	第 23 章 · 外部 IM 工具
Zenova	观测、回放并评估 Agent 执行	第 29 章 · 执行、问题与根因定位
Zenova	了解 Zenova 的产品范围与使用入口	第 1 章 · 产品关系与平台配置流程
Zenova	完成文本、长文和多模态对话任务	第 17 章 · 对话、历史与上下文
Zenova	上传文件并检索企业知识	第 18 章 · 文件与企业知识
Zenova	配置数据连接并执行查询	第 19 章 · 数据连接与查询
Zenova	生成并导出文档、表格和报告	第 20 章 · 文档、表格与报告
Zenova	创建、发布、绑定并授权智能体	第 21 章 · 智能体、工具与审批
Zenova	追溯 AI 全链路审计与合规结果	第 14 章 · 问题诊断与处理
Zenova	同步组织、登录并管理会话	第 35 章 · 常用配置的维护与变更
Zenova	升级、备份、恢复并维护平台	第 34 章 · 平台运行与日常维护
Zenova	查看平台负载并处理容量问题	第 37 章 · 上线检查、稳定运行与管理交接
Zenova	配置和检查平台安全控制	第 37 章 · 上线检查、稳定运行与管理交接
Zenova	检查常用文件与导出操作	第 37 章 · 上线检查、稳定运行与管理交接
Zenova	接入新的模型与外部能力	第 35 章 · 常用配置的维护与变更

适用产品	常用任务	参考章节
Zenova	配置并发布业务智能体	第 36 章 · 配置并发布业务智能体
Zenova	完成平台管理交接	第 37 章 · 上线检查、稳定运行与管理交接
Zenova	完成上线前检查	第 37 章 · 上线检查、稳定运行与管理交接
Zentrix、Zenova	接入 OTLP、SDK 和运行数据	第 28 章 · 可观测数据接入与权限边界
Zentrix、Zenova	查询 Trace 与 Span 并逐层下钻	第 29 章 · 执行、问题与根因定位
Zentrix、Zenova	发现问题、判断回归并定位根因	第 29 章 · 执行、问题与根因定位
Zentrix、Zenova	分析成本、Token 与运行效能	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	创建评分并完成人工评审	第 31 章 · 评分、人工评审与自动评估器
Zentrix、Zenova	创建并运行自动评估器	第 31 章 · 评分、人工评审与自动评估器
Zentrix、Zenova	构建数据集并运行对比实验	第 32 章 · 数据集、实验与 Prompt 版本
Zentrix、Zenova	管理 Prompt 版本并在 Playground 验证	第 32 章 · 数据集、实验与 Prompt 版本
Zentrix、Zenova	创建仪表盘并执行自定义查询	第 30 章 · 用量、成本与自定义看板
Zentrix、Zenova	配置质量告警与通知规则	第 33 章 · 告警、通知、留存、归档与导出
Zentrix、Zenova	配置留存、归档和导出	第 33 章 · 告警、通知、留存、归档与导出
Zentrix、Zenova	观测业务过程并追踪治理动作	第 29 章 · 执行、问题与根因定位
Zentrix、Zenova	按账号权限使用观测能力	第 28 章 · 可观测数据接入与权限边界
Zentrix、Zenova	检查质量评估与观测配置	第 31 章 · 评分、人工评审与自动评估器