

ZENTRIX GATEWAY

企业 AI 产品文档

配置 · 使用 · 结果核对

Zentrix Gateway 路由策略用户手册

流量规则、全局限额、意图集、语义缓存、熔断与故障转移的配置及验证。

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. 入口与生效顺序

2. 创建规则与配置条件

3. 验证规则与真实调用

4. 配置全局限额

5. 处置动作与异常处理

6. 配置并验证熔断

7. 流量分配与故障转移

8. 维护、校验与故障定位

9. 意图集

10. 语义缓存

> 2026.10 · 操作范围以当前许可、账号权限和实际部署为准。
>
> 本手册是《Zentrix Gateway 用户手册》第 15 章的专项版，事实与第 15 章一致，另外列出更细的字段、默认值和取值范围。规则中「内容检测」动作的检测项、模板与判断模型配置见《Zentrix Gateway 用户手册》第 14 章。

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

1. 入口与生效顺序

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 → 全局设置 →「授权许可」页签」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《Zentrix Gateway 用户手册》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
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 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 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。

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。平台不再内置路由规则，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。

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。

2. 创建规则与配置条件

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 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。

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 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 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 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。

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。

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。
- 取值长度：请求路径、请求头、URL 参数和客户端声明的模型名的每个取值不超过 256 字节；数据来源、具体能力等其余取值每个 1–160 字节。「模型服务」「解析后的网关模型」只接受从选择器选出的对象，不能用全零编号占位。
- 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。

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《Zentrix Gateway 用户手册》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 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

请求头条件不能使用的名称（数据面在判定前剥离或改写这些请求头，用它们做条件的规则永远不会命中）：

类别	请求头
凭据	`authorization`、`proxy-authorization`、`cookie`、`x-api-key`
转发来源	`x-forwarded-for`、`x-real-ip`
传输与内容	`host`、`content-type`、`content-length`、`content-encoding`、`accept`、`accept-encoding`、`connection`、`transfer-encoding`
关联标识	`x-request-id`、`x-correlation-id`、`traceparent`、`tracestate`
网关内部	以 `x-lp-`、`x-waner-` 开头的请求头，以及网关用于传递调用标签、终端用户、个人信息类型和数据来源的内部请求头

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”。

3. 验证规则与真实调用

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。



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」。
- 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。

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。

1. 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」。
2. 选择「调用方」，或选择「凭据身份」（API 密钥），两者二选一。
3. 按需填写「请求路径」「试运行时间」（用于时段条件）。内容检测试运行时，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」。
4. 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。
5. 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。
6. 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。

「验证规则」页面字段：

字段	位置	说明
调用类型	基本信息	必填；决定入口与可用的匹配信息
请求的模型	基本信息	模型调用时填写，可输入或从列表选择
能力	基本信息	非模型调用时选择
调用方 / 凭据身份	基本信息	二选一；凭据身份为 API 密钥
请求路径	基本信息	按此路径值模拟匹配；真实请求的路径由调用入口提供
试运行时间	基本信息	用于评估时段条件，无需等待实际时段到达
检测方向、消息、本轮回答	基本信息	按命中规则的护栏配置执行内容检测试运行，结果显示在「护栏试运行」中；消息可选择角色并附图，「检测方向」为「输出」时填写「本轮回答」；结果给出依据与分段耗时（图片转写、内容判定）
工具名、参数或命令、返回结果	基本信息	命中规则配置了工具调用防护时使用；参数能解析为 JSON 对象时原样发送，否则按命令文本发送；填写「返回结果」后，匹配规则同时按「工具返回的结果」位置检测；选择「工具调用防护」后才会调用判断模型

字段	位置	说明
指定模型服务	更多请求事实	回放已知请求，或模型名在多个服务中有歧义时指定；缺少时含「模型服务」条件的规则无法模拟
意图提示词	更多请求事实	用于意图分类，留空时不生成意图信息，判断结果显示在「判断结果」中
客户端 IP	更多请求事实	IPv4 或 IPv6 地址
查询参数、灰度 Cookie、调用标记	更多请求事实	按名称与取值逐项添加
数据来源、PII 类型	更多请求事实	平台判定的事实，填写的是假设值
请求头	更多请求事实	每行一条「名称: 值」
模拟主档失败	底部选项	查看主要目标失败后的候选顺序
按带金额预算的调用方演练	底部选项	链上未定价的档位会被跳过，与运行时判定一致

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，请使用允许送检的测试文本。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。

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 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。状态码与错误码的完整对应关系见错误参考。

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。

4. 配置全局限额

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。

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 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。

新建限额时各类型的预填值（均可修改）：

限制类型	预填值
请求次数	计数维度「按 API 密钥」；每分钟 100 次、即时突发量 20；超限处理「拒绝请求」，「额度用完时」为「直接拒绝」；选择排队时预填最大排队深度 64、最大等待 3000 毫秒、每个组织排队上限 16、每个组织队列容量 1048576 字节；选择固定应答时原因代号预填 `quota_exceeded`
并发限制	并发上限 8，繁忙时的优先级 0，不保留名额
熔断	统计窗口 30 秒、最小调用样本数 10、失败率 50%、慢调用阈值 2000 毫秒、慢调用率 50%、熔断持续时间 30 秒、半开探测 5 次
非法尝试保护	统计窗口 60 秒、次数阈值 5、暂停 300 秒
上游调用超时	无预填，须填写连接超时与请求超时
调用方临时豁免	豁免类型「请求次数」，到期时间为 24 小时后

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 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《Zentrix Gateway 用户手册》第 15.5 节。选择计数维度前确认真实入口会提供该信息。

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。

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。

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。

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。

5. 处置动作与异常处理

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 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《Zentrix Gateway 用户手册》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 5.5 节
单次请求 Token 上限	请求转发前	模型调用	见第 5.4 节
请求体大小上限	请求转发前	模型调用	见第 5.4 节
语义缓存	请求转发前	模型调用	见第 10 节
推理指令	请求转发前	模型调用	见第 5.4 节
会话切换保护	请求转发前	模型调用	见第 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 7.5 节
提示词	请求转发前	模型调用	见第 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 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

动作（画布名称）	阶段	适用调用类型	要点
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《Zentrix Gateway 用户手册》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 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。

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《Zentrix Gateway 用户手册》第 14 章。在规则层面注意：

检测方向、拦截返回与引用就绪

检测方向的作用面。“输入”检测在请求发往目标模型之前执行，“输出”检测在模型回答返回调用方之前执行，“双向”两者都执行。两种接入方式——OpenAI 兼容入口（使用模型 Key 调用 `/api/v1/agent-gateway/openai/v1/*`）与网关原生入口（`/v1/*`）——对输入、输出两个方向都执行，判定口径一致。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。两种入口的差异仅在于流式放行的时机，见下文“流式输出的护栏行为”。

调用方收到的结果。

情形	HTTP 状态	调用方可见内容
输入命中；或非流式回答在输出检测中命中	403	`error.code` 为 `content_policy_violation`，`error.reason` 为 `CONTENT_POLICY_BLOCKED`；`error.message` 是规则配置的拒答话术（未配置时为默认拒答话术）；`error.details` 中 `reason` 为 `safety_intervention` 的条目带命中类别、检测方向与命中的路由策略规则标识
流式回答在输出检测中命中	200（流已开始）	已放出的内容保留，流以所选协议的终结帧结束，拒答话术位于终结帧根上的 `safetyIntervention` 字段
检测服务不可用，且规则的“检测不可用时”选择阻断	403	`error.message` 为“内容检测服务不可用，已按策略拒绝本次请求”；不带 `safety_intervention` 条目，原因码与“命中”不同

同一条命中在两种入口返回相同的 message、code 与 details 结构。调用方应按原因码区分“内容被拦截”（修改请求内容）与“检测不可用”（联系管理员恢复检测服务），不要把后者当作网络故障原样重试。各字段的示例见错误参考的“内容策略命中与检测不可用”一节。“检测不可用时”选择放行的规则，请求照常处理。

引用对象必须先就绪。 内容检测段引用的租户词库、问答集必须已发布成功。引用尚未发布成功的对象时，保存或下发配置时会失败，并在路由策略页指出缺少的对象；已保存的规则若因引用对象失效而不可用，规则列表会标记“当前不可用”并给出原因。不会出现“保存成功但运行时不生效”的情况，也不会因此拒绝不含敏感内容的正常请求。

流式输出的护栏行为

- 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。
- 检测阶段包含输出时，流式响应同样受检：规则层逐段检测，模型判断逐句进行，命中拦截时网关在命中处中止并按调用协议发送终结帧，中止后按上游实际用量计费，见《Zentrix Gateway 用户手册》第 14.10 节。各协议终结帧的形态见错误参考的「流式护栏中止的终结帧」一节。

「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」）。节点显示引用的版本；提示「当前标签为其他版本，未发布」时，应确认发布关系，不能只看提示词名称。「查看提示词」打开对应提示词详情。

工具调用防护、MCP 工具许可与 API 服务匹配规则

- 工具调用匹配规则与自定义类别： 匹配规则只按各自填写的工具模式（精确名或 * 前后缀通配）生效，不经过“检测哪些工具”的范围筛选；按「字段」中填写的路径（点号分隔，如 args.content）取值，留空时遍历全部字符串叶子并先解开一层 JSON 转义再匹配；调用参数非合法 JSON 时对原始字符串做全文匹配并在证据中标记 argsMalformed；例外只豁免与它重叠的那一处命中，其余命中照常处置。语义注入由内容安全负责。
- 模型判断范围与提示词缓存影响： 模型判断（内置类别与租户自定义类别）只作用于模型调用（策略条件只含 MCP/连接器/API 服务时不可配置类别与分类器）；入向工具结果规则命中并改写为“该内容已被拦截”会改变历史消息内容，从而使该位置之后的上游提示词缓存失效。保存时若规则中的工具名尚未出现在目录或近 30 天调用记录中，编辑器会给出工具名核对提示，提醒核对拼写。
- MCP 工具许可与存量迁移： 原 MCP 服务详情页的独立护栏已整体并入路由策略，使用“工具名”“组织成员”“MCP 方法”条件配合拒绝动作或工具调用防护实现。升级时存量配置会自动转换为排在租户现有策略之前的路由策略，命名格式为 MCP 护栏迁移 · {服务名} · {项}（项为「工具许可」「参数内容规则」「默认处置」「请求体上限」或「需人工改写」）；含环视/反向引用正则或含通配符的精确工具名及带 argumentContains 的放行项会转为 MCP 护栏迁移 · {服务名} · 需人工改写 拒绝规则。/v1/execute 扩面后，命中含组织成员条件的拒绝规则且无触发用户（如服务账号）时按 fail-closed 拒绝。
- API 服务匹配规则： 在 API 服务调用类型下，“调用参数”位置对应检查请求体，“返回结果”位置对应检查响应体；JSON 正文按字段指针或全部字符串叶子匹配，纯文本正文按全文匹配，二进制体不检查。

图文联合检测

- 检测流程与远程图片内联： 在规则 YAML 的 action.guard.image 中选择支持视觉输入的图片理解模型（model.gatewayModelId）后，网关将输入图片（及 /v1/images/generations 生成图）转写为结构化描述、图中文字与指令标记，并以 ⟦图片 k⟧描述：…；图中文字：…⟦/图片 k⟧ 插入安全层输入，交由词库、敏感信息与内容安全统一检测（敏感信息命中图片转写且动作为打码时自动升级为拦截）。远程图片（http(s)://）由网关拉取检测后替换为内联 data: URI 再转发给目标模型，确保上游模型收到的就是被检测的那份字节。
- 数据流向、缓存与白名单： 图片数据会发往图片理解模型与目标模型；网关节点在本地保存 24 小时转写缓存（按租户、图片内容 SHA-256、模型与提示词版本索引）。远程拉取默认只允许公网地址（拒绝私网、回环、链路本地与云元数据地址），内网图源需在 allowedFetchHosts（最多 32 项）中显式填写主机名。
- 规格、预算与计费： 单图上限 20 MiB，单次请求最多 8 张未缓存新图（合计 40 MiB），拉取连接超时 3 秒、读取超时 10 秒，最坏额外时延约等于一次视觉模型调用。图片中若检出针对识别或审核的指令，该段转写会强制送越狱检测。图片理解调用以 guard_image_understanding 计入租户账单与用量，在模型服务用量明细中显示为「平台内部调用：图片理解」。
- 图片检测能力边界： 检测结论来自视觉模型的转写，不是确定性 OCR，可能漏识别，也可能被图片里混入的提示注入影响（含指令的转写已强制送越狱检测，见上）。无法读取内容的文件——PDF 等文档类附件、以文件编号引用而不是直接携带字节的图片（Anthropic source.type=file、Responses input_image.file_id/input_file/item_reference、Chat file.file_id 等形态）——不做检测，照常放行，并在策略命中列表中标为「未检测」留痕；这类引用形态可能被用来绕过图片检测，是已知边界，不是实现缺陷。

多轮拼接诱导防护

在规则 YAML 的 action.guard.multiTurn 中可分别开启三项（reassembleShortMessages、cumulativeOutput、sessionEscalation），默认都不开启：

1. 拆分输入检测： 按用户撰写的消息顺序（跳过助手回答与只含工具结果的消息）滑动检查，6 条以内有至少 3 条可见字数不超过 8 个（先做兼容字符归一化，再去掉标点、表情符号、零宽字符与空白）的短消息时，把这一段的可见文字连起来，交给词库、敏感信息与内容安全检测。命中依据显示为“跨消息拼接 · 第 i–j 条用户消息”；拼接内容无法回写，敏感信息配置为打码时按拦截处置。
2. 多轮回答合并检测： 输出方向的内容安全判断带上本轮提问之前的历史对话（从末尾往前截取，最多 4096 个字符），本轮回答与历史回答合并判定，命中即中止本轮。流式输出仍按句扣留、逐句判定，每句判定多出的输入以 4096 个字符为上界；词库与敏感信息不跨轮拼接。
3. 会话多次命中限制： 同一会话在时间窗（10 分钟、30 分钟、1 小时、6 小时、24 小时）内新内容的命中（明确命中、疑似命中、越狱命中）达到设定次数（2–20 次）时，在持续时长内“拒绝该会话的后续请求”（直接返回拒答话术，不调用上游与判断模型）或“将疑似命中按明确命中处置”。同一段内容在时间窗内只计一次：历史里的旧命中随新消息反复送检不会重复累计，拆分输入检测的窗口滑动时已计过的用户消息也不再重复累计。

无论是否开启以上三项，入向检测都会补送历史助手回答。

超长对话的送检口径： 系统文本、全部用户撰写的文本与本轮新增内容（最后一次助手回答之后的消息，含本轮工具结果）一定送检，放不下时按“护栏出错时”处置；历史工具结果与历史助手回答在剩余容量内尽量送检（工具结果优先、新近优先），送不下的部分不算超限，字节数记入命中证据的 historyUncheckedBytes。因此把载荷藏进伪造的早期用户消息再用填充挤出容量的做法不能绕过检测。

会话累计口径：

- 会话累计以调用方提供的会话标识为准，更换标识即视为新会话；应由应用服务端生成不可猜测的会话标识。
- 各网关节点分别在本地累计，不跨节点共享；MCP 调用没有会话标识，不参与会话累计。
- 命中内容的去重摘要由各节点用节点自身的密钥计算，节点密钥轮换后，时间窗内已累计的命中按新内容重新计数（最长一个时间窗）。

处理误伤： 在“策略命中”中可按“会话状态：已限制”筛选。“标记误判”作用于触发会话限制的那条命中、且操作者有路由策略管理权限时，会同时解除该会话的限制，各网关节点在下一次配置下发后恢复。标记的不是触发命中，或操作者没有路由策略管理权限时，只记录误判标注，限制保持。“查看会话”会打开会话详情并定位、展开到命中所在的那一轮。

升级说明与试运行接口

- 滚动升级： 滚动期间，旧版本节点上本租户的 `tenant_config` 与 `mcp_protection` 两类配置段冻结在最后一次可用的版本（控制台显示“正在下发”），旧节点继续按旧规则执行，节点升级完成后自动恢复。网关节点须在控制面可达时滚动。
- 试运行接口： 路由策略的试运行可经管理接口调用。JSON 形态使用 `POST /api/v1/admin/route-policies/simulate`：`direction` 为 `input` 或 `output`，`messages` 为 `[{role, text, images}]`，测试图片以 `` 形式放在 `images` 里，输出方向用 `answer` 填本轮回答。页面带图试运行使用 `simulate/with-images`（表单上传）。试运行与网关走同一条判定路径，规则里关闭的开关在试运行里同样不生效，结果给出依据与分段耗时（图片转写、内容判定）。

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。
  - 不允许改写的请求头：`host`、`set-cookie`、`content-type`、`content-encoding`、`content-length`、`connection`、`keep-alive`、`te`、`trailer`、`transfer-encoding`、`upgrade`。
  - 不允许改写的响应头：`content-type`、`content-encoding`、`content-length`、`authorization`、`proxy-authorization`、`proxy-authenticate`、`www-authenticate`、`set-cookie`、`location` 及连接控制类响应头。
  - 以 `x-lp-`、`x-zr-` 开头的网关内部头不可改写。头字段值不超过 4096 字节，不能包含换行和空字符；删除头字段时不能填写值。
- 「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。

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。

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 4 节和第 6 节。

- 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，且不能使用限额执行时无法判断的条件（工具名、MCP 方法、意图、任务类型、任务难度、业务场景、请求含敏感信息、组织成员、灰度 Cookie、数据来源）；按 MCP 工具限流请改用全局限额的「按 MCP 工具」计数维度。
- 同一请求可能同时受规则内限额和全局限额约束。
- 规则内限额的预填值：请求速率按 API 密钥计数、每分钟 100 次、突发 20；并发上限 8；熔断参数同第 6.2 节表中的默认值；上游调用超时为连接 3000 毫秒、请求 30000 毫秒。「并发」中的「繁忙时的优先级」以「最高」「较高」「普通」三档选择；使用恢复阈值与并发硬上限时，同时开启「启用软硬并发阈值与恢复机制」。

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。

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」，「护栏检测出错」由规则逐条设定，高风险业务设为「拒绝本次调用」；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。

6. 配置并验证熔断

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。

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。

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
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。

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。

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。

7. 流量分配与故障转移

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。

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，价格取自已发布的价目；缺少数据的目标按与最优者并列处理

「目标选择策略」	行为	前提与验证
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。

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。「触发转移的上游状态码」按名称列出常见的 4xx 与 5xx 状态码，例如 401「上游鉴权失败（密钥无效或已过期）」、403「上游拒绝访问（余额不足或无权限）」、429「请求过于频繁」、503「上游暂时不可用」、504「上游网关超时」，以及 Cloudflare 的 520–526 和 529「上游过载（Anthropic）」。选择 4xx 状态码前确认这类错误换一个目标确实可能成功，否则只会增加重复调用。
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。

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《Zentrix Gateway 用户手册》第 15.5 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。

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。

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。

8. 维护、校验与故障定位

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。

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
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

8.3 保存冲突和运行异常

返回 403、429、502、503、504 时的错误码含义与是否可重试，见错误参考。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

现象	排查顺序与处理
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。错误码的含义可先查错误参考。不要提交原始 API 密钥、Cookie、个人信息或上游凭据。

9. 意图集

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。

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。

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。

10. 语义缓存

10.1 开启方式与边界

语义缓存只对命中「开启了语义缓存」的规则的模型请求生效：在规则中添加「语义缓存」动作即开启。命中这类规则的请求先查缓存，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。

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」。

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。