

ZENOVA

企业 AI 产品文档

配置 · 使用 · 结果核对

Zenova 路由策略用户手册

流量规则、全局限额、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

2026.09

核对日期

2026 年 9 月 15 日

目录

1. 入口与生效链路

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

> 2026.09 · 操作范围以当前许可、账号权限和实际部署为准。

> 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. 入口与生效链路

1.1 从哪里进入

打开管理控制台，在侧栏“路由策略”下选择“流量规则”。页面包含“规则清单”和“全局限额”两个标签。菜单与编辑操作受管理权限、所属租户和部署版本影响；看不到入口时，先核对登录账号及权限。

页面	管理对象	适用任务
规则清单	按优先级匹配的流量规则	转发、放行、拒绝、内容检测、改写和附加限额
全局限额	独立的流量保护配置	请求次数、并发限制、熔断、非法尝试保护、上游调用超时
规则详情	一条规则的条件、动作与状态	查看、编辑、启停、查看命中的请求、修改记录、删除
验证规则	输入请求信息并模拟判定	检查命中链、路由目标及未命中原因

“全局限额”表示独立于某条流量规则管理，仍按自己的条件限定范围，并不意味着所有请求共用一份额度。模型服务配额、API 密钥配额和这里的限额也不是同一套计数。

1.2 规则如何参与一次调用



图 1：配置验证链路示意。保存、下发、模拟和真实调用分别回答不同问题。

请求经过身份认证后，平台根据实际调用入口提供的路径、模型和调用身份等信息判定规则。优先级数字越小越先检查。条件不满足时跳过本条，继续检查后续规则。

规则行为	命中后的处理	是否继续检查后续规则
增强规则	叠加内容检测、限额、改写等配置，不指定新目标	是，除非执行被其他保护动作中止
转发到指定目标	根据目标、权重和策略选择去向	否
放行	结束本轮规则判定，继续按默认路由处理	否
拒绝	拒绝请求，不向上游派发	否

放行仍受身份权限、有效额度及目标可用性检查约束。只配置限额或检测、没有指定目标的增强规则，也不会自动替代默认模型。

列表中没有规则，或没有规则命中时，请求继续按网关其他已发布配置处理；这不保证一定存在可用默认目标。平台内置规则如显示为只读，应通过其负责的管理入口维护。

1.3 如何阅读画布

画布从“请求进入”开始，经“条件判断”进入每一个处置动作节点，最后显示“处理结果”和“异常处理”。动作按“请求转发前”与响应阶段区分。

- 绿色箭头“满足条件 / 执行本规则动作”表示进入本规则动作。
- 右侧“不满足条件”说明表示跳过本条，按优先级检查下一条。
- 一个动作对应一个节点。展开节点查看字段；折叠仅改变显示，不会停用动作。
- “处理结果”描述命中后的配置预期，不表示已有请求执行成功。
- “异常处理”只展示当前动作可配置的异常项。增加或移除动作后，可配置项会变化。

使用“收起节点”“画布缩放”“流程总览”定位较长流程；编辑时可撤销、重做尚未保存的修改。“配置检查”用于定位当前草稿中的问题，保存时服务端仍会校验。

2. 创建规则与配置条件

2.1 创建一条模型路由

开始前，在“模型供应商”和“模型服务”中完成目标接入、启用及调用验证；为测试调用方准备有效凭据和模型使用权限。

1. 在“规则清单”选择“新建规则”，填写能说明业务用途的名称，例如“客服问答主路由”。名称为 1–120 个字。
2. 在“条件判断”添加“调用类型”，选择“模型调用”。继续添加需要的调用方、模型或路径条件。
3. 添加“客户端声明的模型名”，填写调用方实际提交的模型名，例如 `customer-support`。这是请求里的名称，不是模型对象 UUID。

4. 选择“添加动作 → 转发到指定目标”，选择目标模型及其归属的模型服务。一个模型可能属于多个服务，必须选清归属。
5. 需要分流时配置同档目标权重；需要故障转移时再增加后备档位，参见第 7 节。
6. 检查条件、动作、依赖和“处理结果”。修复“配置检查”中的问题后选择“保存”。新建成功后，回到列表核对规则名称、启用状态和顺序。
7. 用“验证规则”检查正例与反例，再用实际业务入口发送受控请求，参见第 3 节。

本文示例名称和阈值用于说明操作，不代表推荐容量。不要把示例模型名直接作为已接入模型使用。

2.2 创建一条增强规则

如果只想给客服模型增加限速，不改变路由去向，可添加相同的模型名条件和“模型调用”条件，然后添加“限额”，设置请求速率、计数对象和阈值，保留目标为空。



图 2：真实编辑页面局部。示例为增强规则；满足条件时应用限额，随后继续按后续规则或默认配置选路。

将增强规则放在需要受它保护的终止规则之前。若前面的规则已经指定目标、放行或拒绝，后面的增强规则可能不会参与判定。要设置“允许名单 + 拒绝其他请求”，应使用真正的“放行”动作，并核对它与兜底拒绝规则的先后关系。

2.3 条件之间是什么关系

同一条规则通过“添加一个‘且’条件”组合条件，跨条件按 AND 判定；同一条件中的多个取值通常表示满足任意一个即可。不要把两条互相矛盾的“等于”条件当作“或者”。一条规则最多 16 个条件；多数多值条件最多 32 个取值。

未添加条件的规则范围很广。保存前应明确它是有意设置的兜底规则，并检查其位置。创建全局限额时则必须至少设置一个条件。

条件	填写与选择	注意事项
调用方	先选身份种类，再选择对应对象	相同文字或 UUID 不表示不同身份种类可互换；使用选择器
调用类型	模型、MCP、连接器、A2A 或 API 服务	转发和限额等动作要求限定调用类型；有些动作仅适用于模型
具体能力	选择调用类型及具体能力	先明确能力种类；不能把模型 ID 填作 MCP 工具 ID
客户端声明的模型名	请求中声明的模型名称	可按页面提供的等于、前缀、正则方式匹配；不接受 UUID 代替模型名
模型服务	选择模型服务对象	请求必须具有可识别的服务归属信息
解析后的网关模型	选择网关解析得到的模型对象	与客户端提交的模型名是不同信息
请求包含	选择可用的数据来源，例如 MCP	匹配平台实际提取的信息，不是任意自定义标签
意图	选择意图集与意图项	依赖意图资产和分类结果；未执行分类或分类服务不可用时不能当成命中
个人信息类型	选择需要匹配的类型	依赖内容检测产生的事实，类型由租户检测配置提供
请求路径	填写实际路径，选择比较方式	反向代理前缀和不同入口可能改变可见路径
URL 参数	填写参数名和取值	参数名不得为空；按真实请求的查询参数验证
请求头	填写头名称、比较方式和取值	受保护或被入口剥离的请求头不能作为有效条件
时段	选择时区和时间段	按所选时区判定，不按管理员电脑显示时间猜测
客户端 CIDR	填写合法 IPv4/IPv6 网段	用于网段范围匹配；真实客户端地址受可信代理解析影响
灰度 Cookie	选择允许的 Cookie 名称并填入取值	值为只写敏感字段；重新编辑时按页面提示重新填写，不用摘要代替原值

界面只提供当前可编辑字段。旧配置中的流量标签、数据分级或旧调用方字段，应按页面错误提示重新配置，不沿用旧手册中的入口。

2.4 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。手动填在模拟器里的“调用方”只是用于判定的输入，不会为真实调用授予权限。复制其他对象的 ID 到普通请求头也不能替代认证。

不同调用类型提供的信息不同。例如模型名条件适用于具有模型信息的调用，MCP 能力条件需要对应工具或服务信息。模拟结果如果提示“请求未提供该信息”，应补齐真实请求可提供的信息；如果提示“此调用类型不提供该匹配信息”，应调整条件或调用类型。

配置意图、个人信息类型等条件前，先核对对应分类或检测功能已准备好。未检测、检测失败和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 保存后核对什么

新建规则选择“保存”提交；在规则详情中编辑已有规则，选择“保存”后会展示“确认变更”。核对“字段差异”“关联对象”和“命中顺序”，再选择“确认并应用”。若变更范围与预期不符，选择“取消”返回编辑。



图 3：已有规则的变更确认。确认前核对实际变化和影响范围。

出现“已保存，正在下发到数据面”等提示时，说明配置已经保存，但运行端可能仍在加载。重新打开详情核对保存后的字段，并查看模拟中的版本或下发状态提示；版本落后时等待同步，再做真实调用验证。

3.2 使用“验证规则”

1. 返回“规则清单”，选择“验证规则”。
2. 填入与业务请求一致的调用类型、模型、请求路径及调用方信息；需要时展开“更多请求事实”。

3. 也可以“选择最近的实际调用”作为起点。调用记录不含全部请求信息，页面列出的缺失字段仍需补充。
4. 对时段规则指定评估时间；对意图规则填写意图提示词。涉及内容或操作检测时，按页面要求提供检测文本、工具名和参数。
5. 查看本次模拟命中的规则链、最终目标、期望分配及每条规则的判定原因。修改输入后，等待重新计算完成再读取结果。
6. 分别验证一个应命中的请求、一个不应命中的请求，以及缺字段或边界值的请求。

结果	正确理解	下一步
命中增强规则	本条贡献了配置，后续规则仍可能继续	查看完整命中链和最终目标
命中终止规则	后续规则不再评估	确认它的优先级及动作符合预期
未命中	已评估，但条件不成立或所需信息缺失	展开条件原因，核对输入与真实入口
未求值、未到达	前序规则已结束判定，或本条未进入本次规则集	核对前序规则、启用状态和版本
未命中任何规则	按当前发布的默认配置选路	确认默认目标和权限可用
未提供默认目标	无法展示默认路由落点	按页面提示补充必要请求信息

模拟不会完成真实模型上游调用，也不能证明限额已被计数、响应改写已执行或目标可连通。意图分类、内容检测、操作检测可使用相应检测服务，应使用允许送检的测试文本。

加权轮询的模拟与真实派发使用独立的轮询进度。连续点击模拟不会消耗真实流量的轮询顺序，也不能用模拟次数代替实际分流统计。

3.3 用真实调用完成验证

使用将来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，避免影响其他业务流量。

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，再更换维度值	达阈值后的处理，以及不同调用方是否分别计数
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、结束后释放情况
改写	给上游发送可观察的测试字段，核对返回	上游实际收到的值及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

验证项目	操作	应保留的结果
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	让主档满足已配置的失败条件	后备目标是否收到请求、总尝试次数和耗时

通过规则详情的“查看命中的请求”进入关联记录，也可在“运行总览 → 规则命中”页签及调用记录中核对时间、规则和目标。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。没有查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；记录缺失时提交给运行管理员处理。

4. 配置全局限额

4.1 先确定保护对象

进入“流量规则 → 全局限额”，选择页面的调用类型后新建限额。按画布依次配置限制类型、计数对象、限额参数和超限处理，填写名称并保存。

限制类型	解决的问题	主要字段
请求次数	平滑控制每秒、每分钟或每小时的请求速率	时间窗、周期内请求数、即时突发量、计数维度
并发限制	限制同时执行的调用数量	并发上限、可选软硬阈值
熔断	上游持续失败或变慢时暂停调用	样本、比例、慢调用阈值、熔断时长及半开探测
非法尝试保护	限制被识别为非法尝试的重复行为	统计窗口、尝试次数、暂停时间
上游调用超时	控制建连和请求等待时间	连接超时、请求超时

保护类型、条件字段和计数维度受调用类型约束。禁用选项旁的说明是配置限制，应先调整条件或类型，不通过重复保存绕过。

4.2 匹配条件与计数维度

匹配条件回答“哪些请求受这条配置影响”；计数维度回答“这些请求怎样分别累计”。例如仅匹配客服模型、按 API 密钥计数，表示每枚密钥对客服模型的请求分别受到限制。

“限谁”支持按调用方身份、客户端 IP、路由、模型服务、网关模型、能力、依赖、端点、MCP 服务或工具，以及服务和部署信息等字段选择，实际可用项以所选调用类型为准。会话只能作为计数维度，不能作为匹配条件。

- 全局限额需要 1–16 个条件，跨条件全部满足才匹配。

- 同一条件可按等于、不等于或属于任一值配置。等于只允许一个值。
- 同一字段上两条肯定条件必须有交集，否则永远无法命中，页面会阻止保存。
- 全局限额中的“客户端 IP”是精确地址匹配，不是 CIDR 网段包含；网段匹配应使用规则清单中的“客户端 CIDR”。
- 选择计数维度后，确认真实入口会提供该信息。按会话计数时，只统计网关通过请求头 `x-zenova-conversation-id`、智能体线程或客户端自带的会话编号识别出的会话（Responses 续接请求沿用上一轮的识别方式）；按首条用户消息识别的会话与无法识别会话的请求不计入该维度。识别顺序与取值要求见第 26 章 · 会话识别。

改变匹配条件不代表计数维度自动改变。编辑已有配置时，分别核对这两部分。

4.3 请求次数与临时阈值

配置“时间窗”“周期内请求数”“即时突发量”。时间窗可选秒、分钟、小时；最多同时配置 3 个不同时间窗。同周期不能重复，数值须为正整数。

这是按时间持续补充的速率限制，不应按“每到整点完全清零的日配额”理解。多个时间窗同时配置时，请求需要满足各时间窗限制。

“即时突发量”用于容纳短时峰值。不要把它当作每个周期额外赠送的长期额度。设置前先估算上游吞吐和可接受的峰值，保存后使用实际请求间隔验证。

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。
2. 临时值都不得低于基线，至少一项高于基线。
3. 填写开始和结束时间，结束时间必须晚于开始时间且未过期；窗口最长 30 天。
4. 到期后恢复基线。它只临时放宽流量阈值，不延长凭据有效期，也不授予模型访问权。

4.4 并发、非法尝试与上游超时

并发限制统计同时执行的调用，而不是一分钟内的调用总数。并发上限必须为正整数。若启用软、硬阈值，两者必须同时填写，并满足“软阈值 < 硬阈值 ⩽ 并发上限”；其行为还需按所选入口和超限处理验证。

非法尝试保护需要填写统计窗口、最大尝试次数和暂停秒数，均为正整数。它针对系统归类为非法尝试的事件，不应当作所有 HTTP 失败的通用失败计数。验证时应确认请求确实经过对应认证入口且事件被识别。

上游调用超时包含连接超时和请求超时，单位为毫秒，范围均为 50–600000；连接超时不得大于请求超时。连接超时控制建立上游连接的等待，请求超时控制上游请求时限。慢调用阈值只用于熔断统计，不会直接中断慢请求。

4.5 创建临时豁免

在可豁免限额的“更多操作”中选择“创建调用方临时豁免”。它会新建一条独立豁免配置，并不删除原限额，也不必然继承原限额的全部字段。

1. 重新核对调用方、匹配范围及要豁免的保护类型。
2. 填写明确原因和到期时间。普通豁免最长 7 天。
3. 如页面允许应急越过硬下限，核对授权和用途；此类豁免最长 2 小时。

4. 保存后用豁免范围内与范围外的调用分别验证；到期后再次检查原保护恢复。

可选豁免类型为请求次数、并发限制、熔断和非法尝试保护，不包含上游调用超时。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 选择动作

从规则画布的“添加动作”选择动作。页面会依据调用类型与已有配置显示、隐藏或禁用选项，并给出不满足依赖的原因。按动作节点填写必填字段，查看执行阶段和顺序。

动作入口或节点	用途	配置与边界
转发到指定目标	改变本次调用去向	限定一种调用类型；模型需选择目标及归属服务
放行 / 拒绝	结束规则判定	属于纯判定动作，不与转发、改写、提示词等执行动作混用
内容检测 / 内容检测（护栏）	按模板检测输入、输出或相关操作	选择检测项、模板、命中处置及检测失败策略
限额 / 应用限额	为命中的规则叠加流量保护	可配置请求速率、并发、熔断、上游调用超时；超限仅支持拒绝
故障转移	主档失败时尝试后备目标	模型调用，先配置后备档位，再配置触发条件和预算
提示词	引用已发布的提示词配置	选择提示词、版本标签、前置或后置位置；核对引用版本
添加请求头	修改模型请求的 Header	设置、追加或删除；不能改写被保护的协议或内部头
添加请求参数变换	对指定 JSON 字段进行变换	选择操作、字段路径和所需值，核对字段类型
添加请求体修改	调整请求体内容或类型	按表单选择支持的操作，验证缺字段和非法类型
添加响应头	修改返回给调用方的 Header	面向模型入口，受保护头不可写
添加响应体格式化	对响应 JSON 执行格式调整	最终非流式响应；不要用于修改流式增量片段
添加映射	映射指定上游状态码	不要把上游失败不加区分地包装为业务成功

动作入口或节点	用途	配置与边界
镜像复制	将符合条件的调用复制到对照目标	面向支持的非模型调用；先核对调用类型、目标和重复执行影响

同类可排序字段按编辑器提供的顺序调整。跨阶段或跨类型动作受执行链限制，不应把画布理解成可以任意连线的通用流程编排器。

5.2 内容检测与提示词

按业务需求选择词库名单、敏感信息、必答、风险或操作检测等检测项。模板、类型和判断模型来自相应管理模块；对象不可用时先修复引用，不通过创建同名文本占位。

- 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”等动作与“拦截”“脱敏”具有不同结果，必须分别用样例验证。
- 检测组件异常时，按配置选择继续处理或拒绝。该选择会影响业务可用性与风险边界。
- 多条规则参与时，护栏会按规则链合并，同一项冲突取更严格的配置；应检查完整规则链。
- 引用提示词后，核对发布版本。页面提示标签已经指向其他版本时，重新确认发布与引用关系，不能只看提示词名称。

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

情形	HTTP 状态	调用方可见内容
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。

选择内容安全模型与模型类型

内容检测（风险类目、提示注入）依赖“全局设置 › 默认模型”里的内容安全模型。设置该模型时必须同时选择模型类型，类型需与实际部署的模型一致；选错类型时保存核对或真实检测会失败，并提示“返回的内容不是所选模型类型的格式”。

模型类型	能识别的风险大类	提示注入检测	适合的部署
通义千问安全护栏（Qwen3Guard）	除“高风险领域应用限制”外的 5 类	仅请求方向	本地 / 云端推理服务（需对话接口）
阿里安全推理护栏（YuFeng-XGuard）	6 类	否（模型不识别提示注入）	本地推理服务（需对话接口）
结构化判断模型（SystemOne）	6 类	是	支持 SystemOne 判断线的供应商
自定义内容安全检测服务	6 类	是	自建 / 三方检测服务，按《自定义内容安全检测服务接口》对接

选型提示：

- 编辑器字段下方会静态标注当前模型不识别的大类（如 Qwen3Guard 不识别高风险领域）；选择这些类别的规则不会命中。模型不识别提示注入时，越狱处置按“不做这项检测”处理，不显示拦截。
- 只给出三档判断（安全、有争议、不安全）的模型，阈值语义见编辑器阈值区域下方的说明；概率给分的模型不受此限制。
- 内容安全模型可能识别不到全部风险，选型前先在“试一句”里用业务真实样例验证。
- 判断模型行会把用户内容发送给所选供应商：供应商走公网时，页面会明示“用户的提问会发送给该供应商”。

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；按指针取字段，不填指针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。确定性工具规则不防拼接与编码混淆写法，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}；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： 启用“检测图片内容”并选择支持视觉输入的图片理解模型后，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“允许拉取的图片地址”白名单中显式填写主机名。
- 能力边界、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，但不占用业务配额池。
- 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（员工先把截图上传给供应商、再用编号引用的做法即属此类）——不做检测，照常放行并在策略命中中留下「未检测」记录，可供审计追溯；这类引用形态可能被用来绕过图片检测，发现异常应结合策略命中记录核查。

多轮拼接诱导防护

在“内容安全 › 多轮对话”中可分别开启三项，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。只防字面拼接。

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；把载荷伪造成历史助手回答并超出容量属于已知边界。

防护边界：

- 只防单个会话内的拼接。会话标识由调用方提供，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，冒用他人会话标识属于已知边界。
- 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；命中详情依次显示“已提交解除，等待生效”“已解除”。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。

升级说明与平台 MCP 批量配置

- 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。
- 平台 MCP 批量配置： 路由策略的新建、修改与试运行都能经平台 MCP 调用。试运行使用 JSON 形态的 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的 simulate/with-images（表单上传）不经平台 MCP 暴露。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 改写字段时的约束

请求参数和响应字段使用 JSON Pointer 定位，例如 `/metadata/channel`。路径以 `/` 开头；包含 `/` 或 `~` 的字段名按 JSON Pointer 规则转义。复制、移动操作还需指定来源，字段投影按页面配置保留字段。

客户端声明的模型、计费及其他受治理字段由平台控制。不能通过改写 `/model`、内部模型路由字段或受保护 Header 绕开模型选择和授权。

请求头和响应头名称必须合法。`Host`、`Content-Length`、连接控制头、内部 `x-lp-` / `x-zr-` 头等受保护字段不可随意修改；具体以字段校验为准。凭据只能在受保护的凭据入口配置，不把真实密钥写进规则名、普通文本字段或截图。

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。模拟命中只能证明规则判定，改写后的字段必须从上游请求或调用方响应核对。

5.4 选择超限处理

配置位置	可选择行为	说明
规则内“应用限额”	拒绝	不支持在此配置排队或默认结果
全局“请求次数”	拒绝（429）、排队、返回默认结果	按所选调用类型及页面依赖校验配置
全局其他限制类型	类型对应的处理	熔断暂停、非法尝试暂停、超时等，不统一套用请求次数动作

全局请求次数选择“排队”时，需要有界队列。最大队列深度为 1–512，最大等待为 1–30000 毫秒；组织队列容量为 1–128，组织字节预算为 1–33554432 字节。运行时有效队列容量还受速率参数等约束，可能小于填写的深度。队列满、等待超时或调用方提前超时仍可能失败。

选择“返回默认结果”时，原因码使用小写字母、数字、下划线，长度为 1–64 字节；内容不能为空且不超过 64 KiB。默认结果是预设降级内容，未经过模型生成，应让业务明确识别。

5.5 配置异常处理

模型入口的“改写失败时”可选择拒绝或保留相应原内容继续处理；MCP、A2A、API 服务等能力入口使用对应的改写异常策略。只有条件和动作适用时才可配置。

异常处理不是所有故障的统一兜底开关：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。要测试实际会发生的失败路径，而不是仅确认异常节点存在。

6. 配置并验证熔断

6.1 熔断解决什么问题

熔断根据已经完成的调用样本判断上游是否持续失败或变慢。达到条件后，暂时拒绝后续调用，减少对异常上游的持续请求。等待熔断时长后，允许有限探测；探测失败会再次熔断，成功后恢复正常处理。

熔断与并发、超时是不同控制：并发限制同时执行数，超时限制等待时间，熔断根据历史结果暂停后续调用。它不会自动给请求选择后备模型；需要切换目标时另配故障转移。



图 4：熔断状态示意。半开期间只允许有限数量的并行探测，仍可能拒绝其他请求。

6.2 从哪里配置

可在规则内添加“限额”并启用“熔断”，也可在“全局限额”中新建“熔断”。两处的参数含义一致，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，应说明各自保护对象，并分别验证，避免无法判断由哪条配置触发。



图 5：真实熔断字段。截图显示表单初始值，参数应按业务容量和恢复目标设置。

在规则内配置时，先将“调用类型”限定为一个适用类型。全局限额中部分参数位于“限额配置”，熔断持续时间和半开探测次数位于后续处理区域，应完整展开检查。

参数	含义	可填写范围
统计窗口（秒）	汇总近期调用结果的时间范围	1–3600
最小调用样本数	至少有多少调用结果后才按比例判定	1–1000000
失败率阈值（%）	失败样本比例达到该值触发熔断	1–100
慢调用阈值（毫秒）	用于判定一次调用是否较慢	1–1800000
慢调用率阈值（%）	慢调用比例达到该值触发熔断	1–100
熔断持续时间（秒）	熔断后暂停普通调用的时长	1–3600
半开探测次数	半开阶段允许同时执行的探测调用数量上限	1–1000

以上均为整数。注意“慢调用阈值”用毫秒，其余时间参数用秒。样本达到最小数量后，失败率或慢调用率任一达到阈值即可触发，不要求两者同时达到。失败类别由入口的实际调用结果归类，不能把所有业务 4xx 都假定为上游故障。

6.3 用受控失败验证

先准备独立测试调用方与可控制返回结果的测试上游，避免把正常业务用于制造故障。以下为行为验证示例，不能直接当作生产容量配置。

1. 设置窗口 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、确认应用并核对下发状态。重新打开详情，逐项核对七个字段。
3. 用同一保护对象依次发送成功、失败、成功、失败四次请求，并在窗口内完成。前 3 个样本不足 4，不应仅因样本比例提前触发；第 4 次累计失败率为 50%。
4. 紧接着继续请求，核对熔断拒绝和上游接收计数。拒绝期间不应继续把这些请求发送给被保护上游。
5. 等待 8 秒后发送探测。让探测失败，应重新进入熔断；随后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；释放前两个并核对恢复结果。

隔离环境已按这类配置验证失败率触发、半开限制、失败重新熔断和成功恢复。模型调用在熔断期曾返回 `503 / TRAFFIC_CIRCUIT_OPEN`。实际排查时仍应以所用入口的响应为准。

6.4 用慢调用验证

设置窗口 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。发送一次快速成功请求，再发送一次延迟超过 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 就认为不会触发。

本版本在隔离环境中已验证上述慢调用行为；未据此覆盖所有调用类型、多副本状态一致性、统计窗口过期和流式中途失败。上线前针对实际拓扑与请求类型补充验证。

停用熔断配置后，应等待配置同步并实际调用确认恢复。现有验证没有证明停用能够在当前熔断期限内立即清空所有运行状态，不把停用开关当作即时复位按钮。

7. 流量分配与故障转移

7.1 目标、档位和权重

“转发到指定目标”最多配置 8 个目标。选择有效的模型目标及归属服务，确认服务包含该目标且调用方有权使用。停用或不可用的目标不会因提高权重而恢复。

同一档位按权重分流。主档为第 0 档；后续档位用于故障转移。每档提交的权重合计为 100，页面会按档归一化；权重均为 0 不合法。同一服务、目标和档位组合不得重复。

例如主档中 A、B 分别设置 80、20，表示主档内的期望分配；后备档位的权重在进入该档时使用，不和主档一起计算 100%。

7.2 选择分流策略

策略	行为	验证方法
加权哈希	按租户、模型、调用方、路径等信息及权重选择目标；相同输入和配置通常选择相同目标	使用多个实际调用方或请求特征观察分布，不能只重复完全相同请求
加权轮询	真实派发按权重轮流选择目标	统计足够数量的实际派发；排除提前拒绝和不可用目标影响

权重表示分配策略。短样本比例还受规则变更、目标健康、候选集合和故障转移影响。

7.3 配置故障转移

1. 将调用类型限定为“模型调用”，在转发节点配置主档目标。
2. 增加至少一个后备档位和目标。未配置后备目标时，“故障转移”会不可选并提示原因。
3. 添加“故障转移”，选择触发条件：连不上上游、上游超时、上游返回指定状态码，或响应内容命中指定条件。
4. 指定状态码时填写 100–599 范围内的值；响应内容条件需要有效 JSON Pointer，例如 `/error/type`，以及期望值。
5. 设置总尝试次数 1–10、总时长上限 1–600 秒。总尝试预算包括首次尝试，设为 1 时没有额外尝试空间。
6. 保存并核对应用状态，模拟主档失败查看候选顺序，再制造受控故障检查实际后备调用。

只有配置了触发条件和后备目标，失败时才按规则尝试转移。列表中存在多个目标，或目标排列在后面，都不代表自动故障转移。

“响应内容命中指定条件”只对非流式请求生效。流式调用一旦开始向调用方输出内容，不能假定平台会透明撤回已输出内容再换一个模型；应按实际协议测试失败阶段。

7.4 转移后的结果核对

记录主档与后备目标的接收情况、最终响应、耗时和请求标识；达到次数或时长上限后停止尝试。客户端超时应覆盖网关尝试预算。存在外部副作用时评估重复调用影响；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、启停与删除

规则清单可按名称、规则 ID 或目标模型搜索，并使用状态与筛选条件缩小范围。全局限额使用自己的类型和条件筛选；筛选隐藏了对象，不表示对象已经停用。

- 用拖拽或“移到最前 / 上移 / 下移 / 移到最后”调整规则顺序。调整后验证完整规则链，尤其是增强规则和兜底拒绝规则。
- 使用状态开关启停。停用后配置仍保留，重新启用前核对关联对象是否还可用。
- 多选批量操作后逐项核对结果；只读或不可编辑对象不能按普通规则处理。

- “修改记录”用于追溯配置变化，不等于一键恢复历史版本。需要恢复时依据记录重新编辑并应用，再验证真实请求。
- 删除前确认替代路由和保护已经准备好。删除不可通过页面直接恢复；后续需要该规则时重新创建。
- 编辑尚未保存时退出，按提示选择继续编辑或放弃；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、条件值为空	填写明确名称与完整取值；不要只输入空格
条件或取值数量超限	拆分为可解释的规则并重新核对顺序；不要用互相冲突的条件合并
模型名被填成 UUID	填入客户端声明的名称；对象匹配改用模型服务或网关模型选择器
目标缺少归属服务、已不存在或不可用	重新选择有效服务及其目标，验证调用权限
所有权重为 0、目标重复	按档配置有效权重，去除同一服务/目标/档位重复项
调用类型与动作不匹配	按页面说明限定类型，或移除不适用动作；重新检查自动新增的条件
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
临时阈值低于基线、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	使用新的原始匹配值；不要复制页面中的摘要回填
调用方目录、模板或关联对象加载失败	等待服务恢复并重新加载；保留已有选择，避免用空值覆盖

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
保存出现 409 或版本冲突	先保留待修改内容；刷新读取最新配置，核对他人变更，重新编辑和生成变更预览；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、运行版本和调用时间；版本同步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；确认模拟使用的假设在真实请求中成立

现象	排查顺序与处理
返回 403	结合错误码区分规则拒绝、凭据权限、模型范围或额度问题；检查拒绝规则和授权，不能只调整路由目标
返回 429	识别限流来源；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 / 熔断提示	核对失败比例、慢调用比例、样本量和恢复时长；检查是否正在半开及上游是否仍异常
返回 504 或上游超时	检查连接、请求、网关总尝试及客户端超时预算；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、字段类型及流式/非流式；核对异常策略是否保留了原内容
后备目标没有收到请求	检查后备档位、显式触发条件、预算、目标授权与健康；确认请求确实达到故障转移阶段
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，不以缺日志认定未执行

问题反馈至少包括：规则或限额名称及 ID、配置版本、操作时间及时区、调用类型与入口、脱敏请求信息、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。

8.4 本版说明

本手册对应 2026 年 9 月 15 日新版流量规则页面，覆盖规则编排、条件、动作、全局限额、熔断、故障转移、保存验证和维护流程。本手册是 Zenova 的路由策略专项操作说明。

页面操作与字段说明已按本版界面和配置校验核对；熔断真实验证覆盖隔离环境中的单目标模型调用。具体部署中的权限、模型资源、检测组件、调用入口和多副本拓扑仍需按第 3 节完成验证。保存和模拟结果均不能替代实际业务验收。