

MLMtoken 可信分控策略网关

设备管理配置说明

——简称：边缘 TokenBoss 盒子——

v 1.1

=====

目 录

CONTENTS

=====

一、系统概况	2
二、API 网关	3
三、网络配置	22
四、无线配置 【*仅带无线 WIFI 的设备型号支持】	24
五、定向加速	25
六、智能组网	27
七、网络诊断	28
八、设备信息	30

一、系统概况

1.1 登录管理后台

网关设备通电联网后，电脑在浏览器地址栏输入后台地址即可进入管理后台。

配置项	默认值/说明
后台访问地址	192.168.137.1
登录账号	admin
初始密码	admin123（登录后可自行修改）

1.2 系统概况

登录后首先展示系统概况模块，用于查看设备运行状态与网络参数。



二、API 网关

API 网关用于对接上游第三方 AI 服务厂商，将上游 AI 接口统一接入网关，再对内网下游用户分发代理 API Key，实现 AI 接口统一代理、密钥隔离、调用统计、用量管控。

2.1 API 网关

2.1.1 API 网关服务状态

API 网关服务状态

当前网关运行状态: **运行中 (HTTP)** (授权上限: 12 / 无限制 用户)

内网发布服务地址与端口:

保存配置

停止服务

💡 提示: 直接输入 http://IP:端口 或 https://IP:端口。若使用 HTTPS，客户端需设置启用跳过证书验证。

配置项	默认值/说明
服务运行状态	开启 / 关闭 API 网关转发服务。
内网发布 URL	网关本地监听地址。 内网下游客户端，访问网关机器 IP+端口 进行 AI 接口调用

2.1.2 敏感词过滤拦截

敏感词过滤拦截

敏感词过滤拦截开关

关闭过滤拦截

过滤规则管控模式

☒ 启用提问拦截 ☒ 启用回答熔断

自定义敏感词列表 (一行一词，单条最长40字符)

赌博
涉政
违禁词

自定义规则触发后提示内容

⚠️ 您的提问或 AI 的回答包含不合规内容，已被系统拦截。

保存配置

配置项	默认值/说明
敏感词过滤开关	开启 / 关闭敏感词过滤
提问拦截	对用户发送的提问进行敏感词检测
回答熔断	对 AI 流式返回的回答做敏感词检测
自定义敏感词列表	业务方自行追加的敏感词库，可填多个词，一行一词
触发后提示内容	命中敏感词后，向返回的提示文案，可按需修改

2.2 AI 服务商

2.2.1 第三方 AI 服务提供商（上游）

第三方 AI 服务提供商（上游）

服务商名称

接口协议规范

第三方 API Base URL

第三方真实 API Key (可选)

添加上游

服务商 (协议)

API URL

真实 API Key

用量与Token(Tokens)

操作

天鸿智群
OPENAI

https://api.thtcloud.com

sk-*****KVwu 查看

总请求: 3 次
总 Tokens: 0

限定模型 重置 删除

博伦智汇
OPENAI

https://api.boolemint.com/api

sk-*****4LVt 查看

总请求: 2 次
总 Tokens: 0

限定模型 重置 删除

硅基流动
OPENAI

https://api.siliconflow.cn

sk-*****mxdf 查看

总请求: 0 次
总 Tokens: 0

限定模型 重置 删除

作用：录入真实 AI 服务商接口地址与原始 API Key，网关以此作为上游转发目标。

配置项	默认值/说明
服务商名称	自定义名称，用于区分不同 AI 源
接口协议规范	备选项为 OpenAI 兼容协议 / Anthropic 原生 / Google Gemini
第三方 API	AI 服务商接口地址，兼容 OpenAI 协议接口
第三方 API Key	AI 服务商给到的原始密钥（不直接暴露给内网用户）

操作步骤：

- 依次填入服务商名称、第三方 API Base URL、第三方真实 API Key、选择接口协议规范。
- 点击【添加上游】，添加后出现在服务商列表。

列表操作：

- 查看：点击查看真实 API Key，可以复制。
- 限定模型：设置允许对外提供的模型白名单，避免越权调用。



点击【限定模型】，系统会获取上游服务商 API 中提供的所有模型列表，通过选择可以设置白名单，避免下游或模型池的越权使用。**全不选择定为所有模型可用。**

- 重置：重置用量与词元统计。
- 删除：移除上游 AI 服务商（下游已绑定该服务商的分发 Key 将不可使用）。
- 用量与词元 (Tokens)：统计上游服务商总请求次数、总消耗 Token 数量。
- 一件清零统计：重置全部服务商的请求计数、Token 消耗统计，不会删除配置，不会清除图表统计数据。

注意：支持多上游 AI 源同时配置，内网可选择不同上游服务商做转发。

2.2.2 模型池管理

模型池是把同一个模型名（如 glm-5.2、deepseek-v4.1-flash）对应的**多个上游服务商**组织成一个逻辑池。下游调用时只需指定模型池中的模型名，网关自动在池内多家上游之间做路由、负载均衡与故障切换，对下游完全屏蔽后端具体是哪家服务商。***探测到存在故障的上游服务商系统会标记红色。**



配置项	默认值/说明
故障切换	服务商连接异常临时禁用时长，默认 1800 秒，点 "保存配置" 生效 作用： 当模型池内某个上游服务商连接超时或异常时，网关将无感知地把流量漂移到池内其他可用服务商，同时把故障服务商隔离所设置的时长，期间不再向其转发请求；隔离期过后自动恢复探测。
模型池名称	自定义一个名称，可给下游分发引用
模型名	下游调用时使用的统一模型名

操作步骤：

- 依次填入模型池名称、模型名。
- 点击【新建模型池】，添加后出现在列表。

列表操作：

- 调度编排：进入该池，配置池内挂载上游服务商及权重。

模型池调度编排：111 (目标匹配: glm-5.2) 新增上游服务商

加入池	上游服务商	并发负载策略(权重1-10)	每日Token限额(0为不限)	今日已用 Token
☒	天鸿智算 OPENAI	1	0	0
☐	博伦智汇 OPENAI	1	0	0
☐	轻基流动 OPENAI	1	0	0
☒	阿里云百炼 OPENAI	1	0	0
☒	千帆 OPENAI	1	0	0
☐	UCloud OPENAI	1	0	0
☐	UCloud2 OPENAI	1	0	0

保存模型池编排配置 返回列表

配置项	默认值/说明
智能扫描匹配服务商	自动探测并勾选上游支持当前目标模型（如 glm-5.2），快速推荐可挂载的服务商
加入池	勾选：该上游参与此模型池的转发；不勾：不参与。
并发负载策略	按权重分配流量，权重越大分到的请求越多；默认都为 1，即平均分摊。
每日 Token 限额	给该上游服务商这个模型池里设置每日 Token 消耗上限，0 为不限，次日自动恢复

配置项	默认值/说明
今日已用 Token	展示上游当日在本模型池下已经消耗的 Token 量
保存模型池编排配置	勾选、权重、限额改完后点保存生效

- 删除池：移除该模型池。

2.3 内网分发

2.3.1 分发内网用户 API Key（下游）

分发内网用户 API Key（下游）

一键清零展示统计数据

绑定分发目标（模型池/服务商）：
test (匹配: deepseek-v4.1-flash)

所属工作组：
默认

自定义/生成的 Key：
ed47cd4480cee9b3cd86c3e15b8ff65f

组成员：
例如：张三

添加分发 Key

工作组：默认 (5 个账号)

工作组统计请求: 80 工作组 Tokens: 972,848

内网分发 Key (客户端)	映射服务商/模型池	所属工作组	组成员	累计展示数据	操作
b262bb*****3a55 查看	博伦智汇 (OPENAI)	默认	小强2	累计请求: 2 次 累计 Tokens: 0	重置 删除
dfe31d*****9b1a 查看	111 (匹配: glm-5.2)	默认	小强	累计请求: 3 次 累计 Tokens: 0	重置 删除
59865a*****4107 查看	UCloud (OPENAI)	默认	老王	累计请求: 71 次 累计 Tokens: 963,744	重置 删除
1dee9e*****80e7 查看	UCloud2 (OPENAI)	默认	666	累计请求: 4 次 累计 Tokens: 9,104	重置 删除
bbad70*****3132 查看	111 (匹配: glm-5.2)	默认	6661	累计请求: 0 次 累计 Tokens: 0	重置 删除

工作组：张三组 (1 个账号)

工作组统计请求: 48 工作组 Tokens: 275,329

内网分发 Key (客户端)	映射服务商/模型池	所属工作组	组成员	累计展示数据	操作
44231b*****ca94 查看	test (匹配: deepseek-v4)	张三组	张三	累计请求: 48 次 累计 Tokens: 275,329	重置 删除

保存用户配置

作用：生成内网客户端代理密钥，内网应用无需持有上游真实密钥，仅使用网关分发 Key 访问网关，由网关代为转发请求。

配置项	默认值/说明
绑定分发目标	下拉选择已经添加好的上游服务商 / 模型池
所属工作组	选择工作组（工作组在 API 网关高级中维护）
自定义/生成的 Key	内网客户端使用的 API Key
组成员	标记使用者

操作步骤：

1. 下拉选择绑定目标 上游服务商或模型池。
2. 选择工作组
3. 设置分发 Key（自动生成或手动填写）。

4. 填写组成员。

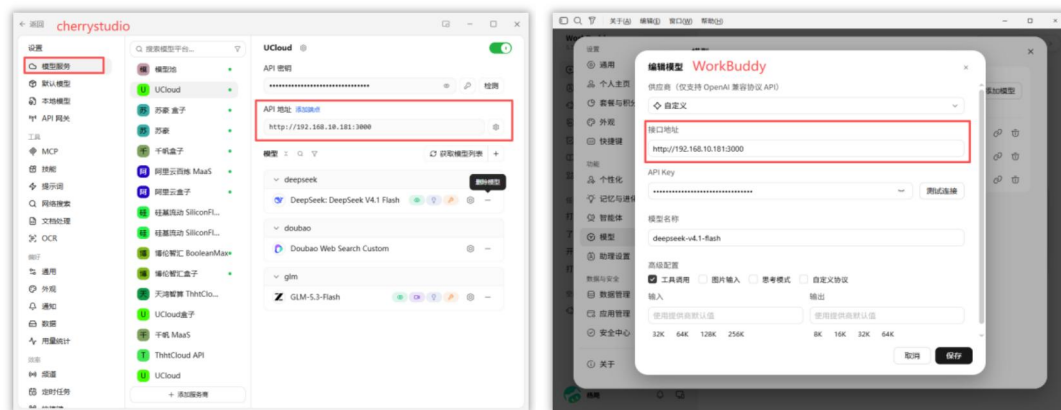
5. 点击【添加分发 Key】，新增下游密钥。

列表操作:

- 内网分发 Key：给内网业务、客户端使用的密钥，点击显示查看完整密钥。
- 映射服务商：该 Key 绑定的上游 AI 源。
- 使用者备注：记录使用人员，支持编辑保存修改。
- 用量与词元 (Tokens)：统计该分发 Key 的请求次数、消耗 Token。
- 重置：重置用量与词元统计。
- 删除：作废该内网分发 Key，客户端将无法继续调用。
- 一键清零所有统计：重置全部下游 Key 的请求计数、Token 消耗统计，不会删除配置，不会清除图表统计数据

注意：客户端（WorkBuddy 、 Cherry Studio 等）只填写内网分发 Key，不要填写上游真实 API Key，网关内部自动替换成上游真实密钥转发。

*WorkBuddy 、 Cherry Studio 的配置截图



2.4 API 网关高级

2.4.1 工作组管理

工作组管理

请输入新工作组名称

添加工作组

工作组名称	组成员总数	操作
默认	5人	默认组不可删
张三组	1人	删除组

保存工作组配置

作用：将内网下游使用者（即分发 Key 的持有人）按部门 / 项目分组，便于按组设置 Token 限额、统计用量，实现精细化的团队级管控。

配置项	默认值/说明
输入框 + 【添加工作组】	输入新组名（如 "研发部" 运营组 "），点添加即可新建工作组
工作组名称	可直接在框内修改组名，点下方 "保存工作组配置" 生效
删除组	默认组显示 "默认组不可删"（系统内置，不能删除）；自建组显示 "删除组"（红字，可删除）

2.4.2 工作组 Token 配额策略

工作组 Token 配额策略

工作组	全组每日总配额	全组每月总配额	组成员单日配额	组成员单月配额
默认	已用: 0 上限: 10000 个	已用: 880,777 上限: 0 个	最高已用: 0 上限: 0 个	最高已用: 871,673 上限: 0 个
张三组	已用: 0 上限: 0 个	已用: 275,329 上限: 0 个	最高已用: 0 上限: 0 个	最高已用: 275,329 上限: 0 个

配额值 0 表示不限制。单位支持 "个"、"千"、"万"。

保存工作组配额策略

作用：用于按组和按成员设置用量上限。

配置项	默认值/说明
全组每日总配额	整组一天内所有成员加起来最多能用多少 Token，超限当天停服
全组每月总配额	整组一个月内所有成员加起来最多能用多少 Token，超限当月停服
组成员单日配额	组内单个成员（单个分发 Key）每天的上限
组成员单月配额	组内单个成员（单个分发 Key）每月的上限

列表操作:

- 上限设置：填数字，旁边单位下拉支持 个 / 千 / 万；填 0 代表不限制
- 用量展示：全组维度显示 "已用"（全组累计用量）；成员维度显示 "最高已用"（组内消耗最多的那个成员用了多少）

修改完成后点底部 【保存工作组配额策略】生效

2.4.3 组成员 Token 配额策略

组成员 Token 配额策略

添加/设置组成员专属策略:

选择或输入组成员:

输入关键字模糊匹配

每日配额: 0

个

▼

每月配额: 0

个

▼

添加策略

组成员	所属工作组	每日专属 Token 配额	每月专属 Token 配额	操作
张三	张三组	已用: 0 上限: 0	已用: 275,329 上限: 0	移除策略

配额值 0 表示继承工作组默认策略。修改后请点击下方保存按钮。

保存组成员配额策略

作用：给某个特定成员（分发 Key）单独设置更严 / 更宽的每日、每月 Token 上限，用于对重点人员、高消耗账号做精细化管控。

配置项	默认值/说明
选择或输入组成员	支持关键字模糊匹配，快速定位某个成员
每日配额	该成员每天上限，单位支持 个 / 千 / 万
每月配额	该成员每月上限，单位支持 个 / 千 / 万

点击【添加策略】加入下方专属策略列表

2.4.4 Token 超额告警与阻断策略

Token 超额告警与阻断策略

超额告警提示

开启客户端超额告警

超额阻断机制

只提示不阻断 (先回复告警内容，再连接AI答案)

告警提示内容

⚠ 您的 Token 使用量已超额!

保存超额策略控制

作用：这是 Token 配额命中后的告警与处置方式配置，决定用户超额时网关怎么响应。

配置项	默认值/说明
超额告警提示	开启 / 关闭客户端超额告警
超额阻断机制	只提示不阻断（先回复告警内容，再连接 AI 答案）； 开启超额阻断（仅回复告警，直接切断不连接 AI）
告警提示内容	自定义返回给客户端的告警文案

2.4.5 远程安全日志审计

远程安全日志审计

远程日志审计开关

启用 Webhook 异步审计日志推送

审计粒度模式

仅记录请求元数据与拦截状态 (推荐)

远端 HTTP Webhook 接收地址

http://192.168.30.187:3001/webhuuk/test

Webhook 鉴权 Token (可选)

磁盘缓存

关闭磁盘缓存 (推送异常直接丢弃新日志)

磁盘缓存路径 (绝对路径)

磁盘缓存最大空间限制 (MB)

50

保存审计与缓存配置

作用：把 API 调用的内容日志异步推送到远端安全审计平台。

配置项	默认值/说明
远程日志审计开关	启用 / 关闭 Webhook 异步审计日志推送
审计粒度模式	仅记录请求元数据与拦截状态（推荐） 完全记录日常所有对话提问与回答全文 ① 只记元数据（调用方、时间、模型、Token 数、是否被拦截），隐私友好、体积小； ② 全量记录（含请求 / 响应内容），排查问题更全，但隐私和存储成本高
远端 HTTP Webhook 接收	接收日志的远端服务器地址

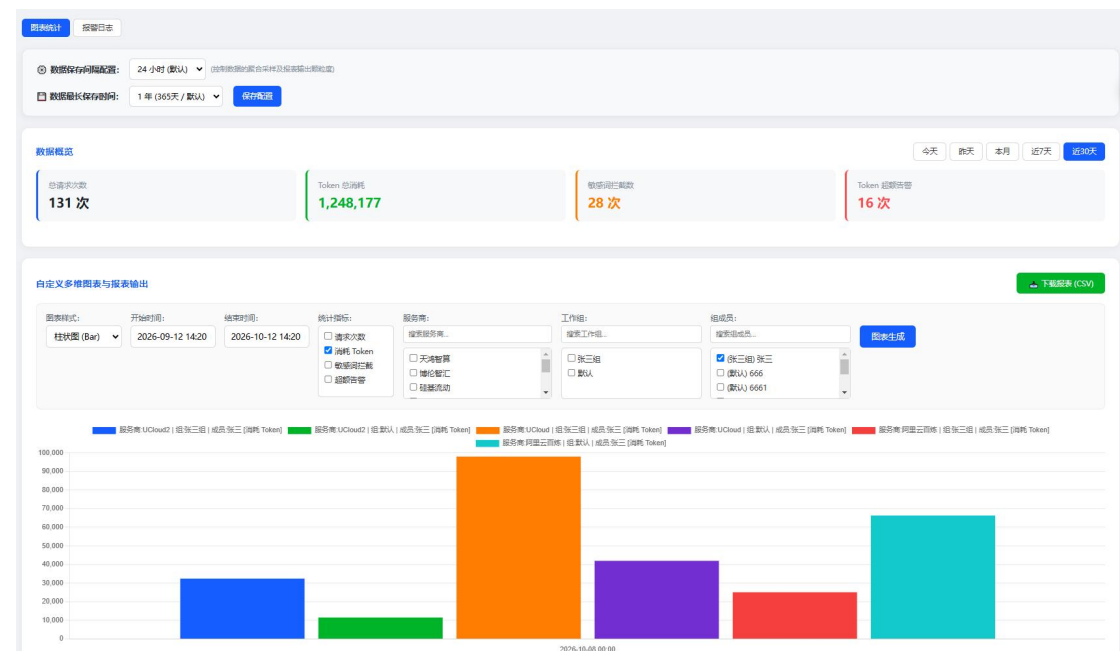
配置项	默认值/说明
地址	
Webhook 鉴权 Token（可选）	远端校验用的 Bearer Token
磁盘缓存开关	远端暂时不可达时的兜底策略： 关闭：推不出去就直接丢，不占本地磁盘； 开启：先落盘缓存，等远端恢复后再补推，不丢日志
磁盘缓存路径（绝对路径）	缓存日志在本机的存放路径
磁盘缓存最大空间限制（MB）	缓存占用磁盘上限，写满后按策略淘汰（防撑爆磁盘）

点底部【保存审计与缓存配置】生效

第三方 Webhook 日志接收服务的开发指南详见 [第 2.6 章节](#)

2.5 图表统计

2.5.1 图表统计



作用：网关自带的多维用量分析看板：看总请求数、Token 消耗、敏感词拦截、超额告警四项 KPI，支持按时间、服务商、工作组、成员自由组合出图并导出 CSV，便于成本核算与安全审计。

2.5.2 报警日志

报表统计

报警日志

报警日志

确认选中项

导出日志 (CSV)

开始时间:
2026-01-01 06:00

结束时间:
2026-10-12 14:28

报警类型 (默认全部):
☐ 服务器
☐ 服务器
☐ token过期

每页显示:
20条

筛选条件

☐	报警类型	内容概要	产生时间	确认时间
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: 42a65896-8497-4f79-bbe2-de68}}, 继续临时封禁	2026-10-12 14:20:22	-
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: 7524fc2e-bb31-4be1-9579-35ea}}, 继续临时封禁	2026-10-12 13:50:21	-
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: 1dcecb18-8e5e-4a2e-8162-aa3a}}, 继续临时封禁	2026-10-12 13:20:20	-
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: 1483cf6e-83f9-4c7e-aed0-e1b6}}, 继续临时封禁	2026-10-12 12:50:19	-
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: a8747a86-2849-45a3-a518-0a04}}, 继续临时封禁	2026-10-12 12:20:18	-
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: d21be554-7fa9-4e96-9c10-104d}}, 继续临时封禁	2026-10-12 11:50:17	-
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: 473bc1d3-a344-4f71-bbe8-97c2}}, 继续临时封禁	2026-10-12 11:20:16	-
☐	服务器	服务器【UCloud】异常阻断, 已无感知, 反馈: HTTP 403 - {"error":{"message":{"trace_id: 1be2dfe9-e878-4851-8877-6e7d	2026-10-12 10:50:14	-
☐	服务器	服务器【UCloud】深度探测失败(状态:403 反馈{"error":{"message":{"trace_id: 80b129d6-761e-4e63-94e7-c553}}, 继续临时封禁	2026-10-08 16:49:31	-
☐	服务器	服务器【千帆】深度探测失败(状态:403 反馈{"error":{"code":"account_overdue","message":"Access denied"}}, 继续临时封禁	2026-10-08 16:48:46	-
☐	服务器	服务器【无海智算】深度探测失败(状态:403 反馈{"error":{"message":"用户额度不足, 剩余额度: 0.000000 (request id: 20)}, 继续临时封禁	2026-10-08 16:48:46	-

作用:网关在对服务商 API 调用时会进行深度探测,发现异常时会生成报警记录,便于用户集中查询各类报警明细。

API 网关功能核心逻辑: 上游配置真实服务商地址 + 密钥, 下游生成内网分发密钥, 内网客户端只使用分发 Key 访问网关, 网关转发到真实 AI 服务商。

2.6 第三方 Webhook 服务开发指南

网关的【远程安全日志审计】模块采用**异步非阻塞**的机制运作，所有请求/拦截日志会在内部队列中缓冲，然后通过 HTTP POST 方式批量推送给接收端。如果你的接收端处理失败，网关会根据配置将日志溢写到外挂磁盘的 SQLite 缓存中。

2.5.1 通信协议规范

- 请求方法: POST
- Content-Type: application/json
- 鉴权头 (Header): 如果你在网关配置了「Webhook 鉴权 Token」，网关会在 HTTP Header 中携带 X-Audit-Token 字段。
- 响应要求: 你的接收服务端必须返回 < 400 的 HTTP 状态码 (通常是 200 OK 或 201 Created)，网关才会认为推送成功。否则判定为失败并触发磁盘缓存机制。

2.5.2 JSON 数据结构 (Payload)

网关每次推送的 JSON 数据结构如下，包含了请求的完整元数据及上下文：

JSON:

```
{
  "timestamp": "2026-09-27T12:34:56.789123", // 记录产生的时间 (ISO 8601 格式)
  "member_id": "user_abc123",                // 实际请求的组成员标识 / Key 前缀
  "group_name": "研发组",                    // 所属工作组
  "provider_id": "provider_123456",          // 命中的上游服务商 ID
  "provider_name": "Claude 官方",           // 命中的上游服务商名称
  "model_name": "claude-3-5-sonnet-20240620", // 客户端实际请求的模型名
  "req_type": "chat",                        // 请求类型 (chat / image_generation / image_to_video 等)
  "total_tokens": 1250,                      // 本次交互消耗的总 Token 数
  "ip_address": "192.168.1.50",              // 发起请求的客户端 IP
  "status_code": 200,                        // 网关最终返回的 HTTP 状态码
  "is_blocked": false,                      // 是否触发了敏感词/配额阻断逻辑 (布尔值)
  "block_reason": "",                       // 阻断原因 (如 "Token 配额超限", "回答包含敏感词熔断")
  "duration_ms": 1500,                      // 整体请求耗时 (毫秒)
  "prompt": "用 Python 写一个快排",        // 用户提问 (仅在开启"完全记录"时有值, 否则为空字符串)
  "response": "好的, 以下是快排代码..."   // AI 回答 (仅在开启"完全记录"时有值, 否则为空字符串)
}
```

注意： 在开启“完全记录日常所有对话提问与回答全文”模式时，为了保护内存，prompt 和 response 的最大长度会被强制截断在 15,000 个字符。

2.5.3 接收端代码示例

下面提供 Python (FastAPI/Flask) 、 Node.js 、 Java (Spring Boot)的标准接收端示例代码。

示例 A: Python (FastAPI) - 推荐

FastAPI 原生支持异步，非常适合处理这种高并发的 Webhook 推送。

Python:

```
from fastapi import FastAPI, Request, Header, HTTPException
from pydantic import BaseModel
import logging

app = FastAPI()
logging.basicConfig(level=logging.INFO)

# 必须与你在网关后台填写的 Token 保持一致
SECRET_TOKEN = "a1b2c3d4-super-secret-token"

class AuditLog(BaseModel):
    timestamp: str
    member_id: str
    group_name: str
    provider_id: str
    provider_name: str
    model_name: str
    req_type: str
    total_tokens: int
    ip_address: str
    status_code: int
    is_blocked: bool
    block_reason: str
    duration_ms: int
    prompt: str
    response: str

@app.post("/api/logs")
async def receive_audit_log(log: AuditLog, x_audit_token: str = Header(None)):
    # 1. 验证 Token
    if SECRET_TOKEN and x_audit_token != SECRET_TOKEN:
```

```

        raise HTTPException(status_code=401, detail="Invalid Audit Token")

# 2. 业务处理 (例如存入 Elasticsearch, MongoDB 或直接写日志)
if log.is_blocked:
    logging.warning(f"🚫 [拦截告警] 用户 {log.member_id} 触发拦截, 原因: {log.block_reason}")
else:
    logging.info(f"✅ [正常请求] 用户 {log.member_id} 调用了 {log.model_name}, 耗时 {log.duration_ms}ms")

# 3. 快速返回 200 OK 告诉网关接收成功
return {"status": "success"}

# 启动命令: uvicorn main:app --host 0.0.0.0 --port 8080

```

示例 B: Node.js (Express)

JavaScript:

```

const express = require('express');
const app = express();

app.use(express.json({ limit: '10mb' })); // 如果开启全文记录, payload 可能较大

const SECRET_TOKEN = "a1b2c3d4-super-secret-token";

app.post('/api/logs', (req, res) => {
    // 1. 验证 Token
    const token = req.get('X-Audit-Token');
    if (SECRET_TOKEN && token !== SECRET_TOKEN) {
        return res.status(401).json({ error: "Unauthorized" });
    }

    const logData = req.body;

    // 2. 处理日志
    if (logData.is_blocked) {
        console.warn(`[拦截] ${logData.group_name}-${logData.member_id} 触发: ${logData.block_reason}`);
    } else {
        console.log(`[正常] ${logData.member_id} 消耗 ${logData.total_tokens} tokens`);
    }

    // 3. 返回状态码小于 400 即可

```

```
res.status(200).json({ message: "Log received" });
});

app.listen(8080, '0.0.0.0', () => {
  console.log("Audit Webhook Server running on port 8080");
});
```

示例 C: Java (Spring Boot) - 企业级后端常用

Spring Boot 非常适合在企业现有 Java 微服务体系中快速集成该审计接口。
以下提供单文件（或拆分包）即可运行的 Controller 与实体类示例。

Java 代码 (AuditWebhookController.java):

```
package com.example.audit;

import com.fasterxml.jackson.annotation.JsonProperty;
import org.slf4j.Logger;
import org.slf4j.LoggerFactory;
import org.springframework.http.HttpStatus;
import org.springframework.http.ResponseEntity;
import org.springframework.web.bind.annotation.*;

import java.util.Collections;
import java.util.Map;
import java.util.concurrent.CompletableFuture;

@RestController
@RequestMapping("/api")
public class AuditWebhookController {

  private static final Logger logger =
    LoggerFactory.getLogger(AuditWebhookController.class);

  // 必须与你在网关后台填写的 Token 保持一致（生产环境建议写在 application.yml 中通过
  // @Value 注入）
  private static final String SECRET_TOKEN = "a1b2c3d4-super-secret-token";

  @PostMapping("/logs")
  public ResponseEntity<Map<String, String>> receiveAuditLog(
    @RequestHeader(value = "X-Audit-Token", required = false) String auditToken,
    @RequestBody AuditLogPayload log) {

    // 1. 验证 Token
```

```

        if (SECRET_TOKEN != null && !SECRET_TOKEN.isEmpty()
            && !SECRET_TOKEN.equals(auditToken)) {
            return ResponseEntity.status(HttpStatus.UNAUTHORIZED)
                .body(Collections.singletonMap("error", "Invalid Audit Token"));
        }

        // 2. 异步处理日志业务（遵循最佳实践：不阻塞网关的 8 秒超时，收到后立刻丢入后台线程写
        库/发告警）
        CompletableFuture.runAsync() -> {
            if (Boolean.TRUE.equals(log.getIsBlocked())) {
                logger.warn("🚫 [拦截告警] 工作组[{}] 用户[{}] 触发拦截, 原因: {}",
                    log.getGroupName(), log.getMemberId(), log.getBlockReason());
            } else {
                logger.info("✅ [正常请求] 用户[{}] 调用了 {}, 消耗 {} tokens, 耗时 {}ms",
                    log.getMemberId(), log.getModelName(), log.getTotalTokens(),
                    log.getDurationMs());
            }
            // TODO: 在此处将 log 存入 MySQL / Elasticsearch / Kafka
        });

        // 3. 立即返回 200 OK 告诉网关接收成功
        return ResponseEntity.ok(Collections.singletonMap("status", "success"));
    }

    /**
     * 审计日志 JSON 映射实体类 (Payload DTO)
     * 使用 @JsonProperty 处理网关下划线字段到 Java 驼峰命名的自动转换
     */
    public static class AuditLogPayload {
        private String timestamp;

        @JsonProperty("member_id")
        private String memberId;

        @JsonProperty("group_name")
        private String groupName;

        @JsonProperty("provider_id")
        private String providerId;

        @JsonProperty("provider_name")
        private String providerName;

        @JsonProperty("model_name")
        private String modelName;
    }

```

```
@JsonProperty("req_type")
private String reqType;

@JsonProperty("total_tokens")
private Integer totalTokens;

@JsonProperty("ip_address")
private String ipAddress;

@JsonProperty("status_code")
private Integer statusCode;

@JsonProperty("is_blocked")
private Boolean isBlocked;

@JsonProperty("block_reason")
private String blockReason;

@JsonProperty("duration_ms")
private Integer durationMs;

private String prompt;
private String response;

// --- Getters and Setters ---
public String getTimestamp() { return timestamp; }
public void setTimestamp(String timestamp) { this.timestamp = timestamp; }

public String getMemberId() { return memberId; }
public void setMemberId(String memberId) { this.memberId = memberId; }

public String getGroupName() { return groupName; }
public void setGroupName(String groupName) { this.groupName = groupName; }

public String getProviderId() { return providerId; }
public void setProviderId(String providerId) { this.providerId = providerId; }

public String getProviderName() { return providerName; }
public void setProviderName(String providerName) { this.providerName =
providerName; }

public String getModelName() { return modelName; }
public void setModelName(String modelName) { this.modelName = modelName; }
```

```

public String getReqType() { return reqType; }
public void setReqType(String reqType) { this.reqType = reqType; }

public Integer getTotalTokens() { return totalTokens; }
public void setTotalTokens(Integer totalTokens) { this.totalTokens = totalTokens; }

public String getIpAddress() { return ipAddress; }
public void setIpAddress(String ipAddress) { this.ipAddress = ipAddress; }

public Integer getStatusCode() { return statusCode; }
public void setStatusCode(Integer statusCode) { this.statusCode = statusCode; }

public Boolean getIsBlocked() { return isBlocked; }
public void setIsBlocked(Boolean isBlocked) { this.isBlocked = isBlocked; }

public String getBlockReason() { return blockReason; }
public void setBlockReason(String blockReason) { this.blockReason = blockReason; }

public Integer getDurationMs() { return durationMs; }
public void setDurationMs(Integer durationMs) { this.durationMs = durationMs; }

public String getPrompt() { return prompt; }
public void setPrompt(String prompt) { this.prompt = prompt; }

public String getResponse() { return response; }
public void setResponse(String response) { this.response = response; }
}
}

```

Spring Boot 配置提示 (application.properties / application.yml): 由于网关在开启“完全记录”时单次推送的 prompt 和 response 文本较大, 建议在配置文件中放宽内嵌 Tomcat 的 POST 请求体积限制 (默认 2MB 一般够用, 但高并发或批量扩展时建议调大) :

```

Properties:
server.port=8080
server.tomcat.max-http-form-post-size=10MB
spring.servlet.multipart.max-file-size=10MB
spring.servlet.multipart.max-request-size=10MB

```

2.5.4 生产环境最佳实践

- 绝对快速响应：

你的 Webhook 接收服务不应该在收到日志后进行耗时的同步操作（例如调用其他慢速 API 或执行复杂聚合计算）。网关发起的请求超时时间设置为 8 秒。正确做法是：收到日志 -> 放入你自己的 Redis 队列或后台线程 -> 立即返回 200 OK。

- Body 大小限制：

如果开启了“全文记录”，单条日志体积可能达到 30KB - 50KB。请确保你的 Nginx 反向代理或 Web 框架（如 Express、Flask）没有把 `client_max_body_size` 设置得过小（建议放宽到 5MB 以上以应对未来可能的批量并发投递）。

- 数据清洗与报表：

利用接收到的 `is_blocked` 和 `total_tokens` 字段，你可以轻易在自己的系统里构建出精美的 Grafana 大盘或者进行企业级的飞书/钉钉违规实时预警。

三、网络配置

3.1 LAN 局域网配置

LAN 局域网配置

IP 地址

192.168.137.1

子网掩码

255.255.255.0

用于设置设备局域网 IP 地址与子网掩码参数。

注意：IP 变更后，本后台的访问地址同步变更

3.2 WAN 上网配置

WAN 上网配置

上网协议

DHCP (自动获取)

DHCP (自动获取)

静态 IP

PPPoE 拨号

上网协议

静态 IP

IP 地址

子网掩码

网关

DNS 服务器

多个DNS用逗号或空格分隔，例如：8.8.8.8 114.114.114.114

上网协议

PPPoE 拨号

宽带账号

宽带密码

配置项	默认值/说明
DHCP	设备从上游网络设备自动获取 IP 地址、子网掩码、网关、DNS 等外网参数，无需手动填写任何网络信息，即插即用，是最常用的上网方式。
静态 IP	手动填写运营商提供的固定 IP 地址、子网掩码、网关、DNS 服务器全部参数。适用于运营商分配固定公网 IP 的场景。
PPPoE 拨号	填写运营商提供的宽带账号、宽带密码，设备主动拨号连接互联网，家用光纤宽带普遍使用该模式。

3.3 路由工作模式



配置项	默认值/说明
路由模式	设备作为网关，可负责拨号、分配内网 IP，普通组网使用。
旁路模式	不充当网关，适配已有主路由的现有网络。需配置下静态 IP。

注意：旁路模式，相连路由器需做一些配置，若有需求可联系客服远程协助。联系电话：400 027 7590

四、无线配置 【*仅带无线 WIFI 的设备型号支持】

2.4G 无线网络

WiFi服务状态 ☒ 开启

WiFi 名称

netbooster_2.4G

加密方式

WPA2/WPA3-PSK(强安全)

WiFi 密码

netbox9527

信道

自动

5G 无线网络

WiFi服务状态 ☒ 开启

WiFi 名称

netbooster_5G

加密方式

WPA2/WPA3-PSK(强安全)

WiFi 密码

netbox9527

信道

自动

配置项	默认值/说明
WiFi 服务状态	开启 / 关闭无线信号
WiFi 名称	设置无线网络的显示名称
加密方式	WPA2/WPA3- PSK 为安全加密模式，保障无线接入安全（建议选择）
WiFi 密码	设置无线连接密码
信道	默认自动，设备自动选择干扰较小信道，可手动指定

2.4G：穿墙好、传输速率较低；5G：速度快、穿墙能力弱。

五、定向加速

5.1 定向加速管理

定向加速管理

服务开关

当前运行: 运行中

开启=立即运行并开机自启 | 关闭=立即停止并禁止自启

选择节点

美国

选择后会同时应用到TCP和UDP协议

加速模式

分流模式(默认)

分流模式(默认)

全局模式

☐ DNS防泄漏

开启后访问国内网络可能会有影响

直连列表

代理列表

屏蔽列表

保存并

直连列表设置

快速应用

☒ 向日葵远程

自定义列表扩展

自定义域名

自定义IP

cn
apple.com
microsoft.com

一行一个IP/CIDR

代理列表设置

快速应用

☐ whatsapp ☐ 小红书 ☐ tiktok ☐ 抖音
☐ wechat ☐ 微信视频号 ☐ linkedin ☐ instagram

自定义列表扩展

自定义域名

自定义IP

一行一个域名

一行一个IP/CIDR

配置项	默认值/说明
服务开关	开启 / 关闭定向加速。
选择节点	选择加速的节点，备选项为购买节点。
分流模式	默认。仅外网代理列表中的地址走节点加速，国内地址走直连。
全局模式	访问的所有地址均经过加速节点。
DNS 防泄露	避免泄露本地网络信息，开启后有可能造成国内网站访问异常。
直连列表	指定域名 / IP 不经过加速节点，直接访问网络。 出现访问不通的情况可在此处配置。
代理列表	指定域名 / IP 走加速节点。 出现访问不通的情况可在此处配置。
屏蔽列表	配置网址黑名单。

5.2 当前定向流量出口

当前定向出口流量

上传流量: 0.07 GB

下载流量: 0.83 GB

总计流量: 0.90 GB

统计当月当前加速节点所产生的流量。

注：数据更新会有延时。

5.3 当前定向出口 IP

当前定向出口IP

IP: 38.34.4.20

城市: Los Angeles

省份: California

国家: 美国

运营商: AS174 Cogent Communications, LLC

查看更多

显示当前定向出口的 IP、所属地址及运营商信息。

查看更多：跳转至第三方平台，可监测当前节点的参数。

PingIP

IP查询

连通测试

网络测速

指纹检测

路由追踪

使用全球最顶尖的IP库，数据每日更新
合作联系：https://t.me/sal_z_z

Neemo

终端模拟 VPS管理软件

网络搭建必备
SSH一键连接
新手小白友好

丽莎主机

双ISP原生住宅VPS 服务全解锁

原生解锁流媒体/海外电商/AI
美国/日本/新加坡/英国/德国/台湾/香港/韩国VPS

覆盖全球的tiktok节点供应商

WePC

提供各国线路: 原生IP / 家宽IP

IP地址

查询

地理位置

us 美国 Los Angeles

经度

-118.2437

纬度

34.0522

ASN

AS174

ASN所有者

ISP Cogent Communications, LLC — cogentco.com

企业

ISP Cogent Communications, LLC

IP类型

家庭宽带IP

原生IP

① 风险

8% 极低

注册国家

us 美国

注册局

ARIN (北美)

CIDR

38.0.0.0/8

截图结果

六、智能组网

可信组网是我公司提供的一项增值服务，能够将用户名下的多台加速宝设备安全、高效地构建专属互联网专网，助力用户实现设备间的稳定互联与数据安全传输。

如有需求，请联系开通相关业务功能。联系电话：400 027 7590

服务状态

✖ 服务已停止

配置参数

设备编码 (UID)

182876083daecc18335476d5c4e2bfd

可信设备ID (deviceId)

182876083daecc18335476d5c4e2bfd@box.netbooster

连接密码 (psd)

.....

查看

配置项	默认值/说明
服务开关	开启 / 关闭智能组网服务。 注意：开启服务前请关闭定向加速，两者不可同时使用。
可信设备 ID / 连接密码	开通业务后由相关业务人员给予分配。

七、网络诊断

Ping 连通性测试

目标地址

www.baidu.com

测试次数

4

开始测试

点击“开始测试”执行Ping连通性检测

Traceroute 路由追踪

目标地址

www.baidu.com

开始追踪

点击“开始追踪”执行路由追踪

Nslookup 域名解析

域名

www.baidu.com

开始解析

点击“开始解析”执行域名解析

互联网测速 (Speedtest by Ookla)

开始测速

点击“开始测速”执行互联网测速
⚠ 测速会消耗流量，建议在非高峰时段执行

内网测速 (iPerf3)

☒ 客户端模式 ☐ 服务端模式

服务端地址

输入iPerf3服务端IP

端口

5201

开始测速

配置项	默认值/说明
Ping 连通性测试	目标地址：填写需要检测的域名或者 IP 地址。 测试次数：设置发送探测包的数量。 功能：检测目标网络是否连通，获取网络延迟、丢包情况。 用途：快速判断网络通断，查看链路时延，简单排查网络是否异常。
Traceroute 路由追踪	目标地址：输入要追踪的域名 / IP。 功能：查看数据包从本设备到目标地址，经过的每一个网络节点。 用途：定位网络卡顿、丢包故障出现在哪一段中间链路，适合排查网络访问慢、访问不通的深层问题。
Nslookup 域名解析	域名：输入需要检测的网站域名。 功能：执行 DNS 解析，返回域名对应的 IP 地址。 用途：判断 DNS 服务是否正常，排查域名解析错误类故障。
互联网测速 (Speedtest by Ookla)	功能：测试外网的上传、下载实际带宽速度。 注意：测速会消耗网络流量，占用带宽，建议避开业务高峰时段。 用途：验证实际上网带宽是否符合预期。
内网测速 (iPerf3)	用来测试局域网内网传输性能，有两种模式： 客户端模式：本机作为测速客户端，填写对端 iPerf3 服务端 IP 与端口（默认 5201），主动连接服务端测试内网带宽。 服务端模式：本机开启 iPerf3 服务，等待其他 iPerf 客户端来连接，本机充当测速服务器。 用途：评估局域网内网传输速度，排查内网带宽不足问题。

八、设备信息

系统在线升级

当前系统版本: **v1.0.0**

在线检查并更新系统

提示: 升级过程中设备可能会短暂断开连接, 完成后将自动重启相关服务。

设备基本信息

设备型号

EDGE Toker50R

设备序列号 (SN)

84e2a618582d55f128a7d2885107eaf0

公司名称

常州万联云网信息技术有限公司

联系电话

400-0277-590

API网关 授权管理

授权状态: **已授权激活**

最大允许内网用户数: 无限制

网关高级功能: **已授权**

网关图表功能: **已授权**

授权到期日期: **2026-12-31**

请输入激活码

ODR1MmE2MTg1ODJkNTVmMTI4YTdkMjg4NTEwN2VhZjB6LTFSMjAyNi0xMi0zMXwxZDEjbFZMZmpSeWFLROFjbkZYR3h6R1FiMlNUeVp

提交授权激活码

配置项	默认值/说明
当前系统版本	当前固件 / 软件版本
在线检查并更新系统	点击自动检查新版本并升级

配置项	默认值/说明
设备型号	硬件型号
设备序列号 (SN)	设备唯一序列号，售后 / 授权绑定用
公司名称	设备厂商
联系电话	厂商客服热线
授权状态	网关是否已开通
最大允许内网用户数	可接入的下游分发 Key 数量上限
网关高级功能	工作组，Token 限额配置，审计日志推送的功能授权状态
网关图表功能	AI 统计报表、报警日志等可视化能力的授权状态
授权到期日期	授权有效期，到期前需续费
激活码输入框	续费 / 变更授权时由厂商提供新激活码，提交后完成激活

以上为本产品完整配置说明，可能会跟软件有差异，一切以软件为准。后续但凡涉及参数调整、版本升级、请记得同步升级本文档，遇到疑难问题欢迎咨询。

联系电话：万联云网 400 027 7590